INTRODUCTION

The combination of the so-called machine learning and causal inference is currently an active area of methodological research. Especially new methods for the estimation of average treatment effects (e.g. Athey, Imbens, & Wager, 2018; Chernozhukov, Chetverikov, et al., 2018; Farrell, 2015; Ning, Peng, & Imai, 2018; Van der Laan & Rose, 2011) and the estimation of heterogeneous treatment effects (e.g. Athey, Tibshirani, & Wager, 2018; Wager & Athey, 2018) are recently proposed. A small but growing strand of literature puts causal machine learning estimators for heterogeneous treatment effects into practice and discusses practically relevant issues (e.g. Bertrand, Crépon, Marguerie, & Premand, 2017; Davis & Heller, 2017; Knaus, Lechner, & Strittmatter, 2020a, 2020b). However, applications that do the same for average treatment effects are currently missing. Still, these methods have the potential to improve causal analysis in observational studies. This paper studies the effects of musical practice on student’s skills through a double machine learning approach.
practice on child development in an observational study and provides novel ideas to fruitfully combine these methods and standard empirical practices.

The analysis is motivated by the relevance of cognitive skills for success at school and in the labour market (Kautz, Heckman, Diris, Weel, & Borghans, 2014). Development of these skills is therefore of fundamental individual, economic and societal importance. Besides schools and families as the main drivers of human capital accumulation of children, the economic literature on child development shows a recent interest in understanding the role of extracurricular activities like sports or music (Bergman Nutley, Darki, & Klingberg, 2014; Cabane, Hille, & Lechner, 2016; Eccles, Barber, Stone, & Hunt, 2003; Felfe, Lechner, & Steinmayr, 2016; Hille & Schupp, 2015). Previous evidence suggests that engagement in these extracurricular activities per se has positive effects on at least some measurable cognitive and non-cognitive skills. However, evidence with respect to the intensity of playing music is missing so far. Shedding light on the dose–response relation of playing music has important implications for individuals, parents and policymakers. It allows to answer at least two important questions: (i) Which level of engagement is required to generate the observed gains? (ii) Is more always better or does very intense musical practice harm human capital accumulation by crowding out other productive activities?

The effects in this and in previous studies are identified using the conditional independence assumption (CIA) that demands usually a large set of control variables to be plausible. A flexible set of potential controls easily leads to a high-dimensional setting where the number of potential controls exceeds the number of observations. Double Machine Learning (DML) addresses the resulting variable selection problem and shows how the estimation of causal effects can be split into several prediction problems while allowing uniformly valid statistical inference (Chernozhukov, Chetverikov, et al., 2018). Thus, it allows to leverage methods from the machine learning literature that are developed for high-dimensional prediction problems (see for an overview, e.g. Hastie, Tibshirani, & Friedman, 2009). Double Machine Learning enables the integration of these methods into causal analysis with observational data and to control for selection bias in an objective and data-driven way.

This paper contributes to two strands of literature. First, it adds to the literature about extracurricular activities and youth development by investigating a potential dose–response relation between musical practice and cognitive skills. To this end, the German National Economic Panel Study (NEPS) (Blossfeld, Roßbach, & von Maurice, 2011) provides unique information on all dimensions of the analysis. Besides measures of music and outcomes of interest, the detailed parental information regarding cultural preferences in the NEPS data allow a more credible identification of causal effects compared to previous studies. Second, the paper contributes to the young causal machine learning literature. While the idea of DML triggered a variety of methodological contributions (e.g. Athey & Wager, 2017; Chernozhukov, Goldman, Semenova, & Taddy, 2017; Chernozhukov, Newey, Robins, & Singh, 2018), applications that are not run for expository purposes in these contributions are missing. This paper provides a first step to combine applied empirical practices and standards with these new methods. Specifically, it builds on the DML estimator of Farrell (2015). In the absence of any established procedures for applications, the paper addresses two practically important questions: (i) How can we check covariate balancing of the estimator? This is standard for estimators based on the propensity score. For estimators that involve variable selection, balancing checks are arguably even more important because potentially only few variables are selected and we still want to assess whether imbalances in other controls drive our effects. To this end, we derive a weighted representation of the DML method that enables the use of standard balancing checks. (ii) How can we assess the sensitivity of our estimators to tuning parameter choices? These tuning parameters are at the core of any machine learning algorithm and control model complexity. Out-of-sample prediction quality heavily depends on their choice. The same might be suspected when using these predictions for causal inference.
This paper proposes a data-driven assessment that is inspired by the one standard error rule (1SE) of Breiman, Friedman, Stone and Olshen (1984).

The paper proceeds as follows. The next section describes the NEPS data. Section 3 discusses the identification and the estimation problem. Section 4 explains DML and the innovations of this paper. Section 5 provides the results and Section 6 concludes. Online Appendices A–F provide additional materials. The accompanying R package `dmlmt` for DML with multivalued treatments is provided at https://github.com/mcknaus/dmlmt.

2 | DATA

The empirical analysis is based on the German National Educational Panel Study (NEPS) (Blossfeld et al., 2011). Specifically, we use the first wave of starting cohort four that was conducted in autumn 2010 with 15,577 students in the 9th grade where students are usually 14 or 15 years old. The student survey and tests were performed in classrooms. Afterwards, parents were surveyed in telephone interviews. The starting cohort of 9th graders is particularly well-suited for the research question at hand because they were exclusively and extensively asked about their extracurricular activities including intensity of musical practice (Frahm et al., 2011).

We focus on the population of students in regular schools (excluding special needs schools) who are extracurricularly active in at least one domain. The former restriction is dictated by data availability and the latter allows for a credible identification of the effects as spelled out in Section 3.1. The resulting number of observations available for the analysis is 5943. Appendix A.1 provides more details about the sample preparation. Steinhauer and Zinn (2016) document systematic non-response in the parental survey and provide sampling weights to adjust for non-response and unequal sampling probabilities. Those are omitted in the main analysis for simplicity. However, the sensitivity analysis in Section 5.3 documents that this does not affect the main conclusions.

Domain-specific cognitive skills are used as outcome variables and obtained from standardized tests in mathematics, reading literacy, information and communication technology (ICT) literacy, and language proficiency (vocabulary test) (Artelt, Weinert, & Carstensen, 2013). Furthermore, self-reported German and mathematics grades can be used to assess whether potential differences in the objectively measured skills are also mirrored in the more subjective evaluation by teachers.

Control variables contain individual characteristics, parental preferences for leisure activities, parental work, household economic conditions, parenting attitudes, household demographics, home possessions, information about social circle of parents and regional information. We extract 54 variables that translate into 328 controls after dummy coding (see Online Appendix A.1 for details).

Cohort four of the NEPS data provides a unique measure for the intensity of musical activities. Students are asked the following question: On how many days in the last month have you made music, e.g. played an instrument or sung in a choir? Making music on the computer does not count. On about … The number of days that are reported serve as our measure of intensity. Figure 1 shows the distribution of the answers for all students (left) and those that report a positive number of music days (right). The left graph shows that the majority (52%) reports no musical practice in the previous month.

