Teaching English as a Second Language Quarterly (TESLQ)
(Formerly Journal of Teaching Language Skills)
45(1) 2026, pp. 1-28
https://doi.org/10.22099/tesl.2025.52065.3370
Online ISSN: 2981-1546
Research Paper

Analyzing Dependability and Bias in WDCT and DSAT using Many-Facet Rasch and G-Theory

Reza Shahi [1*] **Hamdollah Ravand** [2]

Abstract

Pragmatic testing depends on a variety of factors that can impact its dependability. This study intended to examine these factors using a two-phase approach. The first phase examined the impact of test methods, items, raters, and test-takers' characteristics on the variance in pragmatic test scores using generalizability theory, and the second phase explored potential rater bias using the Many-Facet Rasch model. Two test types, including a Written Discourse Completion Test (WDCT) and a Discourse Self-Assessment Test (DSAT), were administered to 110 English language students (98 female, 12 male) aged 17-24 at Vali-e-Asr University of Rafsanjan. Four raters scored the WDCT by using a standardized rubric developed by Lui (2004). The DSAT was self-assessed by test takers based on the same rubric. The findings revealed no significant difference between the WDCT and DSAT test types. However, items and the interaction between items and test takers emerged as substantial contributors to the variance of the scores. This highlights the importance of item calibration and rater training to mitigate bias in pragmatic testing. Finally, the implications were discussed.

Keywords: Pragmatic testing, Generalizability theory, Many-Facet Rasch model, WDCT, DSAT

The assessment of second language (L2) pragmatic competence, as with all domains of language testing, is inherently subject to measurement error due to the complex interplay of various influencing factors. L2 pragmatic competence refers to a learner's ability to use language effectively and appropriately in social contexts (Yang, 2022). It encompasses the knowledge and skills required to understand and produce communicative acts (e.g., requests, apologies, refusals), to interpret meaning beyond the literal level (e.g., implicature), and to navigate discourse in a way that is sensitive to social variables such as power, distance, and imposition (Kecskes, 2014; Taguchi, 2011, Sitorus, 2025).

* Review History:
Received: 01/01/2025 Revised: 28/08/2025 Accepted: 27/09/2025

1. Ph.D. Candidate, Faculty of Foreign Languages, Ilam University, Ilam , Iran (Corresponding Author)
Reza.shahi411@gmail.com
2. Associate professor, Vali-e-asr University of Rafsanjan, Kerman, Iran, Iran; ravand@vru.ac.ir

How to cite this article:
Shahi, R. and Ravand, H. (2026). Dependability and Bias Analysis in WDCT and DSAT: An Application of Many Facet Rasch Model and Generalizability Theory. *Teaching English as a Second Language Quarterly*, *45*(1), 1-28.
https://doi.org/10.22099/tesl.2025.52065.3370

Assessing this multifaceted construct is notoriously challenging (Shahi et al., 2025). A variety of instruments have been developed, each with distinct advantages and limitations. Common methods include Discourse Completion Tests (DCTs), which prompt written or spoken responses to contextualized scenarios; role-play tasks, which elicit more spontaneous spoken interaction; multiple-choice questionnaires that test recognition of appropriate language; and self-assessment measures, which tap learners' metapragmatic awareness (Roever, 2011; Youn, 2020). Among these, DCTs, particularly Written DCTs (WDCTs), have become one of the most frequently utilized tools due to their practicality and efficiency in data collection from large cohorts (Taguchi, 2018). However, their effectiveness hinges on a variety of factors, and both relevant and irrelevant facets can contribute to score variance (Shahi et al, 2025).

One of the most significant factors affecting test score variance is test-takers' characteristics. Test takers' proficiency level, background, and other characteristics affect their performance (Bachman & Lynch, 1994). In addition, the design and characteristics of items play a key role in the effectiveness of a test. The design and selection of test items directly influence a test's ability to measure the intended construct accurately. Poorly designed items may fail to represent the targeted skills adequately and lead to error variance in test scores (Kumar et al., 2021). The test method can also affect the test-takers' performance. Differences in test format can affect how people perform on the test. Such factors may alter test results and lead to inaccurate conclusions about test takers' abilities (Akhavan Masoumi & Sadeghi, 2020; Chappell et al., 2015).

Moreover, raters, who are responsible for scoring tests, also affect test results. Testing should inherently strive for fairness and equity. Rater biases and their interpretation of scoring criteria can lead to error variance (Lui, 2014; Tajeddin et al., 2020). These biases may arise from cultural or linguistic differences that are not considered in scoring procedures (Lozano-Reynolds et al., 2021; Ruiz, 2021). Identifying and mitigating such biases is critical for ensuring the validity and reliability of tests. Ensuring that tests are unbiased across test takers helps produce reliable, pragmatic tests.

In addition, the interactions between these elements are also very critical. How test-takers interpret the task, how evaluators assess their responses, and the intrinsic properties of test items all contribute to shaping test stability and validity. By examining these interactions, we gain valuable insight into the broader construct of pragmatic competence and strengthen the power of the assessment tools. This complexity underscores the importance of addressing these factors to improve the dependability of pragmatic assessments and enhance their effectiveness in measuring true pragmatic competence.

Pragmatic assessment has gained significant research attention due to its key role in effective communication across diverse contexts (Azizi & Namaziandost, 2023; Toe et al., 2020; Youn & Bi, 2019). So far, a large array of research studies has developed and validated different test types for assessing pragmatics (e.g., Ahn, 2005; Bardovi-Harlig & Su, 2023; Cohen, 2020; Farashaiyan et al., 2020; Fussman & Mashal, 2022; Grabowski, 2008; Liu, 2004;

Roever, 2011; Rose, 1994; Rose & Ono, 1995; Taguchi, 2011; Wilson & Bishop, 2022; Xu & Wannaruk, 2018; Zangoei & Derakhshan, 2021). Despite these advances, the key role of these effective factors and their interactions has not been fully investigated.

Existing research acknowledges the influence of individual factors (Roever, 2013; Youn & Brown, 2013), rater variability (Alemi & Rezanejad, 2014; Brown,1995; Choen, 2020; Derakhshan et al., 2020; Dabbagh & Babaii, 2021; Li et al., 2023; Liu & Xie, 2014; Sydorenko et al., 2014; Taguchi, 2011; Sonnenburg-Winkle et al., 2020; Tajeddin & Alemi, 2014; Walters, 2007; Li et al., 2019; Su & Shin, 2024; Youn, 2015), test item design (Cordier et al., 2019; Roever, 2008; Timpe-Laughlin & Choi, 2017), and the interplay between these elements (Brown & Ahn, 2011; Han, 2021). Despite this knowledge, a crucial gap exists. We lack a comprehensive understanding of how these factors interact and how methodological variations within the DCTs (e.g., Written DCT vs. Discourse Self-Assessment) impact test-taker performance and ultimately, the dependability of the DCT in measuring pragmatic competence.

This study intended to address the gap by investigating the impact of test takers, rater bias, test items, and their interactions on the dependability of the DCT tests. WDCT and DSAT are the most common DCT formats. They are widely used tests for pragmatic testing; however, their effectiveness, as mentioned before, hinges on various factors. These factors can potentially compromise the validity and reliability of the tests. This study intends to provide a comprehensive understanding of how these factors affect test results in pragmatic testing by using innovative statistical methods, including Generalizability Theory and the Many-Facet Rasch Model. In addition, this study intends to offer insights into optimizing the assessment of pragmatic competence in the English language as a Foreign Language (EFL) context.

G-theory and Many-Facets Rasch Model

Classical Test Theory (CTT) provides primary approaches to reliability estimation that rely on quantifying measurement error and addressing related problems within specific testing contexts (Steyer, 2001)**.** However, CTT's inherent limitation lies in its inability to account for the complex interactions among these error sources (Khodi, 2021). Researchers have begun using innovative statistical models to address the complex nature of testing and the limitations of CTT in capturing the multifaceted nature of test performance (Wolcot et al., 2022). G-theory and Many-Facets Rasch Model (MFRM) offer complementary methods that overcome these limitations (Anthony et al., 2023). G-theory excels at pinpointing the sources of measurement error, such as examinee background, rater bias, and item difficulty (Han, 2021). By quantifying these components, G-theory elucidates their influence on test reliability. Notably, G-theory's capacity to simultaneously assess multiple sources of variation within a single analysis distinguishes it from CTT (Khodi, 2021).

MFRM builds upon G-theory to statistically examine the impact of different 'facets' on test scores (Gordon et al., 2021). This allows researchers to identify these factors and understand how they influence test scores. In contrast to CTT, MFRM operates within a scaling framework,

aiming to place all facets of the measurement process on a single latent scale (Engelhard & Wind, 2017). MFRM focuses on deriving linear measures of examinee performance that are adjusted for rater and task effects (Iramaneerat et al., 2008). This makes it possible to identify misfitting elements, including rater bias (Li et al, 2021; Lynch & McNamara, 1998). By controlling rater effect, MFRM provides unbiased estimates of examinee ability. As a result, decisions based on MFRM are more objective and impartial.

