April, 2012

A RESEARCH REPORT ON JANUARY 2012 EPE

Prof. Dr. Hüsnü ENGİNARLAR
Director, School of Foreign Languages

I. INTRODUCTION

The new version of METU-EPE was officially administered in January 2012 to approximately 900 graduate students seeking admission to graduate programs at METU. EPE was administered to graduate school candidates in April 2011 and 2010 (about 1,800 test takers) and in January 2009 (about 900 test takers).

Since the previous EPE and new EPE were given to comparable populations, a comparative study on EPE 2012 is well justified

a) to look at success rates and
b) to investigate the reliability and validity of the new components and the overall test.

This is the purpose of this report. The critical reader will, of course, make projections regarding the imminent performance of DBE students, as we all do.

Acknowledgements

This report is actually the joint effort of **Fatma Ataman**, **Gökçen Baskan** and **Naz Dino**. Without the meticulous statistical work by Fatma Ataman, Gökçen Baskan and Naz Dino, and the editing work by Naz Dino, this report would not have been prepared at all. I extend my heartfelt thanks to this hardworking team.

II. A COMPARISON OF AVERAGES AND SUCCESS RATES

Table 1 below displays the arithmetic means of the different components and the overall averages of the 2012 (January), 2011 (April) , 2010 (April) and 2009 (January) EPEs.

When we compare these means, we see that the Reading and Writing sections have produced quite similar results. The Language Use section of the January 2012 EPE is also quite similar to two exams in the past (2010 & 2009) with respect to the averages. We observe a slight decrease in the new Note-Taking section but a more noticeable drop in the Jan. 2012 MC Listening section (compare TOTALS: 68.46 / 80.65 / 87.80 / 76.80).

When we consider the average performance of those = > 24.5 in Language Use & Reading for the Jan. 2012 group to obtain a healthy comparison, the mean goes up to 72.17, which is still the lowest mean in the series of exams.

TABLE 1

JANUARY 2012	**CLOZE TEST (10 pts)**	**DIALOG & SITUATION (10 pts)**	**LANGUAGE USE (Total 20 pts.)**	**READING (30pts)**	**LISTENING (30 pts)**	**NOTE-TAKING (5 pts)**	**WRITING (15 pts)**
TOTAL	5.98	7.13	13.11 (65.55%)	21.13 (70.43%)	20.54 (68.46%) **21.65 (72.17%)**	3.82 (76.40%) **4.01 (80.20%)**	9.92 (66.13%) **10.52 (70.13%)**
MA	6.01	7.22	13.23 (66.15%)	21.34 (71.13%)	20.74 (69.13%) **21.74 (72.47%)**	3.86 (77.20%) **4.03 (80.60%)**	10.09 (67.27%) **10.58 (70.53%)**
PHD	6.64	7.42	14.06 (70.30%)	22.39 (74.63%)	21.38 (71.27%) **22.62 (75.40%)**	3.93 (78.60%) **4.15 (83.00%)**	10.30 (68.66%) **10.82 (72.13%)**
OTHER*	5.1	5.97	11.07 (55.35%)	17.68 (58.93%)	17.76 (59.20%) **19.62 (65.40%)**	3.35 (67.00%) **3.69 (73.80%)**	8.45 (56.33%) **9.42 (62.80%)**
					Figures in RED: **Ave. of students whose L.Use + Reading grades = > 24.50**		

APRIL 2011	LANGUAGE USE (20 pts.)	READING (30pts)	LISTENING (20 pts)	NOTE-TAKING (10 pts)	WRITING (20 pts)
TOTAL	14.43 (72.15%)	21.23 (70.76%)	16.13 (80.65%)	8.31 (83.10%)	14.37 (71.85%)
MA	14.50 (72.50%)	21.36 (71.20%)	16.19 (80.95%)	8.34 (83.40%)	14.37 (71.85%)
PHD	15.25 (76.75%)	23.52 (78.40%)	16.47 (82.35%)	8.04 (80.40%)	14.97 (74.85%)
OTHER*	12.12 (60.60%)	16.50 (55.00%)	13.52 (67.60%)	6.60 (66.00%)	13.45 (67.25%)
			Stage 1 grade = > 24.50		

