Analysis of Statistical Methods on Plurilinguistic Quantitative Data

T Krasnopeeva¹ and M Gromov²

¹ Department of Foreign Languages, Tomsk State University, Russia
² Radiophysics Department, Tomsk State University, Russia

E-mail: tatyllan4@gmail.com, maxim.leo.gromov@gmail.com

Abstract. This multidisciplinary study deals with a population cluster of migrant multilinguals as a complex phenomenon of Social Physics with the help of the mathematical-statistical tools. It enlightens two statistical methods used to analyse the quantitative data on the plurilingual competence of the university students: the Rasch analysis and the T-test. These methods are able to give a more profound picture of the individual migrants’ plurilingual ability in terms of migration as a sociophysical phenomenon. We consider the strengths that each method has as the physical values in order to understand the linguistic behavior of migrant flow as a complex matter and each migrant student as a part of this flow. Our multidisciplinary study empirically investigates whether the positive impact of the statistical methods is relevant by using recent dataset on the example of the multilinguals’ social cluster. The main conclusion of the study shows the positive impact of the Rasch model in measurement of complex sociophysical matters with respect to the ability of a migrant social cluster to the plurilingualism, and results in the statement that this tool is capable to decipher the empirical data in a qualitative manner, in particular, for its application to the group educational trajectories. However, T-test showed the statistical insignificance of the current data set if this data set is used to predict further development of plurilingual competence.

1. Introduction

The statistical methods play a crucial role in absolutely different fields of science from Social Physics to Cognitive Linguistics, and their continuous enhancement has alleviated the insufficient accuracy of the measuring procedures in the multidisciplinary studies. In order to understand the role of these methods for the migrant students’ flow and select the most appropriate of them to measure the plurilingual ability as a social integration characteristic, we carry this study out. The plurilingual ability is essential for the migrant students’ population cluster because the educational institutions such as the university “shape the social atmosphere of the territory, providing their cities with social and intellectual potentials, creating social ladders and new patterns of communication” [1]. T-test in our research serves to discover the students’ diachronic aspect of the plurilingual competence development. The Rasch analysis allows to work out the individualized solutions through the analysis of the whole group, and contributes to the development of the range of methods applied for the evaluation of migrant students’ social cluster linguistic behavior.
2. Research questions
1. What are the mechanisms of the Rasch analysis and the T-test, and which of two is preferable for the evaluation of migrant linguistic behavior as a complex sociophysical matter?
2. Is it possible to construct the group educational trajectories with the help of the methods described?
3. Is it possible to use these methods for the analysis of the plurilingual competence applied to the study of the migrants’ plurilingual ability concerned as a complex physical characteristic?

3. Research methods

3.1. Questionnaire and Data
This study is quantitative in nature with a questionnaire survey as the basis of the research. Participants are 18-23 aged multilinguals speaking three languages each. They fill in the questionnaire on the individual characteristics that influenced their linguistic profile. The questionnaire contains 25 questions of three kinds: social background including personal data (name, gender, parents’ income etc.), level of English as the second language and plurilingual ability, i.e. the ability to switch between the languages. We assume that these three factors influence the students’ plurilingual competence formation. Although one can argue that some personal data may be useful, we have decided to exclude this data from the analysis. The questions are shown in Table 1.

Table 1. The questions of the questionnaire (q4–q10 – Social background; q13–q14 – English as L2 level; q14–q25 – Plurilingual ability)