Both G-theory and MFRM provide distinct approaches to conceptualizing and controlling the impact of various facets (Anthony et al., 2023; Engelhard & Wind, 2017; Han, 2021). However, combining GT with MFRM offers a more comprehensive and powerful statistical framework for better assessment procedures. By combining the strengths of both models, organizations can effectively use data analysis results to select qualified candidates, refine evaluation processes, and optimize rater training (Li et al., 2021). This integrated approach has enormous potential to advance the development of robust and informative pragmatic assessments.

Literature Review

Following the groundwork laid by Hudson et al. (1992) with their six pragmatics test types, a large body of research has explored and validated various tests across diverse contexts. These tests have been translated and adapted for different languages. Yamashita (1996) confirmed that the majority of the tests are effective for Japanese as a Second Language (JSL) learners, except the Multiple-Choice Discourse Completion Test (MDCT). Enochs and Yoshitake-Strain (1999) found issues with the Written Discourse Completion Test (WDCT) and MDCT when administered to Japanese EFL learners. Brown (2001) further examined the tests and compared them across JSL and EFL contexts. He reported that all test types showed reliability in the JSL context, except WDCT and MDCT. In another study, Ahn (2005) generated Korean versions of various tests and evaluated their effectiveness. Similar results were found by Ahn (2005) for the Korean as a Foreign Language (KFL) context.

DCTs are among the most popular tests for assessing pragmatic knowledge in EFL contexts (Taguchi 2018). As a result, DCTs have been used in various contexts to collect data in many research studies (e.g., Budeng & Merza, 2023; Hernández, 2018, 2021; Hernández & Boero,2018; Saleem et al., 2022; Namaziandost et al., 2020; Taguchi et al., 2016). In addition to traditional DCTs, a few studies have developed alternative DCT types, such as self-assessment DCTs. Liu (2004) developed a Discourse Self-Assessment Test (DSAT) for Chinese EFL learners and found it to be highly reliable. Brown (2008) included role-play self-assessment alongside other tests and found it effective for measuring KFL learners' pragmatic knowledge. While WDCTs are well-established, DSAT has not been extensively investigated.

Rater Studies

Investigating the impact of Raters on test scores has been the subject of a large array of research studies in the field of pragmatic testing. Taguchi (2011) noted that native speaker raters differ in their evaluations of speech acts, influenced by their cultural backgrounds and personal norms. This can lead to inconsistency in test results. Sonnenburg-Winkler et al. (2020) found that raters from different linguistic backgrounds showed variability in their assessments of pragmatic performance. This indicates that language background affects scoring patterns. Li et al. (2019) reported that rater variability affects the reliability of pragmatic test scores due to differences in interpretation, proficiency levels, and speech acts.

In addition, studies have consistently demonstrated that raters can introduce bias into the scoring process, leading to unreliable and potentially unfair results (Liu & Xie, 2014; Tajeddin & Alemi, 2014; Youn, 2007). Zhai et al. (2021) reported that judgments of severity and rater scoring sensitivity can significantly affect the validity of constructed-response assessments, particularly in information-rich contexts. In another study, Liu and Xie (2014) found significant differences in rating severity between raters. They reported a general tendency towards severity. In addition, some studies investigated the role of rater training in mitigating rater effects. Tajeddin and Alemi (2014) found that rater training can improve non-native teachers' rating accuracy. Moreover, Rossi and Brunfaut (2020) emphasize the importance of training and monitoring raters to mitigate variability, suggesting that communal rating sessions can enhance reliability.

While rater effects have been a primary focus in pragmatic assessment research, a broader range of factors can significantly influence test scores. These factors include test-taker characteristics such as anxiety and cultural background (Roever, 2013; Youn & Brown, 2013). In addition, rating scales can affect test results, requiring careful development and validation (Chen & Liu, 2016; Derakhshan et al., 2020; Li et al., 2019; Youn, 2015). Furthermore, the test items should efficiently measure the intended pragmatic skills, may have an impact on the test results (Cordier et al., 2019; Roever, 2008; Zhai et al., 2021), and the selected test method can also play a role (Bardovi & Hartford, 1993; Rose, 1995; Youn, 2015). Moreover, the lack of clarity in the construct definition of pragmatic knowledge can be problematic and lead to inconsistencies in testing and test interpretation (Beltran, 2019).

Overall, three limitations were identified in the literature on pragmatic assessment. First, few studies have explicitly utilized innovative statistical methods to evaluate the impact of effective factors on WDCT and DSAT. Second, the dependability of DSAT has been understudied. Third, relatively few studies have investigated the impact of test methods and non-native raters on pragmatic assessments.

This Study

While the DCTs are popular tests for assessing pragmatic competence, their effectiveness depends heavily on different effective factors that can compromise their validity and reliability.

Existing research acknowledges the impact of test takers' individual characteristics, rater variability, test item design, and the interplay between these elements. However, a crucial gap exists in comprehending how these factors interact and how methodological variations (written vs. spoken delivery, contextual framing) affect test-taker performance. This study aims to fill the gap by examining the contributions of test-takers, raters, test items, and their interactions to the dependability of the WDCT and DSAT formats, using the Many-Facet Rasch model and generalizability theory. By applying these statistical models to comprehensively investigate these factors, the study seeks to provide valuable insights into optimizing the assessment of pragmatic competence in language testing contexts. To this end, the following research questions were posed:

To what degree do examinees, raters, item, and their interactions contribute to the dependability of WDCT and DSAT pragmatics tests?

What is the effect of the test method on test-takers' performance?

Is there any significant bias between the interactions of the different variables?

Method

Participants

The participants of this study were 110 English language students (98 female, 12 male) aged 17-24 (M = 20.2, SD = 1.6). A convenience sampling method was employed to recruit participants from the population of undergraduate English translation and literature students at Vali-e-Asr University of Rafsanjan (VRU). All participants had completed at least two semesters of study at VRU and had studied English for approximately five years before attending university. Their English proficiency could be rated as intermediate.

Four raters with at least 6 years of language-teaching experience were invited to score the tests. The raters were MA students in English teaching at Rafsanjan University. Two of them were female (A and B), and two of them were male (C and D). All raters have taught English in high schools. This demonstrated their expertise in evaluating student language skills. While teaching experience in high schools may not directly reflect their capability in pragmatic test ratings, it shows they are proficient enough to understand the nuances of language use and communication. All participants provided written informed consent before data collection began. We ensured strict adherence to ethical principles, including maintaining participant anonymity, to prevent any potential issues.

Instruments

This study employed two test formats to assess pragmatic competence: the Written Discourse Completion Test (WDCT) and the Discourse Self-Assessment Test (DSAT) (detailed descriptions in Appendix A). Both tests used six identical request situations. The WDCT, adapted from Rose (1992), presented these situations and required participants to write appropriate responses (e.g., requesting someone to turn down loud music). The DSAT followed

the WDCT prompts, but participants then rated the quality of their own responses using a provided 5-point scale based on specific criteria. To ensure the tests were suitable, we consulted experts and conducted a pilot study with a small group of participants to identify potential issues. The pilot study was conducted with a separate group of 15 participants who shared characteristics similar to those of the main study participants. Feedback from the pilot participants indicated that the instructions and scenarios were clear and comprehensible. However, the pilot revealed that the original time allotted for the WDCT (60 minutes) was insufficient for most participants to complete their responses thoughtfully. Consequently, the WDCT time limit was extended to 90 minutes for the main study.

A standardized rubric adapted from Lui (2004) was utilized to score WDCT and DSAT (Appendix B). This rubric included five key components, each rated on a 5-point scale ranging from 1 (no evidence of the knowledge) to 5 (complete knowledge of the component).

Data Collection

In a single session at VRU, 110 students were invited to complete both tests. First, test takers were asked to complete the WDCT in 90 minutes. Prior to beginning the test, participants were provided with a brief explanation of the concepts of social power, politeness, and directness. This was done to ensure that they were aware of the cultural and social factors that can influence the appropriateness of requests. They were asked to put themselves in each situation and write down what they think they would do or say. Since all the items measured the request speech act, students were asked to consider the situation and try to make the assumed interlocutor accept the request. To clarify how to complete the tests, students were provided with a Persian example. Next, test takers were asked to do DSAT in 30 minutes. They were asked to assess their own answers for each item using the rating rubric provided (adapted from Lui, 2004; see Appendix B**)**. An explanation of how to score themselves using the 5-point scale, considering grammatical aspects, social power, politeness, and directness, was also provided. To ensure comprehension, the rubric's key criteria (grammatical accuracy, awareness of social power, politeness, and directness) were explained in Persian. Test takers were instructed to read their own written response from the WDCT for each item and then rate its quality on a 5-point scale using the defined criteria**.**

After administration, tests were scored by four raters separately by using the same rubric. Each rater scored all 110 test-takers' responses to all 6 items on the WDCT, resulting in a fully crossed design in which every response was rated by every rater. This amounted to 110 students × 6 items = 660 individual responses per rater. The raters worked independently and were blinded to each other's scores. The order of the test papers was randomized for each rater to prevent order effects. Scoring was conducted over two weeks to avoid rater fatigue. Then scores were collected and entered into an Excel file to organize the data. Finally, data were exported to the EDUG and FACET software programs for analysis of the impact of the aforementioned factors.