One obvious feature of the intensity measure is the rounding pattern at steps of five and especially at steps of ten. Such rounding is frequently observed in surveys when people are asked to recall past frequencies. The analysis is therefore based on discretized intensity measures because the continuous measure might be affected by systematic measurement error and because precision for values between the peaks would be low due to very few observations for these values. The intensity measure is split into four categories for student i depending on the reported days/monthi:
Furthermore, we define a binary indicator $\text{Any}_i$ being one for any positive intensity. As such binary indicators were used in previous studies, we want to check how our findings compare to their results based on a binary indicator before investigating different intensities. Online Appendix A.2 provides a more detailed description about what playing music and the different intensities means.

### 3 | EMPIRICAL CHALLENGES

#### 3.1 | The identification problem

Identification of the effects of music is complicated by the fact that the decision to play music and the intensity are not made at random. To fix ideas, consider the potential outcome framework of Rubin (1974) in a multivalued treatment setting with $T+1$ different treatments $D_i \in \{0, 1, \ldots, T\}$ (Imbens, 2000; Lechner, 2001). Random variables are indicated by capital letters and the realizations of these random variables by lowercase letters. Each individual $i = 1, \ldots, n$ has a potential outcome $Y_{it}$ for each value of the treatment $D_i = t$ but only the potential outcome of the realized treatment value is observed. The observed outcome is therefore $Y_i = \sum 1 \{D_i = t\} Y_{it}$ and the potential outcomes with $D_i \neq t$ remain latent. However, we aim at estimating the average potential outcomes $\mu_t = E[Y_{it}]$ and their differences, where expectations are taken over the population of students in regular schools. For example, the average treatment effect (ATE), $\mu_t - \mu_0 = E[Y_i^1 - Y_i^0]$, is often a parameter of interest in the case of a binary

\[
\text{Intensity}_i = \begin{cases} 
\text{No}_i & \text{if days/month}_i = 0 \\
\text{Low}_i & \text{if days/month}_i \in [1, 7] \\
\text{Med}_i & \text{if days/month}_i \in [8, 22] \\
\text{High}_i & \text{if days/month}_i > 22 
\end{cases} 
\]
treatment variable $D \in \{0, 1\}$. The multivalued treatment setting considered here provides a larger set of average treatment effects by allowing any possible pairwise comparison $\gamma_{m,k} = \mu_m - \mu_k = E \left[ Y^m_i - Y^k_i \right]$ for $m \neq k$. We do not consider different target populations like the average treatment effect on the treated and other alternatives that are discussed, for example, in Lechner (2001).

The selection into treatments in observational studies leads to $Y \neq E[Y|D=t]$ due to selection bias. However, identification of $\mu$ can still be achieved if a vector of exogenous control variables $X_i$ exists such that the following two assumptions are fulfilled:

**Assumption 1**

$$Y_i \perp D_i | X_i = x, \forall t, i \text{ and } \forall x \in \mathcal{X}.$$ 

**Assumption 2**

$$P[D_i = t | X_i = x] > 0, \forall t, i \text{ and } \forall x \in \mathcal{X}.$$ 

Both assumptions can be summarized as the strong ignorability assumption (Rosenbaum & Rubin, 1983). The first assumption means that the treatment status is as good as randomly assigned conditional on covariates $X_i$. The second assumption requires that any unit needs to have a non-zero probability to receive each of the treatments. These assumptions allow the researcher to identify the average potential outcomes $\mu$, and consequently the causal effects $\gamma_{m,k}$. The plausibility of the common support assumption can be assessed and seems to be unproblematic in this application as we show in Online Appendix E. The CIA is however untestable and careful arguments need to be made about the plausibility in each specific application.

The setting in this paper deals with the four treatment levels defined in Equation (1). In this case, the CIA requires that we observe all variables that influence the decision to be musically active and its intensity as well as the outcomes of interest simultaneously. The decision process that leads to the observed intensity of music can be conceptualized as a three-stage process. The first two steps follow Cabane et al. (2016) and consider first the decision to engage in some extracurricular activity or not at all. In the second stage, students decide whether to play music or to engage in a different activity. If they decide to play music, the third step is to choose the intensity of musical practice.

We control for the first stage of selection by considering only students that are at least active in one extracurricular activity. Thus, selection into being active is implicitly controlled for because all remaining students with intensity $N_{ij}$ are active.

The remaining selection to be controlled for concerns the decision to engage in music and not in something else as well as the intensity decision conditional on making music. The discussion in Online Appendix A.3 spells out which variable categories are required to make the CIA plausible and how they are measured in the NEPS data including parental tastes and attitudes, economic household conditions, school and student characteristics. While the NEPS data provide a large set of variables that are needed to make identification via CIA plausible, personality traits and better measures of early ability are potentially missing confounders that still could induce selection bias.

### 3.2 The estimation problem

The reasoning in the previous section and Online Appendix A.3 is helpful to determine the confounders that should be controlled for and need to be balanced to identify the causal effect of
interest. Assumption 1 is non-parametric and we would like to control for all confounders in the most flexible way asymptotically. However, researchers need to trade off bias and variance in finite samples when specifying their models. For this task, no theoretic reasoning is available to choose the set of variables and their functional forms that should finally enter the statistical model. If we are not willing to impose linearity and allow, for example, for second-order interactions and up to fourth-order polynomials for continuous variables, the number of variables that could be considered in this application already rises to about 60,000. In that case, we would end up with nearly 10 times more variables than observations.

The literature provides only limited guidance on how to properly select the estimation models. Hirano and Imbens (2001) propose a systematic way of model selection by keeping only variables that are statistically significant at a pre-determined level in the outcome or propensity score model. However, this is not feasible for a high-dimensional set of potential controls. An alternative for propensity-score-based methods proceeds by iteratively adding interaction terms and polynomials to the propensity score model until the covariate distributions in treatment and control groups are considered as balanced (Dehejia & Wahba, 1999, 2002; Rosenbaum & Rubin, 1984). Other approaches apply machine learning techniques to flexibly estimate the propensity score (Lee, Lessler, & Stuart, 2010; McCaffrey, Ridgeway, & Moral, 2004; Wyss et al., 2014). However, these methods conduct standard statistical inference that ignores the model selection step.

All the reviewed approaches of variable selection share two related problems. First, Leeb and Pötscher (2005, 2008) show that such post-model-selection estimators might lead to invalid statistical inference. They note that inference procedures after model selection are not uniformly consistent. This means that their asymptotic properties provide no reliable approximations in finite samples. As a consequence, statistical inference that ignores the model selection step can be misleading. Second, single-equation approaches that use either the outcome or the propensity score model to select variables can dramatically fail to provide valid statistical inference as illustrated by Belloni, Chernozhukov and Hansen (2014a, 2014b). This problem arises because the CIA requires to control for variables that affect the treatment probability and the outcome. Model selection that is only based on one of the two might miss variables with small coefficients in the considered but a large coefficient in the other model. As a consequence, single-equation approaches might miss relevant controls and can be biased.