| No. | Question | Options |
|-----|----------|---------|
| q4  | How old are you? | (17-20 y.o.) — 1; (20-25 y.o.) — 0 |
| q5  | Do your parents have higher education? | yes — 1; no — 0; |
| q6  | The income of your family is… | (>$ 500 euro) — 1; <$ 500 euro) — 0 |
| q7  | When you were a child, did you have an access to the gadgets, Internet or books? | yes — 1; no — 0; |
| q8  | Do you have a full family? (Mother and father) | yes — 1; no — 0; |
| q9  | Do you have any brothers or sisters? | yes — 1; no — 0; |
| q10 | Do your parents work? | yes — 1; no — 0; |
| q11 | When did you start learning the foreign language (Eng)? | (earlier than 12 y.o.) — 1; (later than 12 y.o.) — 0 |
| q12 | How often do you use the English language? | (< 3 times a week) — 0; (> 3 times a week) — 1 |
| q13 | What was the sequence of language acquisition? | simultaneously — 1; one by one — 0 |
| q14 | What language do you use more often today? | (mother tongue) — 0; (foreign language) — 1 |
| q15 | Does it happen that you use several languages at the same time when you speak? | yes — 1; no — 0; |
| q16 | How many years do you study foreign language? | < 5 — 0; > 5 — 1 |
| q17 | Did you learn the foreign language through some activities or naturally? | (through extracurricular activities) — 0; (naturally) — 1 |
| q18 | Are there any situations when you mediate between two people with no common language? | yes — 1; no — 0; |
| q19 | Do you use the foreign language nowadays for studies or work? | yes — 1; no — 0; |
| q20 | Can you while speaking on one language understand the person answering you on the other? | yes — 1; no — 0; |
| q21 | Are there any situations when you use several foreign languages to understand the text fully? | yes — 1; no — 0; |
| q22 | Can you recognize the foreign words in the international source (e.g. the word “superconductor” in sciences)? | yes — 1; no — 0; |
| q23 | Do you use loan words or non-verbal means to express your thought (mime, gesture, facial expression)? | yes — 1; no — 0; |

All questions are dichotomous (with two possible answers). Answers are marked as increasing plurilingual ability (mark 1) or non-increasing (mark 0). The marks are assigned according to our teaching experience. In this research we consider the teaching experience as non-ambiguous absolute factor. The questionnaire was given to 18 students. Their answers are shown in Table 2.

Table 2. The students’ answers

| q4 | q5 | q6 | q7 | q8 | q9 | q10 | q13 | q14 | q15 | q16 | q17 | q18 | q19 | q20 | q21 | q22 | q23 | q24 | q25 | grds |
|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|    |    |    |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

2
The first column in Table 2 refers to the students. The inner columns contain encoded answers according to our evaluation of the possible answers. The last column (named grds – grades) contains English language study grades that students got in the previous semester. These grades are used as evaluation of the students’ English language level. In this paper the following statement is postulated. The level of only one factor, the level of foreign language, in our example this is the English language studied by the students at the university as the second language, does not absolutely correlate with their plurilingual ability. The hypothesis is proved right according to the results obtained with the help of statistical analyses of the Rasch model and T-test. On the example of this plurilingual dataset, we can answer the three questions of our research concerning the Rasch analysis and the T-test, the effect resulting from their combination and the future prospects for the individual educational trajectories construction using these methods.

3.2. Analysis
In order to answer the first question concerning the mechanisms of statistical methods’ work, we will demonstrate them using the dataset described above. We consider two methods to analyze the data: the Rasch model and the T-test. The aim of use of the Rasch model is to rank the questions of the questionnaire and to rank the students according to the questions’ complexity. The T-test is used to see if there is a correlation between the English grades of students and their plurilingual ability that is implicated in their answers to the questions.

Before considering the results of the analysis, we have to make some notes on the Rasch model.

The Rasch model is usually explained as a probability to give the right answer to a question depending on the difference between student’s abilities and question’s complexity (in the paper this difference is referred as the Latent Dimension since it is not available for direct measurement). The more abilities a student has, the higher the probability to give the right answer is. The more complex the question is, the lower the probability to give the right answer is. However, in our example we do not
suppose that this or that answer is wrong or right. They are just different. For that reason, we suggest to look at the Rasch model from a bit different angle.

The questions of the questionnaire concern certain properties of a student’s background. If the background has these properties, then it is expected that the student has a higher level of the plurilingual ability. The properties are not equally valuable.

So, instead of question complexity the value of the property associated with the question is considered in this paper. Then the Latent Dimension is the difference between the student’s level of plurilingual ability and the value of the property associated with the certain question. And the Rasch model is interpreted as the probability to possess the property, associated with the question, depending on the Latent Dimension.

