Social Indicators Research (2022) 160:287–307
https://doi.org/10.1007/s11205-021-02801-9

ORIGINAL RESEARCH

Social Integration of Second Generation Students in the Italian School System

Francesco Giovinazzi[1] · Daniela Cocchi[1]

Accepted: 5 September 2021 / Published online: 27 September 2021

Abstract
Cultural divides and prejudice complicate the processes of integration and acculturation of migrant families living in a foreign country. Evaluating the impact of such phenomenon can be crucial for social stability and policy making. In this context, the education system has a leading role in fostering and attaining social integration, in particular when it comes to younger sections of the migrant population. In this work, we propose a method for the construction of a quantitative indicator capturing social integration of second generation students in the Italian school system according to areas defined by nationality of the students and administrative region in which they attend school. The indicator, based on survey data, is estimated by means of a 2-step methodology. In the first step, we choose an individual qualitative variable capturing social integration at the unit level, and we compute a first direct estimate of the indicator as the proportion of highly integrated students in each area. Such variable is isolated following alternatively a proxy variable approach or a latent variable model approach. In the second step, we make use of two alternative small area models to improve the estimates, dealing with missing values, low sample size and high variability in smaller domains. At the end, the 2-step methodology results in 4 alternative versions of a synthetic indicator of social integration, that can be used to rank nationalities and administrative regions.

Keywords Immigrant youth · Acculturation · Latent class analysis · Small area estimation

1 Introduction

In the last two decades the public debate in Italy has been focusing on immigration and the many challenges it brings along in an ever-changing multicultural, multilingual and multireligious society (Ambrosini and Molina 2004; Allievi 2010; Thomassen 2010; Armillei 2015). Social and cultural integration of the newcomers is a key factor for a successful management of this phenomenon, and policy makers are particularly interested in knowing whether or not such integration can be transferred from the parental generation to the so-called second generation (Barbagli and Schmoll 2011). Schools across Europe are seeing

✉ Francesco Giovinazzi
francesco.giovinazz2@unibo.it

[1] Dept. of Statistics, University of Bologna, Via delle Belle Arti, 41, 40126 Bologna, Italy

a rise in the number of children born in a country and raised in another, and the Education system has a leading role in fostering and attaining their integration and well-being (Lelie et al. 2012). Giuliani et al. (2018) offer a wide literature review on the impact of low levels of perceived integration on psychological well-being of second generation migrants. Their work focuses on Muslim minorities in Italy and shows how acculturation, which is the process of adapting to a majority or a new cultural context (Berry 1997), can be very problematic for immigrant youth facing discrimination (Kowalczyk and Popkewitz 2005; Levy 2015). Building an indicator of social integration is a very challenging task, not only because social integration is a complex, multidimensional and unobservable phenomenon, as many other social constructs are, but also because of the lack of global and comprehensive available data sources. The present work has been motivated by a survey study carried out in 2015 by the Italian National Institute of Statistics (Istat), and co-financed by the Italian Ministry of Interior and the EU European Fund for the Integration of third-country nationals (EFI). The *Survey on Integration of the Second Generation* (ISG) (Istat 2017) involved lower and upper secondary schools on the whole national territory that were attended by at least 5 foreign students. By means of an extensive questionnaire, Istat investigated many different dimensions of second-generation student's social inclusion, from the use of native and local languages, to their relationship with family, schoolmates and teachers, to how they spend their free time, and how they define their own household conditions. The study produced a rich and complex amount of information that is, as of today, widely underused.

We propose to enhance the potential of ISG data to build an aggregated social integration indicator at the levels of both nationality of the students and Italian administrative region in which they attend school. The indicator reflects the assumption that social integration is the unobservable variable underlying one or more items in the ISG questionnaire. Nationality or citizenship can be seen here not only as a mere administrative information (Bianchi 2011), but as a variable capable of synthesizing a broad variety of personal attributes related to a certain cultural heritage. In particular, we are interested to know if the process of acculturation in the Italian school system con be seen as equally effective on children belonging to different nationalities across different Italian administrative regions.

The combination of nationality and administrative region generates a fixed cross-classification of about 200 cells, that define the areas in which we compute the indicator. Such areas constitute unplanned domains since they have not been considered in the design of the ISG survey. The focus is on the identification of an area-level social integration indicator, taking values in each cell of the cross-classification. At first, using a proxy variable approach to define latent social integration, we select the single item of the questionnaire that asks students a self-evaluation on their feeling about being Italian, then we aggregate the answers according to nationality and administrative region in order to get, in each area, the proportion of students feeling more Italian than foreigner. Such proportion can be interpreted as a very raw estimate of social integration. As a multidimensional alternative, following a latent variable model (Bartholomew et al. 2011) approach, we select a number of items of the ISG questionnaire and perform a latent class model (Hagenaars and McCutcheon 2002) in order to cluster the students into homogeneous groups in terms of latent social integration. The results are aggregated according to the cross-classification and following the same process used for the proxy variable, ending up with a second, and in this case multidimensional, version of the indicator. The dichotomous variables isolated according to both approaches indicate the occurrence of a particular outcome and allow to compute conditional proportions for the subgroups of the population defined by the cross-classification of nationality and Italian administrative region.

The proportions may be affected by very small and unplanned sample size for certain subgroups and, consequently, high variability of the estimates. In order to achieve a reliable estimate for the indicator in all subgroups, we propose to add a second step to the estimation process, considering students' nationalities in each Italian administrative region as unplanned study domains in an area-level small area model (Rao and Molina 2015). In this step, we borrow strength using covariates coming from administrative data sources at a population level, and the direct estimate of the indicator, in one of the two proposed versions, is the response variable of the model. We explore two alternative model formulations for this step: a linear model and a generalized linear model that considers the bounded nature of the response (Ferrari and Cribari-Neto 2004).

The combined use of latent variable models and small area models has been investigated in other works. For instance, Moretti et al. (2020) work with continuous variables and propose a Factor analysis model combined to a unit-level small area model to predict a vector of means of factor scores that can be interpreted as indicators of multidimensional latent well-being in small areas; in Montanari and Ranalli (2010) the latent class model is used to classify the population according to different levels of disability and then local estimates of the number of people belonging to each class are obtained via a small area model. Both examples rely on the use of a 2-step approach, first the estimation of the latent variable and second the small area correction. Fabrizi et al. (2018) propose a one step approach where the small area model is fitted simultaneously together with the latent class model. Such methodology tackles the problem of classifying the population on the basis of a categorical latent variable and getting small area estimates within a global Hierarchical Bayesian framework.

Summarizing what sketched before, the present work proposes a methodology that is developed in two steps: in the first step we compute the indicator in the unplanned domains, with two alternative approaches, and in the second one we refine the proposal via a correction via Small Area Estimation, again with two alternative models. We obtain 4 different versions of the indicator, slightly different in the outcomes, that we use to capture the social integration of foreign students in the Italian school system according to administrative region of their school and their nationality.

The paper is organized as follows: in Sect. 2 we present the ISG survey data and the administrative data used in the construction of the indicator, in Section 3 we discuss the definition of the response variable following the proxy variable approach or the latent variable model approach, in Section 4 we present the small area problem and the models proposed to solve it, in Sect. 5 we present the final results of the estimation.

2 Data

Microdata on individual respondents of the ISG survey are available for the subsequent analysis. The auxiliary information used for the small area models consists in aggregated data from the Ministry of Education. Students are the elementary statistical units, while area level data are aggregated according to the student nationality and the administrative region in which school is attended.