APRIL 2010	LANGUAGE USE (20 pts.)	READING (30pts)	LISTENING (20 pts)	NOTE-TAKING (10 pts)	WRITING (20 pts)
TOTAL	12.93 (64.65%)	21.71 (72.37%)	17.56 (87.80%)	8.74 (87.40%)	12.69 (63.45%)
MASTER	12.90 (64.50%)	21.70 (72.33%)	17.59 (87.95%)	8.75 (87.50%)	12.70 (63.50%)
PHD	14.57 (72.85%)	24.32 (71.06%)	17.27 (86.32%)	8.80 (88.00%)	13.75 (68.75%)
OTHER*	12.55 (62.75%)	20.49 (68.30%)	17.12 (85.60%)	8.25 (82.50%)	11.08 (55.40%)
			Stage 1 grade = > 24.50		

JANUARY 2009	LANGUAGE USE (20 pts.)	READING (30pts)	LISTENING (20 pts)	NOTE-TAKING (10 pts)	WRITING (20 pts)
TOTAL	12.32 (61.60%)	20.22 (67.40%)	15.36 (76.80%)	7.59 (75.90%)	12.48 (62.40%)
MASTER	12.60 (63.00%)	21.10 (70.33%)	15.66 (78.30%)	7.77 (77.70%)	12.61 (63.15%)
PHD	13.72 (68.60%)	21.15 (70;50%)	14.85 (74.25%)	7.42 (74.20%)	11.65 (58.25%)
OTHER*	9.76 (48.80%)	13.38 (44.60%)	11.63 (58.15%)	5.11 (51.10%)	11.29 (56.45%)
			Stage 1 grade = > 24.50		

*** Amnesty, transfer, new students, ÖYP, Article 35, etc**

Table 2 below shows the success rates with respect to the number of candidates who scored = > 64.50/100, which is the minimum requirement for admission to the graduate programs at METU.

What seems to be puzzling between the two tables is that while the arithmetic means are fairly high, there is a decrease of 9-10% in the number of test takers with scores above = > 64.50/100 in the January, 2012, EPE.

TABLE 2

	JANUARY 2012		APRIL 2011		APRIL 2010		JANUARY 2009	
	NO. OF STUDENTS	**SUCCESSFUL***	**NO. OF STUDENTS**	**SUCCESSFUL***	**NO. OF STUDENTS**	**SUCCESSFUL***	**NO. OF STUDENTS**	**SUCCESSFUL***
TOTAL	880	566 (64.31%)	1841	1373 (74.58%)	1864	1381 (74.09%)	711	491 (69.06%)
MASTER	735	491 (66.80%)	1680	1278 (76.61%)	1725	1281 (74.26%)	636	463 (72.80%)
PHD	69	48 (69.57%)	83	64 (77.11%)	75	58 (77.33%)	33	19 (57.58%)
OTHER**	76	27 (35.51%)	78	31 (41.33%)	64	42 (65.62%)	42	9 (21.43%)

*** = > 64.5**

**** "Amnesty", "Transfer", "ÖYP", "Article 35" and "Other" categories**

Except for the noticeable drop in the MC Listening, there seemed to be no concrete reason for the decrease in the success rate. We had to look elsewhere. One other possibility was to look into the relative performances of METU and NON-METU test taker populations applying to graduate programs at METU.

Table 3 below displays numbers of MA and PhD applicants (= > 64.5/100 - pass) and their percentages in 2012, 2011 and 2010 EPE's. It is obvious that there has always been a difference between the two populations but this gap increased in the January, 2012 EPE, considering especially the METU and NON-METU graduate program candidates, respectively (44 %, 57% and 59%).

