To predict and explain. How prediction improves our understanding of models: an application to the study of teacher bias.

Mark D. Verhagen^{*a,b}

^aLeverhulme Centre for Demographic Science, 42 Park End St, Oxford OX1 1JD, UK

^bNuffield College, University of Oxford, 1 New Rd, Oxford OX1 1NF, UK

November 29, 2021

Abstract

The value of prediction is recognized in niche empirical settings within sociology, for instance to generate understanding into effect sizes in categorical outcome models. This paper argues that a predictive perspective is much more broadly relevant to empirical work, and provides a host of unique advantages which can strongly improve our understanding of the models we estimate. In particular out-of-sample prediction gives an intuitive measure of a model's ability to structure an outcome of interest, and complements aggregate statistics obtained from conventional methods. Prediction also provides intuitive insights into substantive effects, and is amenable to interventional reasoning. In this paper, I illustrate the ease with which prediction is incorporated into empirical analysis, and its relevance in providing complementary insights to standard methods. By studying the important topic of bias in teachers' evaluations of students using rich data on high-stakes tracking decisions in The Netherlands, I find a number of nuances to the existing knowledge on the subject. These include under-estimation of bias towards low-SES students and considerable differences in evaluation practices across schools. Important findings which were not picked up in previous work, but are easily identified from a predictive perspective and prompt reflection on the Dutch educational system's equity.

Keywords: Prediction, Education, Tracking, Inequality, Teacher Bias

*All code and publicly available data underlying the analyses in this paper can be found at https://github.com/MarkDVerhagen/Prediction_Teacher_Bias. Author acknowledges funding by Nuffield College and the ESRC. Corresponding email: mark.verhagen@nuffield.ox.ac.uk

21 Introduction

22 Out-of-sample prediction¹ – the process of estimating a model on a (subset of the)
23 dataset and evaluating these models on data not used for estimation – is an under-
24 utilized approach within the social sciences. In this paper, I argue that prediction can
25 and should be used more frequently as an analytical tool, as it is simple to implement,
26 completely complementary to conventional methods, and provides various unique
27 analytical insights. In particular when more convoluted models are appropriate – as
28 is often the case when studying the complexities of social life – prediction becomes
29 increasingly valuable, because it can be used as a tool to generate understanding into
30 models that defy straightforward interpretation. Frequently encountered examples
31 include categorical outcome or multilevel models. The central goal of this paper is to
32 illustrate how prediction complements conventional analyses, and to show both the
33 ease and power of implementing a predictive perspective. As a substantive case study,
34 I analyze the well-studied question of teacher bias in evaluating their students’ ability
35 within a high-stakes context, and find various important nuances to the existing
36 sociological knowledge on the topic when using a predictive perspective.

37 Often when studying social mechanisms, complexities in model estimation arise.
38 These complexities can be a result of the mechanism under study, for example (un-
39 observed) heterogeneity in effects, or the nature of the available data, like censored,
40 categorical or count variables. There are usually econometric solutions to these prob-
41 lems, but these solutions often come at a hefty price: the loss of straightforward
42 model interpretation. For example, addressing the categorical nature of an outcome
43 variable typically leads to coefficient estimates which are difficult to interpret sub-
44 stantively. At the same time, it is well-known that ignoring such complexities can
45 lead to incorrect inference due to misspecification bias (Cameron and Trivedi, 2005).
46 Although model complexities are often present in sociology, and warrant appropriate
47 modelling to avoid biased results, sociologists put a considerable premium on the
48 straightforward interpretation of models as opposed to increasing model complexity.
49 A preference which is evidenced by the continuing dominance of the standard linear
50 additive model throughout the social sciences (Schrodt, 2014; Hindman, 2015) but
51 also the common practice of making simplifying assumptions like modelling an ordi-
52 nal variable as a continuous one (Mare, 1980; Domingue, Kanopka, Trejo, and Freese,
53 2021)

54 Out-of-sample prediction can provide a solution to the trade-off between appropri-

¹I use the terms ‘out-of-sample prediction’ and ‘prediction’ interchangeably throughout this paper, reflecting the holistic view of prediction as simply evaluating an estimated model on data not seen in estimation.

