Analyzing Nonresponse in Longitudinal Surveys Using Bayesian Additive Regression Trees: A Nonparametric Event History Analysis

Sabine Zinn[1*] & Timo Gnambs[2,3*]

[1] German Institute for Economic Research

[2] Leibniz Institute for Educational Trajectories

[3] Johannes Kepler University Linz

[*] Both authors contributed equally

Accepted for publication in the *Social Science Computer Review*.

Author Note

Correspondence concerning this article should be addressed to Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, E-mail: timo.gnambs@lifbi.de.

Acknowledgments

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi:10.5157/NEPS:SC3:8.0.0. From 2008 to 2013, NEPS data was 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.

Abstract

Increasing nonresponse rates are a pressing issue for many longitudinal panel studies. Respondents frequently either refuse participation in single survey waves (temporary dropout) or discontinue participation altogether (permanent dropout). Contemporary statistical methods that are used to elucidate predictors of survey nonresponse are typically limited to small variable sets and ignore complex interaction patterns. The innovative approach of Bayesian additive regression trees (BART) is an elegant way to overcome these limitations because it does not specify a parametric form for the relationship between the outcome and its predictors. We present a BART event history analysis that allows identifying predictors for different types of nonresponse to anticipate response rates for upcoming survey waves. We apply our novel method to data from the German National Educational Panel study including *N* = 4,559 students in grade 5 that observed nonresponse rates of up to 36% across five waves. A cross-validation and comparison with logistic regression models with LASSO (least absolute shrinkage and selection operator) penalization underline the advantages of the approach. Our results highlight the potential of Bayesian discrete time event modeling for the long-term projection of panel stability across multiple survey waves. Finally, potential applications of this approach for operational use in survey management are outlined.

Keywords: panel study, dropout, nonresponse, regression tree, Bayes

Analyzing Nonresponse in Longitudinal Surveys Using Bayesian Additive Regression Trees: A Nonparametric Event History Analysis

Unit nonresponse presents an increasing obstacle for longitudinal social surveys and educational large-scale assessments that threatens the representativeness of samples and compromises the validity of conclusions drawn from these data (Beullens, Loosveldt, Vandenplas, & Stoop, 2018; Kreuter, 2013; Williams & Brick, 2017). Specifically, biased population estimates might result from incomplete data if observed responses differ systematically from responses that could have been theoretically obtained from nonresponding units (e.g., Heffetz & Reeves, 2019; Trappmann, Gramlich, & Mosthaf, 2015). Therefore, longitudinal panel studies strive to prevent nonresponse from the outset and delay panel mortality (i.e., withdrawal of participants) as long as possible. For this purpose, it is important to identify already at an early stage participants with a high nonresponse propensity in order to implement appropriate intervention strategies (e.g., providing more attractive incentives; see Felderer, Müller, Kreuter, & Winter, 2018; McGovern, Canning, & Bärnighausen, 2018). The primary objective of this paper is the introduction of a novel approach for the prediction of future participation behavior in longitudinal surveys that allow implementing countermeasures to prevent nonresponse. So far, most nonresponse research takes a rather short-term perspective and focuses on wave-to-wave participation rates, that is, the share of nonresponders in a given wave as compared to the sample in the previous wave (e.g., Durrant & Steele, 2009; Roßmann & Gummer, 2016; West, 2013). As a consequence, most findings are rather limited in scope and do not allow for long-term projections of panel stability. Moreover, statistical methods commonly used for the analysis of nonresponse (e.g., logistic regression) are ill-equipped to handle large and complex predictor sets (see, for example, van Smeden et al., 2018) that are typically available in longitudinal panel studies. In practice, survey researchers frequently limit their analyses to variable main effects (or, at the most, bivariate interactions) while ignoring higher-order interaction and nonlinear effect patterns, thus, sacrificing a potential gain in precision for statistical efficiency. Therefore, this study proposes a tree-based method that is suitable for handling complex variable sets for analyzing panel attrition (Chipman, George, & McCulloch, 2010) and using nonparametric event history analyses to examine nonresponse across multiple survey waves. Because the relationship between predictors and outcome does not assume a parametric form (as is the case with, for example, linear regression), the adopted machine learning approach places few restrictions on potential effect patterns (e.g., nonlinearity, interactions). As a result, it allows for more valid inferences on important predictors of participation propensities in social surveys. We applied our method to data from the longitudinal German National Educational Study (Blossfeld, Roßbach, & von Maurice, 2011) to predict participation rates across five survey waves and identify relevant predictors for different types of nonresponse.

The Problem of Nonresponse in Longitudinal Surveys

Panel studies require repeated participation of sampling units across long periods of time. However, many respondents are reluctant to invest the sustained effort required for this task and refuse follow-up invitations to surveys (e.g., Kleinert, Christoph, & Ruland, 2019; Williams & Brick, 2017). Unit nonresponse resulting from a refusal to participate in a study is commonly referred to as dropout, breakoff, or attrition (Brüderl & Trappmann, 2017; Peytchev, 2009) and represents an increasing problem in social science research. For example, Zinn and Gnambs (2018) observed a dropout rate in a longitudinal German large-scale assessment across four years of up to 61%. Continually decreasing response rates in social surveys seem to be a global trend. For recent rounds of the European Social Survey the decline in response rates across 36 countries was around 1 to 1.5 percentage points from one wave to another, resulting in overall response rates as low as 35% for some countries despite extensive fieldwork efforts (Beullens et al., 2018). More importantly, sampling units declining survey participation frequently exhibit systematically different characteristics than participants (e.g., Heffetz & Reeves, 2019; Trappmann et al., 2015; Voorpostel & Lipps, 2011). For example, specific life events such as changes in employment status or household composition tend to increase nonresponse in longitudinal social surveys (Trappmann et al., 2015); in contrast, continuous respondents across multiple survey waves tend to report fewer life changes (Voorpostel & Lipps, 2011). Thus, nonresponse can seriously undermine the validity of representative social surveys.

Generally, unit nonresponse refers to the loss of sampling units drawn from the population. In longitudinal studies, unit nonresponse can be further distinguished into two types (Müller & Castiglioni, 2017): respondents refusing participation in a given wave but participating in upcoming waves of the panel study (temporary dropout) and respondents refusing participation in a given wave and any upcoming waves (permanent dropout). Recent studies showed that temporary dropout cases systematically differ from continuous respondents and more strongly resemble permanent dropout cases (Michaud, Kapteyn, Smith, & van Soest, 2011; Watson & Wooden, 2014). Therefore, preventing nonresponds in the first place seems to be a way of improving sample variability and mitigating the biasing effect of permanent dropout in panel studies (Müller & Castiglioni, 2017). This requires identifying candidate nonresponders that might be persuaded to reengage with a panel study in the future and developing respective incentive strategies for hard-to-survey respondents (cf. Adhikari & Bryant, 2018). Ideally, the dropout propensity of sampling units in a panel study is even identified before unit nonresponse actually occurred. Then, respective counterstrategies can be devised that prevent dropout (see Earp, Mitchell, McCarthy, & Kreuter, 2012). Moreover, identifying participation trajectories already early on in a panel study can help planning timelines for sample refreshments and evaluating budgetary requirements. However, this requires accurate prediction models that can estimate the likelihood of temporary and permanent dropout across multiple waves of a longitudinal study.

Regression Trees for Analyzing Survey Participation

Regression relationships in survey data are often complex including many covariates, nonlinear effects, and higher-order interactions. Machine learning methodologies offer flexible modeling techniques that can accommodate these complex data structures and allow studying survey participation without requiring the *a priori* specification of a functional form between nonresponse and its predictors. Contemporary machine learning methods (see Kuhn & Johnson, 2013, for an overview) are also particularly well-equipped to handle large predictor sets that are typically available in social surveys (e.g., respondent characteristics, survey responses, paradata). Thus, recent reviews highlighted the potential of machine learning techniques also for survey research (e.g., Buskirk, Kirchner, Eck, & Signorino, 2018; Kern, Klausch, & Kreuter, 2019; Toth & Phipps, 2014), for example, for adaptive data collection (e.g., identifying additional cases during field time), nonresponse adjustments (e.g., correcting for unequal participation probabilities), and nonresponse analyses (e.g., describing dropout patterns). Particularly, tree-based methods have been shown to outperform commonly used prediction techniques such as logistic regression (Buskirk & Kolenikov, 2015; Phipps & Toth, 2012). Therefore, the next sections introduce the idea of classification and regression trees (CART; Breiman, Friedman, Olshen, & Stone, 1984) and their extension to Bayesian additive regression trees (BART; Chipman et al., 2010). Finally, we show how BART can be used to examine different types of nonresponse using event history modelling.

The Basics of Regression Trees

Tree-based methods such as CART (Breiman et al., 1984) employ a recursive partitioning algorithm to build a tree structure by splitting a sample into mutually exclusive classes (so-called terminal nodes) according to the predictor space (see Figure 1). Let the random variable Y with the realization y_i for respondent $i \in \{1, \dots, I\}$ represent a binary outcome indicating survey (non)participation and x_i a Q dimensional vector of predictors. Starting with the entire sample, the algorithm minimizes the loss function used for model evaluation (e.g., the entropy or the mean squared error) and searches for a decision rule that results in the best split leading to the two most homogenous classes with respect to *Y*. The decision rule is a binary split of a single predictor *s* of all available predictors *X*, $\boldsymbol{s} \in \boldsymbol{X}$, and a cut point *c* such that $\{s < c\}$ versus $\{s \geq c\}$ for a continuous *s* or $\{s = c\}$ versus $\{s \neq c\}$ for a categorical *s*. After the decision rule is found, the resulting classes are themselves considered for splitting. This process is repeated until a prespecified termination criterion (e.g., minimum number of cases per node) is reached. Finally, the tree consists of *B* terminal nodes and associated parameter values $M \in \{\mu_1, ..., \mu_B\}$ that can be used for predicting the outcome for new data. *M* represents the $E(Y \mid x)$ in each terminal node, that is, the modal category of *Y*. Thus, given a tree structure *T* with its nodes and decision rules and the parameter values *M* for each terminal node the regression relationship between y_i and x_i can be formally expressed as

$$y_i = g\left(x_i; T, M\right) + \varepsilon_i \text{ where } \varepsilon_i \sim N\left(0, \sigma^2\right), \tag{1}$$

with *g* as a function describing the tree. Because $g(x_i; T, M)$ in (1) returns the $\mu_b \in M$ assigned to x_i, $E(Y \mid x_i)$ corresponds to the terminal node parameter μ_b given by $g(x_i; T, M)$.

In contrast to single tree-based methods, ensemble methods combine multiple trees and, thus, tend to achieve better predictive performance. A particularly versatile development in this area are Bayesian additive regression trees (BART; Chipman et al., 2010).

Bayesian Additive Regression Trees

The BART model is a sum-of-trees approach with regularization priors on the model parameters. In contrast to CART models, multiple trees are combined in an additive fashion as

$$y_i = \sum_{l=1}^{m} g\left(x_i; T_l, M_l\right) + \varepsilon_i \text{ where } \varepsilon_i \sim N\left(0, \sigma^2\right). \tag{2}$$

Under (2), $E(Y \mid x_i)$ is the sum of all parameter values μ_{bl} for the terminal nodes assigned to x_i by $g(x_i; T_l, M_l)$. Although the number of trees m can be specified as an unknown parameter in the Bayesian implementation of (2), the computational costs are high and do not seem to offer a substantial gain in prediction accuracy as compared to selecting a default value of 200 (Chipman et al., 2010). Typically, the predictive performance of BART strongly increases among smaller numbers of trees and then levels off.

