Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 15, No. 1; 2025, March

ISSN 2476-5880
https://www.ijlt.ir

Examining the Effect of Item Difficulty and Rater Leniency on Iranian Test Takers' Performance on WDCT and DSAT: A Comparative Study

Reza Shahi[1], Hamdollah Ravand *[2], Golam Reza Rohani[3]

ARTICLE INFO

Article History:
Received: April 2024
Accepted: May 2024

KEYWORDS
DSAT
Many Facet Rasch Model
Pragmatic test
WDCT

ABSTRACT

The current paper intends to exploit the Many Facet Rasch Model to investigate and compare the impact of situations (items) and raters on test takers' performance on the Written Discourse Completion Test (WDCT) and Discourse Self-Assessment Tests (DSAT). In this study, the participants were 110 English as a Foreign Language (EFL) students at Vali-e-Asr University of Rafsanjan in Iran. The students were asked to complete the WDCT and rate themselves on that test. Four raters scored the WDCT tests. According to the FACETS results, there were significant differences in students' performance between the two methods. The stable fit statistics and differing levels of difficulty measures for each test method indicated that each test had a unique way of differentiating the test taker's pragmatic ability. Based on the results, both DSAT and WDCT are acceptable measures for pragmatic ability; however, there are some fitness problems in DSAT. This shows the unpredictable pattern of ratings in the DSAT. It is recommended to have rater training to obtain more accurate results from the DSAT. Finally, the implications were discussed.

1. Introduction

Over the last few decades, pragmatics has turned into a field that attracts considerable attention in language teaching research (Mao & He, 2021; Purpura, 2017), and a large array of research studies on pragmatics assessment have significantly contributed to the advancement of researchers' understandings of pragmatics as an assessment construct (Roever ,2011; Youn, 2015). To date, researchers have developed and validated some tests for measuring different aspects of pragmatic competence. However, measures are prone to contamination by factors other than the construct being measured.

Construct-irrelevant factors impair validity by adding systematic error variation to scores. This shows that we can increase validity if we can lessen the influence of construct-irrelevant elements (Wise, 2019). To identify and minimize the impact of the construct irrelevant factors on pragmatics, researchers have started to examine the role of various factors involved in assessing second /foreign language pragmatics, such as the rater role (e.g., Alemi & Rezanejad, 2014; 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), impacts of test takers' characteristics on test functioning (Roever, 2013; Youn & Brown, 2013), development and validation of rating scale (e.g., Chen & Liu, 2016; Derakhshan et al., 2020; Li et al., 2019; Su & Shin, 2024; Youn, 2015), and the function of items (e.g., Cordier et al., 2019; Roever, 2008).

Despite this increasing attention, the research on assessing the impact of construct irrelevant factors is mostly limited to some factors, including raters and rating scale functions (e.g., Brown & Ahn, 2012; Li et al., 2023; Liu, 2006; Liu & Xie, 2014; Tajeddin & Alemi, 2014; Youn, 2007). In addition,

[1] PhD, Ilam University, Ilam, Iran, Email: Reza.Shahi411@gmail.com
[2] Associate professor, Vali-e-Asr University of Rafsanjan, Kerman, Iran, Email: Ravand@vru.ac.ir
[3] Assistant professor, Shahid Bahonar University of Kerman, Iran, Email: r.rohani@uk.ac.ir

Cite this paper as: 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. https://doi.org/10.22034/IJLT.2024.454478.1341

relatively few studies have tried to investigate the impact of rater leniency and item difficulty within and across different test methods in pragmatics assessment (e.g., Brown & Ahn, 2010; Youn, 2007). Most of the above-mentioned studies concerned professional raters and focused on native and trained non-native raters. However, trained raters are not always available to researchers or teachers who want to assess learners' pragmatic knowledge. Additionally, relatively few studies in the area of pragmatics have investigated the leniency or severity with which students assess themselves in DSATs (Brown & Ahn, 2012). Furthermore, to the best of the authors' knowledge, rater effects have not been studied in the Iranian EFL context.

Obviously, more research is still needed to assess the impact of construct irrelevant factors on test takers' performance on pragmatics tests. To contribute to the literature on pragmatic assessment and to provide more insights into studies related to raters' variability, the present study investigates the function of the test method and raters' leniency in the Iranian EFL context.

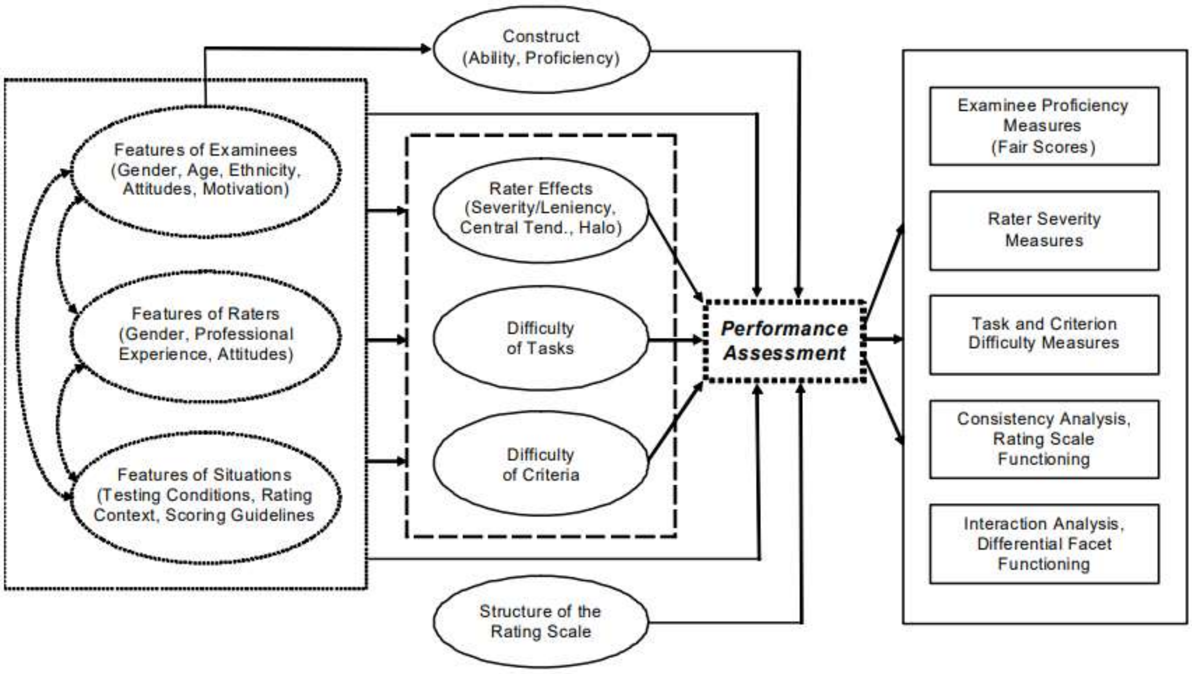
2. Background of the Study

2.1. Effective Factors on Testing

A valid test accurately measures what it claims to assess (Sartori & Pasini, 2007). In order to support the claim of validity, multiple sources of variation in test scores should be investigated (Grabowski, 2008). Test validity hinges on distinguishing between two key sets of factors: construct-relevant and construct-irrelevant (Sireci & Faulkner-Bond, 2015). Construct-relevant factors directly tap into pragmatic competence measurement (Messick, 2012). These include a learner's ability to use language appropriately in different social contexts (Llosa, 2020), demonstrate politeness strategies, and achieve communicative goals clearly and effectively (Schroeder, 2021). Construct-irrelevant factors, on the other hand, can influence scores without reflecting actual pragmatic knowledge (Sherkuziyeva et al., 2023).