Data Analysis

Given the multifaceted nature of pragmatic testing and its susceptibility to various factors, this study employed innovative statistical analyses to ensure measurement precision and assess the effects of multiple factors on participant performance. Generalizability Theory (G-Theory), implemented in the EDUG software, was used to estimate test reliability and explore the impact of various factors on test scores. In addition, the Rasch Multi-Facet Model, implemented in the FACET software, was employed to analyze raters' bias and the interactions among facets, providing a comprehensive understanding of their effects on test scores.

To measure the effects of test takers, raters, items, and the test method on students' performance, multiple generalizability studies (G studies) were conducted. The first G study was related to the WDCT test. It included three facets: *persons*, *raters*, *and items*. In this G study, all facets were fully crossed, meaning all test takers tried all items, and all raters scored all test takers' answers. In this study, 'persons' was the differentiation facet or object of the study, and 'raters' and 'items' were the instrumentation facet. The second G study was conducted to measure the impact of person and item on test takers' performance on DSAT. This G study was a two-factor study in which 'person' was the differentiated factor and the item facet was the instrumentation facet. The final G study was conducted to measure the test method effect. In this G study, items were nested within method and crossed with test takers. Additionally, FACETS software was used on three separate occasions to examine potential rater bias within each test method and its interactions.

Results

EDUG analysis

Estimated variance components

Two different generalizability studies (G study) were conducted to assess the variance components contributing to scores on the WDCT and DSAT. The first G study examined the WDCT, assessing the relative influence of test-takers, raters, and test items on scores. In this analysis, test takers were the object of the study, while raters and items were considered facets. The second G study focused on the DSAT, investigating the variance components attributable to persons and items. Here, persons were again the object of the study, and items were considered the facet. The G studies for each method are presented in Table 1.

Table 1

Component variance

Source	WDCT			DSAT		
	SS	VC	%	SS	VC	**%**
P	642.39545	0.20866	17.1	317.33333	0.39981	**40.5**
R	188.11667	0.06733	5.5	-	-	
I	113.83182	0.04973	4.1	44.03030	0.07540	**7.6**
PR	849.71667	0.38669	31.7	-	-	
PI	482.66818	0.22141	18.1	279.30303	0.51248	**51.9**

Source		WDCT			DSAT	
RI	17.03788	0.00780	0.6	-	-	
PRI	455.12879	0.27837	22.8	-	-	-
Total	2748.89545		100	640.66667		**100**
Coef_G RE	0.85			0.82		
Coef_G AB	0.82			0.80		
AbsoluteE V	0.04519			0.09798		
Relative E V	0.03690			0.08541		
Absolute SE:	0.21258			0.31302		
Relative SE:	0.19210			0.29226		

Note:
Coef_G RE: G-coefficient for relative decisions
Coef_G AB: G-coefficient for absolute decisions
Relative EV: Relative Error Variance
Absolute EV: Absolute Error Variance

Estimated component variance (VCs) for each facet and their interactions are presented in the "VC" column of the table. In addition, the percentage contributions of each facet and interaction are presented in the "%" column. The analysis of the main facets in the WDCT (person [P], raters [R], and items [I]) revealed that the person facet contributed the most (17.1%) to the total variance in scores. The rater facet (5.5%) had the second-highest contribution, followed by the items facet (4.1%), which had the smallest contribution.

Beyond the main effects, the interaction between persons and raters emerged as the most significant variance component in the WDCT, accounting for 31.7% of the total variance in scores. This indicates that raters' evaluations of participants' performance can vary depending on the specific participant being assessed. The interaction of persons, raters, and items (22.8%) also contributes substantially to the variance. This three-way interaction suggests that the impact of a particular item on a test-taker's score may depend on the interaction between the test-taker and the rater. The interaction between persons and items (18.1%) is another noteworthy component, highlighting that some items may elicit different performances from different participants. Finally, the interaction between raters and items has the smallest contribution (0.6%), indicating that the raters' evaluations are relatively consistent across items.

In the DSAT analysis, the person facet (test-takers) again emerged as the primary source of variance in scores, accounting for 40.5% of the variance, compared to items (7.6%). Examining the interactions among the facets in the DSAT, the only possible interaction accounted for the most significant contribution to the total variance, with 51.9 percent of the total contribution. A key observation is the difference in person variance between the two methods. The person facet contributed substantially more to the variance in DSAT scores (40.5%) than to that in WDCT scores (17.1%). The item facet contributed relatively little in both methods, accounting for 4.1% and 7.6% of the variance in WDCT and DSAT scores, respectively. The person-item interaction is the most significant contributor to the DSAT and also makes a significant contribution to the WDCT. Whereas, generally speaking, a higher proportion of variance attributable to the person

facet is preferable, this study found a high contribution of the person-item interaction that may have reduced the relative contribution of the person facet.

G theory emphasizes the importance of analyzing variance components and measurement error to understand the dependability of scores. However, it also offers a generalizability coefficient (G coefficient) that serves a similar purpose to the reliability coefficient in classical test theory (Richard et al., 2009). Building upon CTT, G theory differentiates between relative and absolute G coefficients (Karami, 2010). Relative G coefficients are particularly relevant for norm-referenced tests, where the primary focus is on a test taker's position compared to the entire group (Karami, 2010). In this study, the aim is to assess participants' relative standing, making the relative G coefficient the appropriate choice. This coefficient indicates the degree to which the participants' relative positions within the group are generalizable across different testing situations.

The relative G coefficient is calculated by dividing the estimated variance component attributable to persons by the total variance (Karami, 2010). This coefficient indicates the proportion of the total variance that can be attributed to true differences between participants. A higher relative G coefficient suggests that the test is more dependable in assessing individual differences. Following Cardinet et al. (2010), G confection values exceeding 0.8 are considered acceptable. In this study, both the absolute and relative G coefficients fall within the acceptable range.

Test Method G Study

The final G study was conducted to assess the test method's contribution to score variance specifically. In this design, persons were the differentiating facet, representing the universe of interest for generalization. The test method and items were considered instrumentation facets, which could contribute to the error variance. Importantly, items were nested within the test method. The test method itself was a fixed facet, meaning we were not interested in generalizing its influence. This design isolates the impact of the test method on scores by controlling for variation across items within each method.

Table 2 presents the ANOVA results. As the table shows, the highest contribution to the score variance, about 46.3%, comes from the interaction of all facets (represented in the table as PI: M). The main facets analysis shows that the person facet makes the largest contribution, and the test method makes the smallest.

Table 2

Test method variance

Source	SS	VC	%
P	369.73030	0.23915	21.2
M	3.10303	-0.01325	0.0
I:M	101.73333	0.08774	7.8
PM	239.73030	0.27952	24.8
PI:M	569.26667	0.52226	46.3
Total	1283.56364		100%

G- study table (see Table 3) provides the variance components and their contributions. The first column lists the sources of variance, the second column shows their contribution to the true variance (universe score), and the third column indicates components that do not contribute to either true or error variance. The fourth and fifth columns show the relative contributions of each component to the relative error variance. The last two columns display their contributions to the absolute error variance. Interactions between a focal element (persons, in this case) and a fixed facet (test method) are excluded from the error variance calculations. These interactions are presented in parentheses to show that they do not contribute to either relative or absolute error.

The table reveals that the interaction between person (P) and item (I) within method (M), shown as PI:M, accounts for the highest contribution (100%) to the error variance.

While the interaction of person (P) and item (I) within method (M) dominates the relative error variance, the absolute error variance presents a different picture. In this case, the interaction of person (P), method (M), and item (I), denoted by P:M:I, has the greatest contribution (85.6%). The second-largest contributor to the absolute error variance is the interaction between item and method (I:M), at 14.4%. Despite the above-mentioned findings, the high relative and absolute G coefficients suggest that the test scores have good generalizability.

Table 3

G- Study

% absolute	Absolute error variance	% relative	Relative error variance	Source of variance	Differ-entiation variance	Source of variance
	.....		.....		0.23915	P
0.0	(0.00000)		.....	M	.....	
14.4	0.00731		.....	I:M	.....	
0.0	(0.00000)	0.0	(0.00000)	PM	.....	
85.6	0.04352	100.0	0.04352	PI:M	.....	
100%	0.05083	100%	0.04352		0.23915	Sum of variances
Absolute SE: 0.22546		Relative SE: 0.20862			0.48903	Standard deviation
0.85					Coef_G relative	
0.82					Coef_G absolute	

Bias analysis

The interaction among the study's facets, particularly raters and others, is presented in this section. Rater bias, where scores assigned to a student may deviate from expected values, was investigated. In order to identify and quantify significant biases, four bias analyses were conducted.