TABLE 3

TOTAL PASS / FAIL

		N	Pass	Ave.	%
DR	METU	30	27	80.13	90
	N-METU	39	21	65.88	54
TOTAL		69	48	73	
YL	METU	435	378	75.23	87
	N-METU	301	132	60.65	44
TOTAL		736	510	67.94	

JANUARY 2012

		N	Pass	Ave.	%
DR	METU	50	45	80.99	90
	N-METU	33	19	68.15	56
TOTAL		83	64	74.57	
YL	METU	957	868	78.04	91
	N-METU	725	412	61.71	57
TOTAL		1682	1280	69.88	

APRIL 2011

		N	Pass	Ave.	%
DR	METU	47	41	79.91	87
	N-METU	28	17	64.34	61
TOTAL		75	58	72.13	
YL	METU	960	827	75.56	86
	N-METU	765	454	61.41	59
TOTAL		1725	1281	68.49	

APRIL 2010

When we further investigated the relative performances of the two populations in the different sections of the EPE 2012 (see **Table 4** below), we saw clearly that the greater part of the lower success rate of the NON-METU candidates was caused by the new Language Use section. This was understandable to a large extent. It may be reasonable to expect this gap to narrow down in the future with appropriate training and preparation. What is interesting though is that METU senior students performed quite well in that new section despite their unfamiliarity and lack of training and preparation.

All this lead us to be optimistic about the performance of the DBE students in the upcoming EPE's in the summer (2012).

TABLE 4

LANGUAGE USE

		Ave.	
DR	METU	16.50	overall ave. 13.33
	N-METU	12.48	
YL	METU	14.69	
	N-METU	11.29	

JANUARY 2012

		Ave.	
DR	METU	15.72	overall ave. 14.53
	N-METU	14.55	
	TOTAL		
YL	METU	15.38	
	N-METU	13.33	
	TOTAL		

APRIL 2011

		Ave.	
DR	METU	14.57	overall ave. 12.95
	N-METU	12.64	
	TOTAL		
YL	METU	13.72	
	N-METU	11.98	
	TOTAL		

APRIL 2010

		Ave.	
DR	METU	13.64	overall ave. 12.65
	N-METU	13.81	
		TOTAL	
YL	METU	13.71	
	N-METU	11.32	
		TOTAL	

JANUARY 2009

The second component which might have had a role in lowering the success rate is the MC Listening. This may be true regardless of the respective performances of different groups of test takers since the lowest mean obtained in four years with graduate populations was in the 2012 MC Listening.

A comparison of Reading and the MC Listening in the previous years (2009 / 2010 / 2011) shows higher means for the MC Listening in the range of 10% or so. This year – Jan. 2012 – for the first time, the MC Listening mean is similar to the Reading mean.

Based on the quite high MC Listening means in the past, it was predicted that increasing the weight of this component from 20 pts. to 30 pts. would contribute positively to the pass rate. Decreasing the number of items from 40 to 30 would also take care of the complaints about the length of the MC Listening .With the evidence we now have, the expected advantages may not be realized in the future, or at least not to the extent some of us would hope.

We need to proceed with caution in MC Listening in setting new standards. Curriculum, materials and instructional practices should now pay much more attention to the listening skill.

III. RELIABILITY ESTIMATES OF JANUARY 2012 EPE

Item analysis studies were conducted on the 60-item (Multiple Choice = MC) Listening and Reading sections, the Cloze Test (20 items = 10 pts), using the statistical package Iteman and manually on the Dialogs and Situations section.

Only 4-5 items out of 60 fell in the range of very difficult items (40 % and 50%). The overall difficulty level of the 60-item Listening and Reading sections was computed as P: 0.647, which means, on the average, each item was correctly answered by 65 % of the test takers on average.

The Alpha Cronbach reliability coefficient for this section is : r: 0.959. This value – almost, r: 0.96, indicates a very high level of reliability for this section (Listening & Reading).

In the Cloze Test, out of 20 items, 1 proved to be too difficult and 4 in the range of quite difficult. The overall difficulty level, however, is P: 59.6 (% of average proportion correctly answered). When we consider the recommended range of difficulty for proficiency tests – 40 % - 60 % –, the Cloze Test seems to be within the ideal range. The discrimination power of the

Cloze Test was a little higher than the MC Reading & Listening section, D: 0.65 and 0.55, respectively.

The Alpha Cronbach reliability coefficient for the Cloze Tests was: r: 0.815. For a 20-item rational deletion cloze test, 0.82 is quite a satisfactory level of reliability.