4 | DOUBLE MACHINE LEARNING FOR AVERAGE TREATMENT EFFECTS

4.1 | General idea

Belloni et al. (2014b) and Farrell (2015) offer a constructive solution to the problems of post-model-selection estimators and single-equation approaches described in the previous section. Their methods apply doubly robust scores for treatment effects estimation (Robins & Rotnitzky, 1995) in combination with data-driven variable selection. Chernozhukov et al. (2018) call a generalized framework that extends the ideas to other parameters of interest and other machine learning estimators Double Machine Learning. The ‘Double’ emphasizes that it is not a single-equation approach and ‘Machine Learning’ refers to general supervised machine learning methods to approximate conditional expectations of the nuisance parameters.

We build on the DML estimator of Farrell (2015) because it explicitly allows for multivalued treatments. The estimator is based on the doubly robust score for the average potential outcome under CIA (Robins & Rotnitzky, 1995),
\[ \mu_t = E \left[ \mu_t(X_i) + \frac{d^t_i(Y_i - \mu_t(X_i))}{p_t(X_i)} \right] \forall t, \tag{2} \]

where \(d^t_i = \{ D_i = t \}\) is the treatment indicator. The two nuisance parameters are the conditional expectation of the outcome, \(\mu_t(x) = E [Y_i | D_i = t, X_i = x]\), and the conditional treatment probability, \(p_t(x) = P [D_i = t | X_i = x]\), also referred to as the propensity score.

In contrast to the approaches discussed in Section 3.2 that build on either the outcome or the treatment regression, the doubly robust score in Equation (2) uses both. It is well understood that the resulting doubly robust estimator remains consistent if one of the so-called nuisance parameters \(\mu_t(x)\) and \(p_t(x)\) is misspecified while the other one is correctly specified (see e.g. Glynn & Quinn, 2009). Farrell (2015) shows that this robustness also guards against selection error in both nuisance parameters at the same time. This circumvents the problem of single-equation approaches and allows to apply machine learning methods to estimate the nuisance parameters. As long as these estimators are consistent and converge faster than \(n^{-1/4}\), an estimator based on Equation (2) is root-n consistent and asymptotically normal with variance given by

\[ \sigma_{\mu,t}^2 = E \left[ \left( \frac{d^t_i(Y_i - \mu_t(X_i))}{p_t(X_i)} + \mu_t(X_i) - \mu_t \right)^2 \right] \forall t. \tag{3} \]

The resulting statistical inference about the average potential outcome is uniformly valid and thus circumvents the post-model-selection problem of Leeb and Pötscher (2005, 2008). Building on these results, the pairwise average treatment effects are obtained by subtracting the doubly robust scores of the respective potential outcomes,

\[ \gamma_{m,k} = E \left[ \mu_m(X_i) + \frac{d^m_i(Y_i - \mu_m(X_i))}{p_m(X_i)} - \mu_k(X_i) - \frac{d^k_i(Y_i - \mu_k(X_i))}{p_k(X_i)} \right] \forall m, k. \tag{4} \]

The corresponding variance is given by

\[ \sigma_{\gamma,m,k}^2 = E \left[ \left( \frac{d^m_i(Y_i - \mu_m(X_i))}{p_m(X_i)} + \mu_m(X_i) - \mu_m - \frac{d^k_i(Y_i - \mu_k(X_i))}{p_k(X_i)} - \mu_k(X_i) + \mu_k \right)^2 \right] \forall m, k. \tag{5} \]

### 4.2 | Implementation

The implementation of DML requires three steps: (i) get predictions for the conditional outcome \(\hat{\mu}_t(x)\), (ii) get predictions for the conditional treatment probability \(\hat{p}_t(x)\), and (iii) plug both predictions into the sample analogues of Equations (2)–(5) to estimate the parameters of interest.

We estimate the nuisance parameters using Post-Lasso (Belloni & Chernozhukov, 2013) with cross-validation. This deviates from the expository application of Farrell (2015) that applies group Lasso and asymptotic penalty terms. In the following, we describe the modified implementation before Sections 4.3 and 4.4 explain how these modifications are used to develop novel procedures for balancing checks and sensitivity analyses.
The Post-Lasso is based on the Lasso estimator proposed by Tibshirani (1996). The Lasso solves the following optimization problem:

$$
\min_\beta \left[ \sum_{i=1}^{n} (Y_i - X_i \beta)^2 \right] + \lambda \sum_{j=1}^{p} |\beta_j| ,
$$

(6)

The Lasso can be considered an OLS estimator with a penalty \( \lambda \) on the sum of the absolute coefficients. We obtain the standard OLS coefficients if the penalty term is zero and we have at least as many observations as covariates. For a positive penalty term, some coefficients are shrunken towards zero to satisfy the constraint. Thus, the Lasso serves as a variable selector because some variables have their coefficients set exactly to zero if the penalty is gradually increased. By increasing the penalty term to a sufficiently large number, we obtain a path from a full model to an empty model with all coefficients besides the constant being zero. The idea of this procedure is to shrink those variables with little or no predictive power to zero. The Post-Lasso predictions are then based on a standard OLS regression using all variables with non-zero coefficients at a particular value of the penalty term. This OLS Post-Lasso is used to obtain the outcome predictions \( \hat{\mu}_i(x) \). The propensity scores \( \hat{p}_i(x) \) are estimated in separate logistic Post-Lasso regressions to account for the binary nature of the treatment indicators (see e.g. Belloni, Chernozhukov, & Wei, 2013).

The controls \( X_i \) entering Equation (6) are created in the following way. Starting from all second-order interactions of the 328 control variables and fourth-order polynomials for continuous variables, we drop those interactions that create empty cells or nearly empty cells containing <1% of the observations. We keep only one variable of variable groups that show absolute correlations above 0.99. Finally, we add dummies for states, school track and each school in the sample. This gives a total of 10,066 variables to be considered in the selection process.

Dummies for state and school track are left unpenalized because institutional knowledge tells us that we expect substantial differences across states and school tracks, which should be accounted for. The ‘empty’ model therefore contains already 19 variables and the Post-Lasso is used to find the predictors that should enter on top. The penalty term that determines how many additional controls enter the models of the nuisance parameters is chosen via 10-fold cross-validation, which is described in detail in Section 4.4. The goal of the variable selection is to find the most predictive variables and is not concerned with interpretability of the selected variables. Thus, we do not enforce hierarchical selection that always includes, for example, main effects before interaction effects.