Assuming prior independence, the model uses priors for three components: a prior $p(T_l)$ for the tree structure, a prior $p(\mu_{bl} \mid T_l)$ for the parameter values in the terminal nodes given the tree structure, and a prior $p(\sigma)$ for the variance. The regularization prior $p(T_l)$ determines the depth $d \in [1, \infty)$ of tree l and assigns prior probabilities of $\alpha(1+d)^{-\beta}$ with $\alpha \in (0,1)$ and $\beta \in [0, \infty)$ to each node that it is a non-terminal node and can be used for another split. Chipman and colleagues (2010) recommend using $\alpha = 0.95$ and $\beta = 2$ as default values which give the largest probabilities to smaller trees of depth $d = 2$ or 3 (although larger trees with many terminal nodes can also emerge given the data). The prior $p(\mu_{bl} \mid T_l)$ for the terminal node values is

$$\mu_{bl} \sim N\left(0, \sigma_\mu^2\right) \text{ where } \sigma_\mu = e \big/ \left(k\sqrt{m}\right) \tag{3}$$

which assigns lower probabilities to extreme values and effectively shrinks the μ_{bl} toward zero. In doing so, each individual tree in (2) explains only a small portion of the outcome and can also be interpreted as a ‘weak learner’. For binary outcomes the recommended choice for e is 3.0, whereas a value of 2 has been suggested for k which yields a 95% prior probability that $E(Y, \mid x)$ falls between the observed minimum and maximum values of an appropriately rescaled Y (Chipman et al., 2010). Finally, an inverse chi-square distribution can be used as a prior for the variance σ^2 (see again, Chipman et al., 2010). However, as our discrete-time event models (see below) use a probit regression with latent variables and unit variance ($\sigma^2 = 1$), for our purposes no further prior specification is required.

The BART model is estimated using a Gibbs sampler with a Bayesian backfitting Marko Chain Monte Carlo (MCMC) algorithm embedded (Hastie & Tibshirani, 2000). That is, upon each sequence of draws from the full conditional distributions of the unknown parameters T_l, M_l, and σ the partial residuals are derived as

$$R_l \equiv y - \sum_{k \neq l} g\left(x; T_k, M_k\right) \tag{4}$$

on a fit that excludes the *l*th tree. In each step of the Gibbs sampler (referring to the conditional density of the parameters T_l and M_l of the tree *l*), these are then used as the conditional variables instead of all the trees T_k and parameter sets M_k that exclude T_l and M_l. The conditional tree structure (T_l | R_l, σ) is drawn using a Metropolis-Hastings step (Chipman, George, & McCulloch, 1998), whereas the conditional terminal node values (M_l | T_l, R_l, σ) are drawn from a normal distribution. Finally, conditional on all T_l and M_l, the residual standard deviation (σ | T_1, …, T_m, M_1, …, M_m) is drawn from an inverse gamma distribution. In our case, the last step is dropped since we use a probit specification to fit discrete time event history models (see below). This backfitting algorithm generates a sequence of draws of (T_1, M_1) … (T_m, M_m) that converge to the posterior distribution of the true model

$$p\left(\sum_{l=1}^{m} g\left(\cdot; T_l, M_l\right) | Y\right). \tag{5}$$

The posterior distribution can be used to calculate Bayesian inferential statistics such as posterior means or medians and respective credibility intervals. Further information on the estimation algorithm is given in Chipman et al. (2010) and Kapelner and Bleich (2016).

The predictor importance in tree-based methods is evaluated by focusing on the subset of variables that was used for splitting and growing the trees. Various backward stepwise selection procedures have been suggested that quantify, for example, the reduction in mean square error (Díaz-Uriarte & de Andrés, 2006) or posterior predictive uncertainty within nodes (Gramacy, Taddy, & Wild, 2013) to rank predictors by importance. In BART, Chipman and colleagues (2010) suggested the variable inclusion proportion p_{vi}, that is, the posterior mean of the relative number of times a given variable is used in a tree decision rule. Because predictors with large p_{vi} are likely to be important drivers of the outcome, p_{vi} can be used to rank the predictors in *x* in terms of importance. Note that the p_{vi} are indicators of relative importance and do not reflect whether any given covariate has a "real effect" (Bleich, Kapelner, George, & Jensen, 2014). In situations where all variables are unrelated to *Y*, BART would select covariates randomly to grow the trees. Then, the variable inclusion proportion for each variable would be $p_{vi} = 1 / Q$ for all $x \in \{1,...,Q\}$. Thus, to be considered a relevant predictor p_{vi} should exceed this threshold.

Event History Modeling using BART

Event history modeling aims at the analysis of longitudinal data on the occurrence and timing of events. . As with other regression methods, it models the likelihood that a specific event occurs at a specific point in time dependent on various predictors (see Keiding, 2014, and Mills, 2011, for an introduction). In longitudinal surveys across multiple waves two types of nonresponse can be observed (i.e., temporary and permanent dropout; cf. Müller & Castiglioni, 2017) that can be modeled using discrete time events in BART. Sparapani, Logan, McCulloch, and Laud (2016) initially proposed a BART for survival analyses to examine the risk of experiencing a single event (e.g., dropout versus participation). This approach can easily be extended to also examine competing risks for different types of events (Sparapani, Logan, McCulloch, & Laud, 2019), as long as an independence between competing risks is assumed (as is commonly done in event history modeling; Mills, 2011). In doing so, this model fits to the specific structure of our problem (i.e., temporary dropout versus permanent dropout versus participation).

Let t_j represent the distinct event times (i.e., the *T* survey waves), n_i the number of time points observed for respondent *i*, and $\delta_{ij} \in \{0,1,...,K\}$ the censoring indicator for respondent *i* at time t_j (with $j \in [0, T-1]$) that distinguishes non-events ($\delta_{ij} = 0$) from events ($\delta_{ij} > 0$). In our application, non-events correspond to survey participation, whereas *K* equals 2 and reflects the two types of nonresponse (i.e., temporary and permanent dropout). The response variable *Z* is a stacked vector given by $z_{ijk} = I\left(\delta_{ij} = k\right)$ with $k \in \{0, 1, ..., K\}$ and $j \in \{1, ..., n_i\}$ (see Figure 2). The probability of observing event *k* for respondent *i* between $t_{j\text{-}1}$ and t_j conditional the covariates w_{ij} is modeled using a multinomial probit link as

$$p_{ijk} = P\left(z_{ijk} = 1 \mid w_{ij}\right) = \Phi\left[\sum_{l=1}^{m} g\left(x_{ij} = \left[t_j, w_{ij}, v_{ijk}, N_{i(j-1)k}\right]; T_l, M_l\right)\right] \tag{6}$$

where Φ[·] is the standard normal distribution. Given the independent risks assumption, (6) can also be specified in the form of multiple survival models with a binomial probit link (Sparapani et al., 2019), thus, estimating (6) independently for each event type *k*. The covariates x_{ij} include the event time t_j and (time-invariant) predictor variables w_{ij}. In addition, we also acknowledge the exposure time v_{ijk} of respondent *i* at $t_{j,}$ that is, the number of discrete time points since the beginning of the episode for event *k*. Finally, the covariates also comprise $N_{i(j\text{-}1)k}$, that is, the number of events of type *k* previously observed up to the preceding event time $t_{j\text{-}1}$. For a single (*K* = 1), non-recurring event, v_{ijk} is identical for all observation units and $N_{i(j\text{-}1)k} = 0$, because all individuals have the same study start time and units that have already experienced an event are no longer part of the risk set. Thus, v_{ijk} and $N_{i(j\text{-}1)k}$ are not needed in statistical modelling. Under these conditions, our model reduces to the BART survival model introduced by Sparapani and colleagues (2016). In other words, our event history approach can be viewed as a generalization of the survival model for competing and possibly recurrent events.

The BART specification of (6) places priors $p(T_l)$ on the tree structure and $p(\mu_{bl} \mid T_l)$ on the terminal node values given the tree structure as described above. Then, samples can be generated from the posteriori distribution for any event time t_j and event type *k* to obtain the posterior distribution p_{jk}. These samples can be used to approximate any conceivable distribution of individual and group statistics, such as individual participation probabilities or the average dropout propensity of a group of individuals at a certain point in time.

Prediction of Unconditional Response Rates

From a BART event history model values of *Z* can be predicted from the covariates *w* for future survey waves or for new values of *w*, for example, to evaluate expected response rates in a new panel study. Thus, along with the definition of Rubin (1976) the missing mechanism assumed here is missing at random (MAR). These predictions can be used to estimate nonresponse probabilities for the recurrent or non-recurrent events included in the model at each survey wave t_j. Importantly, these estimates represent conditional probabilities given survival to t_j (i.e., given no permanent dropout). Combining the nonresponse probabilities for recurrent and non-recurrent events using standard rules of probability theory also allows estimating the unconditional response probability at a given wave, thus, estimating the expected sample size in a survey wave.

Formally, survey participation can be understood as a bivariate and time discrete, stochastic process (V_t, Z_t) with V_t representing survey participation resulting from a non-recurrent event (i.e., permanent dropout) and Z_t the respective respondent outcome for a recurrent event (i.e., temporary dropout) at time point *t*. Both processes can result in survey participation (*PA*) or nonresponse (*NR*), that is, $v_t \in \{PA, NR\}$ and $z_t \in \{PA, NR\}$. This process is defined by its start distribution (V_0, Z_0) at the first survey wave and transition probabilities $P(V_{t+1} = v_{t+1}, Z_{t+1} = z_{t+1})$ given the previous states $\left(V_s = v_s, Z_s = z_s\right)_{s\in\{0,\ldots,t\}} = \left(V_0 = v_0, Z_0 = z_0, \ldots, V_t = v_t, Z_t = z_t\right)$. The joint transition probabilities for V_{t+1} and Z_{t+1} can be derived from two submodels by splitting the respective probabilities into conditional probabilities for the non-recurrent event (V_{t+1}) and the recurrent event (Z_{t+1}) in the following way (see the supplement material for the full derivation of this decomposition):

$$\begin{aligned} & P\left(V_{t+1}=v_{t+1}, Z_{t+1}=z_{t+1} \mid \left(V_s=v_s, Z_s=z_s\right)_{s\in\{0,\ldots,t\}}\right)= \\ & \quad P\left(V_{t+1}=v_{t+1} \mid \left(V_s=v_s, Z_s=z_s\right)_{s\in\{0,\ldots,t\}}\right)\times \\ & \quad P\left(Z_{t+1}=z_{t+1} \mid V_{t+1}=v_{t+1}, \left(V_s=v_s, Z_s=z_s\right)_{s\in\{0,\ldots,t\}}\right) \end{aligned} . \tag{7}$$

Note that in this decomposition the transition probabilities for the non-recurrent event (V_{t+1}) depend on the previous states $\left(V_s=v_s, Z_s=z_s\right)_{s\in\{0,\ldots,t\}}$, whereas the respective probabilities for the recurrent event (Z_{t+1}) are additionally conditioned on the current state of the recurrent event ($V_{t+1} = v_{t+1}$). Each of these transition probabilities can be independently estimated using the BART event history approach in (6) and, subsequently, combined to derive the unconditional nonresponse probability in a given wave *t*.