We cannot yet report any reliability indices for the remaining 30% of the test. Manual analysis and interpretation work is almost over on the Dialog and Situations section. Every effort is always made to raise the scoring reliability of these manually scored sections (30 pts.)

IV. CONSTRUCT VALIDITY STUDIES ON JANUARY 2012 EPE

Most of the construct validity studies are based on correlational investigations. 'Correlation', as the term denotes, indicates the extent of relationship between two variables, abilities or two sets of data. For test-takers, this would simply mean if you are good in one ability, you are likely to be good in the other one also.

The magnitude of this relationship is expressed as coefficient of correlation, which varies between r: 0.00 and 1.00. The higher the correlation, the higher the degree of the relationship.

Studies into the predictive validity of EPE in the past few years revealed r's: within the range of .45 and .55 between the EPE scores and the first term GPA's. The correlation between the EPE scores and the ENG 101 term grades was computed to be between r: 0.65 and r: 0.70. The above correlation coefficients provide strong evidence on the role of proficiency in English – as measured in the EPE – on academic success. The degree of this correspondence is certainly higher in the case of ENG 101 and EPE scores. What these coefficients mean is that EPE grades predict future academic success to a reasonable extent.

The constructs in these predictive validity studies are proficiency in English on one hand and academic attainment in all the academic courses (GPA) and performance in an English course, i.e., ENG 101 on the other. The correlation coefficients reported above provide evidence for the validity of EPE.

Further work on the construct validity of EPE involves a breakdown of the exam or rather its components. Correlational work then focuses on the possible existence and degree of relationships between the components studied. What we are doing in such studies is to break

down the overall concept of 'proficiency in English' and try to establish what its constituents are or what skills / abilities make up proficiency in English. This issue is closely related with the question of 'what is each component/section in the exam really measuring?'. It is now commonplace that there is a global proficiency factor underlying all types of language behavior. What this means in practice is that whatever tasks, items or questions we pose to test takers, this global proficiency predicts the rate of success to a certain degree, but only to a certain extent, never wholly in all cases.

When test validators seek answers to the question, 'what is this really testing?', they compare different tests or different parts of the same test.

a) Inter-componential Correlations in Jan. EPE (2012).

One such study involved the computation of correlation coefficients between components and the relationship of each component with the total EPE grade**. Table 5** below displays these correlation coefficients.

TABLE 5

Descriptive Statistics

	Mean	Std. Deviation	N
READING	21.1034	5.4739	880
LISTENIN	20.5182	5.6586	880
NOTETAKI	3.8136	1.0652	880
WRITING	9.9188	3.0909	880
CLOZETES	5.9756	2.0321	880
DIA_SIT	7.1290	2.0424	880
TOTAL	68.4068	16.2494	880

Correlations

	READING	LISTENING	NOTETAKING	WRITING	CLOZETEST	DIA_SIT	TOTAL
READING	1.000	.786	.518	.509	.676	.622	.903
LISTENIN	.786	1.000	.497	.512	.707	.594	.907
NOTETAKI	.518	.497	1.000	.617	.569	.505	.663
WRITING	.509	.512	.617	1.000	.561	.540	.717
CLOZETES	.676	.707	.569	.561	1.000	.627	.819
DIA_SIT	.622	.594	.505	.540	.627	1.000	.759
TOTAL	.903	.907	.663	.717	.819	.759	1.000

** Correlation is significant at the 0.01 level (2-tailed).

The correlation coefficients between components range from r: 0.786 and r: 0.509, the highest relationship (0.786) being between READING and LISTENING and the lowest between WRITING and READING.

These correlation levels between the components are within reasonable limits. Note-Taking and Writing Sections, when examined horizontally in **Table 5** above seem to be displaying lower degrees of correlation, indicating that these two components measure skills and abilities quite different from the other sections.

When we examine the correlation coefficients of each component with the total grade, we see that relatively weaker relationships hold between Note-Taking, Writing and Dialog & Situations Sections and the total grades (r's: 0.667 – 0.759). The components that seem to contribute a great deal to the total EPE grade seem to be LISTENING, READING and the CLOZE TESTS.

In the above analysis, in computing the correlation between a component and the total EPE grade, the grade of that component was included.

b) Correlations between Each Component and the Total Grade in Jan. EPE (2012).

