Comparative Causal Inference within PISA

RESEARCH REPORT

Lauke Stoel (6899544)

Supervisors: Marieke van Onna (CITO) and
Remco Feskens (CITO and University of Twente)

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

Date: 19/12/2021
Word count: 2463

FETC-approved: 21-1939

Candidate journal: Large Scale Assessments in Education

1 Introduction

Myriads of valuable data on pupil's cognitive ability are periodically collected through International Large-Scale Assessment (ILSA) programmes such as the Programme for International Student Assessment (PISA). PISA is a triennial survey of 15-year-old students around the world that assesses their knowledge and skills in three core domains: reading, mathematics and science. Through PISA, the Organisation for Economic Co-operation and Development (OECD) aims to provide internationally comparable evidence on student performance (OECD, 2020). Ideally, these data could be used to compare different education systems to evaluate what features of each system lead to more favourable educational outcomes and inform national policy. This implies a question of causality. A randomised controlled trial would be the golden standard method to expose a causal relationship. However, due to the observational nature of the data, such an experiment is impossible to conduct.

Rubin's potential outcome framework provides a model in which such causal questions could theoretically be addressed in a quasi-experimental setting (Rubin et al., 2004). In the context of ILSAs, it entails a theoretical experiment where we treat a national educational policy of interest as treatment condition. We theorise that a pupil who is subject to one treatment condition could also have been assigned to the other treatment condition of which the data are non-existing. Those data are treated as missing and are estimated based on existing data from pupils with similar background characteristics, who are in the other (observed) treatment condition. In a recent review of research on causal inference with large-scale assessments in education, Kaplan (2016) proposed an approach to causal inference within Rubin's potential outcomes framework. This approach could theoretically expose causal links between educational outcomes and features of education systems. However, his approach comes with a disclaimer: it has not yet been tested on existing data and currently available software is likely insufficient to fully execute the proposed statistical models (Kaplan, 2016).

The present thesis tests Kaplan's approach to causal inference on a question currently relevant in the Netherlands, namely whether it would aid the equality of educational opportunity if the age at which pupils are first selected into educational tracks is postponed. This topic is much discussed, as previous research suggests that early tracking is associated with a stronger positive relationship between socioeconomic status (SES) and ability, which is indicative of inequality of educational opportunity (Bol et al., 2014; Veldhuis & Versteegh, 2021). In the context of Kaplan's approach, the equality of educational opportunity would be the educational outcome and the age of first selection would be the feature of an education system.

Kaplan's approach could provide a comprehensive way of addressing such policy questions, but is expected to run into software limitations. Another, theoretical, limitation is that Kaplan's approach is designed to treat the ability level of a pupil as the outcome variable, whereas the above question requires an analysis with a relationship as the outcome variable, namely the one between SES and ability. Furthermore, Kaplan's approach facilitates analyses with a binary treatment assignment, while many real-life policy questions pertain to multiple treatment categories, in this case the age of first selection. Therefore, the research question the current project addresses is: *how can Kaplan's approach to causal inference be extended and applied to inform educational policy regarding tracking, within the confines of currently available software?*

Thus, the objective of this report is twofold: to develop a methodological extension of an existing approach and to test this approach by means of a practical application. The structure of the paper is as follows. First, I provide the theoretical framework of this research: I position Kaplan's approach within causal inference literature, translate his general approach to a step-by-step methodology and subsequently propose two extensions to his approach. The result is three different methodologies: one directly following from Kaplan's original approach, one extension that facilitates including multiple treatments and a second extension that facilitates treating the relationship between a context variable of interest and ability as outcome variable. This section provides an answer to the first part of the research question, namely how Kaplan's approach can be extended. The subsequent methodology section specifies how I will assess the applicability of each of those methods. The methodology section of the final thesis will contain the specifications of how each method is applied to the question of tracking. Since this report is an intermediate product, that part of the methodology section, the results, discussion and conclusion are not yet included.