As stated in the previous section, DML requires that predictions of the nuisance parameters converge faster than rate \( n^{-1/4} \). At this stage, we need to assume that the cross-validated Post-Lasso achieves this rate because the convergence rate of this particular estimator is not yet available. The assumption that this convergence rate is feasible builds on two theoretical results. First, Chetverikov, Liao and Chernozhukov (2017) show that cross-validated Lasso can reach \( n^{-1/4} \) convergence assuming sparsity of the underlying model. The sparsity assumption means that the number of relevant variables \( s \) is much smaller than the number of observations \( n \). Sparsity might be seen as conceptualization of the empirical practice to include only few controls compared to sample size. Such practices implicitly assume that the chosen variables are sufficient to provide a good approximation of the models of interest. Second, Belloni and Chernozhukov (2013) show that Post-Lasso converges at least as fast as Lasso under data-driven penalty terms based on asymptotic arguments. It seems therefore plausible to assume that a similar relation holds between cross-validated Post-Lasso and cross-validated Lasso such that the required convergence of faster than \( n^{-1/4} \) is feasible in our implementation.
Finally, we need to cluster the standard errors at school level $s$ because the sampling is school based (von Maurice, Sixt, & Blossfeld, 2011). The clustered standard errors are estimated as $\hat{\sigma}_{\mu,s}^2 = \frac{1}{n} \sum_s \left( \sum_{i \in s} \left( \frac{d_i^s (Y_i - \hat{\mu}_i (X_i))}{\hat{p}_i (X_i)} + \hat{\mu}_i (X_i) - \hat{\mu}_i \right) \right)^2$. 

$$\hat{\sigma}_{\mu,s}^2 = \frac{1}{n} \sum_s \left( \sum_{i \in s} \left( \frac{d_i^s (Y_i - \hat{\mu}_i (X_i))}{\hat{p}_i (X_i)} + \hat{\mu}_i (X_i) - \hat{\mu}_i \right) \right)^2.$$  

(7)

The hat notation in Equation (7) indicates estimated sample equivalents of the arguments in Equation (3). Corresponding to, for example, clustered standard errors for OLS, Equation (7) sums first over all students $i$ in the same school $s$ to account for potential within-school correlations before summing over the schools.

4.3 Assesment of covariate balancing

Good practice in treatment effects applications based on propensity scores requires to assess the balancing of the covariates before and after adjusting for selection (Imbens & Wooldridge, 2009). This is even more important in the high-dimensional setting of this paper because only few variables might be selected in the estimation. However, the identification step in Section 3.1 requires that all confounders need to be balanced. At this point, it is important to note that these considerations exclude instruments that are not required for identification and thus neither should be included in the set of confounders, nor are they required to be balanced. In the following, we propose a practical way to check balance and to ensure that estimated effects are not driven by imbalances in confounders that are used to argue for a credible identification.

Existing balancing checks for propensity-score-based methods exploit that the estimate of the average potential outcome of treatment group $t$ can be expressed as a weighted average of the observed treated (Lee, 2013). Formally, this means that $\hat{\mu}_i = Y_i w^p_i$, where $Y_i$ is a $1 \times n_t$ vector containing the $n_t$ observed outcomes in this treatment group and $w^p_i$ is a corresponding $n_t \times 1$ vector containing weights obtained from matching or weighting by the propensity score (see e.g. Huber, Lechner, & Wunsch, 2013; Smith & Todd, 2005). For example, $w^p_i = \begin{bmatrix} w_{t1}^p \ldots w_{tJ}^p \end{bmatrix}$ with $w_{ti}^p = 1/\hat{p}_i (X_i)$ for inverse probability weighting (Hirano, Imbens, & Ridder, 2003; Horvitz & Thompson, 1952). Balancing of the covariates between different treatment groups is then assessed based on the weighted covariates $X_i w^p_i$, where $X_i$ is a $p \times n_t$ matrix containing the $p$ covariates of the observations in treatment group $t$.

So far, balancing tests are not conducted for estimators based on doubly robust scores. Though not naturally appearing in the estimation procedure, the underlying weights can be calculated as soon as a weighted representation of the predicted outcome is available as $\hat{\mu}_i (X_i) = Y_i w^p_{i,t}$. The empirical version of Equation (2) can then be rewritten as

$$\hat{\mu}_i = \frac{1}{n} \sum_{i=1}^n \left( \hat{\mu}_i (X_i) + \frac{d_i^t (Y_i - \hat{\mu}_i (X_i))}{\hat{p}_i (X_i)} \right) = \frac{1}{n} \sum_{i=1}^n \left( Y_i w^p_i + \frac{d_i^t Y_i}{\hat{p}_i (X_i)} \right) = \frac{1}{n} \sum_{i=1}^n \left( Y_i w^p_i + \frac{d_i^t Y_i w^p_{i,t}}{\hat{p}_i (X_i)} \right) = Y_i w^p_i.$$  

(8)

The implementation via Post-Lasso allows us to calculate $w_i$ because the weights for predicting the outcome of unit $i$ are provided by the $n_t \times 1$ vector $w^p_{i,t} = X_i (X'_i X_i)^{-1} X'_i$ and sum to one (Abadie,
Diamond, & Hainmueller, 2015). This shows the benefit of using Post-Lasso instead of standard Lasso as the latter has no weighted representation.

To calculate the weight vector \( w_p \), we need \( w_t^Y = \left[ w_{t,1}^Y, \ldots, w_{t,n}^Y \right] \), where \( j \) is a \( n \times 1 \) vector of ones, the inverse probability weights \( w_p \) defined in the example above, as well as \( w_{t,j}^Y = \left[ w_{t,1}^Y, w_{t,2}^Y, \ldots, w_{t,n}^Y \right] \). The vector \( w_t = w_t^Y + w_p - w_p^Y \) gives the weight that each outcome in the treatment group receives in the estimation of the mean potential outcome. These weights can be used for all weight-based balancing checks that were originally developed for propensity-score-based methods (Lee, 2013).

4.4 Sensitivity analysis regarding penalty choice

The paper of Farrell (2015) concludes by emphasizing the importance of the penalty parameter \( \lambda \) and the lack of knowledge about the proper choice. We propose a data-driven way to check the sensitivity of the results regarding these parameters in the different nuisance parameters. The procedure builds on 10-fold cross-validation, which proceeds in the following way: (i) the sample is randomly split into 10 folds (\( k = 1, \ldots, 10 \)) of similar size, (ii) the Lasso coefficient path is obtained in the subsample leaving out fold \( k \) over a grid of 100 penalty terms, \( \lambda \in \{ \lambda_1, \ldots, \lambda_m, \ldots, \lambda_{100} \} \), (iii) standard OLS or logit coefficients are calculated in this subsample using only the controls with non-zero coefficients at each grid point, (iv) these coefficients are used to predict values in the left out subsample at each grid point, (v) the cross-validated MSE of these predictions is calculated as \( CV_k (\lambda_m) = \frac{1}{n_k} \sum (y_i - \hat{y}_{i,k})^2 \), where \( n_k \) is the number of observations in the \( k \)th fold. Steps (ii)–(v) are repeated 10 times such that each subsample is left out once. This provides 10 series of MSEs over the whole penalty grid. Finally, we take the mean over the 10 folds, \( CV (\lambda_m) = \frac{1}{10} \sum CV_k (\lambda_m) \). For the baseline results, we choose the penalty term that minimizes the average MSE, \( \lambda_{m \text{min}} = \arg \min CV (\lambda) \). \( \lambda_{m \text{min}} \) is then used to estimate the model in the full sample and to get the predictions of the respective nuisance parameter.

Breiman et al. (1984) propose the one-standard-error rule (1SE) to account for uncertainty about the cross-validated \( \lambda_{m \text{min}} \). First, they estimate the standard error of the cross-validated MSEs, \( SE (\lambda_m) = \sqrt{\text{var} (CV_1 (\lambda_m), \ldots, CV_{10} (\lambda_m))} / 10 \). Second, they start from \( \lambda_{m \text{min}} \) and go along the penalty grid into the direction of a smaller model to find the first penalty \( \lambda_{1 \text{SE}} \) with \( CV (\lambda_{1 \text{SE}}) \leq CV (\lambda_{m \text{min}}) + SE (\lambda_{m \text{min}}) \).

