ÖSYM

e-TEP

TECHNICAL REPORT

e-TEP 2025/1 Evaluation Analyses

Üniversiteler Mahallesi İhsan Doğramacı Bulvarı
No: 4D 06800 Bilkent Çankaya/ANKARA

4446796 (OSYM) (Çağrı Merkezi)

www.osym.gov.tr

e-TEP 2025/1 Evaluation Analyses

e-TEP 2025/1 was held on July 19, 2025, with a total of 505 test takers participating. Following the evaluation of the exam results, the necessary analyses regarding the validity and reliability of the e-TEP scores were performed. The means, standard deviation values, reliability coefficients, and standard error of measurement values for each skill and the total scores for e-TEP 2025/1 are presented in Table 1.

Table 1. e-TEP 2025/1 Reliability Estimates and Standard Error of Measurement

Skill	Scale	Mean	Standard Deviation	Reliability Coefficient	Standard Error of Measurement
Reading	0-30	19.47	5.68	.85	2.19
Listening	0-30	16.78	6.07	.84	2.43
Speaking	0-30	13.70	5.78	.84	2.31
Writing	0-30	17.21	5.92	.83	2.44
Total	0-120	66.96	20.96	.91	6.29

DeVellis (2003) suggested that Cronbach's alpha values of between .70 and .80 indicate a respectable level of reliability; between .80 and .90 indicate a very good level of reliability. Reliability of .90 is a determinant level as to making important decisions in individual diagnostic and academic placement processes. As shown in Table 1, the reliability statistics for each skill of e-TEP 2025/1 indicate a very good level of reliability and a higher reliability statistic for the total scores meeting the criterion to make important decisions based on e-TEP total scores.

Relations Between Skills

Test takers are generally expected to demonstrate similar language proficiency levels across different skills. However, receptive skills are typically expected to be more advanced than productive skills. In the context of inter-skill correlations, the performances of test takers exhibiting unusual discrepancies across language skills were reviewed after the exam.

The total scores and correlations among skills for the e-TEP 2025/1 were analysed, and the following results were obtained.

Table 2. Relations Between Skills

	Total Score	Reading	Listening	Speaking	Writing
Total Score	1.00				
Reading	.88	1.00			
Listening	.90	.77	1.00		
Speaking	.86	.65	.69	1.00	
Writing	.90	.71	.73	.75	1.00

According to Dancey and Reidy (2007), correlation values above .70 indicate a strong relationship between two variables. Correlation values between .40 and .70 indicate a moderate relationship. Lastly, correlation values between .10 and .40 indicate a weak relationship. Furthermore, the sign of the correlation coefficient reflects the direction of the relationship between the two variables. As shown in Table 2, there is a strong positive correlation between the e-TEP 2025/1 total scores and the individual scores of the skills. Among the skills, the correlation between scores of the reading and speaking skills indicate moderate relationships as well as the those of the listening and speaking skills. However, all other correlations among other scores based on skills indicate strong relationships, as expected.

Scoring Process of e-TEP Speaking and Writing Skills

The speaking and writing skills are scored independently by two different raters for each task. Scoring is based on a 20-point rubric, and the final score for the task is determined by obtaining the average of the two raters' scores. Each rater scores only one of each test taker's performances.

If the difference between the two raters' scores is 4 points or more, the test taker's performance is re-scored by a "senior rater". The senior rater scores the task independently without seeing the initial scores given by the initial raters. The final score is then obtained by averaging the senior rater's score with the score of the initial rater whose score is closest to that of the senior rater. If the difference between the senior rater's score and at least one of the initial raters' scores is still 4 points or more, the scoring cycle is repeated from the beginning.

Once the rating process is completed, rater reliability and scoring reliability analyses are carried out. The results are released provided that all analyses demonstrate reliability at least at an acceptable level. If rater reliability is low, the rating process may be repeated by senior raters.

Rater Reliability

Infit and outfit values were assessed to determine whether each rater rated each performance on each task and each rubric component in accordance with the measurement model. Infit and outfit statistics below 0.5 indicate artificial conformity to the measurement model, while infit and outfit statistics above 2.0 may indicate rating that distorts the measurement model. Infit and outfit statistics between 0.5 and 2.0 indicate rating in line with the measurement model (Linacre, 2002). The speaking skill ratings fell within these limits at a rate of .95, while the writing skill ratings fell within these limits at a rate of .98.

Scoring Reliability

Scoring reliability refers to the relation between the scores assigned by the raters and the test takers' final scores in the relevant task. Since each test taker's performance is assessed by two different raters, the degree of agreement with the final score obtained from the particular task reflects both rater consistency and the overall reliability of the rating process.

Upon examining the relationships between individual task scores and the test taker's final score for the relevant task, the following results were obtained. Based on this table, strong correlations are observed between the scores assigned by the raters and the test takers' final scores in the corresponding tasks.

Table 3. Relations Between Raters' Scores and Test Takers' Corresponding Task Scores

Task	Correlation Coefficient
Speaking Task 1	.92
Speaking Task 2	.90
Speaking Task 3	.90
Writing Task 1	.90
Writing Task 2	.90

Quality Check

After the rating process is completed, 10% of test takers are randomly sampled for quality check, and their performances are re-rated by senior raters. The relationship between the initial speaking and writing task scores and the scores obtained through re-rating is then examined. The estimated correlation coefficients for this relationship are .92 for the speaking skill and .87 for the writing skill, indicating a strong positive relationship between the initial ratings and the re-ratings for both skills.

REFERENCES

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology* (4th ed.). Pearson Education.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Sage.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.