55 ate model complexity and interpretability. Importantly, a predictive perspective as
56 understood here is meant to *complement* conventional empirical analysis, and should
57 not be understood as supplanting the latter – contrary to the often heated debate
58 regarding the merits of explanation and prediction, which tends to signal an unne-
59 cessary incompatibility between the two (Watts, 2014; Verhagen, 2021). Viewed in
60 such a way, prediction offers two unique sources of insight into the models we typi-
61 cally estimate. First, prediction provides an indication of the ability of a model to
62 structure the outcome of interest on a more detailed and disaggregated level than
63 typical summary fit metrics like the R^2 or goodness-of-fit measures. Second, making
64 predictions using selective subsets of the data to estimate and evaluate models pro-
65 vides a model-agnostic framework to understanding effect sizes in an intuitive way. In
66 addition, many of the typical benefits that motivate the widespread use of prediction
67 in other domains – notably that of machine learning where the approach is common
68 practice – also apply, for example to guard against possible overfitting of the data
69 and to assess a model’s external validity.²

70 To illustrate both the ease and value of implementing a predictive perspective, I apply
71 a predictive analysis to the canonical question of teacher bias in student evaluations.
72 A teacher’s subjective perception of a student’s ability plays an important role in
73 educational systems, both in day-to-day teaching – by driving the amount and type
74 of didactic material offered to the student – but especially when it comes to more
75 comprehensive decision-making, like the choice to have students repeat a year or
76 ability sorting through (informal) tracking. Therefore, when teacher evaluations are
77 biased, educational equity can come under considerable pressure. To study teacher
78 bias, I use the high stakes case study of tracking in the Netherlands, where students are
79 tracked into multiple ability levels. The Dutch case presents an important moment in
80 the life course, as it occurs comparably early – at the age of 12 – and dictates the level
81 of secondary education a student will receive, which then determines access to tertiary
82 education. The case also positions itself as a unique opportunity to study teacher bias
83 in a real-world setting, especially since a 2015 policy change in the Netherlands has
84 made a pupil’s individual teacher the key determinant of his/her assigned track level.³

85 Equally importantly for this paper’s purposes, the study of teacher bias faces a num-
86 ber of hurdles in estimation frequently encountered in the social sciences which make
87 it an excellent case study to illustrate the value of prediction. First, the mechanism
88 under study is complicated since bias manifests itself through individual teachers

²Note that external validity is understood here in a narrow sense; which is verifying that a model performs as well in-sample as out-of-sample on a similar dataset.

³Prior to 2015, a central testing mechanism took precedent.

89 which warrants a hierarchical modelling approach. Second, the outcome variable is
90 categorical in nature thus adding further complexity to model interpretation. In ad-
91 dition, teacher bias itself is a prediction exercise where teachers make forecasts of a
92 student’s ability based on observables, making a predictive approach quite natural.⁴
93 Perhaps most relevant, however, is that the study of teacher bias is a prime example
94 where sociologists have historically tended to prefer interpretability over appropri-
95 ate model complexity in model estimation and as a consequence have made various
96 simplifying assumptions to the models typically used to study teacher bias. As this
97 paper will show, this has led to likely under-estimation of important sources of bias.

98 In this paper, I thus aim to make two distinct contributions. First, to illustrate the
99 value of prediction to improve our understanding of models in an intuitive and easy-to-
100 implement way. To this effect, the first half of the paper will be dedicated to outlining
101 how and why a predictive approach can be relevant to sociologists. Second, to show
102 empirically – using prediction as an analytical tool – that the existing literature on
103 teacher bias in tracking in The Netherlands has under-estimated important sources
104 of heterogeneity in teacher evaluations, providing important pointers for policy and
105 future research. The second half of the paper will be dedicated to this substantive
106 question.

107 The Case for Out-of-sample Prediction

108 *Typical model-building and in-sample diagnostics*

109 The most prevalent functional form throughout the social sciences is the basic linear
110 additive model relating some outcome y with explanatory variables x : $y = \beta_0 +$
111 $\beta_1 x_1 + \dots + \beta_k x_k$. The model is estimated on the entirety of the available data
112 $\mathcal{D}$ and yields a straightforward, in-sample assessment of the associations between
113 explanatory variables and the outcome, captured for each variable x_k by the estimated
114 parameter $\hat{\beta}_k$. Typically, some mechanism of interest is encoded in the model and
115 the sign and significance of some $\hat{\beta}_k$ is used to evaluate the empirical evidence in
116 support of the hypothesized mechanism.

117 This typical approach to empirical analysis has come under severe criticism in recent
118 decades. First, the model-building process is often cited as opaque, allowing for so-
119 called researcher degrees of freedom in estimation and the phenomenon of ‘p-hacking’
120 (Young and Holsteen, 2017). Relatedly, the performance of research findings in struc-

⁴As others have argued forcefully, many explanatory questions typically asked by sociologists directly imply, or can easily be reformulated in a predictive way as well (Watts, 2014).

turing that which we are interested in is rarely assessed, limiting both an appropriate assessment of the generalisability of findings, as well as their practical relevance. Second, the standard linear model is often too simplistic to capture the complexities of social life. There are a number of reasons that may affect the appropriateness of the basic linear model. Data complexities can arise, e.g. when the outcome is categorical or a count variable, or in case the observed data is selected or censored relative to the population of interest. There may also be (unobserved) sources of heterogeneity requiring effects to differ for subsets of the sample, or the mechanism under study may be more complex than the basic model allows due to non-linearities or interactions. Correctly addressing these types of complexities is paramount to valid statistic inference, but often harm the straightforward interpretation of the basic model described earlier.⁵