Consider the example in Figure 3 that describes the potential nonresponse sequences for three waves. In wave 1 (t = 0) no nonresponse is observed. In wave 2 (t = 1), the conditional probability of temporary (p_1^{TD}) or permanent dropout (p_1^{PD}) is estimated using (6), resulting in an overall nonresponse probability of $p_1^{NR} = p_1^{PD} + \left[\left(1-p_1^{PD}\right)\times p_1^{TD}\right]$. Consequently, the unconditional probability of survey participation in wave 2 is $p_1^{PA} = 1 - p_1^{NR}$. For wave 3 (t = 2), the calculation follows comparably, albeit also acknowledging the permanent dropout probability p_1^{PD} from the previous wave. Thus, the unconditional nonresponse probability in wave 3 is estimated as $p_2^{NR} = p_1^{PD} + \left(1-p_1^{PD}\right)\times p_2^{TD} + \left(1-p_1^{PD}\right)\times\left[\left(1-p_2^{PD}\right)\times p_2^{TD}\right]$. This logic can be continued to estimate the participation probabilities for any following survey wave. Moreover, it can also be easily extended to acknowledge additional types of nonresponse. For example, permanent dropout might be a consequence of dwindling motivations and, thus, the respondent's active refusal to continue participation in a panel study. However, it might also result from an inability to contact the respondent because he or she moved without leaving valid contact information. In this case, a sequence of three types of nonresponse could be modeled, resulting in a three-step process.

The Present Study

The study examines nonresponse in the German *National Educational Panel Study* (Blossfeldet al., 2011) that studies trajectories of competence development across the life course. We will demonstrate how to model participation rates across several measurement occasion using event history models that account for competing risks (i.e., participation versus temporary dropout versus permanent dropout). This enables us to elucidate unique predictors for each type of nonresponse. Moreover, using a BART for model estimation (Chipman et al., 2010), we avoid parametric or semi-parametric constraints on the underlying model structure. This allows for the inclusion of large sets of predictor variables with complex interaction patterns and, thus, makes use of substantially more information than typical attrition analyses (e.g., using logistic regression models).

Method

Sample and Procedure

The participants are part of the National Educational Panel Study (NEPS; Blossfeld et al., 2011) that follows representative samples of German students across their school careers. For this study, we focus on a sample of *N* = 4,559 students (48% girls) that were initially surveyed in grade 5 (year 2010). Subsequent measurements occurred each year until grade 9 (year 2014), resulting in five measurement waves. The students attended various schools from rural and urban regions in Germany (see Steinhauer, Aßmann, Zinn, Goßmann, & Rässler, 2015, for details on the sampling procedure): about 38% attended general or intermediate secondary schools ("Hauptschule / Realschule"), 50% went to higher secondary schools ("Gymnasium"), and the remaining 12% encompassed students from several specialized school branches. The mean age in grade 5 was *M* = 15.13 years (*SD* = 0.51). More information on the data collection process including the interviewer selection and training are summarized on the project website (https://www.neps-data.de).

Measures

Survey participation. A respondent’s participation status was recorded at each wave as either participation, temporary dropout, or permanent dropout. In grade 5, all students participated; thus, no dropout was observed. Permanent dropout was defined as an active refusal to further participate in the study or an inability to participate. It occurred for different reasons such as refusal of a school to participate in the NEPS, refusal of a student within an eligible school, or students switching to another school not included in the NEPS. We did not distinguish the different types of permanent dropout in our analyses because the respective numbers of cases for each category were rather small. In contrast, temporary dropout referred to non-participation at a given wave that was not due to permanent dropout.

Conditioning variables. A total of 77 variables were used to predict nonresponse at each survey wave. These variables had been previously used in nonresponse analyses (e.g., Zinn, Würbach, Steinhauer, & Hammon, 2018) or were expected to be related to survey participation. Most variables were included as respondent characteristics (e.g., sex) and also aggregated to the school level (e.g., percentage of female students) to acknowledge individual as well as context effects. All variables were measured in the first survey wave or before. *Respondent characteristics* included the age (in years), sex (0 = “male”, 1 = “female”), migration background (0 = “no”, 1 = “yes”), mother tongue (0 = “German”, 1 = “other”), household size (as number of people), and the number of books at home (1 = “0 to 10 books” to 6 = “more than 500 books”) as an indicator of cultural capital (Sieben & Lechner, 2019). Moreover, as more specific *student characteristics* we recorded whether a student had ever repeated a school year (0 = “no”, 1 = “yes”), the number of missed school days due to being sick, and the grades in German and mathematics (1 = “very good” to 6 = “failing”). We also considered various self-reported *psychological characteristics*: Satisfaction with life, current living standards, health, family, friends, and school were each measured with a single item on eleven-point response scales from 0 (“completely dissatisfied”) to 10 (“completely satisfied”), subjective health was measured with a single item on a five-point scale from 1 (“very good”) to 5 (“very bad”), self-esteem was captured with 10 items ($\omega_{\text{categorical}}$ = .86) from Rosenberg (1965) on five-point response scales from 1 (“does not apply at all”) to (“applies completely”), and self-concept in German ($\omega_{\text{categorical}}$ = .75), mathematics ($\omega_{\text{categorical}}$ = .89), and school ($\omega_{\text{categorical}}$ = .82) were each measured with three items on four-point rating scales from 1 (“does not apply at all”) to 4 (“applies completely”). In addition, six cognitive measures were considered: Perceptual speed (Lang et al., 2014) and reading speed (Zimmermann, Gehrer, Artelt, & Weinert, 2012) were each measured as sum scores across 93 and 51 items, respectively. Figural reasoning was captured with a Raven-type test as a sum score across 12 items ($\omega_{\text{categorical}}$ = .68; Lang et al., 2014). Orthography (*Rel.*[1] = .96; Blatt et al., 2017), mathematical competence (*Rel.*1 = .80; Duchhardt & Gerdes, 2012), and reading competence (*Rel.*1 = .81; Pohl et al., 2012) were each represented by an item response score based on 30, 25, or 33 items, respectively. *School characteristics* included the school type in the form of two dummy-coded variables for “intermediate secondary schools” and “other school types” (reference category: “higher secondary school”), the number of students and classes in grade 8 as indicators of school size, the type of institution (0 = “public”, 1 = “private”), and two dummy-coded variables for the school location as “part urban / part rural” and “rural” (reference category: “urban”). Moreover, all respondent, student, and psychological characteristics were aggregated to the school level to indicate respective contextual influences. Finally, we considered the federal state in Germany as 15 dummy-coded indicators.

Statistical Analyses

Longitudinal nonresponse was analyzed across five waves using the nonparametric event history approach described above. The BART model specified 300 trees using $\alpha = 0.95$

[1] The authors reported marginal reliabilities (Adams, 2005) based on an item response model.

and β = 2 for the regularization prior $p(T_j)$ as well as l = 3 and k = 2 for the prior $p(\mu_{bj} \mid T_j)$. That way we followed the recommendations of Chipman and colleagues (2010) for the specification of the prior distributions. The Bayesian estimation used 500 burnin samples, a thinning of 500 draws in the MCMC algorithm, and 5,000 draws from the posterior distribution for the permanent and temporary dropout propensities at each wave and each respondent. Convergence was evaluated by means of visual inspections of autocorrelation and trace plots. Moreover, the Geweke (1992) statistic was examined by the posterior sample of ten (arbitrarily chosen) individuals. Missing values on the covariates (see supplement material) were imputed with the variable's median. For covariates with missing rates exceeding 5% additional dummy-variables were created and included the prediction models[2]. The analyses were conducted in *R* version 3.5.1 (R Core Team, 2018) using the *BART* package version 2.2 (McCulloch, Sparapani, Gramacy, Spanbauer, & Pratola, 2019).

We validated our approach by conducting two kinds of analyses. First, we performed an out-of-sample validation using two-thirds of the total data as training data and one-third as test data. The observed data was randomly assigned to the two subsets of data, the training and the test data. Then, we estimated the BART models outlined above based on the training data and predicted wave-specific probabilities of permanent and temporary dropout based on the test data. The predicted values were compared to the observed values in the test data. As a measure of accuracy, the percentage of values corresponding to the observed participation status in the test data was used. We repeated this process five times to guard against sampling bias. As a second form of validation, we estimated the BART models using the full data set

[2] In practice, single value imputations are typically not recommended (van Buuren, 2018). However, because few variables exhibited substantial missing rates (see supplement material) and our analyses primarily aimed at demonstrating an application of the BART event history model, we did not resort to more complex missing data models (cf. Zinn & Gnambs, 2018) to simplify the analyses and reporting of results.

but excluding the last survey wave. The last wave was used as test data. As before, we predicted for all individuals who were still at risk in the last wave their probabilities of permanently and temporarily dropping out. Again, the percentage of values that coincided between prediction and observation served as a measure of accuracy. Both types of validation are state of the art when dealing with models of statistical learning such as BART (e.g., Kern et al., 2019).

Finally, as a proof of concept we compared the results of our BART models to results from comparable logistic regression models with LASSO (least absolute shrinkage and selection operator) penalization. Logistic regression models with LASSO are parametric models that are typically used for studying problems as the one addressed in this article (e.g., Pavlou, Ambler, Seaman, Guttmann, Elliot, King, & Omar, 2015). These models were validated in the same way as the BART models.

Data Availability

Most of the data analyzed in this study is provided at https://doi.org/10.5157/NEPS:SC3:8.0.0. Due to German privacy laws, some school variables (e.g., type of institution or school location) cannot be made publicly available. The analyses syntax to reproduce our results can be found at https://github.com/bieneSchwarze/BARTforNonresponse.git.

Results

Nonresponse Rates Across Survey Waves

Across the five survey waves nonresponse rates increased from 11% (wave 2) to 36% (wave 5). The percentage of temporary dropout was rather constant and fell at about 4% at each wave, whereas permanent dropout increased in an approximately linear fashion (see Table 1). In each survey wave the increase in permanent dropout rates varied between 6% (wave 2) and 10% (wave 3). The reasons for permanent dropout were manifold. Many students dropped out involuntarily because they switched to another school or, in later waves, left the school system altogether (e.g., starting a traineeship). The design of the NEPS implements a different (rather limited) survey program for these cases. Thus, we considered them permanent dropouts. Moreover, after the initial wave some schools refused further participation in the NEPS, presumably, to avoid unduly disruptions of the school routines. In contrast, active refusals on part of the students were rather rare.

Prediction of Nonresponse Propensity

Predictors for two types of survey nonresponse (i.e., temporary and permanent dropout) in the NEPS were evaluated using BART event history modeling. The MCMC algorithm for these models converged satisfactorily. We observed no substantial autocorrelations in consecutive draws of the MCMC algorithm and the trace plots indicated good mixing behavior of the distinct parts of the generated chain (see supplement material). Moreover, Geweke (1992) tests showed no pronounced differences between the examined parts of the Markov chain. Thus, at this point our BART approach worked well for the data. As a second step, we validated our models for survey nonresponse. Overall, our BART models performed well in predicting observed temporary and permanent dropout patterns for both types of validation criteria. All accuracy indices for the BART models ranged between 90% and 97% (see Table 2). With accuracies between 89% and 99% the logistic regression models with LASSO performed similarly[3]. These results show that neither approach seemed to outperform the other, at least not concerning the considered accuracy measure.

[3] For the out-of sample cross-validation, we observed a change of less than one percent in the accuracy values over the five random draws of the training and test datasets for both modelling approaches (i.e., BART and logistic regression with LASSO). In other words, the accuracy values obtained were robust to the random drawing of training and test quantities.

