THE ANALYSIS OF MOCK EXAM (JUNE 2022)

1. INTRODUCTION

The Proficiency Exam at ITU SFL is planned to be improved through certain changes in the Fall Term of 2022-23 Academic Year and a Mock Exam for the New Proficiency Exam was administered in the Spring Term of 2021-22 Academic Year. Related comparative statistics were computed so as to see the success rates and evaluate the reliability and validity of the Mock Exam. The results of the analysis are expected to give insights for the new version of the Proficiency Exam. A randomly selected group of students (B1+-42, B1-51, A2-44)* took the Mock Exam. These students also took the Proficiency Exam (June 2022); however, the total number of students taking the Proficiency Exam was used for comparison.

2. AVERAGES AND SUCCESS RATES

Firstly, for the comparison of averages and success rates of both exams, the arithmetic means of different components and overall averages of the Proficiency Exam 2022 (June) and Mock Exam for the New Proficiency Exam were calculated as can be seen in Table 1 and 2. When means of the different sections in the Proficiency Exam are compared, it is seen that Use of English section produced quite similar results (compare Use of Eng: 11.62 /Cloze Test: 3.51 & Rest: 8.53). Comparing the means of the other sections just by their names may not provide valuable information because the grades allocated to those sections vary across the two exams. Therefore, the other sections should be checked in detail and evaluated by those who design the test in line with their testing objectives.

	Use of English	**Reading**	**Listening**	**Writing**	**S1**	**S2**	**Total**
B1+	11.62	28.417	13.61	12.723	28.417	26.54	66.83
B1	9.39	22.440	7.81	8.349	32.10	16.30	48.26
A2+	8.44	19.836	4.55	5.413	28.52	10.06	38.58
All	10.22	24.654	9.83	9.792	35.13	19.79	54.87

Table 1. Arithmetic means of the different components of the Proficiency Exam 2022 (June)

	Use of English		Reading	Listening		Writing		S1	S2	Total
	Cloze	Rest.		NT	W	Ind.	Int.			
B1+	3.51	8.53	19.93	17.46	7.92	14.21	6.33	32.17	46.15	78.31
B1	3.17	6.97	14.34	14.26	5.26	12.57	4.29	24.80	36.63	61.43
A2+	2.79	5.98	11.64	10.85	4.52	9.595	3.01	20.66	28.23	48.90
All	3.27	7.60	16.69	15.23	6.51	12.81	5.09	27.82	39.89	67.72

*Table 2. Arithmetic means of the different components of the Mock Exam 2022.***

Table 3 below shows the success rates with respect to ITU SFL students who took the Proficiency Exam as well as the students who took the Mock Exam.

	Prof June 2022			Mock Exam		
	Number of sts	Pass	Pass %	Number of sts	Pass	Pass %
B1+	526	449	85.36	42	37	88
B1	449	213	47.43	51	26	50.9
A2	203	62	30.54	44	14	31.8**%**
All	1178	724	61.40			56.20

Table 3. Pass & Fail rates of Proficiency Exam 2022 (June) & Mock Exam

3. RELIABILITY ESTIMATES

Firstly, the Alpha Cronbach reliability coefficient was computed for evaluating the reliability of the test. Also, item analysis was conducted, and for item analysis, item facility and discrimination index were checked for each section of the test.

a. Item Analysis

In general, facility values between 30% and 70% are often considered as being acceptable in language proficiency testing by Bachman (as cited in Green, 2019, p. 23), though those which fall between 20% and 80% are thought to be useful by Green on condition that the items discriminate and contribute to the internal consistency (as cited in Green, 2019, p. 23). The table below can be used in order to decide how difficult a specific item is.

Item Facility =*p*	**Interpretation**
≥85%	Easy
Between 51% & 84%	Moderate
≤50%	Hard

Table 4. Facility values

b. Discrimination Index

The discrimination indexes show how well the items separate the stronger test takers from the weaker ones in a positive or negative way. It is measured on a scale of −1 to +1, and in general the figures are expected to be +.3 and above, though in some circumstances +.25 might be considered acceptable by Henning (as cited in Green, 2019, p. 23). When the discrimination is lower, it means that the item is not discriminating in the desired way. In other words, some of the stronger test takers may have completed an item incorrectly while some of the weaker ones may have answered it correctly. Items with weak or negative discrimination indexes must be reviewed.

Ebel and Frisbie suggest the following table for determining the quality of the items, in terms of the discrimination index which shows the values Discrimination Power (D) and their corresponding interpretation as cited in (Khanal, 2020, p.20).