3.2.1. The Rasch Model. First of all, it can be seen, that the questions q10 and q21 do not bring anything into analysis because all students gave the same answers to these questions. For that reason, columns q10 and q21 were excluded from the data set. The Rasch Model was built in R using eRm module [2]. To evaluate the model fit the Andersen’s Likelihood Ratio test (LR-test) was used. When the data set contained items for the questions q7, q13, q22 and q25, the LR-test gave warnings that these columns were excluded due to inappropriate response patterns within subgroups (both for the «median» split criterium and the «mean» split criterium). This happened because LR-test was not able to split the students into the adequate subgroups to check the quality of question (some of the subgroups were too small or empty). That’s why the data set for the final analysis did not contain the items q7, q13, q22 and q25.

After refinement the data set the LR-test (split criterium — «median») ran without warnings and gave the following results: LR-value — 6.91, p-value — 0.94 and degree of freedom of chi-squared — 14. The Item Characteristic Curves (ICC) are shown in Figure 1. Some curves coincide. Let’s list the curves from top to bottom: the 1st curve is for the item q8; the 2nd — for the item q20; the 3rd — q15, q6, q17 and q5; the 4th — q23; the 5th — q18; 6th — q9 and q4; the 7th — q14 and q24; 8th — q19; 9th — q16. The horizontal axis corresponds to the Latent Dimension (i.e. the difference between the student’s level of plurilingual ability and the value of the property associated with the certain question). The vertical axis corresponds to the probability to have the property associated with the certain question. For example, the question q9 is associated with the property “To have a brother or a sister”. When a student has the level of plurilingual ability equal to the value of the property q9, then the probability that this student actually has a brother or a sister equals 0.5.

![Figure 1. The Item Characteristic Curves of the Built Rasch Model](image)

In Table 3 the ranking in descendant order of the questions’ value according to the Rasch Model is given.
Table 3. The Ranking of the Questions by the Rasch Model

| question | value | question | value | question | value | question | value |
|----------|-------|----------|-------|----------|-------|----------|-------|
| q16      | 1.54  | q4       | 0.27  | q5       | -0.45 | q20      | -0.72 |
| q19      | 0.98  | q9       | 0.27  | q17      | -0.45 | q8       | -1.86 |
| q24      | 0.73  | q18      | 0.04  | q6       | -0.45 |          |       |
| q14      | 0.73  | q23      | -0.20 | q15      | -0.45 |          |       |

In Table 4(a) the ranking in descendant order of the students’ plurilingual ability according to the Rasch model is given. For comparison in Table 4(a) the ranking of students according to the English language study grades is given.

Table 4. Ranks of Students According to (a) the Rasch Model; (b) the English Language as L2 Study Assessments

| student | evaluation | student | evaluation | student | assessment | student | assessment |
|---------|------------|---------|------------|---------|------------|---------|------------|
| s11     | 1.47       | s14     | 0.33       | s3      | 5          | s8      | 4          |
| s12     | 1.05       | s1      | 0.01       | s5      | 5          | s11     | 4          |
| s15     | 1.05       | s6      | 0.01       | s14     | 5          | s13     | 4          |
| s16     | 1.05       | s3      | -0.32      | s17     | 5          | s21     | 4          |
| s2      | 0.68       | s4      | -0.32      | s18     | 5          | s6      | 3          |
| s5      | 0.68       | s7      | -0.32      | s1      | 4          | s12     | 3          |
| s8      | 0.68       | s13     | -0.67      | s2      | 4          | s16     | 3          |
| s9      | 0.68       | s17     | -0.67      | s4      | 4          | s19     | 3          |
| s10     | 0.68       | s18     | -1.04      | s7      | 4          | s20     | 3          |

It is clear that these two rankings are even not close. The extreme examples could be the student s18 and the student s12. Being ranked by the Rasch model as the lowest, the student s18 got the highest rank after the assessment. On the contrary, the student s12 is ranked very high by the Rasch Model but the lowest assessment afterwards. It indicates that the ability to the plurilingual competence of the students and their English level do not correlate directly (due to many different reasons one of which can be the overall attitude to the study).

