Is My B2 the Same as Your B1?

Comparing Language Examinations' Levels

Zoltán Kiszely and Gábor Szabó
Hungarian Accreditation Board for Foreign Language Examinations

Introduction

Owing to the local legal requirements, nationally accredited language exam providers in Hungary had to align their exams' levels with those of the CEFR. The Hungarian Accreditation Board for Foreign Language Examinations (HABFLE) prescribed a fairly rigorous procedure (based on the Manual) for exam providers, intending to guarantee that various exams labelled B1, B2 or C1 represent the same CEFR levels. While performing the alignment theoretically results in different exams measuring the same levels, so far no empirical studies have compared the various examinations in this regard directly. HABFLE launched this project in order to gain empirical information on how similar actual exam results are. In other words, to see whether a supposed B2 exam is more like a B1 or a C1 in another system.

Design

The project included the study of the written components of five B2 exams in English and four in German. 20 candidates took the five English exams, while 16 candidates took the German ones. All candidates enrolled in actual live exams over a period of three months. It was assumed that their abilities would not change significantly during this time. For comparability reasons exam results were then converted into percentage figures and were to be analyzed in the following ways:

1. Correlation analyses of candidate performances on the whole of the exam as well as on corresponding subtests;

2. Statistical comparisons (Wilcoxon Signed Ranks Tests) of actual candidate performances on the whole of the exam as well as on corresponding subtests;

3. IRT analyses of tests linked through common test takers; comparison of item difficulty logits (ANOVA)

Results

Exams coded 1 to 5 in English, 1 to 4 in German

1. Correlation analyses of candidate performances on the exams

Whole tests

English

1-2	0,788
1-3	0,917
1-4	0,902
1-5	0,843
2-3	0,829
2-4	0,811
2-5	0,815
3-4	0,967
3-5	0,902
4-5	0,886

German

1-2	0,819
1-3	0,847
1-4	0,820
2-3	0,736
2-4	0,874
3-4	0,710

All correlations significant at p<0.05

Reading

English

1-2	**0,489**
1-3	0,730
1-4	**0,662**
1-5	**0,465**
2-3	0,795
2-4	0,721
2-5	**0,661**
3-4	0,830
3-5	**0,632**
4-5	0,702

German

1-2	0,795
1-3	0,808
1-4	**0,258**
2-3	0,865
2-4	**0,322**
3-4	**0,244**

All correlations significant at p<0.05

Writing

English

1-2	0,721
1-3	0,762
1-4	**0,679**
1-5	0,763
2-3	**0,547**
2-4	**0,682**
2-5	**0,672**
3-4	**0,692**
3-5	0,856
4-5	0,776

German

1-2	**0,503**
1-3	**0,647**
1-4	**0,660**
2-3	**0,696**
2-4	**0,666**
3-4	0,715

All correlations significant at p<0.05

Grammar

English

1-2	**0,697**
1-3	**0,619**
1-4	0,783
1-5	**0,661**
2-3	0,803
2-4	**0,683**
2-5	**0,625**
3-4	0,884
3-5	0,723
4-5	0,732

All correlations significant at p<0.05

2. Statistical comparisons (Wilcoxon Signed Ranks Tests) of actual candidate performances on exams

Whole tests

English

	2 - 1	3 - 1	4 - 1	5 - 1	3 - 2	4 - 2	5 - 2	4 - 3	5 - 3	5 - 4
Z	-3,744	-2,828	-3,443	-3,280	-1,887	-0,785	-0,981	-1,479	-1,333	-0,305
Asymp. Sig. (2-tailed)	**0,000**	**0,005**	**0,001**	**0,001**	0,059	0,433	0,327	0,139	0,183	0,760

German

	2 - 1	3 - 1	4 - 1	3 - 2	4 - 2	4 - 3
Z	-0,848	-1,015	-0,057	-0,138	-0,315	-1,414
Asymp. Sig. (2-tailed)	0,396	0,310	0,955	0,890	0,752	0,158

Reading

English

	2 - 1	3 - 1	4 - 1	5 - 1	3 - 2	4 - 2	5 - 2	4 - 3	5 - 3	5 - 4
Z	-3,323	-1,923	-2,859	-1,771	-2,669	-0,893	-2,810	-1,957	-0,078	-1,329
Asymp. Sig. (2-tailed)	**0,001**	0,054	**0,004**	0,077	**0,008**	0,372	**0,005**	0,050	0,937	0,184

German

	2 - 1	3 - 1	4 - 1	3 - 2	4 - 2	4 - 3
Z	-2,384	-0,739	-1,721	-2,197	-1,016	-0,880
Asymp. Sig. (2-tailed)	**0,017**	0,460	0,085	**0,028**	0,310	0,379

Writing

English

	2 - 1	3 - 1	4 - 1	5 - 1	3 - 2	4 - 2	5 - 2	4 - 3	5 - 3	5 - 4
Z	-3,501	-2,598	-1,873	-3,269	-1,601	-0,915	-1,711	-1,295	-0,785	-2,227
Asymp. Sig. (2-tailed)	**0,000**	**0,009**	0,061	**0,001**	0,109	0,360	0,087	0,195	0,432	**0,026**

German

	2 - 1	3 - 1	4 - 1	3 - 2	4 - 2	4 - 3
Z	-1,521	-1,648	-0,157	-2,205	-1,947	-1,363
Asymp. Sig. (2-tailed)	0,128	0,099	0,875	**0,027**	0,051	0,173

Grammar

English

	2 - 1	3 - 1	4 - 1	5 - 1	3 - 2	4 - 2	5 - 2	4 - 3	5 - 3	5 - 4
Z	-3,920	-2,972	-3,829	-3,399	-2,702	-0,479	-1,449	-3,188	-1,256	-0,501
Asymp. Sig. (2-tailed)	**0,000**	**0,003**	**0,000**	**0,001**	**0,007**	0,632	0,147	**0,001**	0,209	0,616

3. Comparison of item difficulty logits (ANOVA)

Reading, English, 3 exams (coded 1-3)

Descriptives

Exam ID	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
1	2,062	1,319	0,590	0,424	3,700	0,610	4,080
2	0,002	1,108	0,222	-0,456	0,459	-2,630	1,680
3	0,464	0,891	0,230	-0,030	0,958	-1,110	2,220
Total	0,385	1,219	0,182	0,018	0,751	-2,630	4,080

Post-hoc tests

	(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	1	2	2,060	0,521	**0,001**	0,737	3,383
		3	1,598	0,550	**0,021**	0,203	2,993
	2	1	-2,060	0,521	**0,001**	-3,383	-0,737
		3	-0,462	0,348	0,420	-1,344	0,420
	3	1	-1,598	0,550	**0,021**	-2,993	-0,203
		2	0,462	0,348	0,420	-0,420	1,344

Conclusions

Owing to the relatively small sample size, conclusions need to be handled with caution. Few clear patterns emerge though, some of which are the following:

- comparisons of candidate performances seem to imply that certain exams show significant differences;
- item difficulty comparisons indicate that the English reading tasks in Exam 1 were significantly more difficult than those in some other exams;
- English grammar tasks appear to be more difficult in Exam 1;
- English Exam 1, on the whole, seems to yield lower results, even though the tendencies in the results are similar to those in other exams;
- German writing exams seem to yield results showing little harmony across exams, even if average candidate performances do not seem to indicate major differences.

As can be inferred, the most important indication is that differences seem to exist, but more research is needed to clarify how great and how consistent such differences may be.