2 Theoretical framework

2.1 Positioning Kaplan's approach within causal inference research

Center stage in Kaplan's approach is his recommendation to use two-step Bayesian propensity score analysis (BPSA) as developed by Kaplan and Chen (2012). It concerns a Bayesian extension of propensity score analysis (PSA). A main advantage of PSA over other common quasi-experimental methods is its potential to mitigate an endogeneity problem, namely self-selection into schools, by matching the observations in the different treatment arms on important covariates that are correlated with the treatment assignment (Cordero et al., 2018). However, this benefit only materialises if the assumption holds that all covariates that do in fact influence the treatment assignment are observed or at least correlated with the observed measures. Especially in cross-national research, this is a tall order. Furthermore, the success of the method depends on whether the resulting sample of observations that have the same distribution on all covariates in both treatment arms is large enough. The more covariates that are included, the more likely the assumption is to hold, but the smaller the resulting sample size will be. BPSA provides the same potential advantages and faces the same challenges as PSA regarding covariate measurement and sample size. However, an advantage of BPSA over PSA is that the Bayesian extension provides us with the opportunity to include prior information in our estimation of the propensity scores and the treatment effect. The advantage of two-step BPSA over Bayesian joint modelling procedures is that the propensity scores and the treatment effect are estimated separately (Kaplan & Chen, 2012). The two-step approach prevents the estimation of the posterior distribution of the propensity scores from being affected by the estimation of the outcome variable, which is observed after treatment assignment (Kaplan, 2016).

The use of two-step BPSA addresses another relevant methodological challenge. Researchers have recently been drawing more attention to the imbalance that exists between the analytical efforts invested in modelling the cognitive outcomes of PISA and the efforts invested in modelling the context variables. In the former case, measurement error of the latent construct *ability* is taken into account by drawing ten plausible values of ability for each pupil. In the latter case, a simple point estimate is deemed sufficient. Failure to account for measurement error on background variables can lead to biased estimates of ability when comparing across different subgroups of a background variable (Frey & Hartig, 2020; Rutkowski & Rutkowski, 2016). Two-step BPSA addresses this methodological challenge by constructing a posterior distribution of propensity scores for each individual based on a vector of chosen covariates and drawing any large number of values from that distribution per individual. This mimics the way the measurement error of ability is accounted for using plausible values. The posterior standard deviation of that set of propensity scores can be interpreted as the measurement error on the combination of all relevant covariates. So, while it does not quantify the measurement error of each covariate separately, BPSA does give us a measure of uncertainty on the latent construct describing how similar pupils are to each other on a set of covariates. This makes the final estimate of the treatment effect and its comparison across subgroups more accurate.

Nevertheless, selecting the appropriate covariates for this type of analysis remains a major challenge. Kaplan situates his approach within Rubin's potential outcomes framework. One of the assumptions of this framework provides us with at least a theoretical goal post of when the appropriate covariates have been included in the analysis. This is the *strong ignorability* assumption, which holds when the potential outcome under either treatment assignment is completely independent of the treatment assignment mechanism, given a set of covariates. Kaplan refers to the work of Mackie (1974) for further guidance in selecting the covariates that are of immediate concern to the causal question at hand. This and other assumptions are embedded in the four conditions Kaplan specifies must hold. For a further specification of these conditions, please see Appendix A.

2.2 Translating Kaplan's approach to a methodology

The first step in applying Kaplan's approach should always be to perform a check on the four conditions in Appendix A. The next phase in applying Kaplan's approach is performing a two-step BPSA.

Step 1: obtain a propensity score distribution for each pupil.

Start by selecting an appropriate model to compute propensity scores with. A propensity score reflects how likely an observation is to be assigned treatment $T = 1$ given a set of chosen covariates X. Kaplan's approach assumes binary treatment assignment, so a logit model such as

$$\log \left\{ \frac{p(T=1 \mid X)}{1-p(T=1 \mid X)} \right\} = \alpha + \beta X \,, \tag{1}$$

is in place, where α and β are the model parameters that determine how the covariates are weighted to produce the propensity score. An optional step is to include priors on these model parameters in their estimation, such that their posterior distributions read as

$$\begin{aligned} p(\alpha \mid \beta; X, T) &\propto p(X, T \mid \alpha, \beta)\; p(\alpha \mid \beta)\,, \\ p(\beta \mid \alpha; X, T) &\propto p(X, T \mid \alpha, \beta)\; p(\beta \mid \alpha)\,. \end{aligned} \tag{2}$$

Then obtain $m = n.iter$ posterior propensity scores $\hat{e}(x)$ for each pupil, using the expected a posteriori (EAP) estimates for α and β resulting from Equation (1) as follows,

$$\hat{e}(x) = \hat{p}(T=1 \mid X=x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}\,. \tag{3}$$

The result is a distribution of propensity scores for each pupil, of which the EAP and the posterior standard deviation can be calculated. However, instead of taking the EAP here and using that as the final propensity score per pupil, two-step BPSA involves an extra operation to also estimate the treatment effect in a Bayesian manner.