The 1SE rule is motivated by the observation that the cross-validated MSE is often rather similar around \( \lambda_{m \text{min}} \) (see Online Appendix B for a representative example). Breiman et al. (1984) are concerned about overfitting and opt thus for a less complex model to decrease the variance and potentially allowing for more bias. The choice of one standard error and the direction to smaller models is ad-hoc and has no theoretical justification. Still, the 1SE rule is widely applied and taught in machine learning textbooks about prediction (see e.g. Hastie et al., 2009; Hastie, Tibshirani, & Wainwright, 2015).

DML uses prediction methods to estimate causal effects. In this setting, we might be more concerned about bias due to not selected confounders. Thus, we propose to complement the 1SE rule by the 1SE+ rule that considers the more complex model along the penalty grid within one standard error. These rules are particularly useful for estimators with multiple nuisance parameters that are obtained from different machine learners as in our case. The levels of penalties for least squares and logistic Post-Lasso are not necessarily comparable. Thus, running sensitivity checks by changing the penalty terms for all nuisance parameters by a fixed absolute or relative amount is problematic. For example, a 10% decrease in the penalty term could lead to a large number of added variables in the outcome equation but only a few in the treatment equation. Instead, applying the SE rules simultaneously to all nuisance estimators provides a data-driven way to investigate sensitivity of the estimates to the penalty choice. Additionally, this procedure naturally accounts for the possibility that the MSE minimizing
penalty terms might be estimated with different precision. Those nuisance parameters with rather flat and imprecisely measured MSE curves vary more in this procedure than those with a precisely measured global minimum.

The investigation of results obtained using different rules indicates whether or not the method produces stable results for a range of plausible penalty terms. In the ideal case, the estimates should be stable if model complexity is increased beyond the cross-validated minimum but the standard errors should get larger. This would indicate that the confounding is sufficiently controlled for at the cross-validated minimum and all additional variables just decrease efficiency.

Checking the sensitivity with regard to penalty term choice may be also informative about some other issues in the analysis. (i) The Post-Lasso estimator assumes sparsity. If going from the cross-validated model with $\lambda_{\text{min}}$ to more complex models changes the estimated effects substantially, this could indicate a failure of sparsity in a specific application and might be used as an informal check of sparsity. (ii) Cross-validation optimizes the MSE of the treatment and outcome but not of the (unobserved) causal effect. Therefore, the procedure aims to minimize the MSE for the wrong estimand (see Frölich (2005) for a similar argument regarding non-parametric estimators as plug-ins for causal effects). Instability of the estimated effects around the cross-validated model with $\lambda_{\text{min}}$ could indicate that this concern is relevant.

## RESULTS

### 5.1 Variable selection and covariate balancing

Before discussing the estimated effects, we take a look at the variable selection in the machine learning step and the balancing performance of DML. Recall that the Lasso starts with 19 state and school track dummies, leaving 10,047 variables from which Lasso chooses the additional control variables. Panel A of Table 1 shows that on average less than 10 additional variables are selected at the cross-validated minimum of the Post-Lasso. To understand which variables are selected, we count how often a variable is included in all estimated models and divide it by the number of total models.

| TABLE 1 Number of additionally selected variables and covariate balancing |
|---------------------------------------------------------------|-------------------|-----------------|
|                                                               | Binary treatment  | Multiple treatment |
|                                                               | Before          | After            | Before          | After            |
| Mean # of selected variables for treatment equation          | —               | 9                | —               | 4.5              |
| Mean # of selected variables for outcomes equation          | —               | 7.7              | —               | 4.6              |
| Maximum $|SD|$                                               | 32.7            | 8.0              | 35.2            | 11.5             |
| Mean $|SD|$                                               | 3.5             | 1.7              | 3.2             | 2.1              |
| Fraction of variables with $|SD| > 10$                                              | 5.2             | 0.0              | 6.5             | 0.1              |
| Fraction of variables with $|SD| > 5$                                               | 25.1            | 2.5              | 38.3            | 16.2             |

Note: Panel A shows numbers of selected variables at the cross-validated minimum of Post-Lasso additionally to the 19 fixed variables. The numbers for the propensity score of multiple treatments are average over all treatment states. The numbers of outcome predictions are averaged over all treatment states and outcomes. Panel B summarizes the absolute standardized differences ($|SD|$, Yang et al., 2016). The columns before are based on the unconditional differences. The after columns are calculated after DML adjustment using weights of Equation (8) and averaged over all outcomes.
estimated (17 models in the binary case with two outcome regressions for each of the eight outcomes plus one propensity score regression, and 36 in the multiple case with four outcome regressions for each of the eight outcomes plus four propensity score regressions). Online Appendix F shows that recommendations for secondary school as proxy for early ability as well as gender are the most important variables according to this measure. They are on average included more than once which shows that they are parts of different interaction terms. Furthermore, household characteristics with regard to demographics and cultural endowments enter the models frequently.

The weights of Equation (8) allow us to assess whether this rather small number of selected variables successfully balances the distribution of all ten thousand controls. We follow Yang et al. (2016) and check this by calculating standardized differences (SD). These scale the mean difference between one treatment group and the other groups by the square root of the mean variances of all treatment groups and multiply this fraction by 100. We calculate SD for all intensity groups and look at the maximum absolute SD for each variable. Panel B of Table 1 provides summary statistics of the absolute SD for the binary and multiple treatment case before and after DML. The comparison before DML shows that some covariates are highly unbalanced with a maximum absolute SD larger than 30 and thus far above the 20 that are considered as being large by Rosenbaum and Rubin (1985). However, most of the controls are decently balanced as documented by a mean absolute SD between three and four as well as by the fraction of variables with absolute SD above 10% being between 5% and 7%.

DML improves the balancing substantially and the few selected variables suffice to balance also the variables that are not selected. The maximum absolute SD is less than one-third of the before value and far below 20 after DML adjustment. Furthermore, over 500 variables showed an SD above 10 without adjustment. DML reduces this number to zero for the binary case and to less than 10 in the multiple case (Online Appendix C provides a visualization of the balancing improvement). The effect estimates in the next section are thus not driven by large imbalances in the distribution of controls after adjusting for a small set of selected controls. This reassuring insight would not be possible without the weighted representation and emphasizes the value of balancing checks especially in high-dimensional settings.

5.2 Effects of music on youth development

Table 2 shows the results for the comparison of musically active and inactive students in column one as well as comparisons between the different intensity categories in the remaining columns. Pairwise comparisons of intensities always compare the higher with the lower intensity in the respective pair. All outcome variables are standardized to have zero mean and variance one.

The first column reports substantial increases of about 0.1 standard deviations (sd) for objectively measured science, mathematics, vocabulary and ICT skills for students practicing music at least one day per month. Only reading skills show no statistically significant improvement. These results are qualitatively in line with Cabane et al. (2016) and Hille and Schupp (2015). The latter show similar effect sizes for cognitive skill. With their standard errors being four times larger than those obtained in this study, they cannot report statistical significance, though. This might be mainly attributed to their smaller sample size.