Although the importance of appropriate model complexity is crucial to statistical inference (Rüttenauer and Ludwig, 2020; Mare, 1980; Heckman, Humphries, and Veramendi, 2018) a preference for interpretability pervades the social sciences. This phenomenon is for example illustrated by the appeal of linear probability models for categorical outcomes which retain the basic model’s ease of interpretation (Horrace and Oaxaca, 2006) or the often encountered practice of assuming ordered categorical outcomes to be continuous (Timmermans, Kuyper, and van der Werf, 2015; Meisenberg, 2015). Put differently, there is a clear trade-off between model-interpretability and model-complexity, and oftentimes weight lies disproportionately on the former. Fortunately, predictive analysis can provide straightforward solutions to both the risk of an in-sample focus, as well as to the problem of model complexity / interpretability.

A predictive approach to model inference

Incorporating a predictive perspective to an empirical analysis is disarmingly simple. The process consists of splitting the available data $\mathcal{D}$ into an estimation set, $\mathcal{D}_{\text{estimation}}$, and an evaluation set $\mathcal{D}_{\text{evaluation}}$ using a split index $\mathcal{K}$ assigning each observation in the data to either of the two sets.⁶ Typically, 20% – 30% of the total data is

⁵The most well-known example of complexity affecting model interpretability is the simple case of a binary instead of continuous outcome. A logit or probit approach can address the binary nature of the outcome, but no longer retains the straightforward interpretability of coefficients estimates which the continuous outcome model had (Breen, Karlson, and Holm, 2018; Mood, 2010). Another common source of complexity is a nested or hierarchical structure in the data, which can warrant the need for varying coefficients at different levels of clustering in the data. Again, the ease of interpretation of the standard linear additive model is lost since conclusions become conditional on group membership (Gelman and Hill, 2006).

⁶Note that the train- and test-set nomenclature is more common in the fields of machine learning and pattern recognition. I use estimation- and evaluation-set to retain similarity to the terminology more often encountered in the social sciences.

Algorithm 1: Predictive Routine

```
for  $i$  in  $1:K$  do  
    Partition data into an estimation set and an evaluation set;  
    Estimate model  $\mathcal{M}$  using the estimation set;  
    Make predictions using the evaluation set and estimated model  $\hat{\mathcal{M}}$ ;  
    Evaluate predictions (against observed outcome or amongst predictions);  
    Store coefficient estimates in  $\hat{\mathcal{M}}$ ;  
end
```

149 used as evaluation set. Dependent on the substantive research question a stratified
150 split can be used to retain a similar structure across both sets. Then, a model $\mathcal{M}$
151 is estimated on $\mathcal{D}_{\text{estimation}}$ and evaluated by making predictions of the outcome for
152 $\mathcal{D}_{\text{evaluation}}$. This process can be repeated K times for different split indices $\mathcal{K}_k$ to
153 ensure robustness to the split used.⁷ Finally, the predictions $\hat{y}_{\mathcal{M}}$ can be compared
154 with actually observed outcomes or with the predicted outcomes of some other model
155 $\bar{\mathcal{M}}$. Note also that parameter estimates from every estimation cycle can be stored
156 to provide bootstrapped standard errors of coefficients.⁸ The predictive routine is
157 summarised in Algorithm 1.

158 Including such a predictive routine to complement standard inference provides a num-
159 ber of useful benefits. Firstly, predictions are a natural fit metric of the model $\mathcal{M}$
160 to assess its ability to structure the outcome of interest given that predictions are on
161 the same scale as the actual outcome variable. Fit can also be evaluated at any level
162 of aggregation, and improves on summary metrics like the AIC and BIC through its
163 more natural interpretation and compatibility to diverse types of outcome, e.g. cate-
164 gorical.⁹ Secondly, using predictions as a fit metric allows the comparison of widely
165 differing models and enables the researcher to diagnose if and how models differ from
166 one another.¹⁰ Third, by varying $\mathcal{D}_{\text{estimation}}$ and $\mathcal{D}_{\text{evaluation}}$ creatively, intuitive model
167 effects can be obtained for any model $\mathcal{M}$. For example, assume we are interested
168 in the effect of sex on some outcome of interest but our model does not allow for
169 the straightforward interpretation of the coefficient on a dummy-coded sex variable

⁷Many will recognize the similarity to cross-validation with the subtle difference that evaluation of predictions is not summarized across cross validation splits but the estimate / evaluation split is redrawn separately for a single split.

⁸A good practice in general and one that improves confidence into estimated coefficients more generally.

