A pseudo-longitudinal approach to explore educational data

Dr. Clelia Cascella, Dr. Maria Pampaka
Manchester Institute of Education
The University of Manchester

clelia.cascella@manchester.ac.uk – maria.pampaka@manchester.ac.uk

Research aims and questions

- This study is part of a research project aimed at exploring *gender differences in mathematics education*
  - How do they evolve over time?
  - What data can we use to model this?
    - Collecting primary data is expensive and often not effective (sample coverage/size, statistical significance, …).
    - Secondary educational data sets (e.g., national survey or international survey such as OCSE-PISA, TIMMS, PIAAC) are usually available immediately and for free, but
    - combining them poses some methodological challenges due to different reasons such as different sampling design and different conceptual framework (Borgonovi, Choi, Paccagnella, 2018).

What is the problem?

- **At national level** → Many countries do not collect longitudinal data to track students progress
- **At international level** → longitudinal data is not available
  - Trends in International Mathematics and Science Study (TIMSS)
  - Programme in Reading and Literacy Skills (PIRLS)
  - OECD Programme for International Student Assessment (PISA)
  - Assessment of Adult Competencies (PIAAC)
  - **None of these data sets are linked over time**

A pseudo-longitudinal approach

- In the absence of genuinely longitudinal datasets repeated cross-sectional data can be used to create pseudo panels or synthetic cohorts (Deaton, 1985).

Substantially, pseudo-panel (or pseudo-longitudinal) surveys are repeated cross-sectional surveys. Pseudo-longitudinal data is created by pooling comparable cross-section data collected repeatedly over time, using criteria that do not change from one survey to another such as the year of birth.

An empirical example: the INVALSI national sample

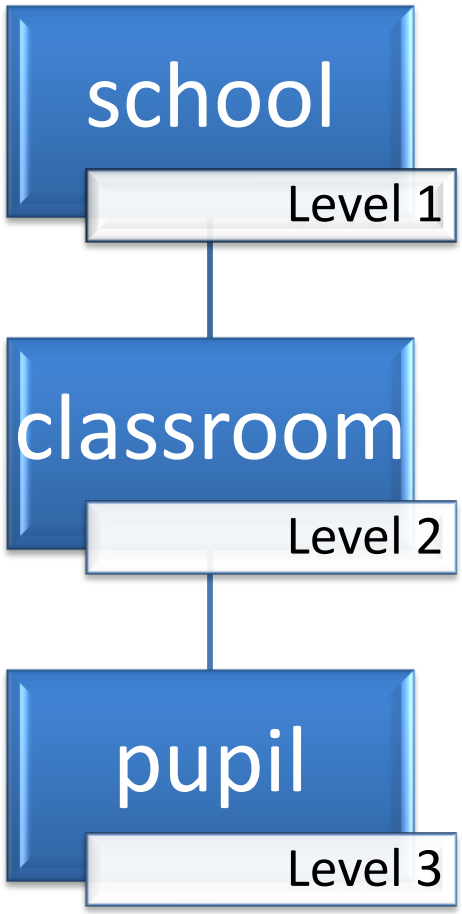
Scholastic year								
Grade	**2009/10**	**2010/11**	**2011/12**	**2012/13**	**2013/14**	**2014/15**	**2015/16**	**2016/17**
G2	30,420	31,842	31,7[illegible]	2[illegible],89	26,356	20,922	24,125	23,001
G5	29,854	31,875	30,843	2[illegible]679	25,331	21,049	23,011	22,014
G6	32,642	40,497	39,668	27,504	-	-	-	-
G8	-	523,111	519,01	520,918	520,917	520,920	519,145	518[illegible]98
G10	-	43,458	41,812	38,060	36,932	27,393	28,635	28,362

Source: Italian National Institute for the Evaluation of Educational System (Istituto Nazionale per la Valutazione del Sistema di Istruzione – INVALSI)

Even without the same "stone", the INVALSI sample is statistically representative of the same birth cohort (and thus exposed for example to the same socio-cultural and economic factors).