2.1 The Survey on Integration of the Second Generation

In 2018 Istat released the ISG microdata for research purposes, consisting of a dataset of $n = 68127$ observations on 255 variables, which are answers to the items of an extensive questionnaire. The population of interest is made by foreign students (sampling units) attending secondary schools in Italy. The study only involves schools attended by at least 5 foreign students. The population of schools is composed by 9386 institutes and has been stratified according to administrative region (21 cases), type of municipality (large or small), type of school (lower or upper secondary school) and incidence of foreign students (3 levels). A simple random sample of schools has been selected from each stratum with equal probabilities. The questionnaire was administered to every foreign student in the sampled schools (Italian students have been selected as control group in the same class of the foreign student). The questionnaire was organized in 6 broad sections:

A. Administrative data and migration history.
B. Use of native and local languages.
C. Relationship with schoolmates and teachers.
D. Relationship with friends, free time and social habits.
E. Composition of the family and relationship with its members.
F. Household conditions.

Table 1 shows an overview of the ISG sample in terms of nationality of the students. Together with Table 3, it reports the occurrence of each nationality and the corresponding unweighted proportions. About half of the sample is composed by Italian students, with a role of control group. As regards foreign students, only 28.4% of them were born in Italy and they can be strictly defined as second generation children. The vast majority migrated to Italy at a young age, then started to attend school. The last column in Table 1 presents the 10 most frequent nationalities in the groups of students born respectively in Italy and abroad. Among children born in Italy, Albanian, Moroccan and Chinese communities are strongly represented, and this is historically related to a long term immigration since the ’90s. Among the children born abroad we can see a primary role of Romanians (more than 25% of the total), together with the appearance of Moldova and Ukraine in the first 5 positions. This emphasizes the strong attractiveness of Italy over families in Eastern Europe during the last decade.

The 21 administrative regions of Italy (considering separately the autonomous provinces of Bolzano and Trento) have been used as study domains in the design of the ISG survey, that is balanced to take into account their heterogeneity. The combination of 10 nationalities and 21 administrative regions produces 210 unplanned domains, some of which have very small sample sizes.

2.2 Official Students Register Data

Small area methods rely on the exploitation of auxiliary information at a population level (such as census data) to borrow strength in the estimation process. The population of interest is composed by all the foreign students in the Italian school system, so we refer to the Anagrafe Nazionale degli Studenti (ANS), the official register kept by the Ministry of Education (MIUR).

Table 1 Nationality of the respondents to the ISG survey: 10 most frequent nationalities (unweighted proportions in brackets)

Citizenship		Birthplace		Nationality	
Foreigner	31687 (46.5)	Italy	9002 (28.4)	Albania	1811 (20.1)
				Morocco	1080 (12)
				China	1015 (11.3)
				Romania	750 (8.3)
				Philippines	581 (6.5)
				Ecuador	244 (2.7)
				Peru	214 (2.4)
				India	104 (1.2)
				Ukraine	87 (1.0)
				Moldova	67 (0.7)
				Others	3049 (33.9)
		Abroad	22685 (71.5)	Romania	5879 (25.9)
				Albania	2852 (12.6)
				Morocco	1555 (6.9)
				Moldova	1361 (6)
				Ukraine	1056 (4.7)
				China	701 (3.1)
				Ecuador	604 (2.7)
				Peru	505(2.2)
				India	487 (2.1)
				Philippines	487 (2.1)
				Others	7198 (31.7)
Italian	36440 (53.5)	–		–	

For each administrative region, data offers the number of Italian students and foreign students attending each of the 8 grades of secondary school. The total number of foreign students in each administrative region and grade is further classified by the Ministry according to the first 10 most frequent nationalities, as in Table 2.

As regards the new auxiliary information we faced two problems:

Table 2 Number of foreign students by nationality attending school in Italy in 2015. Administrative data from ANS (population data)

	Nationality	N_g
1	Romania	65073 (28.8)
2	Albania	45176 (20.0)
3	Morocco	34727 (15.4)
4	China	17655 (7.8)
5	Moldova	13144 (5.8)
6	Philippines	13024 (5.8)
7	Ukraine	11284 (5.0)
8	Peru	9705 (4.3)
9	India	9119 (4.0)
10	Tunisia	7121 (3.2)

- data is not complete with respect to administrative regions, we only have 19 of the 21 in the sample (Bolzano and Valle d'Aosta are missing);
- the 10 most frequent nationalities registered by MIUR are not the same 10 obtained by the survey. Since Ecuador is replaced by Tunisia, we refer only to 9 nationalities when building the model.

We finally work on 171 unplanned domains, obtained by the combination of 19 administrative regions and 9 nationalities.

3 The Response Variable

Social integration is a complex phenomenon, it is not susceptible of direct measurement, and we assume it to be a latent variable (Borsboom et al. 2003) affecting the set of observed responses to the items in the ISG survey. We propose two alternative approaches to the quantification of such phenomenon: a proxy variable approach or a latent variable model approach.

3.1 The Proxy Variable Approach

In order to capture social integration as an unobservable construct we can identify a proxy variable: a manifest variable that is reasonably assumed to have a high correlation with the construct of interest.

We isolated item A11 of the ISG questionnaire as a proxy variable of social integration. It asked directly to students whether they felt more Italian, foreigner or undecided. By choosing this item to assess integration we are making an assumption about the self-evaluation skill of respondents. The answer to A11 is strongly subjective, it incorporates feelings, emotions and it is built on an intimate level. Despite these weaknesses, we consider A11 the most suitable proxy in the questionnaire.

In Table 3 we report the frequencies of answers to item A11 for the first 10 most frequent immigrant nationalities plus an 11th residual group. All proportions can be interpreted as a very raw estimate of social integration. Overall we can see that 38.8% of students with a foreign citizenship asserts to feel more Italian than foreigner. This quantity grows by 10 percent for those students who were actually born in Italy. Albanians, Romanians and Ukrainians are characterized by a proportion greater than 40% of students identifying themselves as Italians. This nationalities are closely followed by Moldova and Morocco. This result is not surprising, particularly for what concerns Albania and Romania, with which Italy has strong economic, cultural and historical relationships. On the opposite side of this ranking of self-asserted integration we find Chinese students. Less than 23% of them declare to feel more Italian than Chinese. This could be explained by a strong influence of Chinese families' cultural roots and traditions, which are for sure the most distant from the western model of society, among the 10 nationalities isolated by Istat.

Similar considerations could be made observing the distribution of the answers to item A11 according to the administrative region in which the school is located. In this case we see that the proportion of students feeling more Italian than foreigner is higher in southern regions with respect to the North. This result shows how our proxy of social integration varies consistently not only across nationalities but also in different administrative regions.

Table 3 Number of students answering to feel more Italian or foreigner by birthplace and nationality in the sample (unweighted proportions in brackets)

		A11 – Do you feel more.		
		Italian	Foreigner	Don't know
Total		12298 (38.8)	10039 (31.7)	9350 (29.5)
By birthplace	Born in Italy	4397 (48.8)	2105 (23.4)	2500 (27.8)
	Born abroad	7901 (37.8)	7934(35)	6850 (30.2)
By nationality	Albania	2031 (43.6)	1347 (28.9)	1285 (27.6)
	Romania	3002 (45.3)	1771 (26.7)	1856 (28)
	Ukraine	505 (44.2)	308 (26.9)	330 (28.9)
	Moldova	484 (33.9)	460 (32.2)	484 (33.9)
	China	390 (22.7)	696 (40.6)	630 (36.7)
	Philippines	330 (30.9)	387 (36.2)	351 (32.9)
	India	187 (31.6)	164 (27.7)	240 (40.6)
	Morocco	954 (36.2)	878 (33.3)	803 (30.5)
	Ecuador	262 (30.9)	332 (39.2)	254 (30)
	Peru	187 (26)	284 (39.5)	248 (34.5)
	Others	3966 (38.7)	3412 (33.3)	2869 (28.0)

We dichotomize item A11 constructing the individual variable $\Xi_i = \xi_i$, $(i = 1, 2, \dots, n)$, which assumes value $\xi_i = 1$ for students answering to feel more Italian, and value $\xi_i = 0$ in the other cases. In this way we emphasizes well integrated students as those answering to feel more Italian than foreigner, in opposition to the rest feeling foreigner or undecided. Considering G subpopulations of size N_g, $(g = 1, ..., G)$, we define the quantity p_g^{ξ} as the proportion of students answering to feel more Italian in each group $p_g^{\xi} = \frac{1}{N_g} \sum_i^{N_g} \mathrm{I}(\xi_i = 1)$, where $I(\cdot)$ is the indicator function. This quantity is bounded in the unit interval, $0 \leq p_g^{\xi} \leq 1$ and it can be interpreted as an indicator of self-assessed social integration of foreign students in the g-th group/area.