Step 2: obtain an estimate for the treatment effect

Select which propensity score method to employ: stratification on $\hat{e}(x)$, weighting or optimal full matching (Cochran, 1968; Hansen & Klopfer, 2006; Hirano & Imbens, 2001). For simplicity's sake, we assume a choice for either stratification or optimal full matching, so that the result is m sets of two matched subsamples of the total population and m corresponding sets of matched propensity scores, η_i.

Then select an appropriate Bayesian outcome model to estimate the treatment effect with. A straightforward, linear model could suffice, such as

$$\hat{y}_i = \alpha + \gamma\, T_i + \sum_{k=1}^{K} \beta_k X_{ik}\,, \tag{4}$$

where $\hat{y}_i$ is the ability score estimate of a pupil i, α is the intercept, T_i denotes a pupil's treatment assignment and γ its coefficient, i.e. the treatment effect, β_k represents the set of regression coefficients of X_{ik}, values for that pupil on the set of covariates.

Now we can obtain the posterior distribution of the treatment effect. It is optional to include a prior for the treatment effect here, such that the posterior distribution of γ is

$$p(\gamma \mid \alpha, \beta; x, y, T) \propto p(x, y, T \mid \gamma, \alpha, \beta)\, p(\gamma \mid \alpha, \beta)\,. \tag{5}$$

Now we estimate the treatment effect by drawing $J = n.iter$ values for γ from Equation 5 *for each set of estimated propensity scores* η_i. See Figure 1 for a visual aid of this process.

$$\mathit{Step}\ 1 : \hat{\mathbf{P}} = \begin{bmatrix} \hat{e}(x)_{11} & \hat{e}(x)_{12} & \dots & \hat{e}(x)_{1m} \\ \hat{e}(x)_{21} & \hat{e}(x)_{22} & \dots & \hat{e}(x)_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}(x)_{N1} & \hat{e}(x)_{N2} & \dots & \hat{e}(x)_{Nm} \end{bmatrix}$$

$$\begin{array}{cccc} \Downarrow & \Downarrow & \dots & \Downarrow \\ \eta_1 & \eta_2 & \dots & \eta_m \end{array}$$

$$\mathit{Step}\ 2 : \mathbf{\Gamma} = \begin{bmatrix} \gamma_1(\eta_1) & \gamma_1(\eta_2) & \dots & \gamma_1(\eta_m) \\ \gamma_2(\eta_1) & \gamma_2(\eta_2) & \dots & \gamma_2(\eta_m) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_J(\eta_1) & \gamma_J(\eta_2) & \dots & \gamma_J(\eta_m) \end{bmatrix}$$

Figure 1: Connection between step 1 and 2.

The first matrix $\hat{\mathbf{P}}$ contains m posterior propensity scores for each of $1 : N$ pupils. For each in m iterations, the propensity score matching results in a different subset of comparable pupils. Their set of matched propensity scores, represented by the green values, are stored in η_i. Then for each η_i, J values for the treatment effect γ are drawn, as presented in the second $J \times m$ matrix $\mathbf{\Gamma}$.

The result is $J \times m$ estimates of the treatment effect. To obtain the final treatment effect estimate Kaplan and Chen (2012) provide the following estimator:

$$E(\gamma \mid x, y, T) = m^{-1} J^{-1} \sum_{i=1}^{m} \sum_{j=1}^{J} \gamma_j \left(\eta_i\right) \ , \tag{6}$$

which takes the average of the posterior sample mean of the treatment effect across all sets of propensity scores. The next step is to check the balance of the propensity score method of choice by examining the variance ratios of both treatment conditions and Cohen's d.