⁹Typical in-sample measures like the R^2 are naturally limited by their aggregate nature, but also don't reflect explained variability well in categorical outcome models. Fit metrics tailored for categorical outcomes like the F1-score or entropy score can easily be derived from predictions.

¹⁰A similar intuition is applied in ensemble methods that combine various models that may be better able to fit specific parts of the dataset than others (Baćak and Kennedy, 2019; Van der Laan, Polley, and Hubbard, 2007).

170 – perhaps because our research question required us to estimate an ordinal outcome
 171 model. By fitting the model using $\mathcal{D}_{\text{estimation}}$ and comparing predicted outcomes for
 172 $\mathcal{D}_{\text{evaluation}}$ when changing the observed sex, an intuitive assessment of the impact of
 173 sex is obtained, which is similar to the interpretability of the basic linear model. In
 174 other words, predictive analysis is amenable to the *do*-style interventional analysis
 175 which is becoming increasingly popular in answering causal questions (Pearl, 2009).
 176 More complex model effects or group differences are also easy to evaluate, for ex-
 177 ample if we suspect the effect of sex to manifest itself throughout the coefficients in
 178 the functional form rather than through a level difference, as would be implied by a
 179 dummy-coded variable. In this case, we could estimate the model on subsets of the
 180 estimation set, e.g. one model fit to females and one to males, and compare the pre-
 181 dicted outcomes for the males or females in the estimation set using both models.¹¹
 182 Both these approaches are schematically visualized in Figure 1.

183 Clearly, a predictive routine provides simple solutions to the problems of not knowing
 184 the external performance of a model, or generating intuitive effect sizes in complex
 185 models. By nature, the ability to fit unseen data is investigated, and any fit metric
 186 of choice can be evaluated using the predicted and actual observations, and on any
 187 level of disaggregation, thus providing a more natural and detailed alternative to
 188 conventional in-sample fit metrics. In addition, intuitive insights can be generated
 189 for models which might otherwise have been difficult to interpret, and prediction
 190 lends itself naturally to evaluating more complex mechanisms and comparing their
 191 ability to structure the outcome of interest. The price of these additional insights
 192 is non-existent, especially in a day and age of computational riches. What is more:
 193 the in-sample results which would otherwise have been analysed remain completely
 194 unchanged, with the added benefit that through the predictive routine bootstrapped
 195 standard errors of the coefficient estimates are automatically obtained.

196 Using prediction as an analytical tool is by no means novel in empirical research. In
 197 fact, it is already used to generate understanding into complex models. The typical
 198 example is calculating predicted probabilities for categorical outcome models, where
 199 the impact of changing some interest variable k is evaluated at representative values
 200 of the rest of the variables or summarised over the sample (Breen et al., 2018; Long
 201 and Freese, 2006). The `margins` routine in Stata reflects the general appeal of using

¹¹The perspective of varying coefficients by group membership has been popularized within the linear additive framework through decomposition methods like the Oaxaca-Blinder decomposition (Fortin, Lemieux, and Firpo, 2011). The method proposed here closely resembles the same intuition: that group differences in an outcome can occur due to compositional differences – e.g. different distributions of independent variables by group – as well as effect differences – different effects of independent variables by group. Prediction using models fit to specific groups is a more general approach applying this same intuition.

Figure 1. Schematic overview of varying estimation and evaluation set to generate insights into effect sizes. Top schema illustrates estimation of a level difference when varying the grouping variable of interest. Lower schema illustrates estimation of group differences throughout the functional form estimated.

prediction as an interpretative tool to understand changes in conditional means in an estimated model. The proposed approach generalizes this intuition and applies it explicitly out-of-sample. In this way, the observations used to make predictions are per definition representative and the predictions obtained reflect the actual sample of interest, and also the model's actual fit of (subsets of) the sample.¹² The novelty lies in the multi-faceted nature of prediction as a holistic tool to evaluate both model fit, and model effects on an intuitive scale. In other words, predictions as proposed in this paper is a much more flexible and broad technique with much more value at various stages of the empirical workflow than current practices where it is relegated to niche empirical settings.

To conclude this section, it is informative to briefly discuss the puzzling fact that prediction as envisioned in this paper has been so under-utilized in the social sci-

¹²Note that the process of setting all observables to representative values needn't lead to a representative observation, or one for which the model performs particularly well in terms of fit. Making predictions addresses these issues.

ences. The acceptance of prediction is often hampered by the (philosophical) debate juxtaposing explanation – the typical mode of inquiry in the social sciences – with prediction (Watts, 2014). Much of this dichotomy stems from a narrow view of prediction, where the goal is to maximize predictive performance at all costs (Shmueli, 2010). As this paper illustrates, prediction needn’t be viewed in this manner at all and the explanatory models typically estimated remain exactly the same. More generally, there is very little reason to dogmatically pick either of the two approaches, as prediction as understood in this paper is clearly designed to *complement* typical explanatory inference. Hopefully, this paper can help address this false dichotomy and pave the way for the inclusion of prediction into the social scientist’s toolkit.