3.2 The Latent Variable Model Approach

The indicator p_g^{ξ} is of immediate understanding, but it has at least two drawbacks: it is a unidimensional measure trying to capture a multidimensional phenomenonen, and it is highly subjective, being the result of a self-assessment process.

A good alternative is to use Latent Class Analysis that, on the basis of a set of selected items, defines a discrete latent variable $\Lambda_i = \lambda_i$, $(i = 1, 2, \dots, n)$. The levels of such variable correspond to latent classes in the population and the variable allows to cluster students into L groups that can be considered homogeneous in terms of latent social integration.

Lets consider a set of M categorical items $\mathbf{Y}_i = (Y_{i1}, \dots Y_{iM})$ $(i = 1, 2, \dots, n$ and $m = 1, 2, \dots, M)$; each item Y_{im} takes values in $1, 2, \dots, r_m$, where r_m is the number of categorical outcomes of the m-th item and may vary with m. We define $\Lambda_i = \lambda_i$ as an unobservable variable indicating the latent class of the i-th student, with $\lambda_i = 1, \dots, l, \dots, L$.

If Λ_i were observable, the joint probability of belonging to the l-th latent class and observing the response pattern $\mathbf{y}_i = (y_{i1}, \dots, y_{iM})$ for the i-th student would be:

$$P(\mathbf{Y}_i = \mathbf{y}_i, L_i = l; \pi_l, \rho_{mk|l}) = \pi_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)} \tag{1}$$

where $\pi_l = P(L_i = l)$ is the probability of belonging to the latent class l, $\rho_{mk|l} = P(Y_{im} = k|L_i = l)$ is the probability of answering k to item m belonging to class l (conditional distribution of the responses Y_i) and $I(\cdot)$ is the indicator function used to point each element.

The likelihood of observing a response pattern $\mathbf{y}_i$ is a function of parameters π_l and $\rho_{mk|l}$

$$P(\mathbf{Y}_i = \mathbf{y}_i; \boldsymbol{\pi}, \boldsymbol{\rho}) = \sum_{l=1}^{C} \pi_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)} \tag{2}$$

The number of parameters to be estimated is

$$\text{npar} = C - 1 + CMr_m. \tag{3}$$

under the following constraints

$$\sum_{l=1}^{C} \pi_l = 1 \quad \text{and} \quad \sum_{k=1}^{r_m} \rho_{mk|l} = 1. \tag{4}$$

The parameters π_l and $\rho_{mk|l}$ are important tools when interpreting the latent classes. The $\pi_1, \ldots, \pi_L$ represent the relative size of each class, while the probabilities $\rho_{mk|l}$, informing on how likely is a certain answer k to a questionnaire item m, given that the respondent belongs to a certain latent class l, allow to characterize the latent classes on the basis of the most probable associated answers. The latent class model can be seen as a categorical mixture model and it can be estimated in a frequentist framework using, for example, the EM algorithm (Linzer and Lewis 2011), or following a full Bayesian approach using a Gibbs sampler (White and Murphy 2014).

Once the model has been estimated we can compute the probability δ_{il} that a student with response pattern $\mathbf{y}_i$ belongs to the l-th latent class

$$\delta_{il} = P(\Lambda_i = l|\mathbf{Y}_i = \mathbf{y}_i) = \frac{\pi_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}}{\sum_{l=1}^{C} \pi_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}} \tag{5}$$

Using the Maximum A Posteriori (MAP) rule, we assign each subject to the latent class for which they present the highest δ_{il}. In this way we define variable $\Lambda_i^\star = l$ indicating the latent class to which the i-th student is allocated.

The number of latent classes is selected using the Bayesian information criterion (BIC):

$$BIC = 2\, log(\text{maximized likelihood}) - (\text{no. of parameters})\, log(n) \tag{6}$$

where n is the number of observations. Each different number of classes defines a different model on the same set of items; we choose the model configuration that minimizes the criterion. After the interpretation, if the BIC leads to a model with $L > 2$, we propose to aggregate the latent classes in 2 groups according to the level of latent social integration. The latent variable is dichotomized by aggregation so that $L = 2$, and the new classes can be directly compared to the proxy variable. One class will be interpreted as the class l_1, Class 1, with high level of social integration and the other l_2, Class 2, will be interpreted in opposition. Once the students have been allocated to the latent classes following the MAP rule, we can measure the proportion p_g^λ of students belonging to the group corresponding to a high level of integration in G subpopulations, each of size N_g. In this way we define a new latent social integration indicator $p_g^\lambda = \frac{1}{N_g} \sum_i^{N_g} \mathrm{I}(\lambda_i = l_1)$, with the same structure of p_g^ξ. This quantity is bounded in the unit interval, $0 \leq p_g^\lambda \leq 1$ and it can be interpreted as an indicator of multivariate latent social integration of foreign students in the g-th group.

3.2.1 The Items Chosen for the Latent Variable Model

In the following version of the latent class model, we selected $M = 9$ items from the ISG questionnaire expressing different dimensions of social integration. The item selection has been performed aiming at a compromise between the theoretical representation of different dimensions of social integration and the technical aspects related to the assumption of local independence. Some of the manifest variables have been transformed in order to better capture the underlying phenomenon, and to reduce the impact of the association among pairs of items on the violation of the assumption of local independence.

The following list defines the columns of matrix $\mathbf{Y}_i$:

- Item $Y_{.1}$ ($r_1 = 3$) is item A11, no longer used as the standalone indicator as in the proxy variable approach, but now one of the many dimensions of social integration. It asks the student to choose between three possible responses: (1) I feel more Italian, (2) I feel more foreigner, (3) I don't know.
- Item $Y_{.2}$, ($r_2 = 3$) measures the amount of time the student has lived in Italy. Answers: (1) born in Italy, (2) born abroad and arrived in Italy one year ago or less, (3) born abroad and arrived in Italy more than one year ago.
- Item $Y_{.3}$, ($r_3 = 2$) asks the students to state in which language they usually think. The answers are: (1) Italian, (2) other language.
- Item $Y_{.4}$, ($r_4 = 4$) asks students to self-assess their own school performance in a scale from (1) I am not so good at school to (4) I am very good at school.
- Item $Y_{.5}$, ($r_5 = 3$) asks students the nationality of the friends with which they spend more time: (1) Italian, (2) foreigner of their same nationality, (3) foreigner of other nationalities;
- Item $Y_{.6}$, ($r_6 = 2$) asks whether or not the student has been bullied for the way she/he talks or appears. This item should incorporate racism and discriminating behaviors. Answers: (1) at least once, (2) never.
- Item $Y_{.7}$, ($r_7 = 7$) intends to summarize the economic status of the student's household. It results from the aggregation of 6 dummy variables of the ISG questionnaire asking if in the student's house the following objects are present: washing machine, fridge, dishwasher, personal computer, television, DVD reader. The responses of the students vary from 0 to 6 according to the number of objects owned by their household: from (1) none of the objects to (7) all the objects.
- Item $Y_{.8}$, ($r_8 = 3$) indicates the composition of the nuclear family. It has three outcomes: (1) the student does not live with her/his parents, (2) the student lives with one parent only, (3) the student lives with both parents.
- Item $Y_{.9}$, ($r_9 = 4$) indicates the composition of extended families, it asks how many people that are not parents or siblings live with the student. This variable includes grandparents, aunts/uncles, relatives of other nature and family friends. It is intended to capture the effect of large family in particular cultures and it varies from (1) zero to (4) three or more.