Rater bias across test takers in WDCT

The results of the bias analysis, focusing on interactions between raters and test-takers' ability, are presented in separate tables for each rater. These tables list only rater-candidate interactions that exhibited significant bias. In other words, they only include interactions where the z-score values are outside the range of ±2.0. Out of a total of 440 interactions (from four raters and 110 test takers; 4 × 110 = 440), 74 exhibited significant bias. Tables 4, 5, 6, and 7 present the bias reports for these interactions between raters and test-takers.

In each table, the first two columns (1 and 2) display the rate numbers and labels in the facet analysis. The third column indicates the number of test-takers. The subsequent five columns present: 1) the total observed scores given by each rater (column 4), 2) the total expected scores for each rater (column 5), 3) the bias size in logits (column 6), 4) the estimated standard error (column 7), 5) the z-scores (column 8), and 6) the infit values (column 9).

The results showed some distinct bias patterns in the four raters' assessment of the examinee's abilities. All raters showed different biases across levels of examinee ability; some assigned higher scores than their average, while others assigned lower scores than expected. This shows complex and idiosyncratic bias patterns across the four raters and underscores the highly individualized nature of rater effects in this study. However, rater B's scoring was particularly more erratic, as evidenced by the most extreme biases in both directions (e.g., a severity bias of +4.56 logits for test-taker 37 and a leniency bias of -2.28 logits for test-taker 27, the latter with a z-score of -5.65). Furthermore, Rater B's significant biases were frequently associated with low infit values (e.g., 0.1), indicating an overly deterministic and inconsistent application of the rubric, which differed from that of the other raters.

Table 4

Bias of rater A across test takers

Rater	label	Test takers	Observed score	Expected score	Bias size	error	Z score	Fit
1	A	27	30	22.40	3.72	1.69	2.21	.1
1	A	70	30	22.65	3.68	1.70	2.16	.1
1	A	71	30	23.15	3.61	1.74	2.07	.1
1	A	54	26	19.66	1.66	.59	2.83	1.3
1	A	57	27	21.66	1.57	.65	2.42	.4
1	A	67	24	1.841	1.33	.52	2.55	.1
1	A	63	24	31.41	-.99	.47	-2.12	.6
1	A	82	17	122.67	-.38	.19	-2.01	.5
1	A	83	17	21.66	-1.5	47	-2.25	.7
1	A	88	12	16.90	-1.10	.50	-2.21	.1
1	A	79	19	23.6	-1.11	.47	-2.34	1.08
1	A	37	12	17.15	-1.15	.50	-2.32	.8
1	A	38	15	20.91	-1.30	.47	-2.79	1
1	A	8	136	152.17	-.60	.19	-3.16	.6
1	A	28	12	21.42	-2.11	.50	-4.23	.1

Table 5

Bias of rater B across test takers

rater	label	Test takers	Observed score	Expected score	Bias size	error	Z score	fit
2	B	37	30	19.29	4.56	1.78	2.57	.1
2	B	18	30	22.16	3.77	1.67	2.25	.1
2	B	38	30	22.86	3.65	1.72	2.12	.1
2	B	36	30	23.32	3.58	1.76	2.04	.1
2	B	66	30	23.32	3.58	1.76	2.04	.1
2	B	82	30	23.32	3.58	1.76	2.04	.1
2	B	86	30	24.21	3.00	1..50	2.01	.1
2	B	34	24	19.29	1.14	.52	2.18	.1
2	B	89	24	19.29	1.14	52	2.18	.1
2	B	106	16	20.05	-.99	.47	-2.13	.3
2	B	103	13	18.29	-1.16	.48	-2.41	1.5
2	B	6	20	24.87	-1.22	.48	-2.56	.5
2	B	26	18	23.32	-1.24	.47	-2.65	.1
2	B	28	18	23.32	-1.24	.47	-2.65	.1
2	B	95	12	18.54	-1.46	.50	-2.93	.6
2	B	69	15	21.69	-1.49	.47	-3.18	.4
2	B	24	12	20.85	-1.98	.50	-3.97	.1
2	B	27	12	24.21	-2.28	.50	-5.65	.1

Table 6

Bias of rater C across test takers

rater	label	Test takers	Observed score	Expected score	Bias size	error	Z score	fit
3	C	29	30	16.32	5.00	1.61	3.10	.1
3	C	69	25	16.84	1.96	.55	3.58	.5
3	C	102	22	13.81	1.83	.49	3.72	.6
3	C	55	23	15.81	1.63	.50	3.24	.5
3	C	28	24	18.68	1.27	.52	2.44	.1
3	C	103	19	13.33	1.25	.47	2.64	1.2
3	C	105	21	16.07	1.09	.48	2.27	.7
3	C	104	22	17.62	.99	.49	2.02	.7
3	C	1	20	15.56	.98	.48	2.05	3
3	C	95	18	13.57	.97	.47	2.07	.6
3	C	52	12	16.58	-1.03	.50	-2.07	.1
3	C	35	12	16.84	-1.09	.50	-2.18	.1
3	C	54	12	16.84	-1.09	.50	-2.18	.1
3	C	70	15	20.03	-1.10	.47	-2.36	1.1
3	C	85	12	17.01	-1.14	.50	-2.29	.1
3	C	53	12	17.89	-1.32	.50	-2.64	.1
3	C	38	12	18.15	-1.37	.50	-2.75	.1
3	C	26	12	18.68	-1.49	.50	-2.99	.1
3	C	66	12	18.68	-1.49	.50	-2.99	.1
3	C	57	12	18.95	-1.55	.50	-3.11	.1
3	C	24	12	20.85	-1.98	.50	-3.97	.1
3	C	25	12	21.42	-2.11	.50	-4.23	.1

Table 7

Bias of Rater D across test takers

rater	label	Test takers	Observed score	Expected score	Bias size	error	Z score	fit
4	D	26	30	20.59	4.00	1.57	2.55	.1
4	D	28	30	20.59	4.00	1.57	2.55	.1
4	D	25	28	23.17	1.69	.76	2.22	1.9
4	D	110	24	17.03	1.63	.52	3.13	.1
4	D	27	37	21.61	1.58	.65	2.44	1.6
4	D	56	23	18.04	1.15	.50	2028	.2
4	D	92	18	22.13	-.94	.47	-2.01	.1
4	D	36	16	20.59	-1.01	.47	-2.18	.2
4	D	30	20	24.21	-1.03	.48	-2.17	.5
4	D	72	23	26.85	-1.18	.50	-2.35	.1
4	D	69	13	18.80	-1.28	.48	-2.65	.1
4	D	32	18	24.21	-1.48	.47	-3.16	.1
4	D	31	18	24.47	-1.55	.47	-3.32	.1
4	D	64	17	23.69	-1.55	.47	-3.34	.1
4	D	81	18	24.74	-1.63	.47	-3.48	.3
4	D	70	14	21.87	-1.75	.47	-3.71	.6
4	D	71	12	22.39	-2.35	.50	-4.70	.6

Raters across items

Table 8 presents the bias calibration report for the raters and items in the WDCT. All rows are organized by the z-score. There were 24 interactions analyzed (4×6= 24). The table lists rater (column 1), items (column 2), the total observed and the total expected score (columns 3 and 4), the bias size in logit (column 5), the estimate of the error (column 6), z-scores (column 7), and fitness (column 8). Following the standard convention of using a z-score of ±2.0 to identify significant bias, only one interaction exhibited a significant bias. This involved the first rater and the second item.

Table 8

Raters' bias over items in WDCT

rater	items	Observed score	Expected score	Bias size	error	Z score	fit
1	2	437	415.36	.33	.13	2.61	.7
3	3	326	312.22	.17	.11	1.52	1.0
2	6	420	408.51	.16	.12	1.36	1.1
1	1	432	422.43	.14	.12	1.16	.8
2	3	405	398.23	.09	12	.79	1.3
3	4	338	332..43	.07	.11	.62	.8
1	4	413	409.03	.06	.12	.47	.8
3	5	335	331.37	0.4	.11	.40	.7
2	5	418	415.18	.04	.12	.34	1.2
2	2	408	405.84	.03	.12	.25	.8
4	4	368	366.79	.02	.11	.14	1.0
4	6	359	358.35	.01	.11	.07	.9
1	4	385	385.59	-.01	.11	-.07	.8
3	6	323	323.73	-.01	.11	-.08	.8

rater	items	Observed score	Expected score	Bias size	error	Z score	fit
4	5	365	365.76	-.01	.11	-.09	.8
1	5	375	380.59	-.07	.11	-.64	1.1
3	1	374	380.67	-.09	.11	-.76	1.1
2	4	410	416.09	-.09	.12	-.73	1.2
4	3	340	347.07	-.09	.11	-.79	1.3
2	1	446	452.64	-.11	.13	-.86	1.0
2	2	442	449.97	-.13	.13	-1.02	.9
1	6	362	373.35	-.14	.11	-1.29	1.0
1	3	345	358.45	-.17	.11	-1.52	1.5
3	2	358	373.61	-.20	.11	-1.77	1.0

Test takers as raters across items in DSAT

This part examines the interaction of the test takers' leniency with item difficulty to show how harshly or leniently items are scored in the DSAT. Table 9 lists the significant bias interactions between raters and items in DSAT. Columns 1and 2 show test-takers' ID and the items. Columns 3 and 4 present the total observed score and the total expected score, respectively. Column 5 displays the bias logit, representing the magnitude of the difference between the observed and expected scores. Column 6 contains the standard error of the estimate of bias. Column 7 contains the z-scores calculated from the estimates of bias in column 6. A negative z-score of less than -2.0 would indicate that this rater has rated this candidate more leniently than his/her pattern of rating other candidates. Similarly, a positive z-score of greater than +2.0 would indicate that this rater has rated this candidate more harshly than other candidates.