Empirical application: Teacher Bias in Tracking

To illustrate how prediction can complement empirical research, I use the case study of teacher bias. Teacher bias is an important topic within educational research and the life course more generally. A teacher’s evaluation of a student determines the attention and type of didactic content offered to a student (Brophy and Good, 1970), and becomes even more relevant with respect to more comprehensive decision-making, like having a student repeat a year. Especially in formally tracked educational systems – where students are assigned to follow a different level of didactic content by their teachers – bias in evaluations can have a profound effect on a student’s life course. However, teacher bias in tracking can be an equally important – and potentially less scrutinized – mechanism in informally tracked systems like that in the United States (Ansalone, 2010; Geven, Batruch, van de Werfhorst, et al., 2018).

Here, I study teacher bias in tracking decisions in The Netherlands. Tracking in The Netherlands is an extremely consequential moment in the life course: when students in the Dutch system transition from primary to secondary education – around age 12 – they get tracked into separate levels of further education. These levels differ in duration and the didactic material taught, and are usually offered at different schools. The track a student is assigned to further dictates a student’s access to tertiary education like university (Checchi and Flabbi, 2006; Bol and Van de Werfhorst, 2013). Importantly, teachers have a crucial say in assigning students to track levels, making the risk of teacher bias to educational equity considerable. In addition, the mechanism of teacher bias poses various modelling complexities often encountered in the social sciences, making it an excellent case to implement a predictive analysis. As is discussed below, this has led to various simplifying assumptions in modelling.

Teacher bias is a frequently studied theme in educational research. Most early work on teachers' ability to accurately evaluate students originated from social psychology and addressed both the overall accuracy, as well as systemic biases in teachers' evaluations of students. Teachers are found to approximate the skill of students well and bias relating to observable student characteristics like gender and race tend to be minor (Jussim and Harber, 2005; Südkamp, Kaiser, and Möller, 2012). However, most of this work makes use of experimental designs – usually in a laboratory setting – where two groups with an equal level of ability differ on other observables (Ready and Wright, 2011). In reality, this so-called 'unconditional neutrality' does not hold and different demographic groups tend to differ in ability, often already at an early age (Ferguson, 2003; Lee and Burkam, 2002). This has led to a second literature that uses observational data to identify systemic biases in teacher evaluations. The observational literature similarly shows that teachers are generally accurate in assigning ability levels to students (Sorhagen, 2013; Tach and Farkas, 2006; Watson, Rubie-Davies, Meissel, Peterson, Flint, Garrett, and McDonald, 2019) although an increasing number of cases has emerged where minority groups in specific regional contexts receive unfavourable bias (Froehlich, Martiny, Deaux, and Mok, 2015; Hinnant, O'Brien, and Ghazarian, 2009; Rubie-Davies, Hattie, and Hamilton, 2006; Rubie-Davies, 2017).

The typical methodological approach throughout the literature is to use an objective measure of ability – often a testing scheme – and compare this measure to teachers' subjective reports pre-dating the test. The typical functional form is as follows:

$$y_i = \beta_0 + \beta_1 x_{\text{test},i} + \beta_2 x_{\text{bias},i} + \epsilon_i \quad (1)$$

Where interest lies in the coefficient β_2 indicating whether students with certain characteristics are evaluated more or less positive net of ability. Note that interpretation is straightforward and the coefficient of interest indicates how a unit increase in the bias factor – often a dummy – affects the ability measure used – often some point scale.

The functional form in (1) pools teachers together and evaluates bias from a system-perspective. However, low systemic biases does not mean that some (or all) teachers aren't highly biased: just that they aren't in a consistent way (Alexander, Entwisle, and Thompson, 1987; McKown and Weinstein, 2008; van den Bergh, Denessen, Hornstra, Voeten, and Holland, 2010; Timmermans, Kuyper, and van der Werf, 2015; Glock, Krolak-Schwerdt, and Pit-ten Cate, 2015). For example, research has shown that institutional characteristics of schools can foster particular perspectives amongst

teachers on the role and optimal functioning of an educational system. Factors which can then manifest themselves in differences in bias at the institutional level (Marcenaro-Gutierrez and Vignoles, 2015; Dronkers, Van Erp, Robijn, and Roeleveld, 1998; Kauppinen, 2008). As a result, the school is often proposed as a more appropriate level to study teacher bias, since interactions amongst teachers within a school are frequent, but often limited towards teachers from other schools (Geven et al., 2018). In other words, without explicit feedback loops there is a risk of group think and drift in evaluation patterns.¹³ In modelling terms, this complexity warrants that some or all of the coefficients in (1) are allowed to differ at either the school or even teacher level:

$$y_i^{[j]} = \beta_0^{[j]} + \beta_1^{[j]} x_{\text{test},i} + \beta_2^{[j]} x_{\text{bias},i} + \epsilon_i \quad (2)$$