Relative Variable Importance

The relative importance of the covariates for the trees of the BART models that predicted the nonrecurrent event (i.e., permanent dropout) are summarized in Figure 4 (complete results for all covariates are given in the supplementary material). Permanent dropout was primarily driven by the measurement occasion, that is, the survey wave: the time variable was chosen in about 26% of all trees. Interestingly, individual-level variables (i.e., respondent, student, and psychological characteristics) were rather uninformative for predicting permanent dropout. In contrast, context variables were more important. For example, the mean number of sick days (p_{vi} = .08), the mean grade in mathematics (p_{vi} = .05), or the mean subjective health (p_{vi} = .04) in the schools were relevant predictors of permanent dropout, whereas the respective respondent information was not. We received a slightly different picture for the prediction of temporary dropout. Figure 5 summarizes the results for covariates with an important contribution to a dropout event (full results are given in the supplementary material). Here we found a strong impact of the students' mean satisfaction with their health status on the school level (p_{vi} = .31), the mean number of students with migrations background at a school (p_{vi} = .14), and the number of books at home (p_{vi} = .10). The number of previous nonresponse events (p_{vi} = .07), the survey wave (p_{vi} = .03), the mean reading speed at school (p_{vi} = .07) , and the number of students in grade 8 (as an indicator of school size; p_{vi} = .06), and the time without a participation event predicted temporary dropout as well, whereas further respondent and school context information did not as much. Together, these results highlight the importance of context information to predict nonresponse in large-scale assessments conducted in schools.

Interestingly, logistic regression models with LASSO identified rather different covariates predicting dropout as BART (see Figures 4 and 5). For example, the BART model found the survey wave to be the most important factor triggering permanent dropout, whereas the LASSO regression identified the proportion of students with migration background as the driving force. BART also uncovered several important covariates that played no role according to LASSO regression (e.g., mean number of sick days in grade 5, mean grade in mathematics in grade 5). On the other hand, the logistic regression model considered a student's German grade to be relevant for his/her permanent dropout propensity, which seemed to be completely irrelevant according to the BART model. Similar discrepancies were observed for the prediction of temporary dropout. Whereas BART identified the mean satisfaction with health and the proportion of students with migration background as the most important factors, LASSO regression highlighted the waves already spend in the study (i.e., the sojourn time per wave) and the total number of waves participating in the study. Again, most factors found relevant according to BART (e.g., number of books at home and mean reading speed in grade 5) were not marked to be essential by LASSO regression (and vice versa). In order to assess whether multicollinearity causes the differences between the predictor sets identified as important by the two approaches, we examined the correlations and variance inflation factors (VIF) between all predictors. As expected, some predictors were substantially correlated (e.g., the cognitive measures or grades). However, none of the important predictors identified by the BART models or the LASSO regressions showed substantial multicollinearity. We conclude from this finding that multicollinearity is not the driving force behind the observed differences. Instead, we find it very likely that the differences are caused by higher level interactions that BART considers, but LASSO does not. Thus, this is a clear argument for the trustworthiness of BART results.

Prediction of Participation Rates

The BART event history models allow predicting the participation rates at any survey wave (potentially, even beyond the observational period of the study) given the observed covariates. Importantly, these represent conditional probabilities dependent that no permanent dropout has occurred in the previous waves. Panel A in Figure 6 summarizes the conditional probability distributions of permanent dropout of all individuals in waves 2 to 5. These results show that the conditional probabilities of permanent dropout were on average about 6.6% with a 95% credibility interval (*CrI*) of [5.9%, 7.2%] in wave 2 and increased to 10.0%, 95% *CrI* [9.1%, 10.9%], in wave 5. In contrast, the conditional probabilities of temporary dropout for the individuals who were at risk to experience such an event (panel B in Figure 6) were nearly identical across the four survey waves. They were about 5.0%, 95% *CrI* [4.4%, 5.6%] in wave 2, 5.2%, 95% *CrI* [4.7%, 5.7%], and 5.5%, 95% [4.8%, 6.0%], in the waves 3 and 4, and 6.6%, 95% *CrI* [5.6%, 7.5%] in wave 5, respectively. Following the two-step process outlined above, we also estimated the unconditional response probabilities, that is, the expected participation rates at each wave (panel C in Figure 6). These results show continually decreasing response rates in successive survey waves. Whereas a response rate of 88.8%, 95% *CrI* [87.8%, 89.6%], was predicted for wave 2, it decreased to 63.7, 95% *CrI* [62.3%, 65.0%], for wave 5. Importantly, these model predicted participation rates closely reproduced the descriptive survey participation rates given in Table 1. Therefore, the BART event history model could be used to predict expected response rates beyond the observational period.

Discussion

Machine learning methods offer intriguing opportunities for survey researchers in various contexts (Buskirk et al., 2018; Kern et al., 2019; Toth & Phipps, 2014). Particularly, tree-based approaches include a bundle of versatile alternatives to parametric regression that allow the modeling of complex relationships with computational efficiency. This paper presented a recently introduced Bayesian ensemble method (Chipman et al., 2010) for the analysis of nonresponse in social surveys. In contrast to wave-to-wave predictions that dominated previous nonresponse research (e.g., Durrant & Steele, 2009), we focused on the analysis of survey participation across multiple waves. We developed a nonparametric BART event history model to analyze competing risks for different types of nonresponse, that is, temporary and permanent dropout. This modeling strategy enables researchers identifying important variables driving the nonresponse process and, more importantly, constructing longitudinal prediction models. We applied our novel approach to data from a German large-scale assessment and showed that nonresponse in longitudinal school-based studies is predominately driven by the school context (e.g., mean number of sick days and mean satisfaction with health) and to a lesser degree by student characteristics. As expected, nonresponse also increased throughout the survey with each wave. The model-implied nonresponse rates highlighted that permanent dropout increased from 6.6% to 10.0%, whereas temporary dropout was nearly constant across the survey waves. Using two types of validity criteria and a model comparison with logistic regression with LASSO, we were able to highlight the advantages of BART. For example, in our application BART was able to predict the last wave's response pattern using only information from previous waves. A notable strength of our BART modeling approach is its flexibility to acknowledge different nonresponse processes. Although our analyses were limited to temporary and permanent dropout, an extension to additional types of nonresponse is straightforward by adapting the outcome coding (see Figure 2). For example, it is advisable to distinguish active refusal to participate in a survey from a failure to contact the respondent. In this case, three competing events could be contrasted. Even structural breaks resulting from different contexts experienced by the respondents could explicitly be modeled. In our data example, a substantial proportion of students left the school system after wave 3 (and, thus, turned into permanent dropout cases by design). Thus, more precise estimates of nonresponse trajectories might be derived by modeling different sequences of nonresponse, that is, before and after the anticipated structural break resulting from the different educational choices.

Implications for Survey Management

The presented nonresponse model can facilitate operational survey management in different ways. For example, prediction models can help gauging expected sample sizes across the course of a panel study. If a valid prediction model can be established, the predicted response rates for upcoming survey waves can be used to plan sample refreshments or preestimate sample mortality (i.e., timeframes for discontinuing further surveying). Alternatively, these analyses might also guide strategies to prevent nonresponse from taking place in the first place. If respondents with a high nonresponse probability can be identified beforehand, incentives might be adjusted accordingly (cf. Tourangeau, Brick, Loh, & Li, 2017). Adaptive incentivisation strategies might even link the size of an incentive to the predicted probability of nonresponse (see Figure 7): respondents that are expected to dropout are promised higher monetary compensations as compared to respondents with lower dropout propensities. Finally, if relevant characteristics of nonresponding units can be identified, these variables might guide sampling strategies for future surveys. Then, oversampling plans might be devised that explicitly target these subgroups with high propensity to dropout.

Limitations and Possible Extensions

Although the BART framework allows for complex modelling approaches, our event history model could be extended in several ways. For example, our predictor set was limited to information collected prior to or at the first measurement occasion. However, panel studies collect new information about respondents in each wave. A fruitful model extension pertains to the inclusion of time-varying covariates that are gathered during the course of a longitudinal study. Moreover, in the present form, our model considers only observed individual-specific heterogeneity. Although it can be assumed that the large number of predictors studied covers individual-specific heterogeneity on its whole, in the future our approach could be extended by a frailty term. It would also be interesting to know how well our BART approach fares as compared to other machine learning methods such as random forests and boosted trees. Because the evaluation of their relative performance is beyond the scope of this study, it is an important undertaking left for future work. In a related vein, it might also be worthwhile to extend the scope of our BART approach and evaluate whether it is suitable for imputing missing values of time-to-event data. So far, studies have already highlighted the potential of BART for imputing MAR covariates (Xu, Daniels, & Winterstein, 2016) or for imputing MAR data in the context of augmented inverse probability estimation and penalized splines propensity prediction (Tan, Flannagan, & Elliot, 2019). The extension of these approaches to longitudinal data seems a worthwhile endeavour for future work. Finally, the generalizability of our findings on the predictors of dropout beyond the studied student sample should be an objective of further research. This would help establishing whether the identified predictors of nonresponse are similar important in different populations (e.g., adults) and contexts (e.g., household surveys).

Conclusion

Modern machine learning techniques such as BARTs augment the statistical toolbox of survey researchers for nonresponse adjustments and the examination of nonresponse patterns. In longitudinal settings, BART prediction models facilitate the estimation of response rates in upcoming survey ways. For this purpose, the present study described a novel event history approach that allows examining competing risks for different types of nonresponse. In an empirical demonstration, we showed that this technique allows identifying important drivers of temporary and permanent dropout across multiple survey waves. Thus, operational survey management might use respective prediction models to gauge sample mortality and plan sample refreshments at an early stage.

Data Availability

Most of the data analyzed in this study is provided at https://doi.org/10.5157/NEPS:SC3:8.0.0. Due to German privacy laws, some school variables (e.g., type of institution or school location) cannot be made publicly available.

Software Information

The analyses were conducted in *R* version 3.5.1 using the *BART* package version 2.2. The analyses syntax to reproduce our results will be available at https://github.com/bieneSchwarze/BARTforNonresponse.git.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162-172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adhikari, P., & Bryant, L. A. (2018). Sampling hard-to-locate populations. In L. R. Atkeson & R. M. Alvaraz (Eds.), *The Oxford Handbook of Polling and Survey Methods* (pp. 155-180). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190213299.013.2

Beullens, K., Vandenplas, C., Loosveldt, G., & Stoop, I. (2018). Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2018-00003

Blatt, I., Jarsinski, S., & Prosch, A. (2017). *Technical Report for Orthography: Scaling results of Starting Cohort 3 in Grades 5, 7, and 9* (NEPS Survey Paper No. 15). Bamberg, Germany: Leibniz Institute for Educational Trajectories.

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, *8*, 1750-1781. https://doi.org/10.1214/14-AOAS755

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). Editorial. *Zeitschrift für Erziehungswissenschaft*, *14*, 1-4. http://doi.org/10.1007/s11618-011-0198-z

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Brooks/Cole Publishing.

Brüderl, J., & Trappmann, M. (2017). Data collection in panel surveys. *Methods, Data, and Analyses*, *11*, 3-6.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press.

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, *11*, 1-10. https://doi.org/10.29115/SP-2018-0004

Buskirk, T. D., & Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2015-00003

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search (with discussion and a rejoinder by the authors). *Journal of the American Statistical Association*, *93*, 935-960. https://doi.org/10.1080/01621459.1998.10473750

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*, 266-298. https://doi.org/10.1214/09-AOAS285

Díaz-Uriarte, R., & de Andrés, A. S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 1-13. https://doi.org/10.1186/1471-2105-7-3

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 19). Bamberg, Germany: Otto-Friedrich-Universität Bamberg.