Table 9

Student bias over items

Test taker	items	Observed score	Expected score	Bias size	error	Z score	fit
106	6	4	2.19	3.78	1.42	2.67	.0
17	4	4	2.54	3.07	1.42	2.17	1.0
87	6	2	3.35	-2.86	1.42	-2.01	.0
11	2	2	3.40	-2.97	1.42	-2.09	.0
69	6	2	3.52	-3.21	1.42	-2.25	.0
6	4	3	4.50	-3.21	1.50	-2.14	.0
24	4	3	4.50	-3.21	1.50	-2.14	.0
27	4	3	4.50	-3.21	1.50	-2.14	.0
29	4	3	4.50	-3.21	1.50	-2.14	.0
25	4	3	4.50	-3.21	1.50	-2.14	.0
89	2	3	4.55	-3.34	1.50	-2.23	.0
20	6	2	3.69	-3.55	1.42	-2.50	.0
13	4	1	3.19	-3.77	1.71	-2.20	.6
13	1	2	3.87	-3.90	1.42	-2.74	.0
99	1	3	4.74	-4.04	1.50	-2.69	.0
7	4	2	4.07	-4.30	1.42	-3.03	.0
78	4	1	3.35	-4.46	1.89	-2.37	.4

The interaction of test takers and the test method

Table 10 presents the results of the bias analysis for the interaction between test methods and test-takers. There were 220 total interactions from the two methods and the test takers. Here again, the table lists the significant biases. The first and second columns indicate the test type and the number of test-takers, respectively. Columns three and four present the total expected scores and the total observed scores, respectively, summed across the test takers (N = 110) and the six items for each test. The average difference between the total observed and expected scores (column 5) is calculated by dividing the difference by 660 (110 test takers × 6 items). In addition, column 6 presents the bias logit, which indicates the size of the residuals, and column 7 presents the standard error of the estimate. Column 8 contains the z-scores. Following the usual conventions concerning z-scores, only those that are outside the range of -2.00 to +2.00 are statistically significant. Column 9 presents the infit mean square, which indicates the consistency of the observed bias pattern. As shown in the table, there were 22 significant biases, of which 11 are related to DSAT and 11 to the WDCT. In most interactions, test takers appeared to find the DSAT easier than the WDCT; however, in some cases, test takers' scores on the DSAT were lower than on the WDCT, suggesting that they underestimate their own ability and score themselves harshly. But in general, most test takers overestimated their ability and assigned themselves higher scores in DSAT.

Table 10

Interaction of test method and test takers

method	Test takers	Observed score	Expected score	Bias size	error	Z score	fit
2	28	30	21.31	4.00	1.49	2.68	.1
1	71	30	23.74	3.64	1.72	2.11	.1
2	29	28	23.27	1.84	.79	2.34	1.9
2	23	26	20.82	1.56	.62	2.52	1.8
2	88	25	19.34	1.56	.58	5.69	1.4
2	22	27	22.78	1.45	.68	2.14	1.5
1	96	24	18.66	1.39	.55	2.52	.1
1	31	26	21.70	1.33	.62	2.15	1.1
1	109	23	18.15	1.21	.53	2.27	.6
2	38	24	19.84	1.12	.55	2.02	1.3
1	22	18	22.21	-1.02	.47	-2.19	2.1
2	109	14	18.85	-1.03	.46	-2.25	.5
2	31	18	22.29	-1.05	.47	-2.23	.7
1	23	15	20.18	-1.13	.45	-2.50	.5
2	96	14	19.34	-1.14	.46	-2.50	.0
1	29	18	22.72	-1.17	.47	-2.49	.1
1	88	13	18.66	-1.20	.47	-2.57	.5
1	19	18	23.74	-1.46	.47	-3.12	.1
1	47	18	23.74	-1.46	.47	-3.12	1.1
2	70	19	24.74	-1.55	.48	-3.24	.3
2	71	18	24.25	-1.62	.47	-3.46	.4
1	28	12	20.69	-1.90	.48	-3.96	.2

Discussion

The current study aimed to provide a comprehensive understanding of how these factors affect pragmatic test knowledge by applying G Theory and MFRM. Regarding the first research question, the analysis of variance components yielded unexpected results. In contrast to our expectations, the magnitude of variance attributable to interactions is higher than that of persons (test takers). Typically, the person facet is expected to account for the largest portion of the variance in test scores, as seen in Brown and Ahn's (2010) study on variance components in pragmatic assessment. They reported that the person was the greatest contributor. The present study, however, does not fit this pattern. Although the person facet (test taker ability) was a substantial source of variance in both tests, accounting for 17.1% of the score variance in the WDCT and 40.5% in the DSAT, in the WDCT, the interaction between persons and raters (P×R: 31.7%) and the three-way interaction between persons, raters, and items (P×R×I: 22.8%) were the largest contributors to variance. This suggests that scoring was highly dependent on which rater evaluated which test taker on which item. Furthermore, the person-item interaction (P×I) was also a significant contributor in both tests (WDCT: 18.1%; DSAT: 51.9%). The prominence of these interaction effects, particularly P×I in the self-assessed DSAT, underscores the complex nature of individual differences in test performance. It suggests that a test-taker's performance is not just a function of their overall ability but is also significantly influenced by their specific interaction with particular test items. This highlights the need for careful item design and calibration to ensure they function consistently across different learners.

Although this was an unexpected pattern in the variance components, the G-coefficients demonstrated acceptable reliability for both DSAT and WDCT in measuring the pragmatic knowledge of EFL learners. This finding of the current study aligns with previous research by Bachman & Palmer (1982), Budeng & Merza (2023), Hudson et al. (1992), Lui (2004), Saleem et al. (2022), and Yamashita (1997). However, some studies showed concerns about the reliability of DCTs and highlighted the need for further exploration of factors influencing DCT reliability (e.g., Brown, 2001; Enochs & Yoshitake-Strain, 1992)

In addition, the findings showed that the WDCT is more reliable than the DSAT. This finding is not in line with those of Brown & Ahn (2011) and Lui (2004). They found that self-assessment tests are more dependable than WDCTs or other DCT types. Furthermore, Brown & Ahn (2011) reported higher validity for self-assessment tests compared with other DCTs, whereas this study suggests that the WDCT (not a self-assessment test) is more reliable than the DSAT.

Contextual variations, differences in participant characteristics, task design, and testing conditions may account for this divergent finding. This shows that the factors that affect pragmatic test effectiveness may vary across contexts and participants. These discrepancies underscore the importance of considering the effects of contextual, cultural, and personal factors on the effectiveness of different DCT tools.

Regarding the second research question in the current study, the single-facet analysis showed no significant test-method effect. This finding contributes to a complex and somewhat divided body of literature. However, it contrasts with other studies that, using similar traditional methods, did find significant differences between test methods (e.g., Bardovi-Harlig & Hartford, 1993; Rose, 1994; Rose & Ono, 1995).

However, the current study is in line with other research studies that utilized traditional statistics and did not identify significant method effects (e.g., Billmyer & Varghese, 2000; Enochs & Yoshitake-Strain, 1992; Rose & Neg, 2001; Rose, 1995). However, the result contrasts with some studies that used traditional statistical methods and reported significant method effects (e.g., Bardovi & Hartford, 1993; Bardovi-Harlig, 2023; Mohammad Hosseinpur et al, 2021; Rose, 1994; Rose & Ono, 1995). This suggests that the choice of statistical method may influence the detection of test-method differences in DCT research.

The G-study results provide a more nuanced understanding that helps contextualize these mixed prior findings. While the test method facet itself (M) made a negligible contribution to the total variance (0.0%; Table 2), its interactions with other facets, particularly the person-method-item interaction (PI: M), were substantial. This suggests that the *choice of method alone* may not systematically alter scores for all test-takers on all items, which is why some studies find no main effect. However, the significant interaction effects indicate that the method can influence scores in complex, unpredictable ways depending on the specific person and item involved. This interaction effect could be a source of the inconsistency in the literature, as its impact might be masked or revealed differently depending on the specific sample of test-takers and items used in a given study, as well as the statistical power of the analysis.