We conducted another analysis in which we subtracted the score of each component from the total EPE grade every time. **Table 6** below displays these correlation values.

TABLE 6

A. READING & TOTAL GRADE

Descriptive Statistics

	Mean	Std. Deviation	N
READING	21.1034	5.4739	880
TOT_REA	47.3034	11.5457	880

Correlations

		READING	TOT_REA
READING	Pearson Correlation	1.000	.797**
	Sig. (2-tailed)	.	.000
	N	880	880
TOT_REA	Pearson Correlation	.797**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).

B. LISTENING & TOTAL GRADE

Descriptive Statistics

	Mean	Std. Deviation	N
LISTENIN	20.5182	5.6586	880
TOT_LIST	47.8886	11.3662	880

Correlations

		LISTENIN	TOT_LIST
LISTENIN	Pearson Correlation	1.000	.799**
	Sig. (2-tailed)	.	.000
	N	880	880
TOT_LIST	Pearson Correlation	.799**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).

C. NOTE-TAKING & TOTAL GRADE

Descriptive Statistics

	Mean	Std. Deviation	N
NOTETAKI	3.8136	1.0652	880
TOT_NT	64.5932	15.5631	880

Correlations

		NOTETAKI	TOT_NT
NOTETAKI	Pearson Correlation	1.000	.624**
	Sig. (2-tailed)	.	.000
	N	880	880
TOT_NT	Pearson Correlation	.624**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).

D. WRITING & TOTAL GRADE

Descriptive Statistics

	Mean	Std. Deviation	N
WRITING	9.9188	3.0909	880
TOT_WRI	58.4881	14.1986	880

Correlations

		WRITING	TOT_WRI
WRITING	Pearson Correlation	1.000	.603**
	Sig. (2-tailed)	.	.000
	N	880	880
TOT_WRI	Pearson Correlation	.603**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).

D. CLOZE TEST & TOTAL GRADE

Descriptive Statistics

	Mean	Std. Deviation	N
CLOZE	5.9756	2.0321	880
TOT_CLZ	62.4313	14.6316	880

Correlations

		CLOZE	TOT_CLZ
CLOZE	Pearson Correlation	1.000	.771**
	Sig. (2-tailed)	.	.000
	N	880	880
TOT_CLZ	Pearson Correlation	.771**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).

E. DIALOG & SITUATION & TOTAL GRADE

Descriptive Statistics

	Mean	Std. Deviation	N
DIA_SIT	7.1290	2.0424	880
TOT_DISI	61.2778	14.7598	880

Correlations

		DIA_SIT	TOT_DISI
DIA_SIT	Pearson Correlation	1.000	.697**
	Sig. (2-tailed)	.	.000
	N	880	880
TOT_DISI	Pearson Correlation	.697**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).

If we were to put in a rank order, the contribution of each component to the total Grade according to these tables, we would get:

1) Listening
2) Reading
3) Cloze Test
4) Dialog & Sits
5) Note-Taking
6) Writing

Correlation coefficients are not meant to be interpreted as percentages but when squared, they are considered as percentages. When we square the coefficients in the **Table 6** above, we would get:

1) Listening _____ 0.6384 %
2) Reading _____ 0.6352 %
3) Cloze Test _____ 0.5944 %
4) Dialog & Sits _____ 0. 4858 %
5) Note-Taking _____ 0.3893 %
6) Writing _____ 0.366 %

We can make a couple of comments on the percentages above. First of all, METU EPE can be said to be more heavily based on text comprehension as evidenced by the fairly high percentages in LISTENING, READING and CLOZE TEST. Secondly, what each percentage means is simply the extent of score information we would get if we gave only that particular component as the whole test: for example, if we were to give only, the LISTENING TEST as the whole EPE, we would be likely to get the same scores, or similar distributions at 63% or 64%. The rest would be missing or liable to chance factors. Since we have no perfect correlation between any two components in the EPE, every component has a role and contribution in the making of the overall construct: language proficiency. If any two sections/components were to display a perfect r:1.00 (100% overlap), we would then have to remove either one of these components. But we don't have any results like this.