The comparison of different intensities shows a clear pattern. The improvements are mainly driven by students with medium and high intensities. For example, science skills improve by 0.14 sd for medium intensity versus inactive students and 0.17 sd for high intensity versus inactive students. In contrast, the increase of 0.04 sd for low intensity versus inactive is not statistically significant. Similar patterns are also observed for mathematics, vocabulary and ICT skills. Columns four and five
document that the differences of the medium or high versus low intensity are also significant at least at the 5% level. The only exception is high versus low intensity for mathematics skills. Column seven shows no further significant improvements for high-intensity versus medium-intensity practice.

The results suggest that the cognitive benefits materialize only for serious practice and not for only occasional music making. This is in contrast to the finding for school performance in the panel below.

In line with previous studies, column one shows significant improvements for German and mathematics grades for musically active students while the improvements of German are more pronounced with a highly significant 0.12 sd compared to a marginally significant 0.05 for mathematics that is not significant in most robustness checks discussed in Section 5.3. Unlike in the case of objectively measured skills, column two shows that even a low intensity of music results in significantly improved German grades. The comparisons between low, medium and high intensities in columns five to seven show no additional significant difference. One potential explanation of this pattern is that the mere signal of playing music is already rewarded by teachers (see for similar results and discussion Hille & Schupp, 2015). A potential explanation for different sizes of the effects is that low-intensity students participate in school-based musical activities like voluntary school choirs and German teachers are

### Table 2: Main results for binary and dose–response treatment effects

|                      | Binary  | Dose–response |
|----------------------|---------|---------------|
|                      | Any–No | Low–No | Med–No | High–No | Med–Low | High–Low | High–Med |
| **Cognitive skills (standardized)** |         |         |         |         |         |         |         |
| Science              | 0.11*** | 0.04   | 0.14*** | 0.17*** | 0.10*** | 0.13*** | 0.03    |
|                      | (0.02)  | (0.03) | (0.03) | (0.04) | (0.04) | (0.04) | (0.04)  |
| Mathematics          | 0.08*** | 0.05   | 0.12*** | 0.10*** | 0.07**  | 0.05    | −0.02   |
|                      | (0.02)  | (0.03) | (0.02) | (0.04) | (0.04) | (0.04) | (0.04)  |
| Vocabulary           | 0.11*** | 0.02   | 0.16*** | 0.18*** | 0.14*** | 0.16*** | 0.02    |
|                      | (0.02)  | (0.03) | (0.03) | (0.03) | (0.03) | (0.04) | (0.04)  |
| Reading              | −0.03   | 0.01   | −0.04   | −0.01   | −0.06   | −0.02   | 0.03    |
|                      | (0.02)  | (0.04) | (0.03) | (0.04) | (0.04) | (0.05) | (0.04)  |
| ICT                  | 0.12*** | 0.06*  | 0.15*** | 0.18*** | 0.09**  | 0.11**  | 0.03    |
|                      | (0.02)  | (0.03) | (0.03) | (0.04) | (0.04) | (0.04) | (0.04)  |
| **School performance (standardized)** |         |         |         |         |         |         |         |
| German grade         | 0.12*** | 0.11*** | 0.13*** | 0.16*** | 0.03    | 0.05    | 0.03    |
|                      | (0.03)  | (0.04) | (0.03) | (0.04) | (0.04) | (0.05) | (0.05)  |
| Mathematics grade    | 0.05*   | 0.04   | 0.08**  | 0.04    | 0.04    | −0.003  | −0.04   |
|                      | (0.03)  | (0.04) | (0.04) | (0.05) | (0.04) | (0.05) | (0.05)  |
| Avg. grade German &  | 0.09*** | 0.09** | 0.13*** | 0.10**  | 0.03    | 0.01    | −0.03   |
| mathematics          | (0.03)  | (0.04) | (0.04) | (0.04) | (0.04) | (0.05) | (0.05)  |

*Note:* This table shows the estimated effects comparing different intensities of musical practice. All outcome variables are standardized to mean zero and variance one. Higher grades are better. ICT stands for information and communication technology. The results are obtained by applying the Farrell (2015) estimator using Post-Lasso with penalty chosen at the minimum of 10-fold cross-validated MSE. State and school track dummies enter the selection unpenalized. Standard errors in parentheses are clustered at the school level. *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.
more receptive to this signal than mathematics teachers. This is plausible for two reasons. First, the subject of German is arguably closer related to arts than mathematics and German teachers potentially care more about artistic activities of their students than their mathematics colleagues. Second, even if German and mathematics teachers care to the same degree, German grades are more subjective compared to the relatively objectively measurable performance in mathematics.

5.3 Sensitivity analyses

Section 4.4 proposes a systematic investigation of the sensitivity of the results to tuning parameter choices. The tables of Online Appendix D.1 show how important parameters of the analysis like number of selected variables, implied weights and covariate balancing vary for different penalty term choices and how these differences affect the results.

Comparing the unconditional skill differences and the baseline estimates in Tables D.1.2 to D.1.7 indicates that selection is substantial. With most differences being more than halved or becoming insignificant after controlling for the relatively small set of controls. Increasing the model complexity beyond the baseline MSE minimizing penalty leads only to marginal changes of the results. This is surprising as the alternative penalty choice rules select up to six times more additional control variables than the baseline penalties. However, the rather sparse specifications of the baseline with not more than 10 additional variables seem already sufficient to control for the selection bias. The estimated effects obtained from more complex models vary only within one standard error of the baseline results and the qualitative conclusions are the same. We find in general that adding more variables than in the baseline model increases covariate balancing of all potential confounders only marginally but decreases efficiency mildly due to more extreme weights.

We apply a variety of alternative estimators to investigate sensitivity to estimator choice. We rerun the analysis with inverse probability weighting (IPW) and regression adjustment (RA) with variables selected by Post-Lasso. The discussion in Section 3.2 shows that uniform statistical inference is not possible for these post-selection and single-equation approaches. However, such approaches might be applied in practice and we are interested to see how their results differ compared to DML. In the absence of any valid inference results for these approaches, the inference ignores the variable selection step. The results are remarkably similar to the ones obtained by DML. This is not surprising as Appendix D.2 documents that the underlying weights of DML and the two single-equation approaches are highly correlated. However, this is not a general result and researchers should not rely only on post-selection and single-equation approaches as it is a priori not known whether the double selection nature of DML makes a crucial difference. Also, the lack of valid statistical inference for those methods is an undesirable feature in practice.

Furthermore, we consider four alternatives with known statistical properties. First, we use standard Inverse Probability Weighting with a parametric propensity score. The model specification of the propensity score resembles as close as possible those of Hille and Schupp (2015) and Cabane et al. (2016). The variable selection is based on exogenous sources and thus not prone to post-model-selection inference problems. The downside of such an approach is that the specification is not adapted to potential peculiarities in our dataset compared to the ones in previous studies, which could lead to biased estimates due to misspecified propensity scores. Online Appendix D.3.1 shows that the point estimates are very similar to the baseline results. However, the standard errors are up to 30% higher. This indicates that the parametric specification might be too rich and the sparser selected models for DML are more efficient while controlling bias in a similar way.
Second, DML is a flexible framework and not restricted to Post-Lasso to estimate the nuisance parameters as for the baseline results. Online Appendix D.3.2 shows that using Random Forest (Breiman, 2001) to estimate the nuisance parameters for DML produces by and large the same results as Post-Lasso in Table 2.