The bias analysis examined interactions among all variables (test method, test taker, item difficulty) and revealed potential rater biases. The result indicated that there was no consistent pattern of rater bias across test takers. Similarly, Youn's (2007) findings revealed that raters' tendency to score did not systematically favor or disfavor some test takers regardless of their ability level. By contrast, Liu (2014) found a different pattern of systematic bias, in which raters granted more lenient scores to lower-scoring test takers. This study again found a different item bias pattern: in the majority of instances, test takers scored themselves more leniently on the DSAT compared with the WDCT. Yet, in some instances, test takers' DSAT scores were lower than their WDCT scores, indicating that they underestimated their own abilities and rated themselves more harshly. Overall, the majority of test takers overestimated their abilities, assigning themselves higher scores on the DSAT. This highlights the importance of item difficulty calibration to mitigate bias in self-assessment tests. In addition, regarding inconsistencies between raters across and within test methods, in line with previous studies (Kang et al., 2019; Neiriz, 2023; Taguchi & Li, 2020), this study suggests rater training to address these issues and enhance test reliability.

Conclusion

This study aimed to detect the effects of test types on students' performance and to determine the variability within each test method, including items, raters, and test takers. The second aim of this study was to detect the potential bias across different facets. The EDUG software analysis indicated that the test type, that is, WDCT versus DSAT, accounted for very little variance in the total error variance. That would thus suggest that both tests may be equally effective and valid for assessing EFL learners' pragmatic knowledge. This undermines the necessarily implicit assumption that differently formatted DCTs yield different results, underscoring the importance of considering the broader context in which these tests are applied. However, large contributions to the error variance came from items and their interactions with test takers, indicating that item quality and its fit with test-taker characteristics warrant further investigation to achieve an accurate assessment. Also, the facet analysis revealed a bias pattern regarding item difficulty.

This study contributes to the field of pragmatic assessment by challenging conventional assumptions about test-method effects and rater behavior. This study suggests that factors such as items and the interaction between test-takers and test items play a more significant role than previously understood. Future research should focus on exploring these factors in greater detail. Addressing these issues can improve the reliability and validity of pragmatic assessments and yield more accurate test results.

This study has implications for teachers and researchers. The significant rater bias identified suggests that rater training should be enhanced with psychometric tools such as the Many-Facet Rasch Model to provide raters with personalized feedback on their unique severity/leniency patterns, thereby mitigating unwanted score variance. Furthermore, the substantial person-item interaction variance underscores the critical need for rigorous piloting and psychometric calibration of test items to ensure they function consistently across learners, a step crucial to improving test reliability. The tendency of learners to overestimate their ability on the self-assessment (DSAT) suggests it may be better used as a metacognitive classroom tool to raise pragmatic awareness rather than for high-stakes assessment. Methodologically, this study demonstrates the superior utility of Generalizability Theory and MFRM over traditional statistics for providing a nuanced understanding of the multifaceted sources of measurement error in pragmatic assessment. Future research should focus on identifying the features of items that cause erratic performance, implementing and evaluating the efficacy of refined rater training protocols, and applying this robust analytical framework to other pragmatic test formats.

This study has three main limitations. First, the study might have been limited by the lack of item-difficulty analysis. Second, the small number of raters and the use of only two test methods might have constrained the generalizability of the findings. A larger sample of raters and different test methods could provide a more comprehensive understanding of the issue. A further limitation concerns the profile of the raters. While the four raters had substantial

experience (a minimum of 6 years) in English language teaching, they did not receive specific training in assessing pragmatic competence prior to this study. Their expertise was rooted in general language teaching at the high school level, which may not have fully equipped them to evaluate the nuanced aspects of pragmatic appropriateness, such as politeness strategies, situational context, and cultural conventions. This lack of specialized rater training could have introduced an additional, unquantified source of variance into the WDCT scores, potentially affecting the severity/leniency patterns and the consistency of their application of the scoring rubric. Consequently, the significant contribution of rater facets and their interactions (e.g., the 31.7% variance from the person-rater interaction in the WDCT) may be partly attributable to this inexperience in pragmatic assessment. This underscores the importance of dedicated rater training in pragmatic research to enhance scoring reliability and reduce measurement error. Future studies would benefit from employing raters with specific training in pragmatics assessment to isolate the effects of test and item design more precisely.

Declarations

Ethics approval and consent to participate: Prior to participation, informed consent was obtained from all individuals involved in the study. All procedures involving human participants were conducted in accordance with ethical standards. These procedures adhered to the principles outlined in the 1964 Declaration of Helsinki and its subsequent amendments.
Informed Consent: Informed consent was obtained from all individuals involved in the study.
Clinical trial number: Not applicable
Ethics Approval Committee: ethical standards were approved by Vali-e-Asr University's research committee.

Acknowledgments

We would like to thank the editorial team of TESL Quarterly for granting us the opportunity to submit and publish the current synthesis. We would also like to express our appreciation to the anonymous reviewers for their careful, detailed reading of our manuscript and their many insightful comments and suggestions. We also acknowledge all the participants who took part in this study.

Declaration of conflicting interests

The authors declare no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for this article's research, authorship, and/or publication.

References

Ahn, R. C. (2005). F*ive measures of interlanguage pragmatics in KFL* (Korean as a foreign language) learners. Unpublished PhD thesis, University of Hawaii at Manoa. https://www.proquest.com/openview/b77e6b2a157cc7f064eef80369123ad8/1?pq-origsite=gscholar&cbl=18750&diss=y

Akhavan Masoumi, G., & Sadeghi, K. (2020). Impact of test format on vocabulary test performance of EFL learners: the role of gender. *Language Testing in Asia*, *10*(1), 2.

Alemi, M., & Rezanejad, A. (2014). Native and non-native English teachers' rating criteria and variation in the assessment of L2 pragmatic production: The speech act of compliment. *Issues in Language Teaching*, *3*(1), 88-65. https://ilt.atu.ac.ir/article_1374.html

Anthony, C. J., Styck, K. M., Volpe, R. J., & Robert, C. R. (2023). Using many-facet Rasch measurement and generalizability theory to explore rater effects for direct behavior rating–multi-item scales. *School Psychology*, *38*(2), 119 –128.. https://doi.org/10.1037/spq0000518

Azizi, Z., & Namaziandost, E. (2023). Implementing Peer-Dynamic Assessment to Cultivate Iranian EFL Learners' Interlanguage Pragmatic Competence: A Mixed-Methods Approach. *International Journal of Language Testing*, *13*(1), 18-43.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449-465. **https://doi.org/10.2307/3586464**

Bardovi-Harlig, K., & Hartford, B. S. (1993). Learning the rules of academic talk: A longitudinal study of pragmatic change. *Studies in Second Language Acquisition*, *15*(3), 279-304. https://doi.org/10.1017/S0272263100012122

Bardovi-Harlig, K., & Su, Y. (2023). Developing an empirically-driven aural multiple-choice DCT for conventional expressions in L2 pragmatics. *Applied Pragmatics*, *5*(1), 1-40. https://doi.org/10.1075/ap.20020.bar

Beltran,j. (2019). A Meaning-Based Multiple-Choice test of pragmatic knowledge: Does It Work?. *Studies in Applied Linguistics and TESOL* 19(1):42-71. https://doi.org/10.7916/salt.v19i1.1407

Billmyer, K., & Varghese, M. (2000). Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Applied Linguistics*, *21*(4), 517-552. https://doi.org/10.1093/applin/21.4.517

Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (Eds.), *Pragmatics in Language Teaching* (pp.301-325). New York: Cambridge University Press.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1–15. https://doi.org/10.1177/026553229501200101

Brown, J. D. (2008). Raters, functions, item types, and the dependability of L2 pragmatics tests. *Investigating pragmatics in foreign language learning, teaching and testing*, *30*, 224-48.

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, *43*(1), 198-217. https://doi.org/10.1016/j.pragma.2010.07.026.

Budeng, R. B., & Merza, H. N. M. (2023). Assessing Interlanguage Pragmatic Competence on Speech Acts in a Filipino ESL Context. *Corpus Pragmatics*, *7*(2), 85-102. https://doi.org/10.1007/s41701-023-00137-y

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language testing*, *32*(3), 385-405. https://doi.org/10.1177/0265532214565386

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. Routledge.

Chen, Y. S., & Liu, J. (2016). Constructing a scale to assess L2 written speech act performance: WDCT and e-mail tasks. *Language Assessment Quarterly*, *13*(3), 231-250. https://doi.org/10.1080/15434303.2016.1213844

Cordier, R., Munro, N., Wilkes-Gillan, S., Speyer, R., Parsons, L., & Joosten, A. (2019). Applying Item Response Theory (IRT) modeling to an observational measure of childhood pragmatics: The pragmatics observational measure-2. *Frontiers in Psychology*, *10*, 408. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00408/full

Cohen, A. D. (2020). Considerations in assessing pragmatic appropriateness in spoken language. *Language Teaching*, *53*(2), 183-202.

Derakhshan, A., Shakki, F., & Sarani, M. A. (2020). The effect of dynamic and non-dynamic assessment on the comprehension of Iranian intermediate EFL learners' speech acts of apology and request. *Language Related Research*, *11*(4), 605-637. https://lrr.modares.ac.ir/article-14-40648-en.html.