One final correlational study on the construct validity of the Jan. 2012 EPE was to investigate the degree of overlap between the receptive (MC Reading & Listening) and productive (Language Use, Note-Taking & Writing) components of the test. **Table 7** and the scatter gram below show the results of this study.

TABLE 7

Receptive: Reading, MC Listening (30+30=60)
Productive: Language Use, Note-taking, Writing (20+5+15=40)

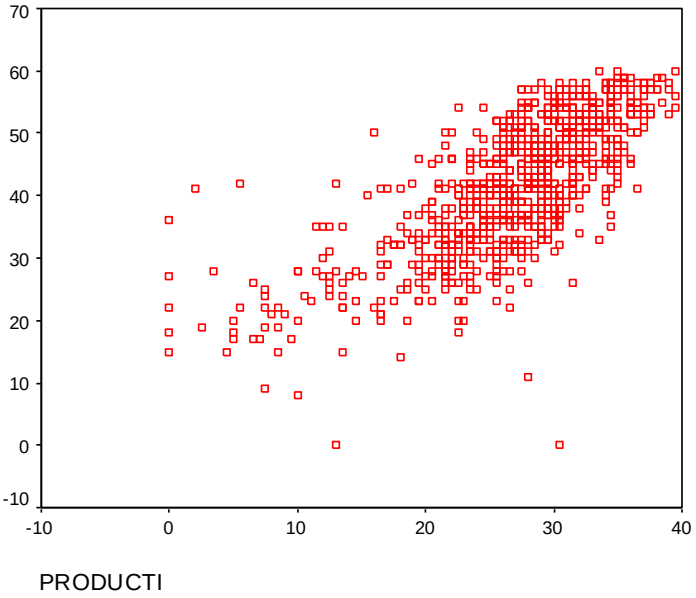
Descriptive Statistics

	Mean	Std. Deviation	N
RECEPTIV	41.6216	10.5189	880
PRODUCTI	26.8369	6.8399	880

Correlations

		RECEPTIV	PRODUCTI
RECEPTIV	Pearson Correlation	1.000	.737**
	Sig. (2-tailed)	.	.000
	N	880	880
PRODUCTI	Pearson Correlation	.737**	1.000
	Sig. (2-tailed)	.000	.
	N	880	880

**. Correlation is significant at the 0.01 level (2-tailed).



When the averages are compared, we see that the MC parts (x:41.62 /60 (69.3 %)) produced a slightly higher average than the production – based parts (x:26.84/40 (67.1%)).

The correlation coefficient between the two parts is r: 0.737. When squared, this corresponds to 54.6%. This means that what can be learned by administering one part instead of both parts would be limited to about 50% of the information provided by the whole test. In other words, these two parts are measuring different abilities, skills and knowledge at a rate of 46% and the common ground between the two is about 54%.

V. CONCLUDING REMARKS

Reliability estimates for the larger portion of the test (70%) are quite high. Construct validity of the whole test seems to have been enhanced. Improvements may be under way in interactivity, authenticity and more favorable backwash effect on instruction.

Slight decreases have been observed in the means of different components, though. Additionally, a drop of 8-10% in the pass rate has been a reason for concern and inquiry. Investigations have revealed that this result largely emanated from two sources:

a) MC Listening and

b) the new version of Language in Use.

However, on the surface, the January 2012 Language Use average was not very revealing. In other words, when compared with previous years, there seemed to be no reason to worry. Further analysis proved, however, that it was the NON-METU population that performed poorly in this component. With more test familiarity and training, performance in this component is likely to increase. Provided that measures are constantly taken to maintain scoring reliability, this new component will contribute to the overall quality of the test.

The second source, MC Listening, which might have impacted the overall scores negatively, deserves more attention in several aspects. Item analysis data over the years must be taken under scrutiny. Perhaps, measures should be taken to approach text selection with more care, considering linguistic and cognitive parameter. Variability and intelligibility in accents is another factor. Reduction of items from 40 to 30 and increasing the weight from 20 to 30 points might be subjected to further statistical tests to gauge the impact on overall scores. In brief, the MC Listening will be under close scrutiny in the upcoming EPEs and we need more data to speak with more certainty.