Introducing heterogeneity through a multilevel structure complicates interpretation of the coefficients estimates. Instead of simply evaluating the estimate $\hat{\beta}_2$, its heterogeneity across schools would have to be assessed. In case multiple coefficients are allowed to vary, these have to be assessed in tandem with one another (Gelman and Hill, 2006; Rüttenauer and Ludwig, 2020). To avoid this complexity in interpretation, empirical work often only allows for heterogeneity in the intercept and retains the population-level estimate of $\hat{\beta}_2$ (Timmermans et al., 2013; Driessen et al., 2007; Marcenaro-Gutierrez and Vignoles, 2015) (cf. (Timmermans et al., 2015)). In this type of setup, the estimated variance of $\hat{\beta}_0^{[j]}$ is typically used as a substantive estimate of school-level heterogeneity.¹⁴

Much of the existing literature studies teacher bias in relatively low-stakes settings. Usually by simply asking a teacher his/her evaluation of a student, often for a specific didactic domain. Teachers might be tempted to suppress biases and, more importantly, there is less room for non-ability related biases to creep in given the direct nature of the question asked. Indeed, it has been found that biases become more pronounced in comprehensive evaluations like deciding a student's track level or retaining a student for an extra year of education (Burkam, LoGerfo, Ready, and Lee, 2007; Tach and Farkas, 2006; Glock, Krolak-Schwerdt, Klapproth, and Böhmer, 2013; Glock and Krolak-Schwerdt, 2013; Lucas, 2001). This literature suggests that teachers incorporate a broader set of characteristics – e.g. home environment, behaviour and

¹³Note that such drift in evaluations can be assumed to be particularly relevant in countries where residential segregation is strong, and student compositions differ strongly, as teachers are less exposed to the full ability distribution of students.

¹⁴Research tends to find between 10%-25% of the variance between a teacher's evaluation and objective ability to be attributable to the teacher or school level, indicating considerable school-level variation (Geven et al., 2018). Note however that comparing proportions of variance across contexts is problematic, because they fundamentally depend on the quality of the testing measure used to proxy for ability.

Figure 2. Example of latent variable approach mapping continuous outcome y^* to five categories using cut-off points c_k , $k = 1, \dots, 4$.

perceived motivation – as higher stakes decisions are made (Becker, 2013; Klapproth, 2018; Böhmer, Gräsel, Krolak-Schwerdt, Hörstermann, and Glock, 2017; Boone, Thys, Van Avermaet, and Van Houtte, 2018). From a modelling perspective, this typically introduces additional complexity since comprehensive decision-making often includes categorical rather than continuous outcomes, in particular in tracking decisions.¹⁵

Modelling the categorical nature of the outcome variable is straightforward, for instance using the ordered probit model (Maddala, 1986; Long and Freese, 2006):

$$y^* = X\beta + \epsilon$$

$$y = \begin{cases} 1 & \text{if } y^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < y^* \leq \mu_2 \\ \vdots & \\ K & \text{if } y^* > \mu_{K-1} \end{cases} \quad (3)$$

Here, the standard linear additive functional form is preserved into a continuous latent variable y^* which is mapped to the categorical outcome space by defining domains of y^* corresponding to each categorical outcome (Figure 2). Random effects can be added in an analogous matter to the standard linear additive model, although they require integrating out the additional heterogeneity in the likelihood due to the non-linearity of the model (Christensen, 2015; Cameron and Trivedi, 2005). The categorical outcome model – arguably even more so than a hierarchical model – complicates interpretation considerably (Breen et al., 2018). The effect of the bias coefficient β_2 depends crucially on a student’s other characteristics because of the unequal spacing of each category’s interval.

The combined complexity in interpretation generated by both the ordered outcome

¹⁵More generally, many have questioned whether continuous ability-related scores should actually be viewed as interval outcomes as well (Mare, 1980). In particular whether the equal-interval assumption is valid (Domingue, 2014).

and hierarchical structure means that in practice the outcome variable $y \in [1, \dots, K]$ is almost exclusively assumed to be an interval variable and treated as a standard continuous outcome (Timmermans et al., 2015; Lek, 2020; Driessen et al., 2008). This approach of assuming ordinal variables to be continuous is often encountered throughout the social sciences (Meisenberg, 2015; Domingue et al., 2021), but can lead to biased inference whenever the equal-interval assumption between observed tracks does not hold Domingue (2014). This bias can become particularly problematic when interest variables tend to cluster at the extremes of the outcome space. As will be outlined below, many such simplifying assumptions have been routinely made in the Dutch context as well.