Engelhard Jr, G., & Wind, S. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Enochs, K., & Yoshitake-Strain, S. (1999). Evaluating six measures of EFL learners' pragmatic competence. *JALT Journal*, *21*(1), 29-50. https://files.eric.ed.gov/fulltext/ED451718.pdf#page=32.

Farashaiyan, A., Sahragard, R., Muthusamy, P., & Muniandy, R. (2020). Questionnaire development and validation of interlanguage pragmatic instructional approaches & techniques in EFL contexts. *International Journal of Higher Education*, *9*(2), 330-342. https://eric.ed.gov/?id=EJ1255710.

Fussman, S., & Mashal, N. (2022). Initial validation for the Assessment of Pragmatic Abilities and Cognitive Substrates (APACS) Hebrew battery in adolescents and young adults with typical development. *Frontiers in Communication*, *6*, 758384. https://doi.org/10.3389/fcomm.2021.758384

Dabbagh, A., & Babaii, E. (2021). L1 pragmatic cultural schema and pragmatic assessment: Variations in non-native teachers' scoring criteria. *TESL-EJ*, *25*(1), 1-17. https://eric.ed.gov/?id=EJ1302438.

Grabowski, K. (2008). Measuring pragmatic knowledge: Issues of construct underrepresentation or labeling. *Language Assessment Quarterly*, *5*, 154-159. https://doi.org/10.1080/15434300801934736

Gordon, R. A., Peng, F., Curby, T. W., & Zinsser, K. M. (2021). An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly*, *55*, 149-164. https://doi.org/10.1016/j.ecresq.2020.11.005

Han, C. (2021). Detecting and measuring rater effects in interpreting assessment: A methodological comparison of classical test theory, generalizability theory, and many-facet rasch measurement. *Testing and Assessment of Interpreting: Recent Developments in China*, 85-113.

Hernández, T. A. (2018). L2 Spanish apologies development during short-term study abroad. *Studies in Second Language Learning and Teaching*, *8*(3), 599-620. https://www.ceeol.com/search/article-detail?id=690173

Hernández, T. A., & Boero, P. (2018). Explicit intervention for Spanish pragmatic development during short-term study abroad: An examination of learner request production and cognition. *Foreign Language Annals*, *51*(2), 389-410. https://doi.org/10.1111/flan.12334

Hernández, T. A. (2021). Explicit instruction for the development of L2 Spanish pragmatic ability during study abroad. *System*, *96*, 102395. https://doi.org/10.1016/j.system.2020.102395

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics* (Vol. 2). Natl Foreign Lg Resource Ctr.

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*, 479-493.

Karami, H. (2012). The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test. Psychological Test and Assessment Modeling, 54(3), 211. https://ptam-journal.com/wp-content/uploads/2025/01/04_Ravand_.pdf

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481-504. https://doi.org/10.1177/0265532219849522

Kecskes, I. (2014). *Intercultural pragmatics* (Vol. 288). Oxford: Oxford University Press.

Khodi, A. (2021). The affectability of writing assessment scores: a G-theory analysis of rater, task, and scoring method contribution. *Language Testing in Asia*, *11*(1), 30. https://doi.org/10.1186/s40468-021-00134-5

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85-S89.

Lozano-Ruiz, A., Fasfous, A. F., Ibanez-Casas, I., Cruz-Quintana, F., Perez-Garcia, M., & Pérez-Marfil, M. N. (2021). Cultural bias in intelligence assessment using a culture-free test in Moroccan children. *Archives of Clinical Neuropsychology*, *36*(8), 1502-1510.

Li, G., Pan, Y., & Wang, W. (2021). Using generalizability theory and many-facet Rasch model to evaluate in-basket tests for managerial positions. *Frontiers in Psychology*, *12*, 660553. https://doi.org/10.3389/fpsyg.2021.660553

Li, S., Li, X., Feng, Y., & Wen, T. (2023). Non-expert raters' scoring behavior and cognition in assessing pragmatic production in L2 Chinese. In *Crossing Boundaries in Researching, Understanding, and Improving Language Education: Essays in Honor of G. Richard Tucker* (pp. 79-102). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-24078-2_4.

Li, S., Taguchi, N., & Xiao, F. (2019). Variations in rating scale functioning in assessing pragmatic performance in L2 Chinese. *Language Assessment Quarterly*, *16*(3), 271–293. https://doi.org/10.1080/15434303.2019.1648473.

Li, S., Wen, T., Li, X., Feng, Y., & Lin, C. (2023). Comparing holistic and analytic marking methods in assessing speech act production in L2 Chinese. *Language Testing*, *40*(2), 249-275. https://doi.org/10.1177/026553222211139.

Liu, J. (2007). Comparing native and non-native speakers' scoring in an interlanguage pragmatics test. *Modern Foreign Languages*, *30*(4), 395-404.

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing*, *4*(1), 50-65. https://www.ijlt.ir/article_114393.html.

Liu, J. (2004). *Measuring interlanguage pragmatic knowledge of Chinese EFL learners* (Doctoral dissertation, City University of Hong Kong). https://www.peterlang.com/document/1100119.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180. https://doi.org/10.1177/026553229801500202

Mohammad Hosseinpur, R., Bagheri Nevisi, R., & Lowni, A. (2021). A tale of four measures of pragmatic knowledge in an EFL institutional context. *Pragmatics*, *31*(1), 114-143. https://doi.org/10.1075/prag.18052.moh

Namaziandost, E., Nasri, M., Rahimi Esfahani, F., Neisi, L., & Ahmadpour KarimAbadi, F. (2020). A cultural comparison of Persian and English short stories regarding the use of emotive words: implications for teaching English to Iranian young learners. *Asian-Pacific Journal of Second and Foreign Language Education*, *5*(1), 7. https://doi.org/ 10.1186/s40862-020-00085-z

Neiriz, R. (2023). Developing and evaluating a contextualized interactional competence rating scale based on a metaphorical conceptualization: A pragmatic mixed-method approach. *Journal of Second Language Studies*, *6*(1), 61-94. https://doi.org/10.1075/jsls.22003.nei

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). The problem of bias in psychological assessment. In *Mastering modern psychological testing: Theory and methods* (pp. 573-613). Cham: Springer International Publishing.

Richard, P. J., Devinney, T. M., Yip, G. S., & Johnson, G. (2009). Measuring organizational performance: Towards methodological best practice. *Journal of Management*, *35*(3), 718-804. https://doi.org/10.1177/0149206308330560

Rose, K. R. (1992). Speech acts and questionnaires: The effect of hearer response. *Journal of Pragmatics*, *17*(1), 49-62. https://doi.org/10.1016/0378-2166(92)90028-A

Rose, K. R. (1994). On the Validity of Discourse Completion Tests in Non-Western Contexts. *Applied Linguistics*, *15*(1), 1-14. https://doi.org/10.1093/applin/15.1.1

Rose, K. R., & Ng, C. (2001). Inductive and deductive teaching of compliments and compliment responses. *Pragmatics in Language Teaching*, *145*(1), 145-170. https://www.researchgate.net/publication/265288342_Inductive_and_deductive_approaches_to_teaching_compliments_and_compliment_responses.

Rose, K. R., & Ono, R. (1995). Eliciting speech act data in Japanese: The effect of questionnaire type. *Language Learning*, *45*(2), 191-223. https://doi.org/10.1111/j.1467-1770.1995.tb00438.x

Roever, C. (2008). Rater, item, and candidate effects in discourse completion tests: A FACETS approach. In E.A. Soler and A.M. Flor (eds.) *Investigating pragmatics in foreign language learning, teaching and testing (pp. 249–266). Clevedon, UK: Multilingual Matters.* https://books.google.nl/books

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing*, *28*(4), 463-481. https://doi.org/10.1177/0265532210394633

Roever, C. (2013). Testing implicature under operational conditions. In *Assessing second language pragmatics* (pp. 43-64). London: Palgrave Macmillan UK. https://link.springer.com/chapter/10.1057/9781137003522_2

Rossi, O. & Tineke, B. (2020). Raters of Subjectively-Scored Tests. English Language Teaching, 1-7. https://doi.org/10.1002/9781118784235.eelt0985

Saleem, A., Saleem, T., & Aziz, A. (2022). A pragmatic study of congratulation strategies of Pakistani ESL learners and British English speakers. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 8. https://doi.org/10.1186/s40862-022-00134-9

Shahi, R., Ravand, H. & Rohani, G. R. (2025). Examining the Effect of Item Difficulty and Rater Leniency on Iranian Test Takers' Performance on WDCT and DSAT: A Comparative Study. *International Journal of Language Testing*, *15*(1), 1-19. doi: 10.22034/ijlt.2024.454478.1341

Sonnenburg-Winkler, S. L., Eslami, Z. R., & Derakhshan, A. (2020). Rater variation in pragmatic assessment: The impact of the linguistic background on peer-assessment and self-assessment. *Lodz Papers in Pragmatics*, *16*(1), 67-85. https://doi.org/10.1515/lpp-2020-0004

Steyer, R. (2001). Classical (psychometric) test theory. *International Encyclopedia of the Social & Behavioral Sciences. 1955-1962.* https://doi.org/ 10.1016. B0-08-043076-7/00721-X.