To determine the stability of the treatment effect, the last step in Kaplan's approach is to perform a sensitivity analysis on the causal estimate. We do this by including plausible values on *unobserved* covariates in the outcome model in Equation 4 that could possibly be of influence on the treatment assignment or the effect. Small changes in these plausible values that inflict a large change in the causal estimate indicate the presence of hidden biases.

2.3 Two extensions of Kaplan's approach

This section introduces the two extensions of Kaplan's approach I propose to make it applicable to real-life policy questions that require 1) the possibility to include a multiple treatments and 2) treating another relationship as an outcome variable.

2.3.1 Extension of step 1: multiple treatment categories

To facilitate the use of a treatment variable with multiple categories, one might consider replacing the logit model, that was used to estimate the propensity scores under step 1, with a multinomial model. However, this would be quite computationally intensive, especially when estimated in a Bayesian manner (Caliendo & Kopeinig, 2008). Based on the work by Lechner (2001), I suggest estimating a series of binomial models instead. Using the example of age, it entails running several analyses with each a different dichotomy that defines *early* and *late* tracking systems, see Table 1. Note that designing the dichotomies this way retains the ordinal nature of the example variable age.

Country	**Age**	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$	$\mathbf{T_4}$	$\mathbf{T_5}$
Germany	10	0	0	0	0	0
$\vdots$	$\vdots$	$\vdots$				
Luxembourg	11	1	0	0	0	0
$\vdots$	$\vdots$		$\vdots$			
Netherlands	12	1	1	0	0	0
$\vdots$	$\vdots$			$\vdots$		
Slovenia	14	1	1	1	0	0
$\vdots$	$\vdots$				$\vdots$	
Portugal	15	1	1	1	1	0
$\vdots$	$\vdots$					$\vdots$
Finland	16	1	1	1	1	1

Table 1: Example dichotomies in treatment assignment T by age of first selection per country

A pair-wise comparison of the treatment effects resulting from each of the analyses will reveal the most informative dichotomy.

2.3.2 Extension of step 2: relationship as outcome

The second extension concerns the possibility of treating a relationship between any covariate of interest, such as SES, and the ability of a pupil as outcome. To that end, I suggest selecting an outcome model with an interaction term between SES and treatment assignment T under step 2 instead of Equation 4, such as

$$\hat{y}_i = \alpha + \gamma\, T_i + \beta_1 SES + \beta_2 T_i SES + \sum_{k=3}^{K} \beta_k X_{ik}\,. \tag{7}$$

Instead of constructing the posterior for the regression coefficient of the treatment effect γ, we solve for β_2 and run through the rest of the steps as before.

3 Methodology

To answer the question of how Kaplan's approach can be extended and applied to inform educational policy regarding tracking, I will run three different analyses using the three versions of the approach to the question at hand: Kaplan's approach as outlined in section 2.2, one including the first extension and one including both proposed extensions. I will assess the applicability of each methodology by evaluating

1. its ability to produce an outcome of interest to the policy question (see condition 1, Appendix A),
2. the stability of the resulting treatment effect as uncovered through a sensitivity analysis, and
3. the ease of application, given the limits of currently available software.

The methodology section of the final thesis will include a specification of how these three methods will be applied to the question of tracking.

We use the data from PISA 2018, which are publicly available. The data set contains the responses of 10,215 pupils from 79 participating countries and background information on the pupils, their parents and schools (OECD, 2020). We enrich this data set with relevant country-level characteristics, such as the age at which pupils are selected into tracks. We perform all analyses in RStudio (RStudio Team, 2020).

References

Bol, T., Witschge, J., Van de Werfhorst, H. G., & Dronkers, J. (2014). Curricular Tracking and Central Examinations: Counterbalancing the Impact of Social Background on Student Achievement in 36 Countries. *Social Forces*, *92*(4), 1545–1572. https://doi.org/10.1093/sf/sou003

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, *22*(1), 31–72.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295–313.

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal Inference on Education Policies: A Survey of Empirical Studies Using Pisa, Timss and Pirls [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12217]. *Journal of Economic Surveys*, *32*(3), 878–915. https://doi.org/10.1111/joes.12217

Frey, A., & Hartig, J. (2020). Methodological challenges of international student assessment. *Monitoring Student Achievement in the 21st Century*, 39–49.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of computational and Graphical Statistics*, *15*(3), 609–627.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, *2*(3), 259–278.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-scale Assessments in Education*, *4*(1), 7. https://doi.org/10.1186/s40536-016-0022-6

Kaplan, D., & Chen, J. (2012). A two-step bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, *77*(3), 581–609.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econometric evaluation of labour market policies* (pp. 43–58). Springer.