D=	**Quality**	**Recommendations**
> 0.39	Excellent	Retain
0.30 - 0.39	Good	Possibilities for improvement
0.20 - 0.29	Mediocre	Need to check/review
0.00 - 0.20	Poor	Discard or review in depth
< -0.01	Worst	Definitely discard

Table 5. Discrimination power of the answers according to their D value

The following tables show the facility values and the discrimination index of the items across all sections of the exam. The majority of the items fall in the "difficult" category for all the sections although the facility values are overall accepted values (0,30 & 0,70) except item 37 and item 41 in the Reading section (0,25 and 0,29, respectively). The distribution of the facility values (0,30-0,51) suggests that nearly all the questions in the test are at a high level of difficulty except item 18 and item 20 in the Restatements part of the Use of English section (0,83 and 0.81%, respectively), which could be considered as moderate items. The discrimination powers of the all the sections are quite high, which shows that the items are good at discriminating.

The Alpha Cronbach reliability coefficient for the Cloze Test part of the Use of English section is: r: 0.907, for Restatement part of the Use of English: .987, for Reading: .997, for Listening & Note taking: .940 and While-Listening: .842. These values indicate a very high level of reliability for all sections of the test.

CLOZE TEST

Mean	Median	Mode	St. Dev
4,11	3	1	7,07

	Facility Values %	Discrimination Index
Item 1	0,42	1.00
Item 2	0,47	0,98
Item 3	0,39	0,97
Item 4	0,45	0,35
Item 5	0,45	0,98
Item 6	0,49	0,96
Item 7	0,31	1,00
Item 8	0,31	1,00
Item 9	0,36	0,85
Item 10	0,46	1,00

RESTATEMENTS

Mean	Median	Mode	St. Dev
5,03	4	2	6,36

	Facility Values %	Discrimination Index
Item 11	0,49	0,98
Item 12	0,41	0,90
Item 13	0,49	0,96
Item 14	0,49	1,00
Item 15	0,36	0,96
Item 16	0,38	0,98
Item 17	0,39	0,91
Item 18	0,83	0,55
Item 19	0,38	0,98
Item 20	0,81	0,63

LISTENING & NOTE-TAKING

Mean	Median	Mode	St. Dev
4,31	3	0	3,99

	Facility Values %	Discrimination Index
Item 1	0,44	1
Item 2	0,46	1
Item 3	0,49	1
Item 4	0,41	1
Item 5	0,44	1
Item 6	0,38	1
Item 7	0,40	1
Item 8	0,42	1
Item 9	0,47	1
Item 10	0,40	1

READING

Mean	Median	Mode	St. Dev
10,38	7	1	17,68

	Facility values %	Discrimination Index
Item 21	0,39	1,00
Item 22	0,45	0,94
Item 23	0,34	1,00
Item 24	0,38	0,91
Item 25	0,45	1,00
Item 26	0,38	1,00
Item 27	0,51	0,96
Item 28	0,33	0,95
Item 29	0,46	0,96
Item 30	0,48	0,98
Item 31	0,41	1,00
Item 32	0,44	1,00
Item 33	0,51	1,00
Item 34	0,59	0,54
Item 35	0,35	0,95
Item 36	0,46	0,89
Item 37	0,25	0,97
Item 38	0,41	0,88
Item 39	0,43	0,94
Item 40	0,45	1,00
Item 41	0,29	0,89
Item 42	0,43	0,98
Item 43	0,45	1,00
Item 44	0,31	0,93
Item 45	0,43	0,90

WHILE-LISTENING

Mean	Median	Mode	St. Dev
1,85	1	0	3,54

	Facility Values %	Discrimination Index
Item 1	0,39	1
Item 2	0,40	1
Item 3	0,35	1
Item 4	0,36	1
Item 5	0,35	1

*Table 6. Item Facility Values and Discrimination Indexes.***

4. CONSTRUCT VALIDITY

Construct validity is about how well a test measures the concept it was designed to evaluate. In research studies, measures of related constructs are expected to correlate with one another. If there are two related scales, people who score highly on one scale tend to score highly on the other as well.

The construct validity of the Mock Exam is based on correlational investigations, which shows how much the two variables are in relation with each other. That means the students who are good at one skill are probably good at another one too. This relationship is called *coefficient of correlation*, which is between r: 0.00 and 1.00. When there is a high correlation, the degree of the relationship is high. Correlational analysis for the Mock Exam focuses on the possible existence and degree of relationships between the components of it.

a. Correlations between Components

Table 7 and 8 show computation of correlation coefficients between components and the relationship of each component with the total exam grade.

	Mean	Std. Dev.	N
Cloze Test	3.279	.9072	183
Restatement	7.607	2.1530	183
Reading	16.699	5.7639	183
Listening & Note-taking	15.235	4.9017	183
While-listening	6.514	3.2579	183
Independent writing	12.817	4.3775	183
Integrated writing	5.093	2.9097	183
Session 1 total	27.825	8.0460	183
Session 2 total	39.896	13.0525	183
Total	67.721	20.1670	183