Eckes (2009) proposed a framework of factors affecting scores, particularly in performance assessment (see Figure 1). According to this framework, construct-irrelevant factors that contribute to measurement error include the rater effect, variability in the difficulty of the test task, and variability in the scoring criteria, which he called proximal factors. Moreover, Eckes called the factors that indirectly affect performance assessment as distal factors. These variables include test-taker traits, characteristics of raters, and characteristics of situations and physical environments.

Figure 1
Proximal and Distal Factors


Construct
(Ability, Proficiency)
Features of Examinees
(Gender, Age, Ethnicity, Attitudes, Motivation)
Features of Raters
(Gender, Professional Experience, Attitudes)
Features of Situations
(Testing Conditions, Rating Context, Scoring Guidelines
Rater Effects
(Severity/Leniency, Central Tend., Halo)
Difficulty of Tasks
Difficulty of Criteria
Performance Assessment
Structure of the Rating Scale
Examinee Proficiency Measures
(Fair Scores)
Rater Severity Measures
Task and Criterion Difficulty Measures
Consistency Analysis, Rating Scale Functioning
Interaction Analysis, Differential Facet Functioning

Assessment of human behavior is associated with measurement error. Language test performance, of which performance on the pragmatics test is no exception, is in turn affected by construct relevant and irrelevant factors simultaneously. While this study acknowledges the importance of construct-relevant factors, it primarily focuses on construct-irrelevant factors, specifically rater leniency and test method.

2.2. Pragmatics Testing

2.2.1. Pragmatics Tests

Following the development and validation of six types of pragmatic tests by Hudson et al. (1992), different studies have developed and validated different tests in different contexts. The tests were subsequently translated into Japanese by Yamashita (1996), who confirmed that all but MDCT performed admirably for Japanese as a second language (SL). However, in order to evaluate the six measures of pragmatics, Enochs and Yoshitake-Strain (1999) administered the tests developed by Hudson et al. (1992) to 25 Japanese EFL learners. They found that the WDCT and multiple-choice discourse completion test (MDCT) were problematic in the Japanese EFL context. In a related study, Brown (2001a) compared different types of pragmatic tests in the Japanese as a Second Language (JSL) and EFL contexts. He used the data that was collected by Yoshitake (1997) and Hudson et al. (1995). The results showed that all test types but MDCT and WDCT were reliable in the JSL context. For all but the MDCT, Ahn (2005) produced Korean versions and tested their efficacy for Korean as a Foreign Language.

In line with the mentioned studies, some other studies have developed other tests and validated them in different contexts (e.g., Birjandi & Soleimani, 2013; Brown, 2001b; Farashaiyan et al., 2020; Grabowski, 2008; Liu, 2004; Matsugu, 2014; Nemati et al., 2014; Roever, 2012; Rose, 1994; Rose & Ono, 1995; Taguchi, 2011; Xu & Wannaruk, 2018; Yamashita,1997).

Most of the mentioned studies investigated MDCT and WDCT. Relatively few studies have been conducted to investigate the use of DSAT and the function of test takers as raters in this format of pragmatics test. Rose and Ng (2001) used a self-assessment questionnaire to investigate the effect of instruction on the use of compliments. They found no significant effect of instruction, while they found instruction effective through WDCT tests. Some other studies tried to investigate self-assessment reliability in different contexts. Bachman and Palmer (1982), applying multi-trait and multi-method, tried to validate a self-assessed test intended to measure grammatical, pragmatic, and sociolinguistic competence. The self-ratings of 116 non-native speakers were analyzed, and it was found that the assessment was reliable.

Liu (2004) developed a DSAT test that was administered to 200 Chinese university students. He found that the DSAT was one of the most reliable tests in the Chinese EFL context. Brown (2008), in his study, used role-play self-assessment, among other types of pragmatics tests such as WDCT, ODCT, and role-play. He found role-play self-assessment is a reliable test to measure a Korean as a Foreign Language (KFL) learner's pragmatic knowledge.

2.2.2. The Role of Construct Irrelevant Factors in Assessing Pragmatics

Considering the impact of constructed irrelevant factors, some research studies have been conducted to investigate the impact of raters on pragmatic test results. Most of the studies investigated the role of raters in assessing test takers' pragmatic knowledge. Liu (2014), using the many-facet Rasch model, found that the raters showed different amounts of leniency in rating. In his study of the rater effects on a WDCT pragmatics test, Liu (2014) measured Chinese EFL learners' interlanguage pragmatic knowledge by administering a DCT that required students to answer questions designed to elicit certain pragmatic functions. By facet analysis, he found that raters had a general tendency toward severity. Additionally, they reported significant differences between the raters regarding the rating severity. Tajeddin and Alemi (2014) investigated the impact of rater training on teachers' rating accuracy and bias. They found that training can improve non-native teachers' ratings, bringing them closer to native speaker judgments and increasing their reliability, although it may not necessarily eliminate all bias.

In another study, Taguchi (2011) investigated the native speaker rater's variability in assessing test takers' performance on pragmatics tests and found significant differences in rating the test takers' performances on the tests regarding the usage of pragmatics and social norms. Roever (2008) investigated the effect of rater, item, and candidate effects on WDCT. He found that the raters were similarly consistent in their ratings. Although the raters used a similar strategy for judging the test takers' performance on the tests, there was no significant bias in their rating.

Relatively few studies were conducted to investigate other construct irrelevant factors, including impacts of test takers' characteristics on test functioning (Roever, 2013; Youn & Brown, 2013), the development and validation of rating scales (e.g., Chen & Liu, 2016; Derakhshan, Shakki, & Sarani, 2020; Li et al., 2019; Youn, 2015), the function of items (e.g., Cordier et al., 2019; Roever, 2007), and test method role (Bardovi & Hartford, 1993; Rose, 1995; Youn, 2015).