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: Evidence from six UK government surveys. *Journal of the Royal Statistical Society: Series A*, *172*, 361-381. https://doi.org/10.1111/j.1467-985X.2008.00565.x

Earp, M., Mitchell, M., McCarthy, J. S., & Kreuter, F. (2014). Modeling nonresponse in establishment surveys: Using an ensemble tree model to create nonresponse propensity scores and detect potential bias in an agricultural survey. *Journal of Official Statistics*, *30*, 701-719. https://doi.org/10.2478/jos-2014-0044

Felderer, B., Müller, G., Kreuter, F., & Winter, J. (2018). The effect of differential Incentives on attrition bias: Evidence from the PASS wave 3 incentive experiment. *Field Methods*, *30*, 56-69. https://doi.org/10.1177/1525822X17726206

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, & J. O. Berger (Eds.), *Bayesian Statistics 4* (pp. 169-193), New York, NY: Oxford University Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Gramacy, R. B., Taddy, M., & Wild, S. M. (2013). Variable selection and sensitivity analysis using dynamic trees, with an application to computer code performance tuning. *Annals of Applied Statistics, 7*, 51-80. https://doi.org/10.1214/12-AOAS590

Hastie, T., & Tibshirani, R. (2000). Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science, 15*, 196-223. https://doi.org/10.1214/ss/1009212815

Heffetz, O., & Reeves, D. B. (2019). Difficulty of reaching respondents and nonresponse Bias: Evidence from large government surveys. *Review of Economics and Statistics, 101*, 176-191. https://doi.org/10.1162/rest_a_00748

Kapelner, A., & Bleich, J. (2016). BartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software, 70*, 1-40. https://doi.org/10.18637/jss.v070.i04

Keiding, N. (2014). Event history analysis. *Annual Review of Statistics and its Application, 1*, 333-360. https://doi.org/10.1146/annurev-statistics-022513-115558

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods, 13*, 73-93. https://doi.org/10.18148/srm/2019.v1i1.7395

Kleinert, C., Christoph, B., & Ruland, M. (2019). Experimental evidence on immediate and long-term consequences of test-induced respondent burden for panel attrition. *Sociological Methods & Research*. Advance online publication. https://doi.org/10.1177/0049124119826145

Kreuter, F. (2013). Facing the nonresponse challenge. *Annals of the American Academy of Political and Social Science, 64*, 23-35. http://doi.org/10.1177/0002716212456815

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Williger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie* [Measurement of fluid cognitive abilities over the life course in the NEPS] (NEPS Working Paper No. 43). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C., & Pratola, M. (2019). BART: *Bayesian Additive Regression Trees*. R package version 2.2. https://CRAN.R-project.org/package?=BART

McGovern, M. E., Canning, D., & Bärnighausen, T. (2018). Accounting for non-response bias using participation incentives and survey design: An application using gift vouchers. *Economics Letters, 171*, 239-244. https://doi.org/10.1016/j.econlet.2018.07.040

Mills, M. (2011). *Introducing Survival and Event History Analysis*. Los Angeles, CA: Sage.

Michaud, P. C., Kapteyn, A., Smith, J. P., & van Soest, A. (2011). Temporary and permanent unit non-response in follow-up interviews of the Health and Retirement Study. *Longitudinal and Life Course Studies, 2*, 145-169. https://doi.org/10.14301/llcs.v2i2.114

Müller, B., & Castiglioni, L. (2017). Do temporary dropouts improve the composition of panel data? An analysis of “gap Interviews” in the German Family Panel pairfam.

Sociological Methods & Research. Advance online publication. https://doi.org/10.1177/0049124117729710

Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., & Omar, R. Z. (2015). How to develop a more accurate risk prediction model when there are few events. *Bmj*, 351, h3868. https://doi.org/10.1136/bmj.h3868

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*, 74-97. https://doi.org/10.1093/poq/nfp014

Phipps, P., & Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, *6*, 772-794. https://doi.org/10.1214/11-AOAS521

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 15). Bamberg, Germany: Otto-Friedrich-Universität Bamberg.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Roßmann, J., & Gummer, T. (2016). Using paradata to predict and correct for panel attrition. *Social Science Computer Review, 34,* 312-332. https://doi.org/10.1177/0894439315587258

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592. https://doi.org/10.1093/biomet/63.3.581

Sieben, S., & Lechner, C. M. (2019). Measuring cultural capital through the number of books in the household. *Measurement Instruments for the Social Sciences, 2*:1. https://doi.org/10.1186/s42409-018-0006-0 van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*. Advance online publication. https://doi.org/10.1177/0962280218784726

Sparapani, R. A., Logan, B. R., McCulloch, R. E., & Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine, 35*, 2741-2753. https://doi.org/10.1002/sim.6893

Sparapani, R.A., Logan, B. R. , McCulloch, R. E., & Laud, P. W. (2019). Nonparametric competing risks analysis using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research*. Advance online publication. https://doi.org/10.1177/0962280218822140

Steinhauer, H. W., Aßmann, C., Zinn, S., Goßmann, S., & S. Rässler (2015). Sampling and weighting cohort samples in institutional contexts. *AStA Wirtschafts- und Sozialstatistisches Archiv,* 9, 131-157. https://doi.org/10.1007/s11943-015-0162-0

Tan, Y. V., Flannagan, C. A., & Elliott, M. R. (2019). “Robust-squared” imputation models using Bart. *Journal of Survey Statistics and Methodology*, *7*, 465-497. https://doi.org/10.1093/jssam/smz002

Toth, D., & Phipps, P. (2014). *Regression tree models for analyzing survey response*. Proceedings of the Government Statistics Section (pp. 339-351). Alexandria, VA: American Statistical Association.

Tourangeau, R., Brick, M. J., Lohr, S., & Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society: Series A*, *180*, 203-223. https://doi.org/10.1111/rssa.12186

Trappmann, M., Gramlich, T., & Mosthaf, A. (2015). The effect of events between waves on panel attrition. *Survey Research Methods 9*, 31-43. https://doi.org/10.18148/srm/2015.v9i1.5849

Voorpostel, M., & Lipps, O. (2011). Attrition in the Swiss Household Panel: Is change associated with later drop-out? *Journal of Official Statistics, 27*, 301-318.

Watson, N., & Wooden, M. (2014). Re-engaging with survey non-respondents: evidence from three household panels. *Journal of the Royal Statistical Society: Series A*, *177*, 499-522. https://doi.org/10.1111/rssa.12024

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A, 176*, 211-225. https://doi.org/10.1111/j.1467-985X.2012.01038.x

Williams, D., & Brick, J. M. (2017). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *6*, 186-211. http://doi.org/10.1093/jssam/smx019

Xu, D., Daniels, M. J., & Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17, 589-602. https://doi.org/10.1093/biostatistics/kxw009

Zimmermann, S., Gehrer, K., Artelt, C., & Weinert, S. (2012). *The Assessment of Reading Speed in Grade 5 and Grade 9*. Bamberg, Germany: University of Bamberg,

Zinn, S., & Gnambs, T. (2018). Modeling competence development in the presence of selection bias. *Behavior Research Methods, 50*, 2426-2441. https://doi.org/10.3758/s13428-018-1021-z

Zinn, S., Würbach, A., Steinhauer, H. W., & Hammon, A. (2018). *Attrition and selectivity of the NEPS Starting Cohorts: An overview of the past 8 years* (NEPS Survey Paper No. 35). Bamberg, Germany: Leibniz Institute for Educational Trajectories.

Table 1.

Nonresponse Across Survey Waves.

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Survey participation	4,559 (100%)	4,064 (89%)	3,606 (79%)	3,275 (72%)	2,905 (64%)
Temporary dropout	-	202 (4%)	201 (4%)	187 (4%)	214 (5%)
Permanent dropout:	-	293 (6%)	752 (16%)	1,097 (24%)	1,440 (32%)
- Student switched school		144 (3%)	432 (9%)	6 (0%)	-
- Student left school system		-	-	924 (20%)	1,145 (25%)
- Student refused		-	-	53 (1%)	203 (4%)
- School refused		111 (2%)	219 (5%)	-	-
- Unknown / other reasons		38 (1%)	96 (2%)	114 (3%)	92 (2%)
Total	4,559 (100%)	4,559 (100%)	4,559 (100%)	4,559 (100%)	4,559 (100%)

*Not*e: Due to rounding the percentage of survey participation may not equal 100%.

Table 2.

Accuracy measure of validation studies.

Type of Validation	Wave-specific accuracy[†]	Model for permanent dropout		Model for temporary dropout	
		BART	Logit with LASSO	BART	Logit with LASSO
Out of sample: over all waves	Wave 2	0.95	0.94	0.95	0.94
	Wave 3	0.91	0.89	0.96	0.99
	Wave 4	0.93	0.91	0.96	0.99
	Wave 5	0.90	0.89	0.96	0.98
Out of sample: only last wave	Wave 5	0.90	0.90	0.97	0.95

Note: [†] Wave 1 only consists of participants and constitutes the reference set for all further waves. Therefore, per definition in Wave 1 no dropout had occurred and is thus not modeled/predicted.

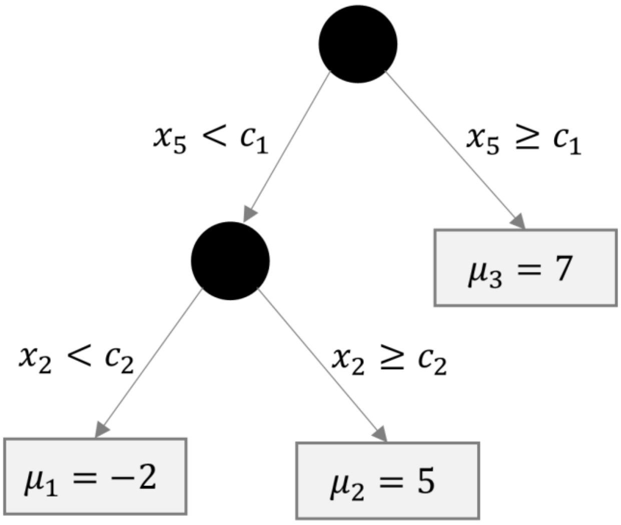

Figure 1. Example of a regression tree of depth *d* = 2 with two splits (including cut scores c_1 and c_2) and *m* = 3 terminal nodes.

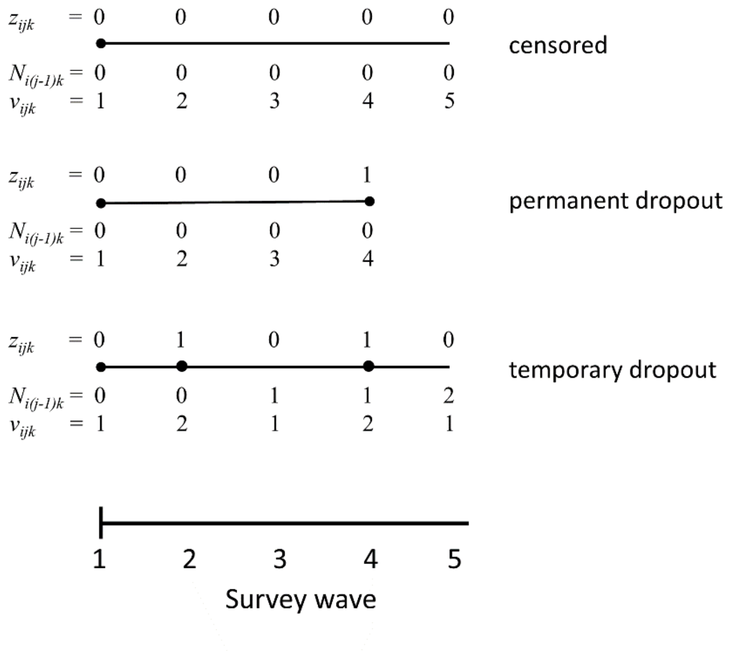

Figure 2. Outcome coding for different event types.