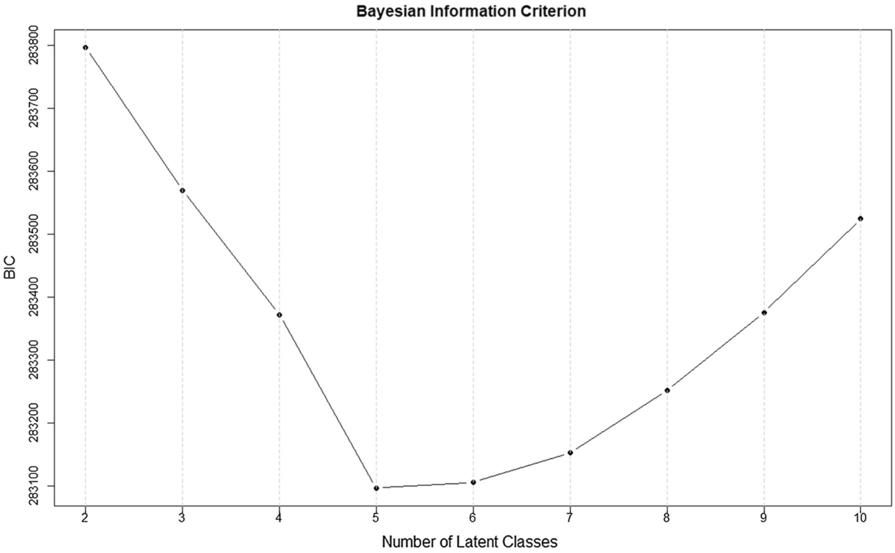

Fig. 1 Values of the BIC for increasing number of latent classes

Table 4 Estimates of parameters π_l

Class 1	Class 2	Class 3	Class 4	Class 5
0.3545	0.0651	0.3030	0.1532	0.1242

We assume that the above listed items expand the information contained in A11, adding migration history ($Y_{.2}$), use of language ($Y_{.3}$), school performance ($Y_{.4}$), relationship with other children ($Y_{.5}$ and $Y_{.6}$), economic status ($Y_{.7}$) and family composition ($Y_{.8}$ and $Y_{.9}$).

3.2.2 Interpretation of the Latent Classes

The BIC criterion leads to select a model with 5 latent classes, as reported in Fig. 1, were the curve reaches its minimum at $K = 5$.

Tables 4 and 5 report the estimates of the model parameters. Table 4 shows the estimates $\hat{\pi}_l$ representing the relative size of the classes in the population as the a priori probability of belonging to a class. Class 1 is the biggest (35.45%), followed by Class 3 (30.40%). Table 5 reports the estimates $\hat{\rho}_{mk|l}$ representing the conditional probabilities of answering category m to the k-th item given the membership to the l–th latent class; the columns correspond to the 5 classes of the selected model, the rows organized in 9 blocks correspond to all the possible answers to the 9 items, as described in the previous section. In the following, we use the estimated conditional probabilities to interpret the latent classes, browsing Table 5 block-wise. In any row within each block of Table 5 a high value of $\hat{\rho}_{mk|l}$ characterizes the class corresponding to the column.

Starting with the first item ($Y_{.1}$), the estimated parameter $\hat{\rho}_{11|1}$ is the probability of answering to feel more Italian to item A11 if the respondent belongs to Class 1. Such probability is higher in column 1 than in any other column, taking value of 75.93%, thus

Table 5 Estimates of parameters $\rho_{mk|l}$

		Class 1	Class 2	Class 3	Class 4	Class 5
$Y_{\cdot 1}$	1	0.7593	0.3997	0.1272	0.3644	0.0326
	2	0.0240	0.2791	0.4642	0.2719	0.7690
	3	0.2167	0.3212	0.4086	0.3637	0.1984
$Y_{\cdot 2}$	1	0.6716	0.7654	0.8529	0.4075	0.7153
	2	0.0026	0.0934	0.0237	0.0005	0.1082
	3	0.3258	0.1413	0.1234	0.5920	0.1765
$Y_{\cdot 3}$	1	0.9395	0.6004	0.5103	0.8126	0.0261
	2	0.0605	0.3996	0.4897	0.1874	0.9739
$Y_{\cdot 4}$	1	0.1175	0.0874	0.0741	0.0746	0.0575
	2	0.5720	0.4728	0.5423	0.4295	0.4174
	3	0.2702	0.3357	0.3269	0.3693	0.3694
	4	0.0404	0.1041	0.0566	0.1265	0.1557
$Y_{\cdot 5}$	1	0.9725	0.8611	0.8358	0.6726	0.3643
	2	0.0148	0.0550	0.1183	0.2337	0.5381
	3	0.0127	0.0839	0.0460	0.0936	0.0976
$Y_{\cdot 6}$	1	0.2367	0.4234	0.3414	0.5106	0.4487
	2	0.7633	0.5766	0.6586	0.4894	0.5513
$Y_{\cdot 7}$	1	0.0004	0.0129	0.0026	0.0005	0.0067
	2	0.0006	0.0107	0.0000	0.0000	0.0059
	3	0.0006	0.0053	0.0016	0.0081	0.0235
	4	0.0154	0.0361	0.0265	0.0539	0.0852
	5	0.0882	0.1827	0.1546	0.1903	0.2157
	6	0.4321	0.3261	0.4529	0.4475	0.3851
	7	0.4625	0.4262	0.3618	0.2997	0.2779
$Y_{\cdot 8}$	1	0.0028	0.1761	0.0000	0.0010	0.0388
	2	0.1102	0.3383	0.1009	0.1019	0.1686
	3	0.8869	0.4857	0.8991	0.8971	0.7926
$Y_{\cdot 9}$	1	0.8662	0.2973	0.9049	0.7218	0.6469
	2	0.1161	0.5399	0.0941	0.2274	0.2742
	3	0.0162	0.1422	0.0000	0.0498	0.0680
	4	0.0015	0.0206	0.0011	0.0010	0.0109

characterizing Class 1 as the class of students feeling more Italian than foreigner. Similarly, Class 5 is characterized by students answering to feel more foreigner than Italian, while answer 3 ”Don’t know” is ambiguous, showing similar probabilities in classes 2 and 3. Item $Y_{\cdot 2}$ at a first sight appears less discriminant than $Y_{\cdot 1}$, but it shows how the probability of living in Italy by less than one year is higher for Class 5 ($\hat{\rho}_{22|5} = 10.82\%$), and that being born in Italy is not too relevant to discriminate among Class 1 and Class 5. On the other hand item $Y_{\cdot 3}$ is strongly polarized, with a probability of 97.39% to think in a language different from Italian if in Class 5, opposite to a probability of thinking in Italian equal to 93.95% if in Class 1. For item $Y_{\cdot 4}$, note how students in Class 1 tend to perceive their school performance as good, while students in Class 5 behave oppositely ($\hat{\rho}_{14|1}$ is higher than any other probability in the first row, while $\hat{\rho}_{14|5}$ is the lowest, and $\hat{\rho}_{44|1}$ is the lowest in the fourth row while $\hat{\rho}_{44|5}$ is the highest). Moving to $Y_{\cdot 5}$, students in Class 1 have the highest probability to spend time with Italian friends $\hat{\rho}_{15|1} = 97.25\%$, while students in Class 5 have the lowest value $\hat{\rho}_{15|5} = 36.43\%$. Item $Y_{.6}$ reports the impact of bullying on social integration of children; students in Class 1 have a high probability $\hat{\rho}_{26|1} = 76.33\%$ of not experiencing bullying at all. Item $Y_{.7}$ describes the economic situation of the student's household, Class 1 has the highest probability and Class 5 the lowest of having all the 7 objects used as proxy of economical well-being. Finally, items $Y_{.8}$ and $Y_{.9}$ summarize the student's family composition; Class 2 shows the highest probabilities of students not living with their parents or living with only one parent, and at the same time the highest probability of living with 3 or more relatives that are not parents.