2.3. This Study

Overall, three limitations were identified in the literature on pragmatic assessment. First, relatively few studies have investigated the impact of test methods and non-native raters on pragmatic assessments. Second, no study has investigated how linearly or severely students can be in assessing themselves on a DSAT. Third, few studies have explicitly utilized all of Linacre's (1999, 2002) guidelines as the basis for evaluating the function of the test method (non-native and untrained), which reflects leniency within and across WDCT and DSAT.

Therefore, in the present study, Multifaceted Rasch Measurement (MFRM) is used to investigate whether test methods exercise a systematic influence on pragmatic test results. Moreover, this study intends to investigate to what degree person ability, situation difficulty, and rater leniency are different within and across two test methods. To this end, the following research questions are posed:

1. To what extent are person, raters, and situation measures relatively higher or lower on WDCT and DSAT?
2. How well are the five-point scales functioning on the WDCT, and DSAT?

3. Method

3.1. Participants

In The current study was conducted with 110 students studying at Vali-e-Asr University of Rafsanjan. The participants ranged in age from 17 and 24 (M=20.2, SD=1.6), selected conveniently from Translation Studies and English literature students. They had studied English for approximately five years, and their English proficiency could be rated basically as intermediate. In addition, four raters participated in the present study. They 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 possessed more than six years of experience teaching English as a Foreign Language (EFL) in high school settings. While high school teaching experience may not directly translate to pragmatic test rating expertise, it demonstrates experience in evaluating student language skills and applying instructional practices relevant to language use.

3.2. Instrumentation

Describe In this study, WDCT and DSAT were employed (see Appendix A). The WDCT was the test that was used by Rose (1992). There were six situations of request portrayed in the WDCT to which the test takers were supposed to react. For example,

> 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? (Rose, 1992)
>
> YOU:………………

For the DSAT, test takers were asked to rate their own performance on each situation on a 5-point scale ranging from 1 (very bad) to 5 (very good). For example,

> 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? (Rose, 1992)
>
> YOU:…………………

How well do you think that you answer this question?
Very bad <--> very good
1 2 3 4 5

A scoring rubric used by Liu (2004) was employed to score the DSAT and WDCT tests. The scoring scale ranged from 1 (no evidence of considered component knowledge) to 5 (complete knowledge of that component). The scoring process involved raters evaluating test takers' responses based on several criteria, including (a) Ability to employ the correct speech act, (b) Levels of politeness, directness, and formality, (c) Amount of information given, and (d) Appropriate expressions and wording (see Appendix B).

3.3. Procedures

The study began with the administration of the WDCT. In the WDCT, participants were instructed to write short paragraphs in response to six different scenarios. Importantly, they were informed that there were no specific word limitations. The emphasis was on conveying the message clearly and concisely within a short paragraph format. To minimize the influence of time pressure on performance, participants were given ample time (about 2 hours) to complete the WDCT. Following each scenario, participants were asked to self-assess their responses based on provided criteria. After participants completed the self-assessment section and returned their WDCT booklets, the evaluation process began. Four independent raters, all Master's students in English Teaching with extensive experience (over six years) in EFL high schools, were enlisted to assess the participants' responses. To guarantee consistent scoring across raters, they received some instruction on the WDCT rubric prior to evaluation. There was no formal training. The instruction only focused on familiarizing them with the specific pragmatic aspects being assessed in the scenarios, the detailed scoring criteria outlined in the rubric, and applying the rubric consistently by evaluating sample responses. Finally, each rater independently scored all WDCT responses using the standardized rubric, ensuring reliable evaluation of participants' pragmatic competence.

3.4. Data Analysis

To analyze the data, the FACETS software program (Linacre, 2006) was used. Here, Facet was run three times to conduct the analyses for the present study: First, rater leniency, item difficulty, and person ability in WDCT were estimated. Next, student leniency and item difficulty in the DSAT were estimated, and in the final run of the FACETS, the effect of the test method was estimated.

4. Results

4.1. General Results

Table 1 shows the general results on the facets of the study: person's ability (referred to as person), rater's severity (referred to as rater), and situation difficulty (referred to as situation) in both WDCT and DSAT. Five statistics are provided for each facet in both tests: fit, error of measurement, separation, chi-square (fixed), and reliability.

In Rash analysis, all of the facets should meet the expectations of the model. Elements that do not meet the theoretical expectations are called misfits. Here, the first column of the table displays how many people, raters, and situations were not fit for the model or considered misfits. McNamra (1996) suggests that values exceeding the mean by plus or minus two standard deviations should be considered outliers. There were four misfitting examinees for the WDCT. This indicates that the responses of these individuals deviated from the expected pattern for the WDCT. Notably, there were no misfitting raters or situations identified for the WDCT, suggesting good consistency in scoring and assessment context. On the other hand, 6 test takers were found to be misfit for the DSAT. Moreover, there were 4 misfitting self-assessors and one misfitting situation for the DSAT. The presence of misfitting self-assessors suggests some individuals might have provided inaccurate self-evaluations, while the misfitting situation indicates one specific context might not have been well-captured by the DSAT.

It is important to acknowledge that the estimates are not without error. RMSE (Root of Mean Square Standard Error) is an index of measurement error. Here, it is shown by RMSE, which refers to all non-extreme measures. While there is not a single acceptable range for RMSE in Rasch analysis, generally, lower RMSE values indicate a better model fit. Values closer to zero suggest a stronger fit between the observed data and the model predictions. The RMSE values (.63 and.15, for person and situations, respectively) on DSAT are relatively higher than values on WDCT (.25 and.06, for person and situations, respectively). The RMSE for raters on WDCT is low enough to indicate the rater's fitness to the model. However, the high RMSE for student ratings on the DSAT shows that they may not align as well with the model's predictions as desirable. The separation index is a measure of how many distinct levels of persons, items, raters, and so on exist in the sample (Myford & Wolfe, 2004). Higher separation indices are desirable. A separation index of 1.5 or higher is generally considered acceptable for distinguishing between examinees' abilities (person separation). This indicates a good spread of person and item locations on the latent variable being measured. For example, a separation index of 3.98 and 13 for persons and raters, respectively, indicates that there are about 4 distinct levels of persons and about 13 distinct strata of raters. As Table 1 shows, all the separation indices for the WDCT, except for situation, are higher than those of the DSAT.

The reliability index is an indicator of how reliably the program software is able to differentiate between the elements of the facets. According to Linacre (1999), this indicates the differences in reproducibility of the measures. In addition, this shows how ‘‘good’’ the test is in other respects (Brown & Ahn, 2011). High (near 1.0) reliability is preferred. In this study, all measures seem to be reliable or consistent for both tests. However, the reliability index of measures in the WDCT is higher than the reliability index in the DSAT.