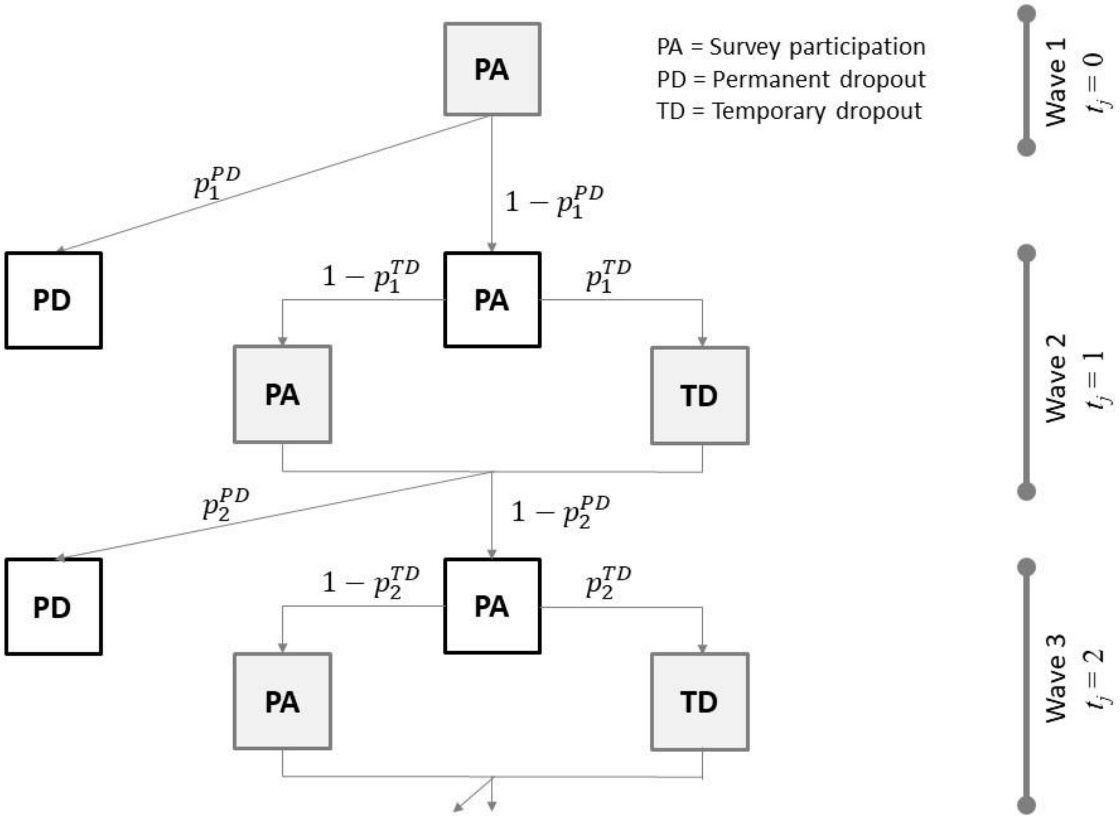

Figure 3. Two-step process of survey participation for two types of nonresponse across three waves.

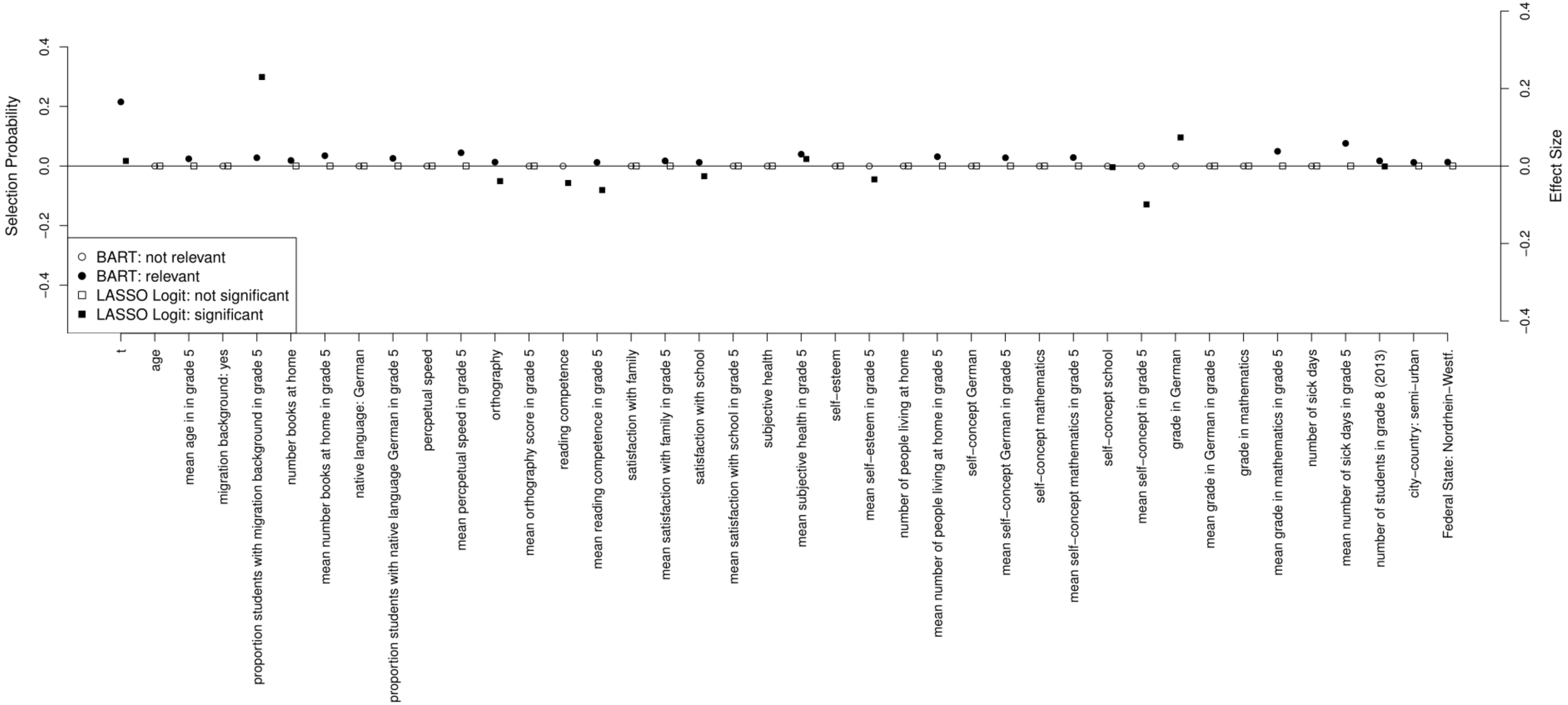

Figure 4. Relative importance of selected covariates for predicting permanent dropout with *t* as survey wave using BART and LASSO regression. The solid line represents the threshold for nonignorable importance, filled dots mark variables of nonignorable importance with significant impact ($p < .05$), and empty dots mark variables of ignorable importance with non-significant impact ($p > .05$). Full results are given in the supplement material.

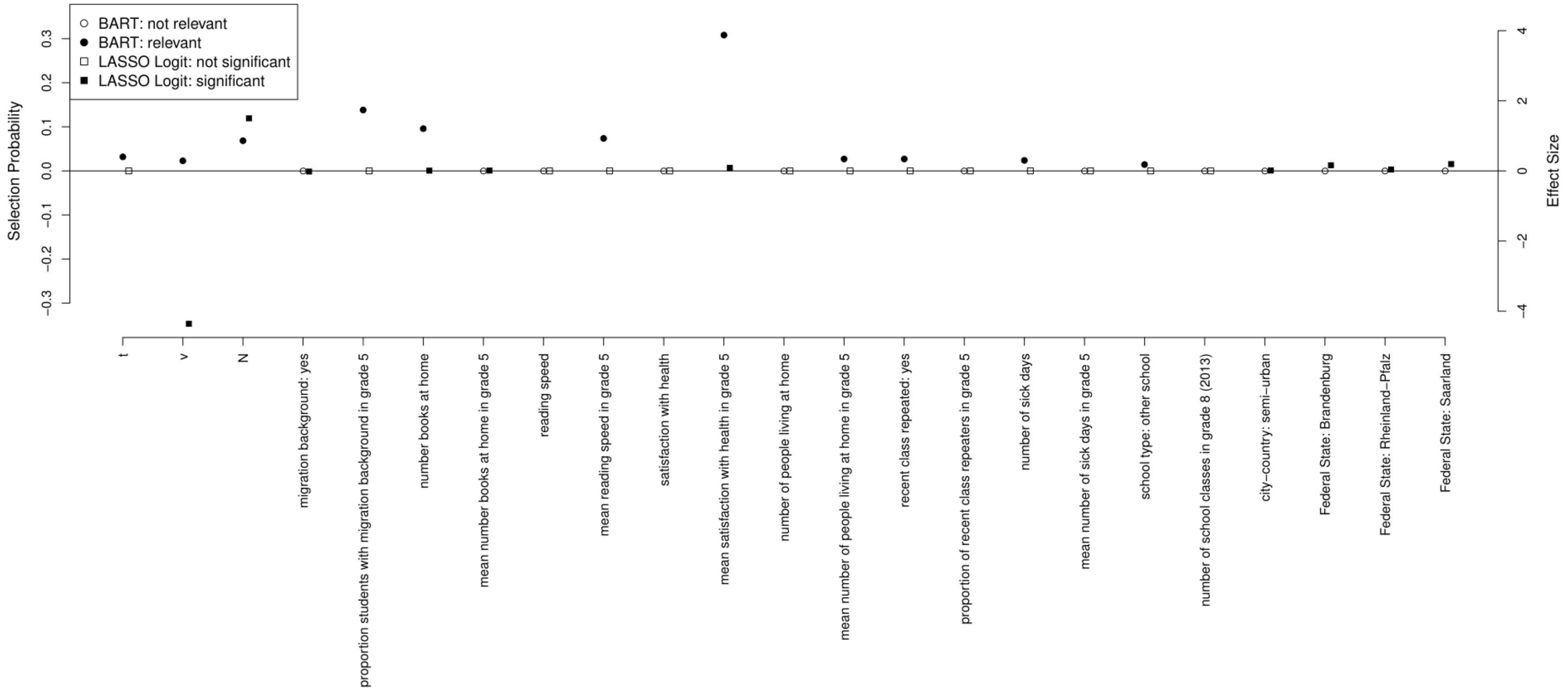

Figure 5. Relative importance of selected covariates for predicting temporary dropout with *t* as survey wave, *v* as the event time, and *N* as the number of previous dropouts for BART and LASSO regression. The solid line represents the threshold for nonignorable importance, filled dots mark variables of nonignorable importance with significant impact ($p < .05$), and empty dots mark variables of ignorable importance with non-significant impact ($p <$ .05). Full results are given in the supplemental material.