To build the indicator of social integration under the proxy variable approach, we aggregated the students answering ”Foreigner” and ”Don't Know” to item A11, then calculated the proportion of students answering ”Italian”. Analogously, under the latent variable approach we are interested in identifying a reasonably large well characterized class of highly integrated students, the remaining latent classes can be aggregated in a second class of less integrated students. According to the above interpretation of Tables 4 and 5, we identify Class 1 as the one with a high level of social integration. These students identify themselves as Italians, think in Italian, have a good self-evaluated school performance, spend time with Italian friends, have a lower probability to experience bullying, have a good economic status and tend to live in nuclear families. On the other hand, Class 5 is strongly defined as the class of less integrated students, self-identifying as foreigners, thinking in another language, with complicated educational, relational and familiar situations. The interpretation of the remaining classes is less straightforward: Class 2 is characterized by students living within a non-standard family composition, and Classes 3 and 4 are quite similar but differ when looking at migration history and at the use of language. Following the same process as in the proxy variable approach, we treat classes from 2 to 4 as the ”Don't know” answer to item A11, aggregating the students together with Class 5 in the broader class of less integrated students.

After the classes have been interpreted, students are assigned to the latent classes following the MAP rule. Then classes from 2 to 5 are aggregated, generating a single dichotomous latent variable $\Lambda_i = \lambda_i, (i = 1, 2, \dots, n)$, which assumes value $\lambda_i = 1$ for students that are highly integrated, and value $\lambda_i = 0$ in the other cases.

4 Inference via Small Area Estimation

In the previous section we proposed two alternative definitions of the response variable, $\Xi = \xi_i$ and $\Lambda_i = \lambda_i$ with $i = 1, 2, \dots, n$. No inference is performed from the ISG survey to the population of students in Italian schools. The object of inference are the proportions of students for which $\xi_i = 1$ or $\lambda_i = 1$ in the study domains.

Direct estimates of the population proportions p_g^{ξ} and p_g^{λ} can be computed in G subpopulations or domains, representing combinations of nationality and administrative regions. The 210 direct estimates are obtained using the Horvitz-Thompson (HT) estimator (Särndal et al. 2003). This traditional estimation method requires sufficiently large domain-specific sample sizes (n_g). Unfortunately, when the research interest lies, as in the present case, in estimates valid for specific domains, facing small unplanned sample sizes is rather common. The so-called small area problem arises when sample data are not large enough for all domains to provide adequate statistical precision of the estimates (sample size n_g may even be zero for some small areas). In this case the traditional estimator will have low precision, leading to useless too wide confidence intervals for the direct estimates. In this case, it is necessary to borrow strength from external data sources, incorporating auxiliary information from other neighboring areas by means of statistical models able to link the response to a set of predictor variables that are known for small areas at a population level. The auxiliary information we employ comes, as mentioned in Sect. 2.2, from the national archive ANS. We split the population into G unplanned domains, and consider p_g^{ξ} and p_g^{λ} as alternative response variables in a small area model. The number of subpopulations restricts to G=171, coming from 9 nationalities and 19 administrative regions. The auxiliary information will enter the model in the form of ratios. We compute the two following population auxiliary information:

- the proportion of students attending lower secondary school in each domain;
- the proportion of foreign students attending school in each administrative region.

With the first variable we capture the age effect (younger students appear to be more inclined to feel integrated), with the second to capture the differences across the administrative regions in terms of impact and visibility of immigrant children in school.

Small area models are usually classified as area level models or unit level models. In the first type of models, information on the response variable is available only at the small area level, while in the second type, data are available at the unit or respondent level. Since we are interested in an aggregated comparison between domains, we refer to area level models, proposing two alternative formulations: the Fay-Herriot model (Lahiri 2003) and the Beta regression model (Figueroa-Zúñiga et al. 2013). Both models allow to mitigate the occurrence of outliers, to reduce the variability of the estimates and to correct for missing values; the first one is a linear mixed effect model, the second one is non-linear and takes into account the bounded nature of the response assuming that the dependent variable is Beta-distributed (Gupta and Nadarajah 2004).

4.1 The Fay-Herriot Model

The Fay–Herriot model is a widely used area-level linear mixed model. Fay and Herriot (1979) proposed it for the first time as a two-level Bayesian model to estimate the per capita income of small areas with the population size less than 1000. Under this model, the dependent variable is a direct estimator calculated by using the survey data; the covariates are true population domain means obtained from external data sources. The Fay-Herriot model can be basically expressed as:

$$\tilde{p}_j = p_j + e_j = \mathbf{x}_j^T \boldsymbol{\beta} + v_j + e_j, \quad j = 1, \dots, J \tag{7}$$

where $\mathbf{x}_j$ is a vector of known covariates, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, v_j's are area specific random effects and e_j's represent sampling errors, assuming that $v_j \sim N(0, \psi)$ and $e_{j'} \sim N(0, D_{j'})$ are independent for all pairs (j, j').

Model (7) can be specified as a Bayesian hierarchical model:

$$\tilde{p}_j | p_j, A_j \sim N(p_j, A_j) \tag{8}$$

$$p_j | \boldsymbol{\beta}, \sigma_e^2 \sim N(\mathbf{x}_j^T \boldsymbol{\beta}, \sigma_e^2) \tag{9}$$

where (8) is the data model, (9) is the process model, the variances A_j are known, and the prior distributions on σ_e^2 and $\boldsymbol{\beta}$ are:

$$\sigma_e^2 \sim \text{Unif}(0, \sigma^2_{\max}) \quad \text{and} \quad \boldsymbol{\beta} \sim MVN(\mathbf{0}, \Sigma_{\beta}). \tag{10}$$

4.2 The Beta Regression Model

Linear mixed effects models, such as the Fay-Herriot model, are very popular and have been used to estimate all sorts of survey data. However, when the data are restricted to a bounded interval, as in the case of proportions, the linear model and the assumption of normality may be inadequate. This is particularly true when the data are near the boundary. Janicki (2020) illustrates the use of a Beta distribution as an alternative to the normal distribution as a sampling model for survey estimates of proportions which take values in (0, 1). Inference for small area proportions based on the posterior distribution of a Beta regression model ensures that point estimates and credible intervals take values in (0, 1). Other examples can be found in Fabrizi et al. (2016); Fabrizi et al. (2016), where hierarchical Beta regression models are used for the small area estimation of poverty and inequality rates.

In this spirit, we propose to fit a small area Beta regression model with mixed effects:

$$\tilde{p}_j | v_j, \boldsymbol{\beta}, \psi \sim \text{Beta}(\mu_j \phi, (1 - \mu_j \phi) \tag{11}$$

where ϕ is a common precision parameter and we can write

$$\log \left\{ \frac{\mu_j}{1 - \mu_j} \right\} = \eta_j = \mathbf{x}_j^T \boldsymbol{\beta} + v_j \tag{12}$$

with

$$\mu_{ij} = E(\tilde{p}_j | v_j) = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)}. \tag{13}$$

The effects v_j are assumed to be independent and normally distributed. A model based on equations (11), (12) and (13) allows to model a wide range of continuous random variables that assume values in the unit interval such as rates, proportions, and concentration or inequality indices (Fabrizi and Trivisano 2016). The Beta regression model is very flexible, since the Beta density can take different shapes depending on the combination of parameter values (Cribari-Neto and Zeileis 2010).

