PLAN OVERVIEW

A Data Management Plan created using DMPonline.be

Title: Phantom compositional effects in value-added models: investigating the effects of the type of ability estimators

Creator: Elodie Pools

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Project abstract:
Many countries use value-added(VA) estimates as indicators of the effectiveness of schools. Because VA models adjust for prior achievement, they offer information about the effect of prior academic composition of schools on students' current achievement. However, these compositional effects might be biased, resulting in phantom effects. One of the characteristics that drives the bias in these compositional effects is the measurement error on prior achievement. A challenge is therefore to prevent measurement error from biasing school VA estimates.

This project focuses on the relationship between phantom effects and the way prior achievement is estimated (i.e. ability estimator). Different ability estimators account differently for measurement error on prior achievement, some estimators leading to more phantom effects. Using simulated and real data, this project will identify which ability estimators are best when estimating compositional effects. Several ability estimators, including traditional methods of scoring and IRT estimation, are compared in their capacity to produce adequate prior achievement estimates for VA models. More precisely, we will assess the recovery (1) of school differences in prior achievement, (2) of compositional effects, and (3) of school VA estimates. We will include recommendations regarding the estimation of ability for the analysis of compositional effects and VA-estimates, providing practitioners with adequate and practicable statistical routines

ID: 211614

Start date: 01-10-2024

End date: 30-09-2027

PHANTOM COMPOSITIONAL EFFECTS IN VALUE-ADDED MODELS: INVESTIGATING THE EFFECTS OF THE TYPE OF ABILITY ESTIMATORS

FWO DMP (Flemish Standard DMP)

1. RESEARCH DATA SUMMARY

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: ● Generate new data ● Reuse existing data	*Please choose from the following options:* ● Digital ● Physical	*Please choose from the following options:* ● Observational ● Experimental ● Compiled/aggregated data ● Simulation data ● Software ● Other ● NA	*Please choose from the following options:* ● .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … ● NA	*Please choose from the following options:* ● <100MB ● <1GB ● <100GB ● <1TB ● <5TB ● <10TB ● <50TB ● >50TB ● NA	
PISA	PISA survey data	Reuse existing data	Digital	● Observational	.sas7bdat	<100GB	
PIRLS	PIRLS 2016 survey data	Reuse existing data	Digital	● Observational	.sas7bdat	<1GB	
PIRLS repeat	PIRLS repeat (2018 data)	Reuse existing data	Digital	● Observational	.csv	<100MB	
Simulation data	Created with data-generating code using R	Generate new data	Digital	● Simulation data	.csv (data) .r (code)	<100GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

PISA data: data retrieved from: PISA data and methodology | OECD
PIRLS 2016 data: data retrieved from: PIRLS 2016 International Database
PIRLS repeat data: data received upon request via the research center (Centre for Educational Effectiveness and Evaluation)

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

An ethical application has been submitted and approved for the re-use of PIRLS-repeat data (ref. nr. G-2024-8872). Simulated data have no ethical issue while the PISA and PIRLS 2016 data are publicly available.
Concerning the PIRLS-repeat data, the dataset of interest consists in a re-assessment of the Flemish students who participated in the PIRLS 2016 assessment 2 years later, in 2018. The data thus allow to see the evolution in time in the reading achievement of Flemish primary school students between the 2016 assessment (public data) and the 2018 re-assessment (data owned by the Research Center). The PIRLS-repeat data are already pseudonymized and the variables that will be used in this study are the pseudonymized identification variables (to link the students' 2018 record with their corresponding record in the PIRLS 2016 public database) and the variables related to their performance on the reading test in the 2018 reading re-assessment (to be able to estimate the evolution in time of reading proficiency).

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

PISA and PIRLS 2016 data (public databases) contain personal data. These databases (pseudomized) are publicly available.
PIRLS-repeat data is a re-assessment of the students who participated to PIRLS 2016 in Flanders, 2 years later, in 2018. This dataset is not publicly available. In this dataset, we mostly need the student pseudomized id (to link with the participating students' records in the PIRLS 2016 public database) and information regarding their achievement in 2018. The data are already pseudonymized.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

The PISA, PIRLS 2016 and PIRLS-repeat data belong to, respectively, the OECD (Organisation for Economic Co-operation and Development), the IEA (International Association for the Evaluation of Educational Achievement) and the Centre for Educational Effectiveness and Evaluation. It is thus not possible for me to disseminate these data. Researchers interested in working with the PISA and PIRLS 2016 data can download them from their respective institutional websites; those interested in the PIRLS-repeat data can contact the Research Centre.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

PIRLS-repeat data were previously collected by the research center, PIRLS 2016 by the IEA, and PISA by the OECD. Proper acknowledgement and reference to these datasets will be made in this research outputs.