Sitorus, T. A. P., Siregar, D. Y., Aulia, D. N., Zahra, N. A., Parinduri, A. I., Lubis, D. N. A., & Wardiah, F. D. (2025). A Systematic Review of Pragmatic Competence in Second Language Acquisition. *Sintaksis: Publikasi Para ahli Bahasa dan Sastra Inggris*, *3*(1), 142-152. https://doi.org/10.61132/sintaksis.v3i1.1291

Su, Y., & Shin, S. Y. (2024). Comparing two formats of data-driven rating scales for classroom assessment of pragmatic performance with role-plays. *Language Testing*, *41*(2), 357-383. https://doi.org/10.1177/02655322231210217

Sydorenko, T., Maynard, C., & Guntly, E. (2014). Rater behavior when judging language learners' pragmatic appropriateness in extended discourse. *TESL Canada Journal, 32*(1), 19–41. https://doi.org/doi:10.18806/tesl.v32i1.1197

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *21*(3), 453-471. https://doi.org/10.1075/prag.21.3.08tag

Taguchi, N., & Li, S. (2020). Contrastive pragmatics and second language (L2) pragmatics: Approaches to assessing L2 speech act production. *Contrastive Pragmatics*, *2*(1), 1-23. https://brill.com/view/journals/jocp/2/1/article-p1_1.xml

Tajeddin, Z., & Alemi, M. (2014). Pragmatic rater training: Does It affect non-native L2 teachers' rating accuracy and bias?. *International Journal of Language Testing*, *4*(1), 66-83. https://www.ijlt.ir/article_114394.html

Tajeddin, Z., Alemi, M., & Khanlarzadeh, N. (2020). Rating Criteria and Norms for Pragmatic Assessment in the Context of EIL. In *Pragmatics Pedagogy in English as an International Language* (pp. 212-231). Routledge.

Timpe-Laughlin, V., & Choi, I. (2017). Exploring the validity of a second language intercultural pragmatics assessment tool. *Language Assessment Quarterly*, *14*(1), 19-35. https://doi.org/10.1080/15434303.2016.1256406

Toe, D., Mood, D., Most, T., Walker, E., & Tucci, S. (2020). The assessment of pragmatic skills in young deaf and hard-of-hearing children. *Pediatrics*, *146*(Supplement_3), S284-S291.

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, *24*(2), 155-183. https://doi.org/10.1177/0265532207076362

Wolcott, M. D., Olsen, A. A., & Augustine, J. M. (2022). Item response theory in high-stakes pharmacy assessments. *Currents in Pharmacy Teaching and Learning*, *14*(9), 1206-1214.

Wilson, A. C., & Bishop, D. V. (2022). A novel online assessment of pragmatic and core language skills: An attempt to tease apart language domains in children. *Journal of Child Language*, *49*(1), 38-59

Xu, L., & Wannaruk, A. (2018). Reliability and validity of WDCT in testing interlanguage pragmatic competence for EFL learners. *Journal of Language Teaching and Research, 6*(6), 1206-1215. https://doi.org/10.17507/jltr.0606.07

Yang, H. (2022). Second language learners' competence of and beliefs about pragmatic comprehension: Insights from the Chinese EFL context. *Frontiers in Psychology*, *12*, 801315. https://doi.org/10.3389/fpsyg.2021.801315

Youn, S. J. (2007). Rater bias in assessing the pragmatics of KFL learners using facets analysis. S*econd Language Studies* 26(1): 85–163. http://hdl.handle.net/10125/40691

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199-225. https://doi.org/10.1177/026553221455711

Youn, S. J. (2020). Interactional features of L2 pragmatic interaction in role-play speaking assessment. *TESOL Quarterly*, *54*(1), 201-233. https://doi.org/10.1002/tesq.542

Youn, S. J., & Bi, N. Z. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. *Intercultural Pragmatics*, *16*(2), 185-218.

Youn, S. J., & Brown, J. D. (2013). Item difficulty and heritage language learner status in pragmatic tests for Korean as a foreign language. *Assessing second language pragmatics* (pp. 98-123). London: Palgrave Macmillan UK. https://link.springer.com/chapter/10.1057/9781137003522_4.

Yamashita, S.O. (1996). *Comparing six cross-cultural pragmatics measures*. Unpublished doctoral dissertation, Temple University, Philadelphia, PA. https://www.proquest.com/openview/a45390785a21b1a799ba10f4e346bced/1?pq-origsite=gscholar&cbl=18750&diss=y

Yamashita, S. O. (1997). Self-assessment and role play methods of measuring cross-cultural pragmatics. *Pragmatics and Language Learning*, *8*(1), 129-162. https://scholarspace.manoa.hawaii.edu/collections/abc81e47-c948-4d15-9284-783942d637cd

Zangoei, A., & Derakhshan, A. (2021). Measuring the predictability of Iranian EFL students' pragmatic listening comprehension with language proficiency, self-regulated learning in listening, and willingness to communicate. *Journal of Applied Linguistics and Applied Literature: Dynamics and Advances*, *9*(2), 79-104.

Zhai., X, Kevin, C., Haudek., Chris, H., Wilson., Molly, Stuhlsatz. (2021). A Framework of Construct-Irrelevant Variance for Contextualized Constructed Response Assessment. Frontiers in Education, 6 doi: 10.3389/FEDUC.2021.751283

Appendix A: *Pragmatic Tests*

DSAT

Name: ……

Below are six situations. Score yourself based on the scales that were provided after each item.

1 **You are trying to study in your room and you hear loud music coming from another student's room down the hall. You don't know the student, but you decide to ask them to turn the music down. What would you say?**

YOU:

How well do you think that you answer this question?

Very bad <----------> very good

1 2 3 4 5

2 **You missed class and need to borrow a friend's notes. What would you say?**

How well do you think that you answer this question?

Very bad <----------> very good

1 2 3 4 5

3 **You need a ride home from school. You notice someone who lives down the street from you is also at school, but you haven't spoken to this person before. You think they might have a car. What would you say?**

How well do you think that you answer this question?

Very bad <----------> very good

1 2 3 4 5

4 A student in the library is making too much noise and disturbing other students. A librarian decides to ask the student to quiet down. What will the librarian say?

LIBRARIAN:

How well do you think that you answer this question?

Very bad <----------> very good

1 2 3 4 5

5 Your term paper is due, but you haven't finished it yet. You want to ask your professor for an extension. What would you say?

YOU:

How well do you think that you answer this question?

Very bad <----------> very good

1 2 3 4 5

6 **A professor wants a student to present a paper in class a week earlier than scheduled. What would the professor say?**

PROFESSOR

How well do you think that you answer this question?

1 2 3 4 5

WDCT Questionnaire

Name: …….

Below are six situations. Read the description of each situation and write down either what you would say in that situation or what you think the person in the situation would say.

1 **You are trying to study in your room, and you hear loud music coming from another student's room down the hall. You don't know the student, but you decide to ask them to turn the music down. What would you say?**
YOU:

2 You missed class and need to borrow a friend's notes. What would you say?

3 You need a ride home from school. You notice someone who lives down the street from you is also at school, but you haven't spoken to this person before. You think they might have a car. What would you say?
YOU:

4 A student in the library is making too much noise and disturbing other students. A librarian decides to ask the student to quiet down. What will the librarian say?
LIBRARIAN:

5 Your term paper is due, but you haven't finished it yet. You want to ask your professor for an extension. What would you say?
YOU:

6. A professor wants a student to present a paper in class a week earlier than scheduled. What would the professor say?
PROFESSOR:

Appendix B: Scoring Rubric

Grade	Criteria
5 (Demonstrates excellence)	– Correct speech act is elicited. – Expressions and wording are completely appropriate. – The amount of information given is completely appropriate. – Levels of formality, directness, and politeness are completely appropriate.
4 (Demonstrates good command with only limited difficulties)	– Correct speech act is elicited. – Expressions and wording are mostly appropriate. – The amount of information given is appropriate. – Levels of formality, directness, and politeness are mostly appropriate.
3 (Demonstrates adequate command with some weakness)	– Correct speech act is elicited. – Expressions and wording are generally appropriate. – The amount of information given is generally appropriate. – Levels of formality, directness, and politeness are generally appropriate.
2 (Falls below expectations)	– Intended speech act is vaguely implied but may cause misunderstanding. – Expressions and wording are non-typical but still acceptable. – The amount of information given is inappropriately much or little but still acceptable. – Levels of formality, directness, and politeness are not very appropriate but still acceptable.
1 (Unacceptable)	– Incorrect speech act or no speech act is elicited. – Expressions and wording are not appropriate. – The amount of information given is either too much or too little. – Levels of formality, directness, and politeness are not