Methods

Multilevel modelling



- The same birth cohort:
  - INVALSI sample statistically representative of the same students population over time
  - Data selected for this chapter is statistically representative of the same birth cohort (**pupils born in 2000-2001**)

Variables

- **Pre-primary school attendance**
  - Yes
  - Not
  - [No information about “dosage” (i.e., number of years attended)]
- **Sex**
  - Male
  - Female
- **Regularity** of students' pathways through the education system
  - Regular
  - Anticipated enrolments
  - Retained students
- **Citizenship**
  - Italian student
  - First generation foreign student (born in Italy, but parents born abroad)
  - Second generation foreign student (born abroad, and parents born abroad)
- **Macro-geographical area** (i.e., place of residence), each consisting of 4 regions:
  - North West (Liguria, Lombardia, Piemonte, and Valle d’Aosta);
  - North East (Emilia-Romagna, Friuli Venezia Giulia, Provincia Autonoma di Trento e Bolzano, and Veneto);
  - Centre (Lazio, Marche, Toscana, and Umbria);
  - South (Abruzzo, Campania, Molise and Puglia);
  - South and islands (Basilicata, Calabria, Sardegna, and Sicilia).
- **Socio-economic background (SES)**
  - A continuous variable based on highest parental education and occupation.

Multilevel models

- 3-level (i.e., pupil, classroom and school) model has been estimated at grade 5 (primary school), grade 6 and 8 (lower intermediate school), and at grade 10 (secondary school) to estimate the effect of both individual and contextual variables on math test scores.
- Math test scores have been estimated via a Rasch model.
- Individual scores have been transformed on a scale with mean 200 and standard deviation 40.

Multilevel models - Random effects
Grade 5 (data collected in 2012)

The null-model

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	Coef.	SE	Coef.	SE	Coef.	SE
intercept	201.101	0.239	200.54	0.47	200.331	0.585
	variances					
$v_{0k\ (school)}$					172.292	13.601
$u_{0jk\ (classroom)}$			247.498	11.865	76.97	7.824
$e_{ijk\ (student)}$	1564.666	13.343	1326.892	11.627	1326.55	11.662
	1564.666		1574.39		1575.812	
	variance partition coefficient (Davis et al., 1995)					
$v_{0k\ (school)}$					11%	
$u_{0jk\ (classroom)}$			16%		5%	
$e_{ijk\ (student)}$	100%		84%		84%	

Multilevel models - Random effects

Grade 5 (data collected in 2012)

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	Coef.	SE	Coef.	SE	Coef.	SE
Random intercept	199.79	2.248	194.67	2.414	194.53	2.465
Sex (ref. boy)						
Girl	-7.325	0.434	-7.334	0.37	-7.341	0.37
Preschool attendance (ref. no)						
Yes	-3.087	0.911	2.336	1.223	2.498	1.234
Regularity (ref. In advance)						
Regular	1.335	1.828	2.338	1.585	2.335	1.585
Retained	-7.752	2.202	-6.155	1.909	-6.128	1.909
Citizenship (ref. II generation foreign)						
Italian	7.436	0.959	6.899	0.873	6.906	0.873
First generation foreign	-3.687	1.42	-2.508	1.243	-2.449	1.243
Macro-geographical area (ref. North East)						
North West	-1.932	0.682	-2.275	1.69	-2.462	1.813
Centre	-0.055	0.661	-0.328	1.646	-0.14	1.768
South	3.301	0.688	2.65	1.69	2.335	1.804
South and Islands	-4.751	0.708	-4.55	1.725	-4.936	1.812
SES	8.924	0.224	8.37	0.211	8.366	0.211
v_{0k} (school)	1449.2	11.672	443.81	17.121	175.71	24.484
u_{0jk} (classroom)			1007.9	8.355	269.49	22.878
e_{ijk} (student)					1007.9	8.355

Multilevel models - Random effects
Grade 6 (data collected in 2013)