5 Results

In this Section we present the results of the 2-step methodology, structured in the 4 combinations of models: the Fay-Herriot and the Beta regression model respectively on p_g^{ξ} and p_g^{λ}. The models have been estimated in R using packages sae (Molina and Marhuenda 2015) and rstanarm (Goodrich et al. 2020). By means of these tools, we fulfill the double aim of our work: on one hand we use Small Area Estimation on proportions to build a quantitative area-level indicator measuring the intensity of a qualitative unit-level latent variable, on the other we apply the methodology to the estimation of social integration of

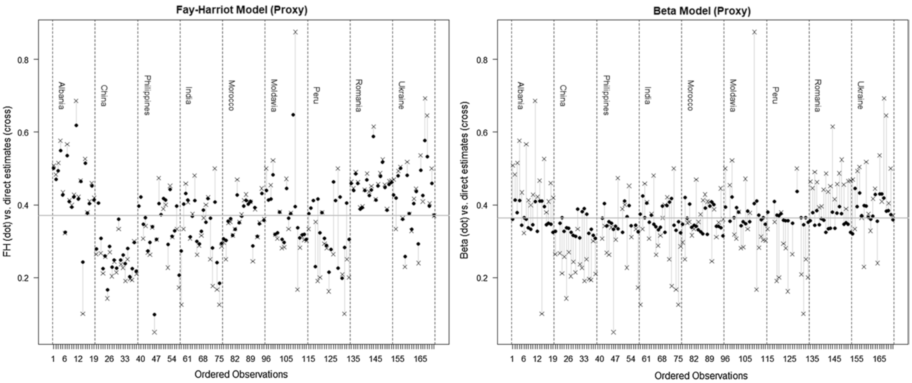

Fig. 2 Direct (crosses) and model (points) estimates of the level of social integration for the unplanned domains under the proxy variable approach: Fay-Herriot model and Beta regression model

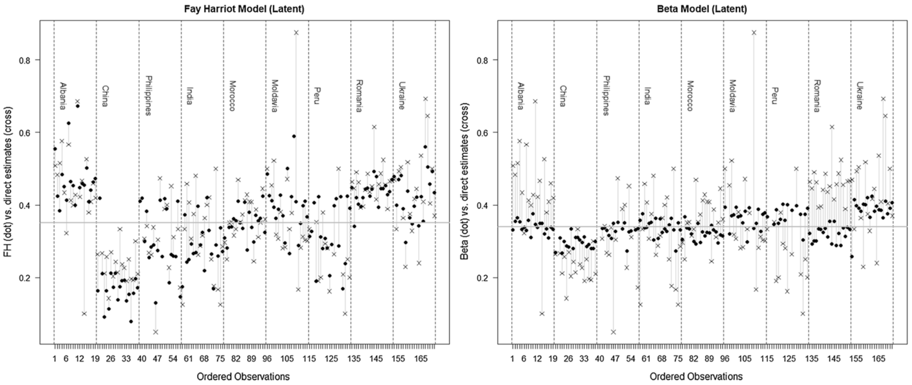

Fig. 3 Direct (crosses) and model (points) estimates of the level of social integration for unplanned domains under the latent variable approach: Fay-Herriot model and Beta regression

second generation children in Italy, making good use of a unique and underused survey data source.

The small area estimates of the proportion are aggregated in macro-domains and are interpreted as indicators of social integration within that domain. The 4 sets of results do not lead to the same direct conclusions, due to the difference in the small area models and in the definition of the response variable. The models are not intended as competing, the 4 sets of results are rather presented with the aim to look at the differences in behavior with respect to the construction of the indicator; no selection is performed. For each set, in Figs. 2, 3 we display the model estimates as ordered points in a plot, emphasizing their dispersion around the mean (horizontal gray line). The crosses represent the corresponding Horvitz-Thompson direct estimates. The plots are organized in 9 vertical sections, mirroring the 9 nationalities; within each section, points and crosses are ordered according to the same sequence of 19 Italian administrative regions. Both nationalities and administrative regions are reported in alphabetical order. No trend can be deduced from the inspection of the 4 panels of the Figures, they can only be inspected for comparison.

5.1 Small Area Estimation of Social Integration under the Proxy Variable Approach

Under the proxy variable approach we model p_g^{ξ} using as response variable the Horvitz-Thompson estimate $\hat{p}_g^{\xi(HT)}$ of the proportion of students answering "Italian" to item A11. Figure 2 reports the results of the model estimates for the Fay-Herriot model $\hat{p}_g^{\xi(FH)}$ and the Beta model $\hat{p}_g^{\xi(Be)}$ compared to the benchmark $\hat{p}_g^{\xi(HT)}$. In both cases, the model estimates improve the results of the direct estimates producing an overall lower variability, mitigating the occurrence of outliers and filling the missing values for smaller domains (those domains in which the size was too small to produce a reliable Horvitz-Thompson estimate). The lower variability can be worked out in Figs. 2, 3 by the distance between the crosses (direct estimates) and the black points (model estimates). The mean squared error of the fitted values for the Fay-Harriot model (here not reported) are computed by a specific R function in the package sae; the equivalent quantities for the Beta regression model would need a bootstrap approximation for building confidence intervals as proposed in Appendix B of Ferrari and Cribari-Neto (2004).

Figure 2 shows how the Fay-Herriot model tends to preserve the structure of the direct estimates, while the Beta regression model produces a higher level of shrinkage towards the global mean of the direct estimates themselves. According to the proxy variable approach, the most integrated nationalities are Albania, Romania and Ukraine, nationalities for which the black points are mostly above the horizontal line. On the other hand, Chinese students appear to be those feeling less integrated. The graphical interpretation of the domain-specific estimates in the Beta regression model is made difficult by the shrinkage effect, suggesting a more homogeneous behavior of foreign students across nationalities.

5.2 Small Area Estimation of Social Integration under the Latent Variable Approach

Under the latent variable approach we assume the level of social integration as a latent variable, underlying the items of the ISG questionnaire. We recall to have selected 9 items, aiming to cover different dimensions of social integration, and have performed a Latent Class Analysis, obtaining a model with 5 classes, then aggregated in 2. The response variable of the small area models is, in both cases, the direct estimate $\hat{p}_g^{\lambda(HT)}$ of the proportion of students falling into the latent class corresponding to the highest level of social integration. Using such quantity as an indicator of social integration in the different study domains, we estimated the Fay-Herriot model and the Beta regression model. When fitting the Fay-Harriot model on the proportion of the latent variable, we assume the same estimated variances of domain direct estimator as in the HT direct estimates of the proxy variable. Results are shown in Fig. 3.

The new definition of the response variable according to the latent variable model produces a result that is coherent to the proxy variable approach in terms of global mean value. The estimates in the Fay-Herriot model, shown in the left panel of Figure 3, have a high variability, and nationalities like Chinese, Filipino and Indian have a strong negative deviation from the mean. This effect is mitigated by the Beta model in the right panel. Shrinking the values towards the global mean, this latter model produces a more concentrated pattern of points. In this scenario Chinese students remain the less integrated, while the most integrated are students from Ukraine.

