The German Language Worldwide: Data – Course Participation Abroad (Technical Description and Codebook)[*]

Silke Uebelmesser — University of Jena and CESifo

Matthias Huber[†] — University of Jena

Severin Weingarten — University of Jena

March 26, 2018

When using this data set, please always cite:

The official data presentation article

Silke Uebelmesser, Matthias Huber, and Severin Weingarten (2018). “The German Language Worldwide: a New Data Set on Language Learning”. *CESifo Economic Studies* 64 (1), 103–121.

and the data repository

Silke Uebelmesser, Matthias Huber, and Severin Weingarten (2018). “The German Language Worldwide: Data – Course Participation Abroad”. *Harvard Dataverse*, V1. doi: 10.7910/DVN/XVNUY8

[*] This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, UE 124/2-1) – 270886786.

[†] Corresponding author: matthias.huber@uni-jena.de

0 Background

We provide data from the Goethe Institut (GI), a German cultural association with institutes worldwide and in Germany. Since 1965, the GI has continuously published annual reports in which activities of each institute including data about language course and exam participation are reported. These annual reports are publicly available. We digitised this information in order to construct three datasets[1]. In the following, the technical description and the codebook for the dataset about language course and exam participation at institutes worldwide (Uebelmesser, Huber, and Weingarten 2018b) are presented. From the annual reports of the GI, we construct three variables for different time periods. First, from 1986 to 2014 the GI reported for each institute and year the number of participants in standardised exams ("zentrale Prüfungen") which are widely recognised, e.g. for language requirements in universities. Furthermore, there is information on course participation by two variables per institute and year: the number of registrations in language courses for the period 1990 to 2014 and an indicator, which we construct, for sold course units for the periods 1972 to 1989 and 1997 to 2014. Total course and exam participation at institutes worldwide reached their maximum for all three variables in 2014 with 287,630 exams, 229,702 registrations and 17,113,040 sold course units.

1 Technical description

The dataset only provides observations for institutes for which at least one of the three variables, i.e. exams, registrations or sold course units, has a positive entry. Each variable is available for different time periods. Therefore, the dataset is an unbalanced dataset on the city-level from 1972 to 2014.[2] In the following, we describe the variables in more detail.

[1] See Uebelmesser, Huber, and Weingarten (2018c,d) for the technical description of the other two datasets, and Uebelmesser, Huber, and Weingarten (2018a) for a comprehensive presentation of the data.

[2] The assignment of the cities to the standardised 3-letter country codes by ISO 3166 alpha-3 to existing countries (and the 4-letter codes provided by ISO 3166-3 for Yugoslavia) did not require any adjustments.

1.1 Exam participation

From 1986 onwards, the GI reported numbers for exam participation per year and institute. The GI offers different types of standardised exams ("zentrale Prüfungen"), which can be categorised in language exams for adults, children and adolescents, and for professional use and which are further differentiated by the level of language skills. Variation and differentiation of the exams have increased over the years. Only in very few years numbers were reported separately for each type of exams, while for most years we only have aggregate information. Hence, we only report aggregate numbers for exam participation per year and institute in the variable "exams".

1.2 Registrations

The variable "registrations" is the first indicator for course participation and contains the number of registrations for language courses per year and institute for the years 1990 to 2014. Courses are organised in course periods, mostly two periods (semesters) or three periods (trimesters) per year.

While from 2000 onwards the annual number of registrations are reported in the annual reports, in the years 1990 to 1999 only average numbers of students per course period were published. For this period, we construct the number of registrations per year by multiplying the average number of students per course period with the number of course periods.

1.3 Sold course units

A further indicator for language course participation at the GI is the number of sold course units per year and institute ("units_sold"). We construct this measure for the years 1972 to 1989 and 1997 to 2014 as follows:

$$sold\,course\,units\;=\;total\,number\,of\,lecture\,units\;*\;average\,course\,size \tag{1}$$

where the number of lecture units[3] is the sum of the units all teachers taught at an institute within a year. That variable was reported in the year 2006 and from 2009 onwards. For the years 1972 to 1989 and the remainder of the years between 1997 and 2014, we calculate the variable with equation (1), where the average course size is the number of students divided by the number of courses.

1.4 Joint reporting

In some cases the GI reported numbers of course and exam participation jointly for two (or more) institutes. If this is the case, it is not clear whether not-first-named institutes offered language services.[4] If so, we do not know whether the numbers are included in the jointly reported numbers nor the relative size of the jointly reported institutes.

In this dataset, we constructed five flag variables which indicate different cases of joint reporting[5] which we describe in more detail:

- "jr_case1": The dummy variable indicates whether the institute in the respective year has reported jointly with one or more other institutes for which numbers on course and exam participation have never been reported separately. Additionally, we also flag years without joint reporting where in the years before and afterwards the case just described was observed. The flagging of these gaps is a conservative way of preventing contamination of the data due to misreporting, as it could be the case that numbers are for more than one institute also in the gap year(s). E.g., the annual reports present participation numbers for London jointly with Birmingham in the years 1983, 1984 and 1986 to 1989. We flag the observations for London in the years of joint reporting as well as the observation of 1985. Birmingham cannot be found in any year in this dataset on course and exam participation, but only in the dataset on the presence of the GI.

[3] A lecture unit has 45 minutes.

[4] In this dataset, you can find only first-named institutes. For information on not-first-named institutes please refer to Uebelmesser, Huber, and Weingarten (2018d).