Fixed (all same) chi-square tests address the hypothesis that a set of elements shares the same measure after allowing for measurement error (Linacre, 1996). The fixed hypotheses tested in the present studies are:

All persons have the same level of ability
All raters have the same level of leniency
All items have the same level of difficulty

The analysis revealed that all of the chi-square (fixed) statistics were significant ($p < .01$) in this study, leading to the rejection of all null hypotheses for both the WDCT and DSAT. This shows that the elements being compared are statistically different in all cases.

Table 1
General Results of Facet Analysis

Test method	**Misfit**	**RMSE**	**Separation**	**Reliability**	**Chi-square**	**Probability**
WDCT						
Person	4	0.25	3.98	0.88	858.8	.00
Raters	0	0.05	13	99	270.0	.00
Situations	0	0.06	8	0.97	165.67	.00
DSAT						
person	6	0.63	2.36	.70	80.06	.00
Student Raters	4	0.63	2.36	.70	80.06	.00
situations	1	0.15	8.72	.95	102.5	.00

4.2. Facet Ruler for WDCT

The facet vertical map (Figure 2) visually represents the relationship between the levels of the analyzed factors. The first column represents the interval scale of all facets in Logit. So, this column can be used as a frame of reference to compare the facets. The second column displays Person ability and each of the asterisks (*) represents two students, and each dot (.) represents one student. Conventionally, the mean of the facets is set at 0 in Rasch-related analyses. So, values above 0 indicate higher personability, item difficulty, rater severity, and so on. A student with zero Logit ability would have a 50 percent chance to answer an item with an average level of difficulty correctly. The last column displays the severity of the categories on the scale used to score WDCT. As the column shows, the distance between the categories is not the same. The implication is that treating the categories of the scoring scale as intervals, a common practice in non-IRT procedures would be misleading. Therefore, in all the columns of Figure 2, the most able persons, the severest raters, the most difficult situations (items), and the most difficult categories fall at the top of the respective columns.

Figure 2
WDCT Facet Ruler

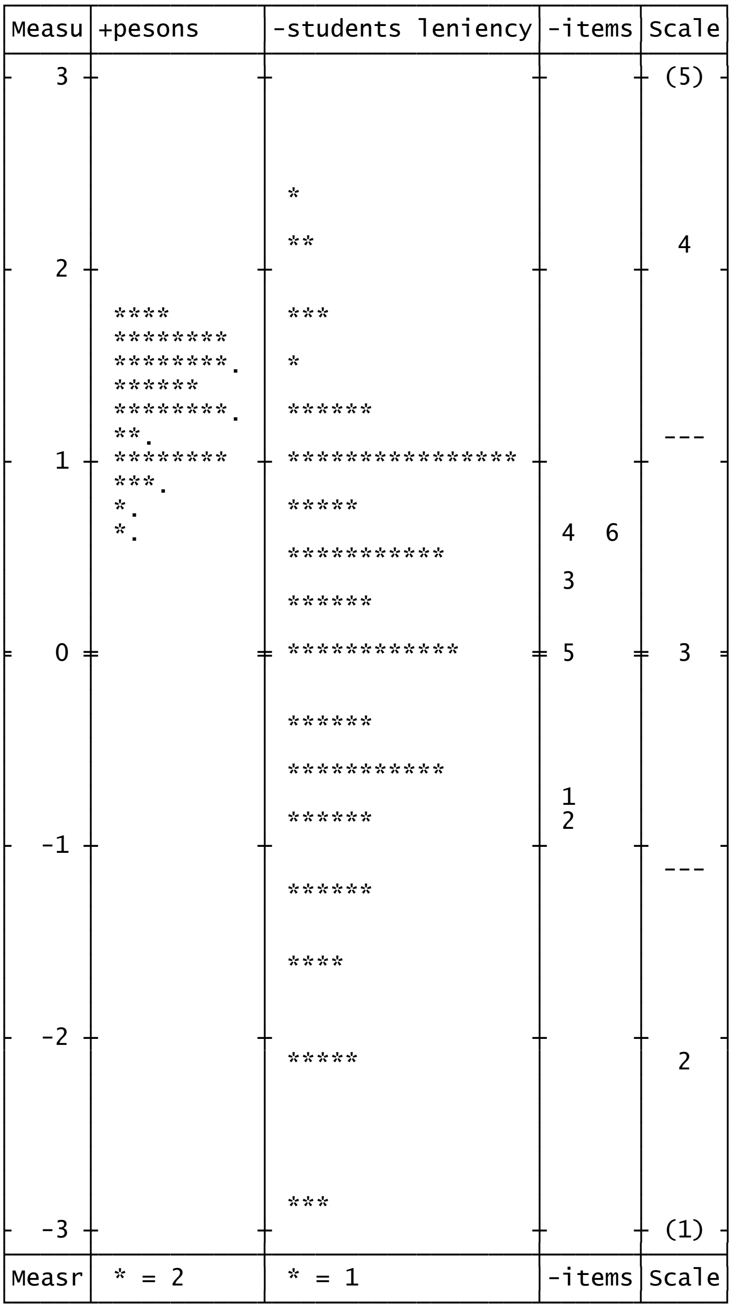
```
+-----+--------+-------+------+-----+
|Measr|+persons|-raters|-items|Scale|
|-----+--------+-------+------+-----|
+   3 +        +       +      + (5) +
|     |        |       |      |     |
|     |        |       |      |     |
|     |        |       |      |     |
|     |        |       |      |     |
|     | .      |       |      |     |
|     | .      |       |      |     |
|     |        |       |      | --- |
|     | .      |       |      |     |
|     |        |       |      |     |
+   2 + *      +       +      +     +
|     | .      |       |      |     |
|     |        |       |      |     |
|     | .      |       |      |     |
|     | **     |       |      |     |
|     | ***    |       |      |     |
|     | ***.   |       |      |  4  |
|     | **     |       |      |     |
|     | *.     |       |      |     |
|     | **.    |       |      |     |
+   1 + **     +       +      +     +
|     | **.    |       |      |     |
|     |        |       |      |     |
|     | **.    |       |      |     |
|     | ****   |       |      | --- |
|     | **.    | C     |      |     |
|     | *.     |       | 3    |     |
|     | **     |       |      |     |
|     | **.    |       | 6    |     |
|     | **.    | D     | 4  5 |     |
*   0 * ****** *       *      *     *
|     | .      | A     |      |  3  |
|     | *.     |       |      |     |
|     | ****   |       |      |     |
|     | *      |       | 1  2 |     |
|     | *.     |       |      |     |
|     |        | B     |      |     |
|     | .      |       |      | --- |
|     |        |       |      |     |
|     |        |       |      |     |
+  -1 +        +       +      + (1) +
|-----+--------+-------+------+-----|
|Measr| * = 2  |-raters|-items|Scale|
+-----+--------+-------+------+-----+
```

As Figure 2 shows, students' abilities range from -.70 to 2.51, and most of the students were between -.50 and 2 logits. The second column represented the rater's leniency in Logit. The raters' leniency was between .55 and .53. The third column indicates the item's difficulty. In line with the interpretation of the other columns, items with negative logit values are easier than average, while those with positive logit values are more difficult than average. In this study, situations 1 and 2 were the easiest with -.44 logit difficulty, and situation 3 was the most difficult one with. 37 logit difficulty.