Third, DML of Farrell (2015) is currently the only machine-learning-based method that explicitly considers multivalued treatments. However, for the part considering binary treatments exist more readily available alternatives. One competing framework to DML is Targeted Maximum Likelihood Estimation (TMLE) (see for an overview, Van der Laan & Rose, 2011). Both are based on the efficient influence function for treatment effects and require the estimation of the same nuisance parameters. Both allow for a variety of supervised machine learning approaches but differ in their implementation (see for TMLE, Luque-Fernandez, Schomaker, Rachet, & Schnitzer, 2018). Online Appendix 3.3 shows that the differences between DML and TMLE are marginal using the same nuisance parameters as inputs. However, the standard implementation of TMLE uses an ensemble of machine learners called Super Learner (van der Laan, Polley, & Hubbard, 2007). Using the Super Learner to estimate the nuisance parameters reproduces by and large the baseline point estimates. However, it leads to a substantial reduction of the standard errors by up to 40% for both DML and TMLE. The choice of the estimators for the nuisance parameters seems to be more important than the choice between DML and TMLE, at least in this application. As the Super Learner has in general no weighted representation, the balancing checks introduced in this paper are not possible. The smaller standard errors point thus to a well-known trade-off between performance and the ability to open the black-box of machine-learning-based estimators.

Fourth, Ning et al. (2018) propose to estimate binary treatment effects based on a High-Dimensional Covariate Balancing Propensity Score (CBPS). This estimator uses Lasso to select control variables for both nuisance parameters. In a second step, the resulting propensity score is numerically recalibrated to explicitly balance also the variables selected for the outcome nuisance. This procedure overcomes the single-equation problem and the estimator remains square-root-n consistent if either the propensity score or outcome model is not correctly specified, while this is not the case for DML. This robustness as well as the explicit targeting of covariate balancing make CBPS an attractive alternative for the binary treatment case. Online Appendix D.3.4 compares DML and CBPS. It shows that the point estimates of CBPS are slightly more positive than the DML baseline. Furthermore, the goodness of balancing is very similar which indicates that the explicit balancing approach of CBPS is not superior to the implicit one of DML in our specific application. Thus, DML remains the dominant choice as it covers both binary and multiple treatments.

Overall, the results are remarkably stable over a variety of estimators. This indicates that the choice of the particular estimator is not a critical part of the analysis in this application. However, DML is an attractive choice because it allows for valid inference compared to the post-selection and single-equation IPW and RA, it is more adaptive to new datasets compared to taking parametric specifications from previous studies, and it is computationally attractive compared to CBPS that allows also for valid inference after variable selection. TMLE shares the attractive properties of DML and they are known to be asymptotically equivalent. The fact that they are very similar in this application indicates that the sample size is large enough for finite sample differences to become negligible.

Appendix D.4 discusses further sensitivity analyses in detail. Restricting the comparison to music versus sports instead of music versus any kind of extracurricular activities, different common support procedures, the removal of state and school track dummies and the use of sampling weights do not alter the qualitative findings. However, using the full sample and comparing music versus all non-musicians produces several more significant positive effects that might be explained by the failure to
control for selection into any extracurricular activity. This emphasizes the importance of the approach advocated in Cabane et al. (2016).

6 | CONCLUSION

This study investigates the effect of playing music on the cognitive skills and grades of 9th-grade students in Germany. The results are in line with previous studies showing significantly positive effects of musical practice per se. Going beyond the mere comparison of musicians and non-musicians, the study assesses the effects of different intensity levels of practice. It is shown that standardized and objectively measured cognitive skills require at least a medium level of practice to show notable benefits. However, substantial improvements in teacher assessed German grades are already observed for low intensity practice. This is in line with similar observations in Hille and Schupp (2015) who argue that playing music might affect school grades also through a positive signal to teachers. Overall, we find no evidence that a high intensity of making music could be harmful by crowding out other important activities.

The estimation of the effects is implemented via recent DML estimators that allow a flexible and transparent way to obtain causal estimates in observational studies. One concern regarding these methods is that they might depend heavily on the specific parameter choice. This paper proposes a systematic way to address these concerns based on cross-validation of Post-Lasso. The procedure provides important insights and is of general use for DML with all machine learners that are tuned via cross-validation. The sensitivity analysis finds stable results for a range of plausible penalty terms. Maybe surprisingly, very small models that include only about 10 variables suffice to obtain stable effects and moving to substantially richer model specifications leads only to mildly decreased efficiency.

The paper derives a weighted representation of the DML estimator that has proven to be useful to incorporate standard empirical practices regarding covariate balancing checks in applications. It generalizes to all machine learners where the predictions can be written as weighted averages of the outcomes. Thus, practitioners might face a trade-off between the possibility of checking covariate balancing and the use of sophisticated machine learners with unknown weighted representation.

On the methodological side, further research is required to investigate how different choices made throughout the paper are sensible. The goal should be to find good practices for DML in empirical applications. These are needed for all details of the implementation, especially regarding different choices of predictors and penalty terms, the dimension of the covariate matrix (order of interactions and polynomials), and common support enforcement. Furthermore, a more systematic comparison of DML with alternative estimators is an interesting direction for future research. In particular, it would be interesting to compare balancing and overall performance of DML to estimators that directly target balancing in high dimensions (e.g. Athey, Imbens, et al., 2018; Ning et al., 2018; Tan, 2018). However, they are currently not available for multiple treatments and thus not considered in more detail in this application.

Regarding identification, the available data about parental tastes seem to be crucial. However, future investigations should add better measures of early ability and personality traits to check whether those factors are driving the mostly positive results of extracurricular activities in the literature.

ACKNOWLEDGEMENT

Michael C. Knaus is also affiliated with IZA, Bonn. Financial support from the Swiss National Science Foundation (SNSF) for the project “Causal Analysis with Big Data” (grant number SNSF
407540_166999), which is part of the Swiss National Research Programme ‘Big Data’ (NRP 75) is gratefully acknowledged. This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:7.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. Chris Hansen, Michael Lechner, Jeff Smith, Martin Spindler, Stefan Wager and Michael Zimmert are greatly acknowledged for their helpful comments and suggestions. The usual disclaimer applies.

REFERENCES

Abadie, A., Diamond, A. & Hainmueller, J. (2015) Comparative politics and the synthetic control method. American Journal of Political Science, 59(2), 495–510.

Artelt, C., Weinert, S. & Carstensen, C.H. (2013) Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. Journal for Educational Research Online, 5(2), 5–14.

Athey, S., Imbens, G.W. & Wager, S. (2018) Approximate residual balancing: Debiasd inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4), 597–632.

Athey, S. & Wager, S. (2017). Efficient policy learning. Working Paper. Retrieved from https://arxiv.org/abs/1606.02647.