Table 7.Descriptive Statistics

	Cloze	Rest	R	L&NT	W-L	Ind-W	Int-W	S1_T	S2_T	Total
Cloze	1	.523	.582	.510	.394	.524	.428	.671	.563	.632
Rest	.523	1	.735	.670	.490	.557	.637	.850	.703	.794
R	.582	.735	1	.758	.645	.598	.646	.976	.791	.901
L&NT	.510	.670	.758	1	.654	.657	.637	.778	.902	.894
W-L	.394	.490	.645	.654	1	.538	.456	.635	.778	.757
Ind-W	.524	.557	.598	.657	.538	1	.649	.637	.862	.812
Int-W	.428	.637	.646	.637	.456	.649	1	.681	.795	.786
S1_T	.671	.850	.976	.778	.635	.637	.681	1	.817	.928
S2_T	.563	.703	.791	.902	.778	.862	.795	.817	1	.973
Total	.632	.794	.901	.894	.757	.812	.786	.928	.973	1

*Table 8. Correlations of Each Section of Mock Exam with Each Other.****

The correlation coefficients between components range from r: 0.758 and r: 0.394, the highest relationship (0.758) being between Reading and Listening & Note-Taking, and the lowest between While-Listening and Cloze Test.

When the correlation coefficients of Session 1 and 2 are examined, it is seen that there is a good correlation level (0.817). As for the correlation between each component with the total grade, a relatively weaker relationship is seen between Cloze Test and the total grade (.632). The components that seem to contribute a great deal to the total grade seem to be Reading, Listening Note Taking and Independent Writing.

b. Correlations of the Components and the Total Grade for the Mock Exam

Table 6 below displays the correlation values of each section and the total grade from which the score of that specific component was subtracted.

**P1, P2, P3, P4 levels were previously called A1, A2, B1 and B1+ successively.*

*** After the feedback, some changes were made in both the content and number of questions in the exam; therefore, the question numbers in this report may not match up with the ones in the sample exam.*

****The exact names of the sections can be seen in Table 7.*

Cloze Test & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
ClozeTest	3.28	.907	183
Total_Cloze	64.443	19.6063	183

Correlations

		ClozeTest	Total_ClozeTest
ClozeTest	Pearson Correlation	1	.604**
	Sig. (2-tailed)		<.001
	N	183	183
Total_ClozeTest	Pearson Correlation	.604**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

Restatement & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
Restatement	7.61	2.153	183
Total_Restatement	60.11	18.503	183

Correlations

		Restatement	Total_Restatement
Restatement	Pearson Correlation	1	.749**
	Sig. (2-tailed)		<.001
	N	183	183
Total_Restatement	Pearson Correlation	.749**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

Reading & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
Reading	16.70	5.764	183
Total_Reading	51.02	15.179	183

Correlations

		Reading	Total_Reading
Reading	Pearson Correlation	1	.818**
	Sig. (2-tailed)		<.001
	N	183	183
Total_Reading	Pearson Correlation	.818**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

Note-Taking & Listening & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
Listening & Note-taking	15.23	4.902	183
Total_ Listening & Note-taking	52.49	15.937	183

Correlations

		Listening & Note-taking	Total_ Listening & Note-taking
Listening & Note-taking	Pearson Correlation	1	.824**
	Sig. (2-tailed)		<.001
	N	183	183
Total_ Listening & Note-taking	Pearson Correlation	.824**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

While- Listening & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
While- Listening	6.51	3.258	183
Total_While- Listening	61.21	17.830	183

Correlations

		While- Listening	Total_ While- Listening
While- Listening	Pearson Correlation	1	.673**
	Sig. (2-tailed)		<.001
	N	183	183
Total_ While- Listening	Pearson Correlation	.673**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

Independent Writing & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
Independent Writing	12.82	4.378	183
Total_ Independent Writing	54.904	16.8084	183

Correlations

		Independent Writing	Total_ Independent Writing
Independent Writing	Pearson Correlation	1	.714**
	Sig. (2-tailed)		<.001
	N	183	183
Total_ Independent Writing	Pearson Correlation	.714**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

Integrated Writing & Total Grade

Descriptive Statistics

	Mean	Std. Deviation	N
Integrated Writing	5.09	2.910	183
Total_ Integrated Writing	62.628	17.9699	183

Correlations

		Integrated Writing	**Total_ Integrated Writing**
Integrated Writing	Pearson Correlation	1	.720**
	Sig. (2-tailed)		<.001
	N	183	183
Total_ Integrated Writing	Pearson Correlation	.720**	1
	Sig. (2-tailed)	<.001	
	N	183	183

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

5. CONCLUSION

The reliability estimates of the Mock Exam show that although nearly all questions are in the accepted range, many of the items fall in the range of 'difficult'. This may have been caused by the number of students who did not answer many questions, or simply made a random choice. On the other hand, pass/fail rates of the Mock Exam seem in line with the Prof 2022 (June). For all sections of the test, an expected level of correlation is observed, which shows a high construct validity of the test.

References

Green, R. (2019). Statistical analyses for language testers. In V. Aryadoust & M Requel (Eds.). *Test development Reliability and generalizability (pp. 15-29). Routledge.*

Khanal, P. (2020). Key considerations in test construction, scoring and analysis: A guide to pre-service and in-service teachers. *International Journal of Research*, *9*(5), 15-24.