The null-model

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	coef	SE	coef	SE	coef	SE
intercept	201.101	0.239	200.54	0.47	200.331	0.585
	variances					
$v_{0k\ (school)}$					172.292	13.601
$u_{0jk\ (classroom)}$			247.498	11.865	76.97	7.824
$e_{ijk\ (student)}$	1564.666	13.343	1326.892	11.627	1326.55	11.662
	1564.666		1574.39		1575.812	
	variance partition coefficient (Davis et al., 1995)					
$v_{0k\ (school)}$	0%		0%		11%	
$u_{0jk\ (classroom)}$			16%		5%	
$e_{ijk\ (student)}$					84%	

Multilevel models - Random effects

Grade 6 (data collected in 2013)

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	coef	SE	coef	SE	coef	SE
Random intercept	205.503	2.393	203.359	2.508	202.582	2.58
Sex (ref. boy)						
Girl	-7.723	0.438	-7.786		-7.799	0.423
Preschool attendance (ref. no)						
Yes	1.616	0.704	3.723	0.956	4.373	1.016
Regularity (ref. In advance)						
Regular	0.311	2.087	0.132	2.043	0.068	2.042
Retained	-19.55	2.267	-18.566	2.22	-18.46	2.219
Citizenship (ref. II generation foreign)						
Italian	7.075	0.958	7.673	0.964	7.767	0.967
First generation foreign	-3.52	1.339	-2.884	1.318	-2.746	1.319
Macro-geographical area (ref. North East)						
North West	0.926	0.672	1.016	1.128	1.223	1.325
Centre	-4.582	0.686	-4.5	1.153	-4.389	1.352
South	-12.94	0.66	-13.465	1.114	-13.27	1.309
South and Islands	-19.796	0.725	-20.27	1.194	-20.036	1.394
SES	9.854	0.23	9.358	0.237	9.313	0.238
	variances					
v_{0k} (school)					65.772	7.761
u_{0jk} (classroom)			131.265	7.26	67.408	6.893
e_{ijk} (student)	1306.163	11.158	1177.821	10.338	1177.495	10.334

Multilevel models - Random effects
Grade 8 (data collected in 2015)

The null-model

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	Coef.	SE	Coef.	SE	Coef.	SE
intercept	200.872	0.23	198.908	0.365	198.908	0.365
	variances					
$v_{0k\ (school)}$					0	0
$u_{0jk\ (classroom)}$			788.642	18.737	788.642	18.737
$e_{ijk\ (student)}$	1593.307	13.349	827.027	8.437	827.027	8.437
Total	1593.307		1615.669		1615.669	
	variance partition coefficient (Davis et al., 1995)					
$v_{0k\ (school)}$					0%	
$u_{0jk\ (classroom)}$			49%		49%	
$e_{ijk\ (student)}$	100%		51%		51%	

Note. Socio-economic background not available for this grade

Multilevel models - Random effects

Grade 8 (data collected in 2015)

	Model 1 [student]		Model 2 [classroom]	
	Coef.	SE	Coef.	SE
Random intercept	203.719	2.618	204.523	2.396
Sex (ref. boy)				
Girl	-5.854	0.498	-5.811	0.413
Preschool attendance (ref. no)				
Yes	3.315	0.775	2.188	0.968
Regularity (ref. In advance)				
Regular	0.369	2.193	-2.34	1.867
Retained	-17.703	2.363	-21.029	2.031
Citizenship (ref. II generation foreign)				
Italian	6.674	1.185	7.668	1.017
First generation foreign	0.857	1.66	2.532	1.455
Macro-geographical area (ref. North East)				
North West	1.415	0.748	2.286	1.141
Centre	-3.9	0.789	-4.044	1.191
South	-17.217	0.768	-15.907	1.162
South and Islands	-19.232	0.799	-18.392	1.247
SES	-	-	-	-
	variances			
v_{0k} (school)				
u_{0jk} (classroom)			745.751	19.332
e_{ijk} (student)	1593.307	13.349	806.829	8.849

Multilevel models - Random effects
Grade 10 (data collected in 2017)