4.3. Facet Ruler for DSAT

Figure 3 depicts the DSAT facets. Student abilities (Column 2) ranged from -1 to 3 logits. Student leniency (Column 3) varied between -3 and 2 logits. Situation difficulty (Column 4) revealed that Situation 2 was the easiest, while Situations 4 and 6 were the most difficult.

Figure 3
DSAT Facet Ruler



```
|Measu|+pesons   |-students leniency|-items|Scale|
|-----+----------+------------------+------+-----|
+  3  +          +                  +      + (5) +
|     |          | *                |      |     |
|     |          | **               |      |  4  |
+  2  +          +                  +      +     +
|     | ****     | ***              |      |     |
|     | ******** |                  |      |     |
|     | ********.| *                |      |     |
|     | ******   |                  |      |     |
|     | ********.| ******           |      |     |
|     | **.      |                  |      | --- |
+  1  + ******** + **************** +      +     +
|     | ***.     |                  |      |     |
|     | *.       | *****            |      |     |
|     | *.       |                  | 4  6 |     |
|     |          | ***********      |      |     |
|     |          |                  | 3    |     |
|     |          | ******           |      |     |
*  0  *          * ************     * 5    *  3  *
|     |          | ******           |      |     |
|     |          | ***********      |      |     |
|     |          |                  | 1    |     |
|     |          | ******           | 2    |     |
+ -1  +          +                  +      +     +
|     |          |                  |      | --- |
|     |          | ******           |      |     |
|     |          | ****             |      |     |
+ -2  +          +                  +      +     +
|     |          | *****            |      |  2  |
|     |          | ***              |      |     |
+ -3  +          +                  +      + (1) +
|-----+----------+------------------+------+-----|
|Measr| * = 2    | * = 1            |-items|Scale|
```

4.4. Examining Central Tendency Effect for WDCT

Conceptually, rater's central tendency is defined as the overemployment of the central categories of a rating scale by raters. According to Myford and Wolfe (2004), the Rater central tendency can manifest itself in one of the following ways: (1) A rater is able to accurately assess the highest and lowest performing test takers; however, he tends to inaccurately assign a middle category rating to those falling between the two extremes. (2) A rater cannot differentiate the categories along the entire scale and assigns all the test takers middle-category ratings. Raters' central tendency is displayed in Table 2.

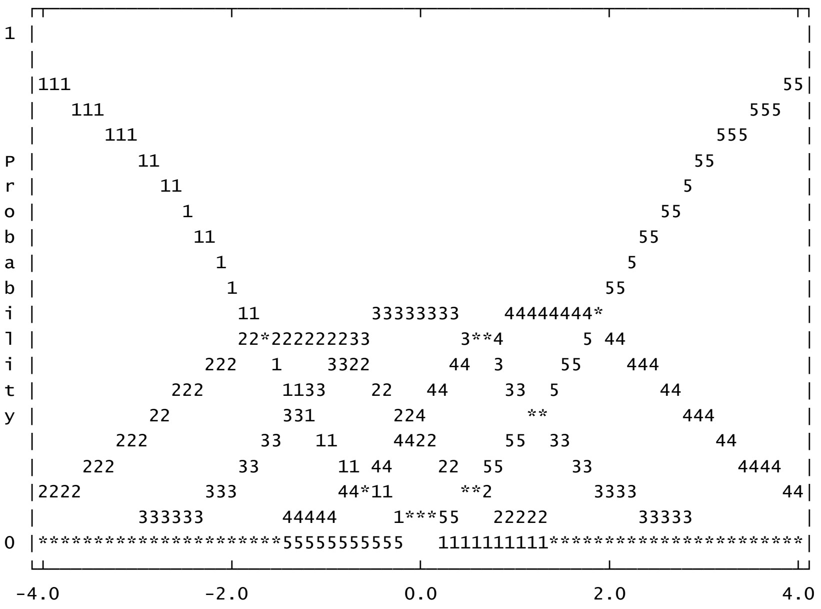
Table 2
Category Statistics for WDCT

Data					Quality Control				
		Category Counts		Cum.	Avge	Exp.	Outfit		
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.
1	83	83	3%	3%	-.16	-.43	1.3		
2	345	345	13%	16%	-.11	-.08	1.1	-1.69	.12
3	885	885	34%	50%	.32	.37	.8	- .81	.06
4	879	879	33%	83%	.94	.90	1.0	.64	.05
5	448	448	17%	100%	1.47	1.46	1.0	1.86	.06

Table 2 shows the categories used to rate the test takers. The second, fourth, and fifth columns display the frequency, percentage, and cumulative percentage of ratings assigned by raters, respectively. The data reveals a clear tendency for categories 3 and 4, with 34% and 33% of all ratings assigned to these categories, respectively. Therefore, the distribution of ratings is not evenly spread across all categories of the rating scale. However, the raters used the lower categories (1 and 2) 16% of the time and the higher rating categories (4 and 5) 50% of the time. This shows that despite some tendencies toward central categories, the function of the scales is acceptable.

In addition, Figure 4 shows the probability curve for WDCT. A bell-shaped probability curve is desirable. In this case, the curves are not completely bell-shaped, indicating that the scale functioning is not perfect. However, their near bell shape suggests that the scale functioning is acceptable.

Figure 4
Probability Curve for WDCT



```
      ┌┬────────────────┬────────────────┬────────────────┬────────────────┬─┐
    1 |                                                                      |
      |                                                                      |
      |111                                                                 55|
      |   111                                                          555   |
      |      111                                                    555      |
    P |         11                                                55         |
    r |           11                                             5           |
    o |             1                                          55            |
    b |              11                                      55              |
    a |                1                                    5                |
    b |                 1                                 55                 |
    i |                  11          33333333    44444444*                   |
    l |                  22*222222233        3**4       5 44                 |
    i |               222   1    3322       44  3     55    444              |
    t |            222       1133    22   44     33  5         44            |
    y |          22          331       224         **            444         |
      |       222          33   11     4422      55  33             44       |
      |    222           33       11 44    22  55      33             4444   |
      |2222           333         44*11      **2         3333              44|
      |         333333       44444     1***55   22222        33333           |
    0 |**********************55555555555   1111111111************************|
      └┴────────────────┴────────────────┴────────────────┴────────────────┴─┘
      -4.0             -2.0              0.0              2.0              4.0
```

4.5. *Examining Central Tendency Effect for DSAT*

The categories that were used by students to rate their own performance in DSAT, based on frequency count (%) and percentage of rating (cum-%), were category 3 (36%), followed by category 4 (32%), category 5 (20%), category 2 (10%), and category 1 (1%). The students used the lower categories (1 and 2) 12% of the time and the higher rating categories (4 and 5) 52% of the time (Table 3). The width of the probability curve (Figure 5) and the whole frequency presented in category statistics (Figure 6) suggest that the category usage in the DSAT was resealable.