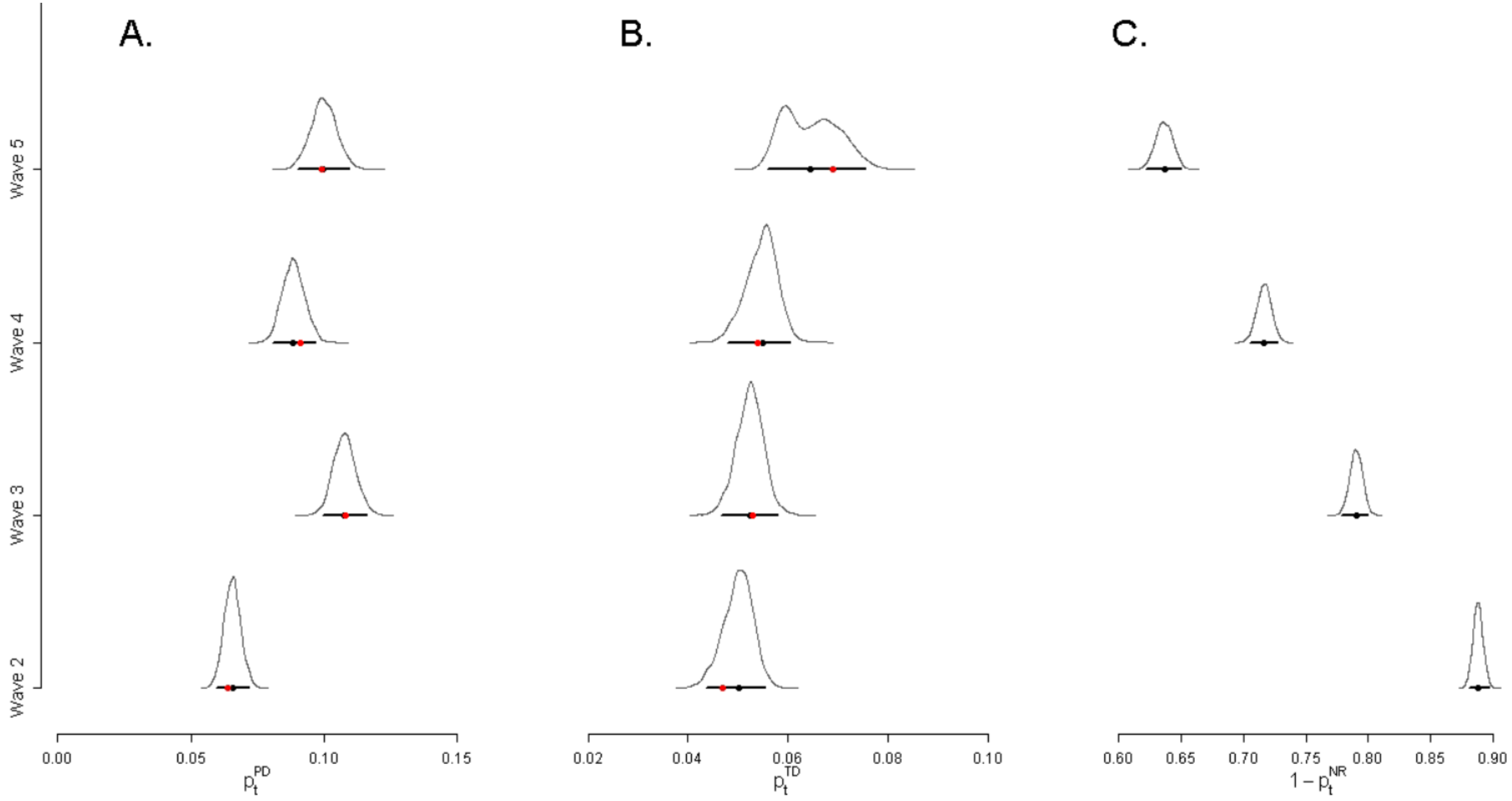

Figure 6. Predicted participation status across waves. A: Conditional permanent dropout rates, B: Conditional temporary dropout rates, C: Unconditional participation rates. The black curves are the posterior densities of the estimated probabilities. The black dots mark their median, the horizontal lines their 95% credibility intervals, and the red dots in panel A and B the empirical frequencies of the number of observed events.

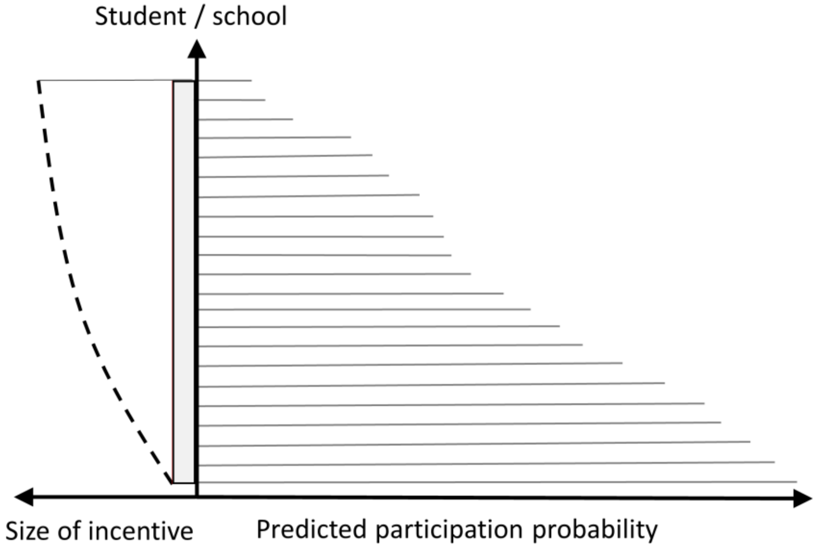

Figure 7. Example of an adaptive incentivisation scheme dependent on predicted participation probabilities. The gray box represents the baseline incentive and the dashed line the additional incentive.

Supplement material for

Analyzing Nonresponse in Longitudinal Surveys Using Bayesian Additive Regression Trees:

A Nonparametric Event History Analysis

S1. Derivation of Decomposition of Formula for Survey Participation at Wave $t + 1$

In the stochastic process (V_t, Z_t), V_t represents the survey participation resulting from a non-recurrent event (i.e., permanent dropout) and Z_t the respective respondent outcome for a recurrent event (i.e., temporary dropout) at time point *t*. The joint transition probabilities for V_{t+1} and Z_{t+1} can be derived from two submodels by splitting the respective probabilities into conditional probabilities for the non-recurrent event (V_{t+1}) and the recurrent event (Z_{t+1}) in the following way:

$$P(V_{t+1} = v_{t+1}, Z_{t+1} = z_{t+1} | (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}) =$$

$$= \frac{P\big(V_{t+1} = v_{t+1}, Z_{t+1} = z_{t+1}, (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}\big)}{P\big((V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}\big)}$$

$$= \frac{P\big(V_{t+1} = v_{t+1}, Z_{t+1} = z_{t+1}, (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}\big)}{P\big(Z_{t+1} = z_{t+1}, (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}\big)}$$

$$* \ \frac{P\big(\ Z_{t+1} = z_{t+1}, (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}\big)}{P\big((V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}}\big)}$$

$$= P(V_{t+1} = v_{t+1}\ | Z_{t+1} = z_{t+1}, (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}})$$

$$* \ P(Z_{t+1} = z_{t+1} |\ (V_s = v_s, Z_s = z_s)_{s\in\{0,\dots,t\}})$$

Table S1.

Conditioning Variables for Nonresponse Analyses

		M	*SD*	Range	MV
	Respondent characteristics				
1.	Sex	0.5	0.5	[0, 1]	0%
2.	Age	15.1	0.5	[12.7, 18.2]	0%
3.	Mother tongue	0.1	0.3	[0, 1]	0%
4.	Migration background	0.2	0.4	[0, 1]	0%
5.	Household size	4.5	1.7	[2, 35]	5%
6.	Number of books at home	4.0	1.4	[0, 6]	1%
	Student characteristics				
7.	Repeated school	0.1	0.3	[0, 1]	2%
8.	Number of days sick	2.0	3.8	[0, 50]	21%
9.	Grade in German	2.3	0.9	[1, 6]	7%
10.	Grade in mathematics	2.3	0.9	[1, 6]	7%
	Psychological characteristics				
11.	Satisfaction with life	8.2	2.3	[0, 10]	5%
12.	Satisfaction with current living standards	8.7	2.2	[0, 10]	4%
13.	Satisfaction with health	8.8	2.2	[0, 10]	3%
14.	Satisfaction with family	9.1	2.0	[0, 10]	4%
15.	Satisfaction with friends	8.9	2.0	[0, 10]	3%
16.	Satisfaction with school	7.8	2.5	[0, 10]	3%
17.	Subjective health	1.7	0.7	[1, 5]	1%
18.	Self-esteem	4.0	0.7	[1.0, 5.0]	15%
19.	German self-concept	3.0	0.6	[1, 4]	7%
20.	Mathematical self-concept	2.9	0.8	[1, 4]	7%
21.	School self-concept	3.2	0.6	[1, 4]	7%
22.	Perceptual speed	44.0	13.0	[1, 93]	0%
23.	Reading speed	21.5	6.9	[0, 51]	0%
24.	Reasoning	7.0	2.6	[0, 12]	0%
25.	Orthography	0.0	1.3	[-7.2, 4.6]	0%
26.	Mathematical competence	0.1	1.2	[-4.4, 4.0]	0%
27.	Reading competence	0.0	1.2	[-4.2, 4.1]	0%
	School characteristics				
28.	School type: intermediate secondary	0.2	0.4	[0, 1]	0%
29.	School type: other	0.3	0.4	[0, 1]	0%
30.	Number of students in grade 8	102.8	50.7	[12, 346]	2%
31.	Number of classes in grade 8	3.9	1.7	[1, 12]	2%
32.	Institution type	0.1	0.3	[0, 1]	0%
33.	School location: part urban / part rural	0.4	0.5	[0, 1]	0%
34.	School location: rural	0.1	0.3	[0, 1]	0%

Note. Respondent, student, and psychological characteristics aggregated to the school level and dummy-indicators for federal states and missing values are not included.

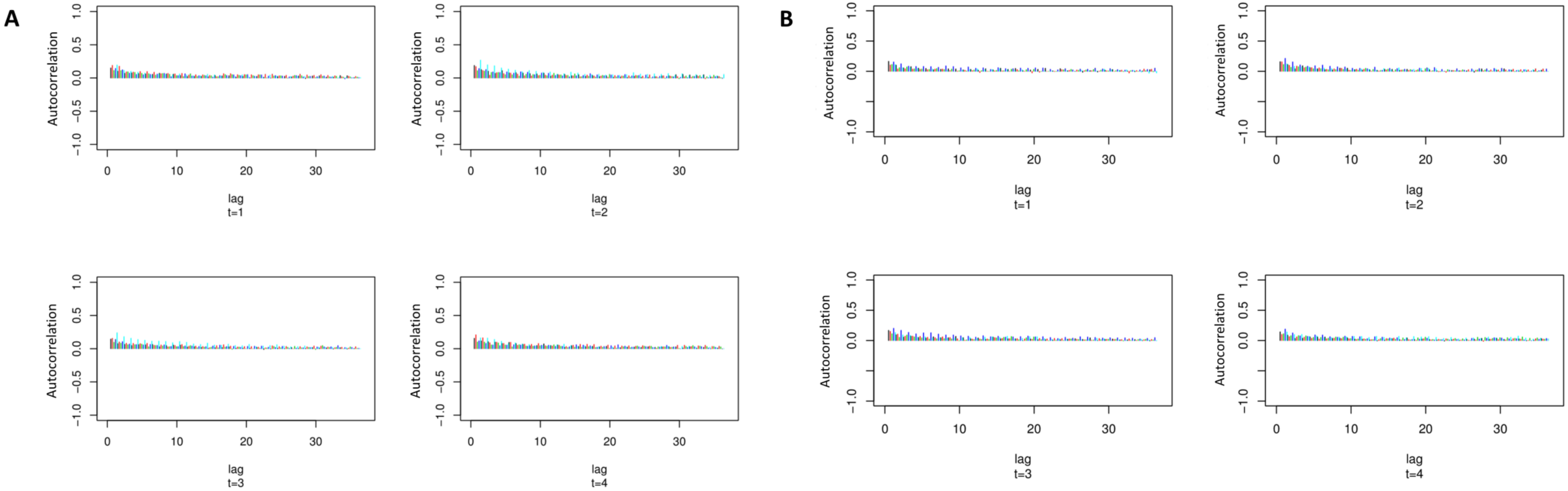

Figure S1. Autocorrelation plots for BART event history model. A = Permanent dropout model, B = Temporary dropout model.