2. DOCUMENTATION AND METADATA

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The following files will be saved in the same folder as the dataset:
· Project documentation will be provided in a readme file. This file will include the project name, keywords, name of involved researchers and their ORCID ID, name of funder, funding code, start- and end date of the project, references of datasets, creative common license, approval of ethical committee, links to publications.
· The ethical application will be saved as a PDF document and the approval will be added to the project documentation.
· The data management plan will be provided as a pdf file.
· The data preparation and statistical analyses will be documented in an annotated analysis code file (R-script). The version of R and R packages will be documented the code file.
For the simulation study, beside the R code, a codebook will also detail the meaning of the variables in the produced datafiles.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

The PISA, PIRLS2016 and PIRLS-repeat datasets have been created and documented by their respective Authors. I will not share these data but refer to their respective source.
For the simulation study, as this dataset was built in the context of large-scale assessments in education, metadata will be in the continuity of the metadata describing the most prominent studies in the field (PIRLS and PISA for instance), for consistency with the practices in the scientific community working with these educational survey data.

3. DATA STORAGE & BACK-UP DURING THE RESEARCH PROJECT

Where will the data be stored?

The data will be stored on the secured OneDrive for Business provided by KU Leuven. PISA, PIRLS and PIRLS-repeat data will not be shared with other researchers, as I do not own these data.
If simulation study data needed to be shared, they would be shared via Belnet.

How will the data be backed up?

There will be standard automatic back-up provided by KU Leuven ICTS for my storage solution.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

As the size of all data files do not exceed the available individual storage space of 250GB (OneDrive), there is sufficient storage and backup capacity during the project.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Multi-factor authentication is activated for the KU Leuven login of all researchers having access to the data.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

There are no costs expected as the size of the data files does not exceed the available storage space.

4. DATA PRESERVATION AFTER THE END OF THE RESEARCH PROJECT

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Data will be preserved for 10 years according to KU Leuven RDM policy.

Where will these data be archived (stored and curated for the long-term)?

After the study, we will transfer the data to Prof. Koen Aesaert's OneDrive for Business for at least 10 years.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

none

5. DATA SHARING AND REUSE

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Other, please specify:

-PISA and PIRLS 2016 data: no: data are already publicly available and can be downloaded from the OECD's and IEA's websites, respectively.
-PIRLS-repeat data: no: I'm not owner of these data, accessing this dataset requires permission of the Research Center (Centre for Educational Effectiveness and Evaluation)
-Simulated data: Data will be shared upon request.

If access is restricted, please specify who will be able to access the data and under what conditions.

Request to access the simulation study should be done.
PISA, PIRLS and PIRLS-repeat data cannot be made available; PISA data can be freely downloaded from the official OECD website; PIRLS data can be freely downloaded from the official IEA website; PIRLS-repeat data sharing is only possible upon request to the administrator of the PIRLS-repeat project (Centre for Educational Effectiveness and Evaluation).

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Other

PIRLS-repeat data: these data are not public. Sharing this dataset is only possible upon request to the administrator of the PIRLS-repeat project.
No concern for PISA, PIRLS 2016 (public data) or simulation data.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

The data (simulation study only) will be made available via the KU Leuven Research Data Repository (RDR).

When will the data be made available?

The data will be made available at the end of the research project.

Which data usage licenses are you going to provide? If none, please explain why.

- for the simulation study data: CC-BY-NC-ND
- for the PISA, PIRLS and PIRLS-repeat datasets: I don't own the data.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes, a DOI reference number will be added, referring to the RDR location of the data

What are the expected costs for data sharing? How will these costs be covered?

No costs are expected.

6. RESPONSIBILITIES

Who will manage data documentation and metadata during the research project?

The principal investigator will manage data documentation and metadata during the project supported by the other involved researchers.

Who will manage data storage and backup during the research project?

The principal investigator will manage data storage and backup during the project supported by the other involved researchers.

Who will manage data preservation and sharing?

The principal investigator will manage data preservation and sharing during the project supported by the other involved researchers.

Who will update and implement this DMP?

The principal investigator will update and implement this DMP supported by the other involved researchers.