Table 3
Category Statistic for DSAT

Data					Quality Control				
	Category Counts			Cum.	Avge	Exp.	Outfit		
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.
1	8	8	1%	1%	-.04	-1.17	1.8		
2	64	64	10%	12%	-.68	-.41	.8	-2.88	.37
3	227	227	36%	48%	.60	.54	1.1	- .21	.15
4	202	202	32%	80%	1.64	1.71	.8	1.23	.10
5	159	123	20%	100%	3.12	3.05	1.0	2.87	.13

Figure 5
Probability Curve for DSAT

```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ┌┬──────────┬──────────┬──────────┬──────────┬──────────┬──────────┬┐
  1 |                                                                   |
    |1111                                                           5555|
    |    111                                                     555    |
    |       11                                                 55       |
    |         11                                             55         |
  P |           1                                           5           |
  r |            1                                         5            |
  o |             1                                       5             |
  b |              1               3333333               5              |
  a |               1     22      3       3      44     5               |
  b |                1 222  222  3         3  444  444 5                |
  i |                 *        23           3*        *                 |
  l |               22 1      3322          4 3      5 44               |
  i |              2    1    3    2       44   3    5    4              |
  t |            22      1  3      2     4      3  5      4             |
  y |           2         13        2   4        35        44           |
    |         22          311        2*4        553          44         |
    |       22          33   1       4 2       5   33          44       |
    |    222          33      11  444   222  55      33          444    |
    |2222         3333        44**11     55**22        3333         4444|
  0 |*************************555555*****111111*************************|
    └┴──────────┴──────────┴──────────┴──────────┴──────────┴──────────┴┘
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

4.6. Test Method

4.6.1. Facet Ruler of Test Methods

The effect of the two different methods on students’ performance in the six situations is illustrated in Figure 6. As the figure reveals, student abilities range from -1 to 4 logits, considering the testing methods. Additionally, the two methods themselves exhibit varying difficulty levels centered around zero logits. The scenarios themselves also present differing levels of difficulty, ranging from -1 to 1 logits. It is important to note that the unequal distances between categories in the last column of Figure 6 indicate the scale categories are not on an interval scale. Rasch analysis assumes an underlying interval scale for the latent variable being measured. This means the difference between any two categories on the scale should be consistent and reflect equal differences in the underlying variable. Unequal distances between categories suggest the scale is not truly capturing equal intervals in the latent variable.

Figure 6

Ruler for Test Method