Mackie, J. L. (1974). *The Cement of the Universe: A Study of Causation*. Oxford University Press. https://doi.org/10.1093/0198246420.001.0001

OECD. (2020). *PISA 2018 Results (Volume V)*. https://doi.org/https://doi.org/https://doi.org/10.1787/ca768d40-en

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education [Publisher: American Educational Research Association]. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116. https://doi.org/10.3102/10769986029001103

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting pisa results. *Educational Researcher*, *45*(4), 252–257.

Veldhuis, P., & Versteegh, K. (2021). Onderwijscrisis: Zijn brede brugklassen de oplossing? Retrieved October 4, 2021, from https://www.nrc.nl/nieuws/2021/04/16/onderwijscrisis-zijn-brede-brugklassen-de-oplossing-a4040183

Appendix A

The first condition is the presence of a well-defined causal question. A question is well-defined when all relevant stakeholders of the ILSA agree on the priority of obtaining an answer to the question, which is articulated through the entire ILSA framework.

Condition 1: the causal question must be well-defined and stem from a theoretical framework that is presumably of interest to governing bodies responsible for policy priorities.

The second condition Kaplan specifies is that the question should be framed as a counterfactual question that is capable of yielding a real-life manipulation or intervention. Specifically in the case of ILSAs, the form of the question must have cross-cultural comparability. To properly define a counterfactual question, one must define the context in which the causation takes place. However, to meet the first assumption of Rubin's causal model, strong ignorability of treatment assignment, we would theoretically have to account for every covariate that could possibly be of influence on the treatment assignment. Especially when trying to answer causal questions on an international scale, this is virtually impossible, since infinitely many covariates can differ on the country level that may affect treatment assignment. To help select which covariates are relevant to the treatment and the effect, Kaplan puts forward Mackie's theory on causation.

Mackie states that a factor that can be identified as the cause of an effect under some conditions, might not be the cause of that same effect under different conditions. He suggests that the issue of distinguishing between causes and conditions is addressed by properly defining the *causal field* in which the causal relationship takes place (Mackie, 1974). In Mackie's theory, one defines a causal field by isolating the set of *conjunctions*. Conjunctions are sets of factors, where each set sufficient but not necessary to the effect, and the whole set of conjunctions form a condition that is both necessary and sufficient to the effect. Our interest lies with the properties of the individual factors that make up these conjunctions, since those factors are the variables we can measure. When a conjunction is sufficient, but not necessary to an effect and each individual factor in that conjunction is not sufficient to the effect itself, but without that factor, the conjunction is not sufficient to the effect anymore, we speak of a factor with the INUS property: an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition. To satisfy this second condition, we must identify the factors that serve as an INUS condition to the effect of interest.

Condition 2: the causal question is framed as a counterfactual question, capable of yielding a real-life manipulation or intervention within the framework of a randomized experiment

While Mackie's notion of a causal field helps narrow down the number of covariates that need to be collected, the satisfaction of the strong ignorability assumption still depends on the availability of those covariates. This is captured in the third condition. The agency to determine which questions are incorporated in the questionnaire of course lies with the content experts designing the ILSA questionnaires.

Condition 3: the collection of ancillary covariate information is relevant to the causal question of interest

The last condition pertains to the choice of statistical model and the robustness of the resulting estimand. Naturally, the statistical model used for the analysis must be chosen such that it actually yields an estimand that is of interest to the question at hand. More intricately, the statistical model must allow us to evaluate the obtained estimand against violations of causal assumptions. It is, for example, quite plausible that not all relevant covariates are captured in the questionnaire, violating conditions 2 and 3. To test against such violations, Kaplan suggests performing a series of sensitivity analyses by incorporating several reasonable values on possible unobserved confounders and measuring their effect on the estimand, thus controlling for hidden biases.

Condition 4: the choice of statistical model provides the appropriate causal estimand accounting for the ancillary covariate information and allows that estimand to be tested in a sequence of sensitivity analyses