In synthesis, as expected, in Figs. 2, 3 we can graphically appreciate a reduction in the variability of the direct estimates of the latent variable by means of the small area model proposed as a second step. It is possible to identify nationalities that coherently exhibit the same type of deviation from the mean in the four schemes and across Italian administrative regions. The impact of the Beta model is very severe on the estimates, reducing the overall variability so that the differences among regions and nationalities are difficult to appreciate. For a better look at the behavior of the point estimates across domains, in Table 6 we report the numeric values of the Fay-Herriot estimates of the indicator built under the latent variable approach, that are the points in the left panel of Figure 3. Columns and rows of Table 6 show the marginal distributions of the cross-classification in the 171 unplanned domains.

The indicator $\hat{p}_g^{\lambda(FH)}$ reaches its maximum, with scores above 0.6, for Albanian students attending school in the southern region of Molise, immediately followed by Albanians in the Lazio region. On the other hand, with a scores below 0.1, the lowest level of social integration is attained by Chinese students in the regions of Tuscany and Campania. The indicator for Albanians and Romanians takes values above 0.34 in all regions, making this nationalities the most uniformly integrated over the whole Italian peninsula.

However, the main interest of this work lies in the aggregated comparison among nationalities and administrative regions, which are the marginals of the cross-classification, and that are object of the next paragraph.

Table 6 Fay-Herriot estimatesof the level of social integration for unplanned domains under the latent variable approach (in percentage)

Region	Nationality								
	ALB	CHN	PHL	IND	MAR	MDA	PER	ROU	UKR
Abruzzo	55.52	16.40	41.11	17.52	30.88	48.51	31.38	44.84	47.90
Basilicata	42.39	41.87	41.88	37.33	28.12	42.62	32.68	34.17	39.97
Calabria	38.37	21.05	29.91	24.31	33.79	41.22	40.59	41.98	47.04
Campania	48.35	9.11	38.31	24.98	34.01	39.35	18.99	40.27	48.11
Emilia Rom.	45.08	16.41	25.54	29.25	36.21	36.35	42.34	39.56	39.03
Friuli V.G.	41.29	11.35	28.43	26.62	35.61	38.79	31.23	42.01	29.74
Lazio	62.53	21.19	28.98	26.94	41.07	42.68	28.05	44.39	43.80
Liguria	46.49	17.30	12.97	39.63	35.14	37.04	30.73	44.59	36.51
Lombardy	45.37	21.26	27.50	28.99	33.57	29.59	28.17	42.50	34.79
Marche	39.89	13.90	41.27	32.50	40.73	50.11	20.59	44.37	39.72
Molise	67.26	17.48	25.83	21.96	27.61	26.68	29.14	49.20	42.16
Piedmont	44.86	19.20	41.79	42.04	37.95	42.35	39.75	47.78	34.25
Apulia	45.85	19.29	38.97	33.05	33.90	58.93	41.68	39.45	42.29
Sardinia	45.51	13.56	18.56	26.46	29.79	40.84	28.68	44.45	35.47
Sicily	50.18	15.41	26.29	16.94	37.12	28.87	42.27	44.46	56.05
Tuscany	40.91	7.89	25.90	28.96	35.69	34.94	16.90	45.30	50.49
Trento	44.07	15.69	25.86	25.91	36.36	39.83	23.79	42.79	45.73
Umbria	46.37	19.63	41.05	33.76	42.47	41.11	42.40	43.73	49.24
Veneto	47.28	17.22	14.71	27.14	32.50	36.73	39.07	47.03	43.42

Table 7 Estimated social integration indicators aggregated by nationality of the student

Nationality	$\hat{p}^{\xi(HT)}$	$\hat{p}^{\xi(FH)}$	$\hat{p}^{\xi(Be)}$	$\hat{p}^{\lambda(HT)}$	$\hat{p}^{\lambda(FH)}$	$\hat{p}^{\lambda(Be)}$
Romania	0.4600	0.4527	0.3554	0.4600	0.4331	0.3200
Albania	0.4481	0.4436	0.3667	0.4481	0.4724	0.3407
Ukraine	0.4355	0.4226	0.3902	0.4355	0.4241	0.3872
Moldova	0.3807	0.3785	0.3732	0.3807	0.3982	0.3629
Morocco	0.3609	0.3616	0.3570	0.3609	0.3487	0.3221
India	0.3180	0.3400	0.3654	0.3180	0.2865	0.3384
Philippines	0.3088	0.3385	0.3554	0.3088	0.3026	0.3291
Peru	0.2653	0.3329	0.3709	0.2653	0.3202	0.3713
China	0.2339	0.2613	0.3398	0.2339	0.1764	0.2944

Table 8 Estimated social integration indicators aggregated by macro-region of school attendance

Region	$\hat{p}^{\xi(HT)}$	$\hat{p}^{\xi(FH)}$	$\hat{p}^{\xi(Be)}$	$\hat{p}^{\lambda(HT)}$	$\hat{p}^{\lambda(FH)}$	$\hat{p}^{\lambda(Be)}$
South	0.4150	0.4170	0.3962	0.4150	0.3603	0.3385
Islands	0.3809	0.3729	0.4021	0.3809	0.3338	0.3347
Centre	0.3776	0.3761	0.3517	0.3776	0.3652	0.3425
North	0.3116	0.3317	0.3440	0.3116	0.3400	0.3405

5.3 Aggregate Comparison

The 171 domains come from a combination of two distinct variables: the nationality of the students and the administrative region in which they attend school. We aggregate the estimates by computing their means within these macro-domains, in order to have a synthetic view on how social integration is distributed across nationalities and among the administrative regions (and marco-regions) of Italy in Tables 7, 8, 9.

In the following tables, the Horvitz-Thompson direct estimates of $\hat{p}^{\xi(HT)}$ and $\hat{p}^{\lambda(HT)}$ (in the first and fourth column) represent the benchmark for comparison with the estimates obtained via small area models. The first column is used to order the labels of the nationalities.

In Table 7 we present the values of the indicator aggregated by nationality. Looking at the results, we can see that the indicators $\hat{p}^{\xi(FH)}$, $\hat{p}^{\xi(Be)}$, $\hat{p}^{\lambda(FH)}$ and $\hat{p}^{\lambda(Be)}$ produce slightly different rankings when it comes to nationalities. It is however possible to observe several recurring regularities: Chinese students are always at the last position in the ranking of social integration, with the indicator corresponding to Chinese nationality taking always the minimum value. On the other hand Albanians and Ukrainians occupy stably high positions in each column.

In Table 8 we propose to aggregate the indicators according to Italian macro-regions: North, Centre and South. Moreover, we separate the two extended Italian islands (Sicily and Sardinia) from the greater South macro-region. Here, when considering the direct estimates $\hat{p}^{\xi(HT)}$, we can see that the macro-region South dominates the others; when the dependent variable is $\hat{p}^{\lambda(HT)}$ the macro-regions tend to exhibit similar values, with a slight prevalence of the Centre. This indeterminacy may be attributed to the multidimensional nature of $\hat{p}^{\lambda(HT)}$, which includes different items from the ISG survey and captures more aspects of social integration.