The null-model

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	Coef.	SE	Coef.	SE	Coef.	SE
intercept	198.738	0.201	195.632	0.633	195.773	0.853
	variances					
$v_{0k\ (school)}$					640.326	33.752
$u_{0jk\ (classroom)}$			830.076	26.518	205.966	10.588
$e_{ijk\ (student)}$	1544.786	11.189	707.839	5.281	707.947	5.281
Total	1544.786		1537.915		1554.239	
	variance partition coefficient (Davis et al., 1995)					
$v_{0k\ (school)}$					41%	
$u_{0jk\ (classroom)}$			54%		13%	
$e_{ijk\ (student)}$	100%		46%		46%	

Note. Preschool attendance not YET available for this grade

Multilevel models - Random effects
Grade 10 (data collected in 2017)

	Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	Coef.	SE	Coef.	SE	Coef.	SE
Random intercept	224.15	2.14	215.874	2.035	216.104	2.279
Sex (ref. boy)						
Girl	-9.276	0.392	-6.565	0.355	-6.742	0.353
Preschool attendance (ref. no)						
Yes	-	-	-	-	-	-
Regularity (ref. In advance)						
Regular	-6.261	1.922	-3.471	1.493	-3.322	1.491
Retained	-27.878	1.981	-13.072	1.546	-12.64	1.544
Citizenship (ref. II generation foreign)						
Italian	4.349	0.873	4.317	0.686	4.246	0.684
First generation foreign	-2.41	1.226	-2.45	0.955	-2.487	0.952
Macro-geographical area (ref. North East)						
North West	-5.762	0.591	-6.152	1.728	-6.132	2.273
Centre	-14.247	0.607	-14.413	1.738	-14.819	2.287
South	-23.854	0.61	-25.111	1.715	-25.171	2.259
South and Islands	-32.831	0.639	-33.672	1.778	-34.044	2.322
SES	8.489	0.206	1.773	0.175	1.618	0.175
v_{0k} (school)					431.273	24.641
u_{0jk} (classroom)			600.227	19.862	181.158	9.922
e_{ijk} (student)	1228.134	9.663	689.95	5.62	689.825	5.615