[5] Some observations are flagged with more cases, as joint reporting could refer to more than one institute.

- “jr_case2”: The dummy variable flags observations with cases of joint reporting where for the not-first-named institute numbers on course and exam participation are reported separately after a period of joint reporting. Between these two periods is a gap, where there is no information on course and exam participation for the not-first-named institute during joint reporting. E.g. Tehran reported numbers on sold course units jointly with Shiraz in 1975 and 1976, but from 1978 Shiraz reported separately. However, it is not clear, what happened in 1977. Therefore, we flag observations of Tehran for 1975 to 1977.

- “jr_case3”: The dummy variable flags observations with cases of joint reporting where for the not-first-named institute numbers on course and exam participation are reported separately directly after a period of joint reporting. The difference to the second case of joint reporting is that there is no gap between joint and separate reporting. E.g. numbers for Barcelona are reported jointly with Zaragoza in the years 1972 to 1977, while from 1978 onwards numbers for Zaragoza are reported separately. Therefore, we flag observations for Barcelona in the years 1972 to 1977. The difference to the second case of joint reporting is that in the third case it is more likely that language courses also took place in the joint reporting period, while in the second case the gap might indicate that during the joint reporting period no language courses were offered.

- “jr_case4”: The dummy variable flags observations with cases of joint reporting where for the not-first-named institute numbers on course and exam participation are reported first separately, followed by a period of joint reporting before numbers are separately reported again. In these cases the numbers of course and exam participation when jointly reported (roughly) match the sum of the numbers with separate reporting. E.g before 2002, numbers of course and exam participation in Porto are reported separately from Lisbon, while from 2002 to 2006 the GI reports the numbers jointly. In the numbers of course registrations there is an obvious change due to the aggregation: the numbers of registrations for Porto (972 registrations) and Lisbon (2 010 registrations) in 2001 add up approximately to the jointly reported numbers in 2002 (2 804 registrations). The same pattern holds when the numbers are reported separately again: 1395 registrations in Lisbon and 728 in Porto in 2007 add up approximately to 2 158 registrations reported for both institutes jointly in 2006.

- "jr_case5": The dummy variable flags observations with cases of joint reporting where for the not-first-named institute numbers on course and exam participation are reported first separately, followed by a period of joint reporting before numbers are separately reported again. Different from the fourth case of joint reporting, the numbers with separate reporting do not add up to the numbers with joint reporting and the changes seem to be more complex. One reason might be that more institutes are involved in the joint reporting with several changes.

2 Codebook

region	Goethe-Institut region			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	1	Central and Eastern Europe	143	2.86
	2	East Asia	246	4.92
	3	Eastern Europe and Central Asia	138	2.76
	4	North America	424	8.47
	5	North Africa and Middle East	475	9.49
	6	Northwest Europe	478	9.55
	7	South America	674	13.47
	8	Southeast Asia, Australia and New Zealand	402	8.03
	9	South Asia	499	9.97
	10	Southeast Europe	381	7.61
	11	Sub-Saharan Africa	411	8.21
	12	Southwest Europe	733	14.65
Missing	0/5,004			

country	Country of origin
Format	Character
Comment	ISO 3166 alpha-3 for currently existing countries
	ISO 3166-3 for dissolved countries
Missing	0/5,004

city	Name of city with institute
Format	Character
Comment	City names
Missing	0/5,004

year	Year
Format	Numeric
Range	1972 – 2014
Missing	0/5,004

exams	Number of exams		
Format	Numeric		
Range	1 – 33,900		
	Mean	Median	Std. Dev
	580.91	110	1,930.12
Missing	1,679/5,004		

registrations	Number of registrations to language courses		
Format	Numeric		
Range	2 – 8,824		
	Mean	Median	Std. Dev
	1,475.06	1,036	1,318.85
Missing	2,185/5,004		

units_sold	Number of sold course units		
Format	Numeric		
Range	20 – 828,300		
	Mean	Median	Std. Dev
	100,582	67,469.8	97,253.6
Missing	990/5,004		

jr_case1	Joint reporting case 1			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No	4,846	96.99
	1	Yes	158	3.01
Missing	0/5,004			

jr_case2	Joint reporting case 2			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No	4,951	98.94
	1	Yes	53	1.06
Missing	0/5,004			

jr_case3	Joint reporting case 3			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No	4,991	99.97
	1	Yes	13	0.03
Missing	0/5,004			

jr_case4	Joint reporting case 4			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No	4,991	99.96
	1	Yes	21	0.04
Missing	0/5,004			

jr_case5	Joint reporting case 5			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No	4,846	97.48
	1	Yes	126	2.52
Missing	0/5,004			

References

Uebelmesser, Silke, Matthias Huber, and Severin Weingarten (2018a). “The German Language Worldwide: a New Data Set on Language Learning”. *CESifo Economic Studies* 64 (1), 103—121.

— (2018b). “The German Language Worldwide: Data – Course Participation Abroad”. *Harvard Dataverse,* V1. DOI: 10.7910/DVN/XVNUY8.

— (2018c). “The German Language Worldwide: Data – Course Participation in Germany”. *Harvard Dataverse,* V1. DOI: 10.7910/DVN/V5BWMT.

— (2018d). “The German Language Worldwide: Data – Language Learning Opportunities Abroad”. *Harvard Dataverse,* V1. DOI: 10.7910/DVN/VSQBM7.