```
+--------------------------------------+
|Measr|+students  |-method|-items|Scale|
|-----+-----------+-------+------+-----|
|   4 + *         +       +      + (5) |
|     |           |       |      |     |
|     | *         |       |      |     |
|     |           |       |      |     |
|     |           |       |      |     |
|     |           |       |      |     |
|     | *         |       |      |     |
|     |           |       |      |     |
|     |           |       |      |     |
|   3 +           +       +      +     |
|     |           |       |      |     |
|     |           |       |      |     |
|     | *         |       |      |     |
|     |           |       |      |     |
|     | ****      |       |      |     |
|     |           |       |      | --- |
|     | *         |       |      |     |
|     | ***       |       |      |     |
|   2 +           +       +      +     |
|     | *         |       |      |     |
|     |           |       |      |     |
|     | *****     |       |      |     |
|     | ***       |       |      |     |
|     | ********* |       |      |  4  |
|     |           |       |      |     |
|     | *******   |       |      |     |
|     | *******   |       |      |     |
|   1 +           +       +      +     |
|     | *****     |       |      |     |
|     | *****     |       |      |     |
|     | ******    |       |      |     |
|     |           |       |      |     |
|     | ********  |       |      | --- |
|     | ****      |       | 3  6 |     |
|     | ******    |       | 4    |     |
|     | ******    |       |      |     |
|     | ****      | WDCT  | 5    |     |
*   0 * *******   *       *      *     *
|     |           | DSAT  |      |     |
|     | *         |       |      |  3  |
|     | *****     |       |      |     |
|     | **        |       |      |     |
|     | **        |       |      |     |
|     | ***       |       | 1    |     |
|     | *         |       | 2    |     |
|     | *         |       |      | --- |
|     |           |       |      |     |
|  -1 +           +       +      + (1) |
|-----+-----------+-------+------+-----|
|Measr| * = 1     |-method|-items|Scale|
+--------------------------------------+
```

4.6.2. Test Method Measure

Table 4 details the method's difficulty. The reliability index of .80 indicates the consistency of test method measures. As Table 4 shows, the fixed hypothesis that the two methods had the same level of difficulty is rejected by a significant chi-square (chi square= 5, *p=0.02*). Thus, the two methods are significantly different in terms of difficulty.

Table 4

Test Method Measure

Methods	Measure	Error	Fit	
WDCT	.08	0.5	.99	0.99
SDCT	-.08	0.5	1.0	1.02
Mean	.00	0.5	.01	1.00
SD	.11	.00	0.2	.02

Fixed (all same) chi-square: 5.0; significance (probability): .02; Reliability .80, RSM .05; separation 2.01

5. Discussion

Regarding the first research question, the FACETS analyses showed that there are differences in the average ratings assigned to different persons and situations. Person and situation measures show that the elements within both WDCT and DSAT are significantly different. Moreover, they are different across the test methods. These differences might result from the rating that each test method requires. The stable fit statistics for raters on WDCT, which show raters' predictable behavior, indicate raters reliably assessed the test takers' performance on WDCT. This finding suggests that raters had internal consistency and that their actions were not by chance. However, a significant chi-square and a high separation index (13) show substantial variability across the rater measures in terms of leniency. Fit statistics for raters (students as raters) on DSAT are not as stable as they might be. This finding shows that students' performance in assessing their own performance is not predictable, and there are some signs of inconsistencies. In addition, significant Chi squares suggest that they have meaningful differences in their leniency. The rater finding of this study is in line with most of the previously conducted studies that found raters' severity or leniency levels were different (e.g., Brown &Ahn, 2010; Li et al, 2023; Grabowski, 2008; Liu & Xie, 2014; Sonnenburg-Winkler et al., 2020; Youn, 2007). However, Rover (2006) didn’t find any difference among raters’ performance in DCT.

The FACETS general results showed that the situations' difficulties are significantly different within the WDCT and DSAT. The situations on the WDCT stably differentiated between varying degrees of the 110 test takers’ pragmatic abilities. However, the fit statistics of DSAT are not stable. This shows that there is no predictable pattern for the assigned scores by raters (students as raters) to the situations on the DSAT. This might result from the test taker's lack of experience, which leads them to underestimate or overestimate their own abilities and assign scores to themselves by using the rating scale that was selected by chance.

The method measure and the method vertical map showed that the mentioned differences are significant across the test methods. The significant chi-square for the difficulty differences between the test methods indicates that test methods exercise a systematic influence on test takers' performance on pragmatic tests. However, a stable infit finding shows that both methods firmly distinguished between varying degrees of the test taker’s pragmatic ability. The stable fit statistics and divergent levels of difficulty of test method measures indicate that both tests have a unique way of differentiating the test taker's pragmatic ability.

While the findings suggest systematic differences between the two test types, the facet analysis in this study does not provide sufficient evidence to definitively recommend one test method as a superior measure of Iranian EFL learners' pragmatics ability. Despite the shortcomings that the tests have in terms of their fitness and variations in difficulty measures, both of the tests are acceptable in terms of their functioning level, and they exercise a unique way of measuring learners' ability. However, in line with Tajeddin and Alemi (2014), this study would like to suggest that for obtaining an accurate result from the mentioned test methods, it's better to have rater training, especially for the

DSAT. Moreover, this study would like to suggest that DSAT should not be used for decision-making purposes. If the test takers know the decision that will be made by the result of the test, they will overestimate their own performance.

Regarding the second research question, the functioning of the five-point scales and raters' tendency to use the middle categories can be measured by person measures. At the group level, if raters exhibit a central tendency, there should be a lack of variation between persons in the level of performance; therefore, a non-significant chi-square value shows a group-level central tendency (Myford & Wolfe, 2004). The findings of this study (chi-square =858.8, p<.05, and chi-square = 80.06, p<.05, respectively for WDCT and DSAT) showed that there is no group-level central tendency in the test methods. And relatively high separation reliabilities (.88 and.70, respectively, for WDCT and DSAT) suggest that raters reliably differentiated test takers in their level of performance. These findings don't suggest the existence of group-level central tendencies in the test methods. We can conclude that the usage of middle categories that are used in rating is not due to the rater's inability to distinguish between scale categories, or what Myford and Wolfe (2004) put as resorting to "middle-of-the-road" rating. Moreover, all of the raters finished the rating procedure at a convenient time (both in WDCT and DSAT, by considering the fact that enough time was allocated to them for rating); therefore, the usage of middle categories is not due to fatigue. According to Myford and Wolfe (2004), "If a rating procedure requires raters to work for several days, there may be raters who are prone to fatigue or boredom" (p. 391) and show lower levels of accuracy in distinguishing between the scales and try to use the middle categories.

Although the scale functioning at the group level was acceptable, and there was no significant central tendency at the group level, there were some signs of central tendencies at the individual level. Individual levels of central tendency can be detected by rater fit statistics. A fit statistic lower than 1 implies a central tendency. In WDCT, all raters' fit statistics are around 1. This suggests that there is little variation between raters observed and expected scores. Therefore, there is a small degree of central tendency at the individual level. However, in DSAT, there is a larger variation in fit statistics. This shows that there is a greater degree of central tendency.

The probability curves that are hill-like with little overlap show good scale functioning. Although the individual central tendency of DSAT is somehow greater than that of WDCT (as mentioned, the variation of the fit statistic from 1 was larger in the DSAT), the probability curve of DSAT is much steeper and hill-like, and the curve of WDCT is much flat with somehow more overlap. This can be explained by considering the fact that the number of raters in the DSAT was higher than in the WDCT, which only had four raters. Overall, we can say the scale functioning on DSAT and WDCT at the group level is reasonably acceptable. However, the scale functioning seems better in the DSAT.