Advantages of pseudo-longitudinal approach

	Grade 5 (2012)						Grade 6 (2013)						Grade 8 (2015)				Grade 10 (2017)					
	Model 1 [student]		Model 2 [classroom]		Model 3 [school]		Model 1 [student]		Model 2 [classroom]		Model 3 [school]		Model 1 [student]		Model 2 [classroom]		Model 1 [student]		Model 2 [classroom]		Model 3 [school]	
	Statistic	SE	β_{0j}	SE	β_{0jk}	SE	Statistic	SE	β_{0j}	SE	β_{0jk}	SE	Statistic	SE	β_{0j}	SE	β_{0jk}	SE	β_{0j}	SE	β_{0jk}	SE
Random intercept	199.789	2.248	194.667	2.414	194.534	2.465	205.503	2.393	203.359	2.508	202.582	2.58	203.719	2.618	204.523	2.396	224.15	2.14	215.874	2.035	216.104	2.279
SES	8.924	0.224	8.37	0.211	8.366	0.211	9.854	0.23	9.358	0.237	9.313	0.238	-	-	-	-	8.489	0.206	1.773	0.175	1.618	0.175
Sex (ref. boy)																						
Girl	-7.325	0.434	-7.334	0.37	-7.341	0.37	-7.723	0.438	-7.786		-7.799	0.423	-5.854	0.498	-5.811	0.413	-9.276	0.392	-6.565	0.355	-6.742	0.353
Preschool attendance (ref. no)																						
Yes	-3.087	0.911	2.336	1.223	2.498	1.234	1.616	0.704	3.723	0.956	4.373	1.016	3.315	0.775	2.188	0.968	-	-	-	-	-	-
Regularity (ref. In advance)																						
Regular	1.335	1.828	2.338	1.585	2.335	1.585	0.311	2.087	0.132	2.043	0.068	2.042	0.369	2.193	-2.34	1.867	-6.261	1.922	-3.471	1.493	-3.322	1.491
Retained	-7.752	2.202	-6.155	1.909	-6.128	1.909	-19.55	2.267	-18.566	2.22	-18.46	2.219	-17.703	2.363	-21.029	2.031	-27.878	1.981	-13.072	1.546	-12.64	1.544
Citizenship (ref. II generation foreign)																						
Italian	7.436	0.959	6.899	0.873	6.906	0.873	7.075	0.958	7.673	0.964	7.767	0.967	6.674	1.185	7.668	1.017	4.349	0.873	4.317	0.686	4.246	0.684
First generation foreign	-3.687	1.42	-2.508	1.243	-2.449	1.243	-3.52	1.339	-2.884	1.318	-2.746	1.319	0.857	1.66	2.532	1.455	-2.41	1.226	-2.45	0.955	-2.487	0.952
Macro-geographical area (ref. North East)																						
North West	-1.932	0.682	-2.275	1.69	-2.462	1.813	0.926	0.672	1.016	1.128	1.223	1.325	1.415	0.748	2.286	1.141	-5.762	0.591	-6.152	1.728	-6.132	2.273
Centre	-0.055	0.661	-0.328	1.646	-0.14	1.768	-4.582	0.686	-4.5	1.153	-4.389	1.352	-3.9	0.789	-4.044	1.191	-14.247	0.607	-14.413	1.738	-14.819	2.287
South	3.301	0.688	2.65	1.69	2.335	1.804	-12.94	0.66	-13.465	1.114	-13.27	1.309	-17.217	0.768	-15.907	1.162	-23.854	0.61	-25.111	1.715	-25.171	2.259
South and Islands	-4.751	0.708	-4.55	1.725	-4.936	1.812	-19.796	0.725	-20.27	1.194	-20.036	1.394	-19.232	0.799	-18.392	1.247	-32.831	0.639	-33.672	1.778	-34.044	2.322
	Variances explained by variables at each level						Variances explained by variables at each level						Variances explained by variables at each level				Variances explained by variables at each level					
$v_{0k\ (school)}$	1449.182	11.672	443.806	17.121	175.711	24.484					65.772	7.761									431.273	24.641
$u_{0jk\ (classroom)}$			1007.876	8.355	269.491	22.878			131.265	7.26	67.408	6.893			745.751	19.332			600.227	19.862	181.158	9.922
$e_{ijk\ (student)}$					1007.85	8.355	1306.163	11.158	1177.821	10.338	1177.495	10.334	1593.307	13.349	806.829	8.849	1228.134	9.663	689.95	5.62	689.825	5.615

Results and conclusions

- Individual variables (at pupil level) explain most of the individual math test score (unless at grade 10!)
- On average, female students are disadvantaged compared to males. Such disadvantage is constant at grade 5 and grade 6, slightly decreases at grade 8, and reaches a quarter of standard deviation on the Rasch scale at grade 10 (N.B. gender coefficient at grade 10 is double than that at grade 8).
- Moreover, at grade 10, gender effect is smaller at classroom or school level than at individual level. In contrast to this, at previous grades, gender coefficients are constant across hierarchical levels.
- Consistently with our on-going studies, contextual variables play a critical role that affect test scores increasingly over time:
  - For example, compared to living in northern Italy, living in southern country has a negative effect on math test scores.
  - Students living in south have a math test score similar to students living in northern Italy at grade 5 (-4.751 points on the ability scale). Such disadvantage reaches half standard deviation at grade 6 and at grade 8, and increases up to ¾ of standard deviation at grade 10.
- In the absence of panel data, cross-sectional designs are generally carried out.
- INVALSI has started to provide properly longitudinal data only in recent years. Therefore, such longitudinal data does not cover the entire span reported in this study.
- Moreover, compared to a purely cross-sectional perspective, the approach employed here offers a representativeness of the same birth cohort: the answers provided by students belonging to different samples over time are statistically representative of the answers provided by the same students population over time.
- Pseudo-panel approach even though not properly longitudinal actually provide “longitudinal” results, at least at systemic (rather than at properly individual) level.

Thank you for coming!

Discussion time!!!