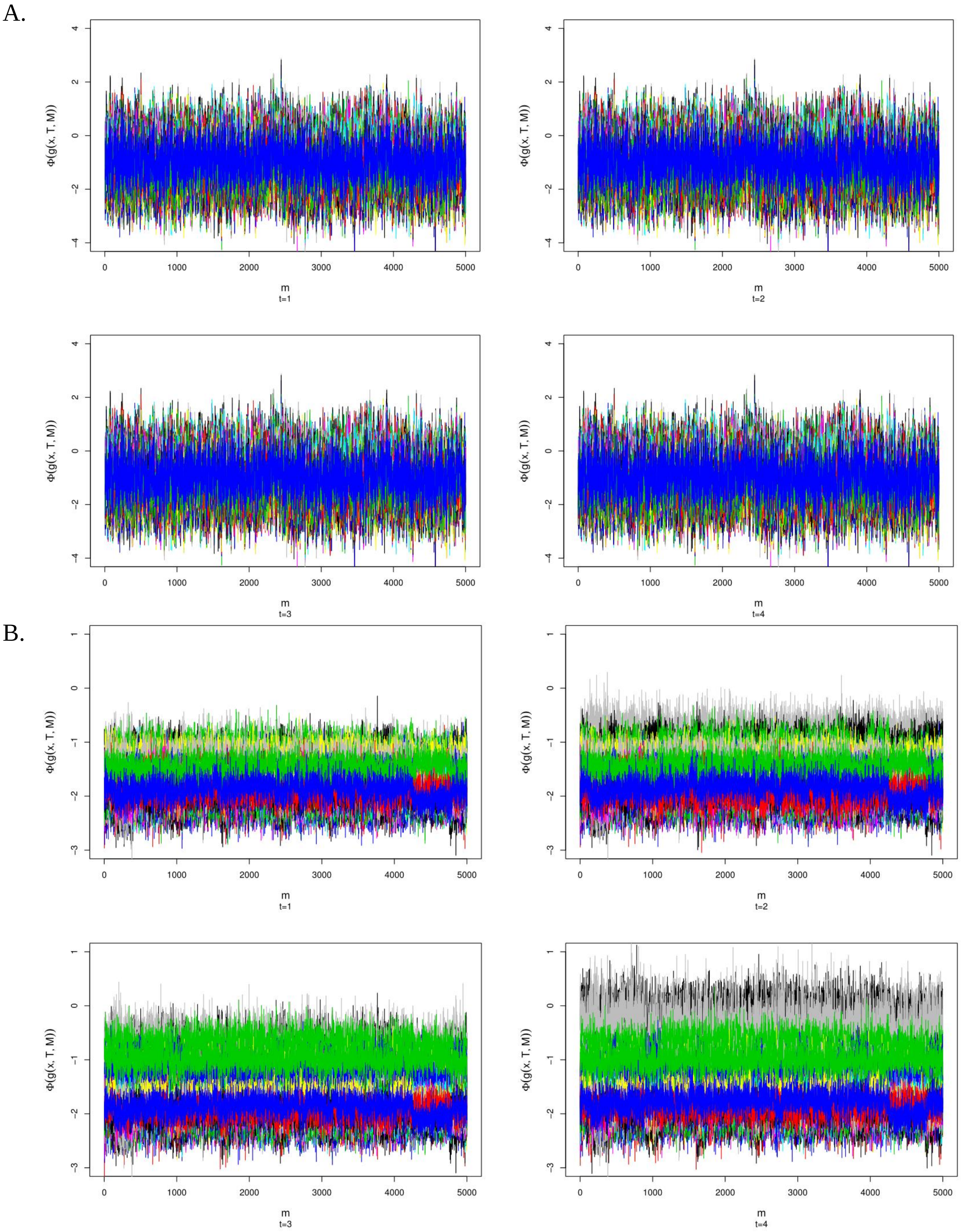

Figure S2. Trace plots for BART event history model. A = Permanent dropout model, B = Temporary dropout model.

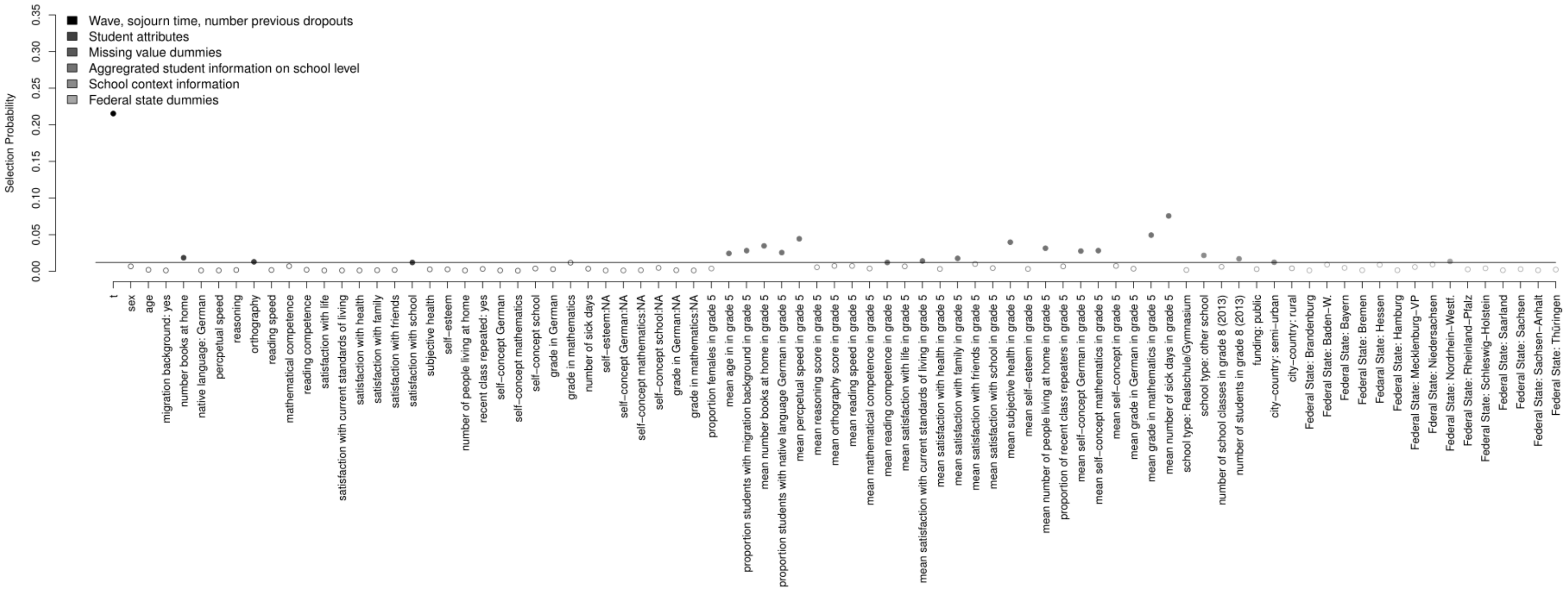

Figure S3. Relative covariate importance for predicting permanent dropout with *t* as survey wave, *v* as the event time, and *N* as the number of previous dropouts for BART. The solid line represents the threshold for nonignorable importance, filled dots mark variables of nonignorable importance, and empty dots mark variables of ignorable importance.

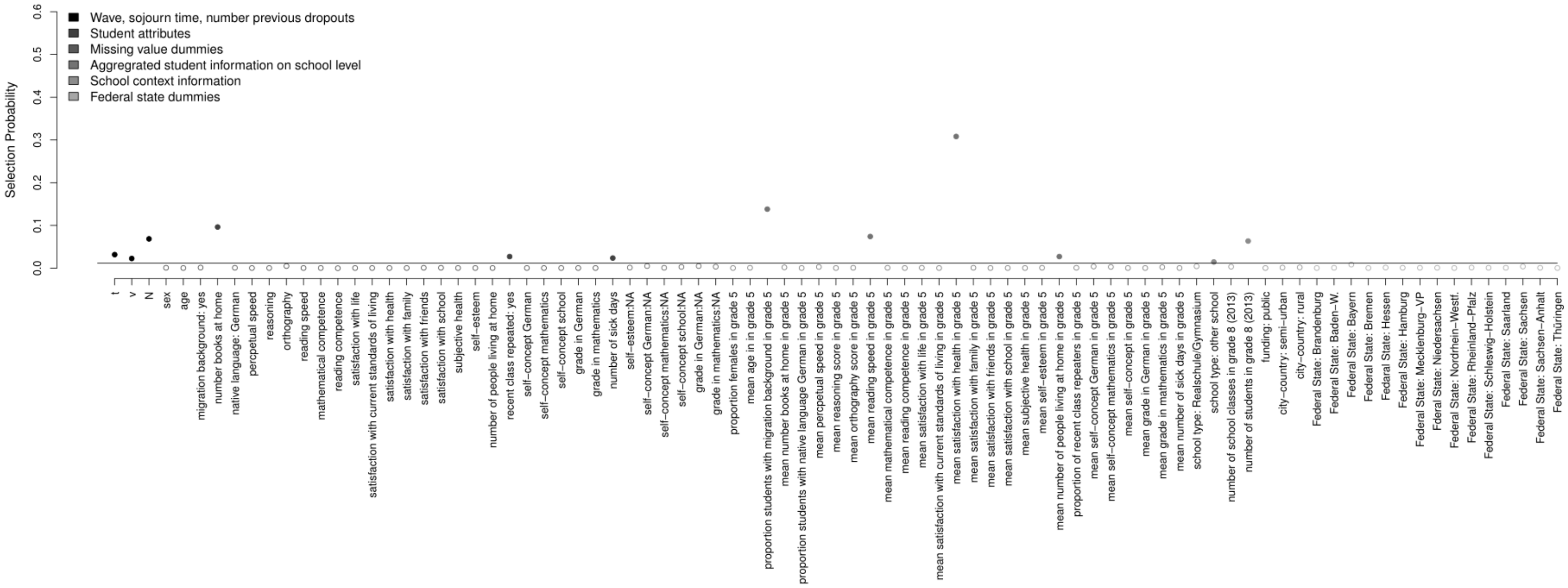

Figure S4. Relative covariate importance for predicting temporary dropout with *t* as survey wave, *v* as the event time, and *N* as the number of previous dropouts for BART. The solid line represents the threshold for nonignorable importance, filled dots mark variables of nonignorable importance, and empty dots mark variables of ignorable importance.