Belloni, A. & Chernozhukov, V. (2013) Least squares after model selection in high-dimensional sparse models. Bernoulli, 19(2), 521–547.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014a) High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29–50.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014b) Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies, 81(2), 608–650.

Belloni, A., Chernozhukov, V. & Wei, Y. (2013) Post-Selection Inference for Generalized Linear Models with Many Controls. Available at: http://arxiv.org/abs/1304.3969

Bergman Nutley, S., Darki, F. & Klingberg, T. (2014) Music practice is associated with development of working memory during childhood and adolescence. Frontiers in Human Neuroscience, 7, 1–9.

Bertrand, M., Crépon, B., Marguerie, A. & Premand, P. (2017). Contemporaneous and post-program impacts of a public works program: Evidence from Côte d’Ivoire. World Bank Working Paper.

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (2011) Education as a lifelong process – The German National Educational Panel Study (NEPS). [Special Issue] Zeitschrift Für Erziehungswissenschaft, 14.

Breiman, L. (2001) Random forests. Machine Learning, 45(1), 5–32.

Breiman, L., Friedman, J., Stone, C.J. & Olshen, R.A. (1984) Classification and regression trees. Boca Raton: CRC Press.

Cabane, C., Hille, A. & Lechner, M. (2016) Mozart or Pelé? The effects of adolescents’ participation in music and sports. Labour Economics, 41, 90–103.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & et al. (2018) Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68.

Chernozhukov, V., Goldman, M., Semenova, V. & Taddy, M. (2017). Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels. Available at: http://arxiv.org/abs/1712.09988

Chernozhukov, V., Newey, W., Robins, J. & Singh, R. (2018) Double/de-biased machine learning of global and local parameters using regularized Riesz representers. Available at: http://arxiv.org/abs/1802.08667

Chetverikov, D., Liao, Z. & Chernozhukov, V. (2017). On cross-validated Lasso. ArXiv:1605.02214. Retrieved from http://arxiv.org/abs/1605.02214

Davis, J.M.V. & Heller, S.B. (2017) Using causal forests to predict treatment heterogeneity: An application to summer jobs. American Economic Review, 107(5), 546–550.

Deheja, R.H. & Wahba, S. (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448), 1053–1062.
Dehejia, R.H. & Wahba, S. (2002) Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151–161.

Eccles, J.S., Barber, B.L., Stone, M. & Hunt, J. (2003) Extracurricular activities and adolescent development. *Journal of Social Issues*, 59(4), 865–889.

Farrell, M.H. (2015) Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1), 1–23.

Felfe, C., Lechner, M. & Steinmayr, A. (2016) Sports and child development. *PLoS One*, 11(5), e0151729.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, Let al. (2011). Transition and development from lower secondary to upper secondary school. In Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)*. Zeitschrift für Erziehungswissenschaft [Special Issue 14]. Wiesbaden: Springer VS, pp. 217–232.

Frölich, M. (2005) Matching estimators and optimal bandwidth choice. *Statistics and Computing*, 15(3), 197–215.

Glynn, A.N. & Quinn, K.M. (2009) An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1), 36–56.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning - Data mining, inference, and prediction*, 2nd edition. New York: Springer.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015) *Statistical learning with sparsity: The Lasso and generalizations*. Boca Raton: CRC Press.

Hille, A. & Schupp, J. (2015) How learning a musical instrument affects the development of skills. *Economics of Education Review*, 44, 56–82.

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.

Hirano, K., Imbens, G.W. & Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Horvitz, D.G. & Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Huber, M., Lechner, M. & Wunsch, C. (2013) The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1), 1–21.

Imbens, G.W. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3), 706–710.

Imbens, G.W. & Wooldridge, J.M. (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86.

Kautz, T., Heckman, J.J., Diris, R., ter Weel, B. & Borghans, L. (2014) *Fostering and measuring skills* (OECD Education Working Papers No. 110). Paris: OECD Publishing.

Knaus, M.C., Lechner, M. & Strittmatter, A. (2020a) Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*. utaa014. https://doi.org/10.1093/ectj/utaa014

Knaus, M.C., Lechner, M. & Strittmatter, A. (2020b) Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*. 0718-9615R. https://doi.org/10.3368/jhr.57.2.0718-9615R1

Lechner, M. (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M. and Pfeiffer, E. (Eds.) *Econometric evaluation of labour market policies*. Heidelberg: Physica, pp. 43–58.

Lee, B.K., Lessler, J. & Stuart, E.A. (2010) Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3), 337–346.

Lee, W.S. (2013) Propensity score matching and variations on the balancing test. *Empirical Economics*, 44(1), 47–80.

Leeb, H. & Pötscher, B.M. (2005) Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1), 21–59.

Leeb, H. & Pötscher, B.M. (2008) Sparse estimators and the oracle property, or the return of Hodges’ estimator. *Journal of Econometrics*, 142(1), 201–211.

Luque-Fernandez, M.A., Schomaker, M., Rachel, B. & Schnitzer, M.E. (2018) Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16), 2530–2546.

McCaffrey, D.F., Ridgeway, G. & Morral, A.R. (2004) Propensity score estimation with boosted regression for evaluating causal effect in observational studies. *Psychological Method*, 9(4), 403.

Ning, Y., Peng, S. & Imai, K. (2018) Robust estimation of causal effects via high-dimensional covariate balancing propensity score. Available at: http://arxiv.org/abs/1812.08683
Robins, J.M. & Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129.
Rosenbaum, P.R. & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
Rosenbaum, P.R. & Rubin, D.B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516–524.
Rosenbaum, P.R. & Rubin, D.B. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.
Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.
Smith, J.A. & Todd, P.E. (2005) Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2), 305–353.
Steinhauer, H.W. & Zinn, S. (2016) *NEPS Technical Report for Weighting: Weighting the Sample of Starting Cohort 4 of the National Educational Panel Study (Wave 1 to 6)* (NEPS Survey Paper No. 2). Bamberg.
Tan, Z. (2018) Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. Retrieved from http://arxiv.org/abs/1801.09817
Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
van der Laan, M.J., Polley, E.C. & Hubbard, A.E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 1–21.
Van der Laan, M.J. & Rose, S. (2011) *Targeted learning: Causal inference for observational and experimental data*. Berlin: Springer Science & Business Media.
von Maurice, J., Sixt, M. & Blossfeld, H.-P. (2011) *The German National Educational Panel Study: Surveying a Cohort of 9th Graders in Germany* (NEPS Working Paper No. 3).
Wager, S. & Athey, S. (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
Wyss, R., Ellis, A.R., Brookhart, M.A., Girman, C.J., Jonsson Funk, M., LoCasale, R. & et al. (2014) The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology*, 180(6), 645–655.
Yang, S., Imbens, G.W., Cui, Z., Faries, D.E. & Kadziola, Z. (2016) Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4), 1055–1065.

**SUPPORTING INFORMATION**
Additional supporting information may be found online in the Supporting Information section.
Supplementary Material

---

**How to cite this article:** Knaus MC. A double machine learning approach to estimate the effects of musical practice on student’s skills. *J R Stat Soc Series A*. 2021;184:282–300. [https://doi.org/10.1111/rss.a.12623]