The current study's findings on scale functioning align with previous research. Liu (2014) observed that Chinese raters using the WDCT performed well, with no significant central tendency. However, Brown and Ahn (2010) suggest that scales used in self-assessment role-play contexts function better than the WDCT and ODCT. This suggests that the effectiveness of a scale might depend on the assessment context. Building on these findings, Su and Shin (2024) proposed that the format, discriminatory power, and pedagogical value of rating scales all contribute to their effectiveness. Additionally, they emphasized the importance of rater perceptions regarding the ease of use and perceived effectiveness of the scales. These insights highlight the need to consider not only the statistical properties of scales but also their usability and rater acceptance for optimal functioning.

6. Conclusion

The current study intended to detect the impact of test type on students' performance and also to determine the variability of the elements within each test method, including items, raters, and testers. Two test formats, written and self-assessment, were used in this study. The analysis showed that students' measures varied significantly. It served as a marker for various levels of ability, demonstrating that the tests were able to distinguish between test takers of different levels of ability.

The findings showed that test items and situations had different levels of difficulty in both tests. Although the items were the same in both tests, there were some differences in their difficulty. It may be related to the scoring. Students, as raters in the DSAT, may overestimate or underestimate their own ability to score themselves, and because of that, they may use some extreme scale in rating themselves.

The analysis revealed significant differences in rater leniency, suggesting that raters did not consistently apply the rating scale. However, this finding does not necessarily imply a lack of internal consistency. Other factors, such as individual rater tendencies, could also explain these differences. Furthermore, the findings indicate systematic variation in how the two test types (WDCT and DSAT) measured test-taker performance. This suggests that the tests might capture different aspects of performance, or there might be differences in the difficulty levels of the tasks used in each test. To improve the accuracy of future test administrations, particularly for the DSAT, it might be beneficial to implement rater training beforehand. This training could focus on ensuring consistent application of the rating scale and fostering a shared understanding of the performance characteristics being assessed by each test.

Limitations

This study only used two test methods to find out the interaction between the test method and other elements, such as the rater on pragmatic tests. Other test types could have contributed to this study. Because of the nature of the DSAT, which requires students to assess themselves, and the nature of the multi-faceted software, it was impossible to compare group raters. Rates only were investigated in their own group.

Implications

This study has implications for teachers and researchers. This study suggests that DSAT can be used as a valid test type for measuring learners' pragmatic knowledge. In addition, this study encourages researchers to use multi-facets in their studies. Multi-facet has great potential to measure the impact of different factors on the test, and because pragmatic assessment is new in the language testing field, it is suggested to employ multi-facet more in further pragmatic studies.

Declaration of Conflicting Interests

All authors declare that they have no conflicts of interest.

Funding

There is no funding in this study.

References

Ahn, R. C. (2005) *Five 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

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

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

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

Birjandi, P., & Soleimani, M.M. (2013). Assessing language learners' knowledge of speech acts: A test validation study. *Issues in Language Teaching, 2* (1), 1-26. https://journals.atu.ac.ir/article_1349.html

Brown, J. D. (2001a). 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, J. D. (2001b). Pragmatics tests: Different purposes, different tests. In K. Rose & G. Kasper (eds.), *Pragmatics and language teaching* (pp. 301-326). Cambridge: Cambridge University Press, https://cir.nii.ac.jp/crid/1360292620986551168.

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, H. D. (1994). *Teaching by principles: An interactive approach to language pedagogy* (Vol. 1, p. 994). New Jersey: Prentice Hall Regents. https://thuvienso.hoasen.edu.vn/handle/123456789/11518

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.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47. https://www.uefap.com/tefsp/bibliog/canale_swain.pdf.

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

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.

Eslami.R, Zahra. (2014). Introduction from the Guest Editor. *Iranian Journal of Language Testing*. *4*(1) ,1-4. https://www.ijlt.ir/article_114388_50b1ca6add838ff71c9c41e0aff0c217.pdf.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. ed. S Takala, Section H. Strasbourg, Fr.: Council of Europe http://www.coe.int/t/dg4/linguistic/Source/CEFrefSupp-SectionH.pdf.

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.

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

Hudson, T., Brown, J. D., & Detmer, E. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Vol. 7). Natl Foreign Lg Resource Ctr.

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

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 nonnative speakers' scoring in an interlanguage pragmatics test. *Modern Foreign Languages*, *30*(4), 395-404.

Linacre, J.M., 2006. A user's guide to FACETS. Downloaded from www.winsteps.com.

Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement*, *3*(4), 382-405. https://www.researchgate.net/profile/Alfred-Stenner/publication.

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.

Llosa, L. (2020). Revisiting the role of content in language assessment constructs. In G. Ockey & B. Green (Eds.), *Another generation of fundamental considerations in language assessment: A festschrift in honor of Lyle F. Bachman* (pp. 29–42). Springer. https://link.springer.com/chapter/10.1007/978-981-15-8952-2_3.

Mao, T., & He, S. (2021). An integrated approach to pragmatic competence: Its framework and properties. *Sage Open*, *11*(2), 1–13. https://doi.org/10.1177/21582440211011472.

Matsugu, S. (2014). Developing a pragmatics test for Arabic ESL learners. *Arab World English Journal*, *5*(3), 3-14.

Messick, S. (2012). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In *Construction versus choice in cognitive measurement* (pp. 61-73). Routledge.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227. https://d1wqtxts1xzle7.cloudfront.net.

Nemati, M., Rezaee, A. A., & Mahdi Hajmalek, M. (2014). Assessing pragmatics through MDCTs: A case of Iranian EFL learners. *Iranian Journal of Applied Language Studies*, *6*(2), 59-80. https://ijals.usb.ac.ir/article_2190_7f34dd23c881a05b064b19579fdcaa83.pdf

Purpura, J. E., (2017). Assessing meaning. In E. Shohamy, L. Or, & S. May (Eds.), *Language testing and assessment: Encyclopedia of language and education* (pp. 33–61). New York, NY: Springer International Publishing. doi: 10.1007/978-3-319-02326-7_1-1

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. (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., & Ono, R. (1995). Eliciting speech act data in Japanese: The effect of questionnaire type. *Language Learning*, *45*(2), 191-223.

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. (2012). Roever, C. (2013). Assessment of pragmatics. In C. Chapelle (ed.), *The Encyclopedia of applied linguistics*. Oxford, UK: Blackwell Publishing.

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

Sartori, R., & Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity*, *41*(1), 359-374. https://doi.org/10.1007/s11135-006-9006-x

Schroeder, H. (2021). A pragmatic view on clause linkages in Toposa, an Eastern Nilotic language of South Sudan. *Ghana Journal of Linguistics*, *10*(1), 329-352. https://www.ajol.info/index.php/gjl/article/view/211856

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

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, *39*(1), 215-252. https://doi.org/10.3102/0091732X14557003

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

Sherkuziyeva, N., Imamutdinovna Gabidullina, F., Ahmed Abdel-Al Ibrahim, K., & Bayat, S. (2023). The comparative effect of computerized dynamic assessment and rater mediated assessment on EFL learners' oral proficiency, writing performance, and test anxiety. *Language Testing in Asia*, *13*(1), 15. https://link.springer.com/article/10.1186/s40468-023-00227-3.

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

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

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

Wise, S. L. (2019). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry*, *10*(1), 21-33. https://doi.org/10.1080/20004508.2018.1490127

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

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., & 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.

Yoshitake, S. S. (1997). Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation. *Unpublished doctoral dissertation, Columbia Pacific University, Novata, CA.*

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?

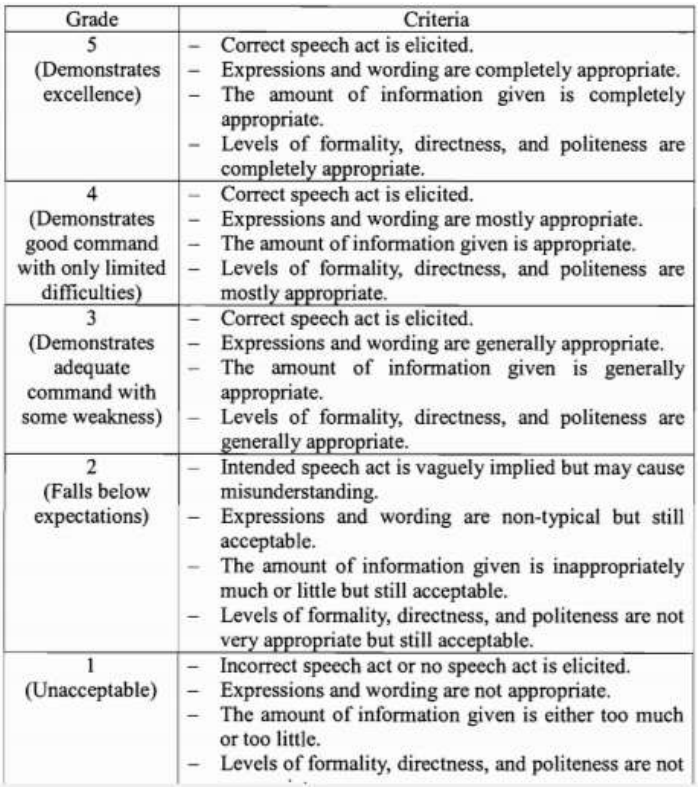
3 You need a ride home from school. You notice someone who lives down the street from you is also at school, but you haven't spoken to this person before. You think they might have a car. What would you say?
YOU:

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