The Dutch context

The high stakes nature of the Dutch tracking system, and the accompanying risk teacher bias poses could not have been conveyed any better than in the Dutch Education Inspectorate’s 2016 report, stating that generally, the Dutch system was of a high level, “[but] at the same time [the Inspectorate] finds that not all students have the chance to obtain the level of education their ability prescribes” (Onderwijsinspectie, 2016). The report goes on to mention that these differences in “chances” – the ability to reach one’s potential – have always been unequal across demographic groups, but that this inequality was now on the rise. For example, Figure 3 shows how students with parents who enjoyed high educational attainment increasingly get tracked into higher ability levels than their objective performance would prescribe. The Inspectorate goes on to provide the telling example of two hypothetical students, both at the end of primary school and of similar overall ability. One gets tracked into a slightly higher level than the other: a relatively small difference at face value, but which means that both students will enter different schools with different incentive structures, didactic environments and expectations. When the two hypothetical student leave secondary education, both have ended up in different educational strata, with different skill-sets and access to educational opportunities (Onderwijsinspectie, 2016). A self-reinforcing effect of small but early educational selection which has been documented more generally (Lüdemann and Schwerdt, 2013; Checchi and Flabbi, 2006; Matthewes, 2021), and in The Netherlands specifically (Tolsma and Wolbers, 2010; Van de Werfhorst and Mijs, 2010; Bol, Witschge, van de Werfhorst, and Dronkers, 2014).¹⁶

Importantly, there are a number of structural features of the Dutch system that exac-

¹⁶Although students are allowed to switch tracks during secondary education, evidence shows that students tend to stay within the tracks they are assigned to (Tieben and Wolbers, 2010).

Figure 3. Vertical axis indicates the difference between the definitive track level assigned to a student and the track level as suggested by the student's objective performance on standardised tests for the period 2009-2017. Source: CBS Microdata (calculations by Onderwijsinspectie).

erbate the downstream risks of tracking, leading international bodies like the OECD to criticize the Dutch tracking system as running the risk of leading to downward spirals and the waste of talent (Field, Kuczera, and Pont, 2007). First, there are five educational tracks and seven educational certifications to be obtained, ranging from practical / vocational education to preparatory academic education. This is more than most countries with formal tracking systems. The moment of tracking is also unusually early in The Netherlands: around the age of 12. Only Germany, Austria and Switzerland have similarly early tracking moments (Prokic-Breuer and Dronkers, 2012).¹⁷ Second, the Dutch system is characterized by rigorous, between-school tracking rather than within-school tracking. This means that students in The Netherlands follow a specific didactic program and have limited interaction across tracks which has been showed to further pronounce the negative consequences of tracking (Betts, 2011; Houtte and Stevens, 2009; Chmielewski, 2014; Bol et al., 2014).¹⁸ Third, the role of individual teachers in tracking is considerable. Until 2015, track assignment was primarily governed by a central testing scheme, the school leaver's test, but

¹⁷France, England and Italy do not track until the ages 14-16 (Hillmert and Jacob, 2010) and most Scandinavian countries and the United States do not have formalized tracking schemes at all (van der Werfhorst et al., 2015) – early tracking is even prohibited by law in Sweden (Rudolphi and Erikson, 2016).

¹⁸Schools are allowed to offer multiple tracks, but students follow their own set program within each school. In some cases, schools may offer multi-track classes for the first one or two years of secondary education but this is at their own discretion and the number of schools offering multiple tracks has been steadily declining (Onderwijsinspectie, 2019).

Figure 4. Educational structure and tracking moment in The Netherlands. The pre-2015 situation is illustrated with dashed lines, the post-2015 law change is illustrated with solid lines. The five track levels are depicted together with the formal qualifications on offer at the end of secondary education (top right per track level). Subsequent access to tertiary education is illustrated using the double-edged squares at the bottom.

teachers were allowed to propose an alternative assignment in case they deemed necessary. This possibility of course already allows for some degree of biased tracking (Timmermans et al., 2015), but as of 2015, the importance of these two criteria has been reversed. The teacher’s evaluation of a student now has precedent, with the test acting as a second opinion. The date of the central test was also pushed back until after the start of the official enrolment process into secondary education, making it even less consequential (Smeets, Kuijk, and Driessen, 2014) (see Figure 4).¹⁹ Finally, the Dutch system permits considerable liberty for schools to follow their own didactic philosophy (Patrinos, 2011; Scheerens, Ehren, Slegers, and de Leeuw, 2012; Ritzen, Van Dommelen, and De Vijlder, 1997). Coupled with the fact that teachers tend to discuss tracking decisions amongst one another further exacerbates the risk of ‘group think’ mentioned earlier, where schools might collectively differ in terms of their evaluation of students due to drift (Geven et al., 2018).