Table 9 Estimated social integration indicators aggregated by administrative region of school attendance

Region	$\hat{p}^{\xi(HT)}$	$\hat{p}^{\xi(FH)}$	$\hat{p}^{\xi(Be)}$	$\hat{p}^{\lambda(HT)}$	$\hat{p}^{\lambda(FH)}$	$\hat{p}^{\lambda(Be)}$
Molise	0.4578	0.4241	0.3966	0.4578	0.3415	0.3245
Calabria	0.4170	0.4143	0.3642	0.4170	0.3536	0.3379
Apulia	0.4164	0.4414	0.4005	0.4164	0.3927	0.3316
Campania	0.4146	0.3984	0.4037	0.4146	0.3350	0.3377
Basilicata	0.3692	0.4070	0.4162	0.3692	0.3789	0.3610
Sicily	0.3969	0.4182	0.3989	0.3969	0.3529	0.3286
Sardinia	0.3649	0.3275	0.4053	0.3649	0.3148	0.3407
Lazio	0.4271	0.4089	0.3602	0.4271	0.3774	0.3305
Marche	0.3718	0.3648	0.3400	0.3718	0.3590	0.3656
Abruzzo	0.3701	0.3841	0.3573	0.3701	0.3712	0.3255
Tuscany	0.3654	0.3446	0.3512	0.3654	0.3189	0.3470
Umbria	0.3535	0.3780	0.3497	0.3535	0.3997	0.3440
Piedmont	0.3504	0.3758	0.3332	0.3504	0.3889	0.3518
Lombardy	0.3264	0.3281	0.3391	0.3264	0.3241	0.3233
Emilia Romagna	0.3165	0.3413	0.3505	0.3165	0.3442	0.3455
Veneto	0.3098	0.3257	0.3309	0.3098	0.3390	0.3471
Liguria	0.3020	0.3160	0.3396	0.3020	0.3338	0.3647
Trento	0.2909	0.3256	0.3491	0.2909	0.3334	0.3096
Friuli V.G.	0.2855	0.3096	0.3658	0.2855	0.3167	0.3412

Finally, Table 9 reports the values of the 19 administrative regions that were used to produce Table 8; the horizontal blocks correspond to the macro-regions. Also in this case the results confirm, as a whole, how social integration is generally perceived as more difficult in the Northern part of the country and is considered better elsewhere: there is a tendency to have higher values of social integration in the southern regions whether the indicator is built according to the proxy variable or latent variable approach.

6 Conclusions

The aim of this work was to explore the possibility of building an indicator of social integration using different approaches and working within the limitations of the available information. We investigated social integration of foreign students in Italy, proposing a 2-step methodology for the estimation of a synthetic indicator in a small area perspective. We worked on data from the first Italian survey on "Integration of the second generation" (ISG) and, using additional information from the official register of the Italian Ministry of Education, we showed how it is possible to improve the estimates of two variants of a social integration indicator, built respectively following a proxy variable approach and a latent variable approach. In the first case we selected a single item as a good proxy of the unobservable variable of interest, in the second case we used a latent class model to group students into latent social integration levels. The indicators are area-level and are built as proportions of students belonging to the most integrated group. The results suggest that the level of social integration of foreign students in Italy varies consistently according to their nationality and administrative region of school attendance. Chinese students seem to be the ones with the most problematic integration, while the North of Italy hosts the administrative regions in which the process of social integration of second generation students appears to be more difficult.

With this work we faced the problem of defining and measuring social integration of second generation students in the Italian school system using a proxy or a latent variable. We arrived at a coherent conclusion, knowing that the definition of the latent variable is based on the assumption that the latent construct captured by the model is indeed social integration. However, we are aware that the resulting latent construct may express instead other latent attributes, for instance well-being or privilege. Concerning the small area correction, the Beta model is to be taken in great consideration since it tackles the bounded nature of the response variable. When applied to the ISG data it produced an excessive amount of shrinkage on the estimates, making the Fay-Harriot model preferable to interpret. Further developments include a deeper study of the latent structure beyond the ISG data, a comparison of similar studies across countries, the evaluation of alternative formulations of non-linear small are a models, and the implementation of the one-step approach where the small area model is fitted simultaneously with the latent variable model.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Allievi, S. (2010). Immigration and cultural pluralism in Italy: Multiculturalism as a missing model. *Italian Culture, 28*(2), 85–103.

Ambrosini, M., & Molina, S. (2004). *Seconde generazioni: un'introduzione al futuro dell'immigrazione in Italia*. Fondazione Giovanni Agnelli.

Armillei, R. (2015). A multicultural Italy?. In *Cultural, Religious and Political Contestations* (pp. 135-151). Springer, Cham.

Barbagli, M., & Schmoll, C. (2011). La generazione dopo. *Il Mulino*

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Berry, J. W. (1997). Immigration, acculturation, and adaptation. *Applied psychology, 46*(1), 5–34.

Bianchi, G. E. (2011). Italiani nuovi o nuova Italia? Citizenship and attitudes towards the second generation in contemporary Italy. *Journal of Modern Italian Studies, 16*(3), 321–333.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review, 110*(2), 203.

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software, 34*(2), 1–24.

Fabrizi E., Ferrante M.R., & Trivisano C. (2016). Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas. In: Pratesi, M. (Ed.): *Analysis of poverty data by small area methods*, John Wiley and Sons, ISBN: 978-1-118-81501-4, 299-314.

Fabrizi, E., Montanari, G. E., & Ranalli, M. G. (2016). A hierarchical latent class model for predicting disability small area counts from survey data. *Journal of the Royal Statistical Society Series A, 179*(1), 103–131.

Fabrizi, E., & Trivisano, C. (2016). Small Area Estimation of the Gini concentration coefficient. *Computational Statistics & Data Analysis, 99*, 223–234.

Fay, R. E., & Herriot, R. A. (1979). Estimation of income from small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association, 74*, 269–277.

Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B., & Ferrari, S. L. (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics & Data Analysis, 61*, 137–147.

Giuliani, C., Tagliabue, S., & Regalia, C. (2018). Psychological well-being, multiple identities, and discrimination among first and second generation immigrant muslims. *Europe's journal of psychology, 14*(1), 66.

Goodrich B., Gabry J., Ali I., Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. *R package version 2.21.1*, https://mc-stan.org/rstanarm.

Gupta, A. K., and Nadarajah, S. (Eds.). (2004). *Handbook of beta distribution and its applications*. CRC press.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge University Press.

Istat (2017). Indagine sull'integrazione delle seconde generazioni: Obiettivi, metodologia e organizzazione. *Letture statistiche - Metodi*.

Janicki, R. (2020). Properties of the beta regression model for Small Area Estimation of proportions and application to estimation of poverty rates. *Communications in Statistics - Theory and Methods, 49*(9), 2264–2284.

Kowalczyk, J., & Popkewitz, T. S. (2005). Multiculturalism, recognition and abjection:(Re) mapping Italian identity. *Policy Futures in Education, 3*(4), 423–435.

Lahiri, P. (2003). A review of empirical best linear unbiased prediction for the Fay-Herriot small-area model. *The Philippine Statistician, 52*, 1–15.

Lelie, F., Crul, M., & Schneider, J. (2012). The European second generation compared: Does the integration context matter?. Amsterdam University Press.

Levy, C. (2015). Racism, Immigration and New Identities in Italy. In: Mammone, A.; E. G. Parini and G. A. Veltri, eds. Routledge *Handbook of Contemporary Italy*. London: Routledge, pp. 49-63. ISBN 978-0415-60417-8

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software, 42*(10), 1–29.

Molina, I., & Marhuenda, Y. (2015). sae: An r package for Small Area Estimation. *The R Journal, 7*(1), 81–98.

Montanari, G. E., & Ranalli, M. G. (2010). Uno studio della non autosufficienza a partire dai dati dell'Indagine Multiscopo: il caso dell'Umbria. *Rivista di statistica ufficiale, 12*(1), 53–71.

Moretti, A., Shlomo, N., & Sakshaug, J. W. (2020). Multivariate small area estimation of multidimensional latent economic Well-being indicators. *International Statistical Review, 88*(1), 1–28.

Rao, J. N., & Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology.

Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Germany: Springer Science & Business Media.

Thomassen, B. (2010). 'Second generation immigrants' or 'Italians with immigrant parents'? Italian and European Perspectives on Immigrants and their Children. *Bulletin of Italian Politics, 2*(1), 21–44.

White, A., & Murphy, T. B. (2014). BayesLCA: An R package for Bayesian latent class analysis. *Journal of Statiscal Software, 61*(13), 1–28.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.