3.2.2. The T-test. Welch two sample T-test is used to evaluate if there is a difference in mean values of a parameter of interest for two different populations one of which has a given property and another one has not. In our case populations are students, the given property is the property associated with this or that question and the parameter of interest is the grades, i.e. we would like compare the mean value of the grades when q4 = 0 against the mean value of the grades when q4 = 1. And this way we one for every question. The T-test assumes that the data for estimates of both mean values are normally distributed. It is the usual problem for small data sets such as we have in our research. And it is easy to check (we did it in R using the Shapiro-Wilk test) that our data do not follow the restriction whatever question is considered. However the danger here is to reject the null-hypothesis in more than 5% of cases when it is actually holds (just as a reminder – the null hypothesis for the T-test is that there is no significant difference between two mean values). Knowing that, we would be very careful in conclusions if the T-test showed the high predictive power of our data set.

The T-test was done for all items (questions). Only answers on q8 (t-value = 6.26 and p-value << 0) and q18 (t-value = -3.09 and p-value = 0.007) showed to be statistically significant (and again, keeping
in mind, that Shapiro-Wilk test failed). This means that using the data set of this paper, the language study grades based on the level of plurilingual ability of a student cannot be predicted (and therefore the development of the ability cannot be predicted). However, we believe that combining different items (e.g. q8 with q18), one can give such a prediction, using Analysis Of Variance (ANOVA) and regressions. But for this kind of analysis, more comprehensive data sets are needed.

4. Results and Discussion
In our multidisciplinary study we considered the strengths of each method in order to understand the linguistic behavior of migrant flow as a complex matter, and each migrant student as a part of this flow. The results of the work are ambiguous. On the one hand, the built Rasch model showed to be of a high quality according to the LR-test. Also, in our analysis it helped to eliminate irrelevant questions from the questionnaire (i.e. q7, q10, q13, q21, q22, q25). It means that the Rasch model works well for the evaluation of the questions as the physical variables affecting migrants’ plurilingual ability as the sociophysical parameter of a migrant flow. On the other hand, according to the T-test the questionnaire is not able to predict the further plurilingual ability development of students learning English as foreign language. We believe that this situation could be improved with more comprehensive data set which is obligatory for such kind of sophisticated sociolinguistic studies [3].

The second question of this study seeks to answer if the educational trajectories can be created with the help of the statistical methods described earlier. The ranking of a student according to the Rasch model is a convenient numerical measurement tool of the student’s competence profile. We consider that it is possible to create an appropriate program for the migrant students grouping them with close values of this measurement. It is also possible to consider the didactic recommendations for the development of the individual trajectories using LMS Moodle in relation to the students’ personal linguistic profiles controlled by the Rasch model in the future studies. It will affect the dynamics of migrant flows which reveals a solid ground for the further multidisciplinary studies.

Answering the last question of this study, we can state that the Rasch model proved to be useful in analysis of the current plurilingual competence (both in terms of most valuable properties for plurilingual ability and in terms of latent students’ plurilingual ability). But according to the results of T-test, answers to the questions did not allow us to make up a relevant model for prediction of the development of the plurilingual competence in multilinguals’ cluster. However, we would evaluate the results of the study as positive and the proposed framework suitable for the analysis of complex sociophysical matters, which are constituted by a set of elements with different types of interactions [4].

References
[1] Rykun A Yu, Yuzhaninov K M and Vychuzhanina E V 2015 Universities and local communities: problems and perspectives of interaction, Elsevier Procedia – Soc. and Behav. Sci. 200 359-365
[2] Baayen R H 2008 Analyzing Linguistic Data: a Practical Introduction to Statistics Using R. vol 2, ed G Krajewski and D Matthews (Cambridge: Cambridge University Press) pp 368
[3] Raducha T, Gubiec T 2018 Predicting language diversity with complex networks. PLoS One 13 0196593
[4] Carpi L C, Schieber T A, Pardalos P M, Marfany G, Masoller C, Diaz-Guilera A and Ravetti M G 2019 Assessing diversity in multiplex networks NLM 9