Unsurprisingly, there is a wealth of empirical work on the Dutch tracking system and teacher bias in particular (Luyten and Bosker, 2004; Timmermans et al., 2015; Lek, 2020; van der Werfhorst et al., 2015). Typically, demographic bias factors on the student level have been studied, like sex, SES and ethnicity. Consistent evidence is found of positive biases towards girls relative to boys in assigned track levels net of ability (Lek, 2020; Timmermans et al., 2015).²⁰ General bias against ethnic minorities tends to be small but has developed from positive bias to negative bias over the course of the early 2000s and 2010s (Jungbluth, 2003; de Boer et al., 2010; Timmermans et al., 2013; Roeleveld et al., 2011; Geven et al., 2018; Driessen et al., 2007). There is considerable evidence that high-SES students are tracked into higher track levels net of ability and it is hypothesised that parents from high-SES families tend to put more pressure on teachers (Dronkers, Van Erp, Robijn, and Roeleveld, 1998; Onderwijsinspectie, 2016), while low-SES families are less likely to attempt to influence the tracking process (Hillmert and Jacob, 2010). Some suggest that an increase in average track level over time can be largely attributed to increased parental involvement (Timmermans and Rubie-Davies, 2018).

The existing work on The Netherlands has three important limitations, some of which are general to the study of teacher bias in tracking. First, the existing literature exclusively concerns the pre-2015 period, prior to teachers becoming the main mechanism

¹⁹In 2019, the Secretary of Education announced plans to return the school leaver’s test to an earlier date to coincide with the timing of a teacher’s advice (Note to parliament, June 21st 2019 – reference number: 7850864).

²⁰Sex bias has been suggested to be driven by same-sex bias between teachers and students although evidence dismisses this hypothesis (Neugebauer et al., 2011; Coenen and Van Klaveren, 2016).

through which tracking is done. Second, most work focuses on system-level biases and might underestimate the true size of teacher bias when evaluated at a lower level of aggregation – e.g. at the school-level. An important exception to the system-level focus is the work by Timmermans et al. (2015), where teacher-level variability is explicitly modelled. Tellingly, the authors conclude that evaluating biases on the system-level might be insufficient to grasp the full extent of teacher bias in The Netherlands. Third and finally, the existing work almost entirely ignores the categorical nature of the outcome variable and assumes assigned track levels to be an interval variable or collapses the outcomes into binary variables (Timmermans et al., 2015; Lek, 2020; Driessen et al., 2008, 2007; Dronkers et al., 1998; Brandsma and Doolaard, 1999).²¹

Data & Descriptive statistics

In order to study teacher bias in The Netherlands, I use data on 746 primary schools covering nearly 50,000 students who finished primary school in either 2018 or 2019. All student-level information comes from the CITO LVS, which is a nationally implemented Pupil Tracking System (PTS) (van de Werfhorst, 2021). The dataset has been fully anonymised and was developed for research purposes in partnership with the Mirror Foundation, a research institute in The Netherlands. The data concerns about 15% of the total population. Representativity of the data is discussed in more detail below.

Assigned track level and ability test scores Each student in the dataset is assigned one of five track levels (Figure 4) which is the outcome of interest. As a measure of observed ability of the student, I use test scores from the PTS with standardized scores (de Wijs, Kamphuis, Kleintjes, and Tomesen, 2010; Feenstra, Kamphuis, Kleintjes, and Krom, 2010; Janssen, Verhelst, Engelen, and Scheltens, 2010). These tests are meant to evaluate both new and past material across three didactic areas: reading, maths, and language. The reading test evaluates students' ability to understand written texts. The math test evaluates students' ability to solve problems concerning addition, subtraction, multiplication and division. The language tests evaluate students' ability to correctly spell words. Tests are taken twice yearly and are uploaded to the PTS by the schools, as is mandated by law.²²

²¹To the best of this author's knowledge, the only case where the ordinal outcome of track decisions has been modelled is a study of the Amsterdam area where neighbourhood effects are evaluated (Kuyvenhoven and Boterman, 2020).

²²All tests are developed and psychometrically evaluated by CITO. Each test score is represented by a raw score, reflecting the student's performance on the test, which is transformed to a percentile score by CITO based on representative samples of the Dutch population. The norming by CITO ensures accurate assignment of test results to relatively broad percentile bins – specifically the 0%-

Figure 5. Left panel shows density plots of test scores for each of the three didactic areas included in the sample. For every student, the average percentile test score across all tests made in the final two grades is shown. Right panel shows a box-plot indicating average percentile score per didactic area for each track level. A random sample of 2.5% of the dataset is used to illustrate the relationship further through scatter-plots.

I average percentile tests for each didactic domain during the final two grades prior to the tracking moment to use as proxies for ability. Usually, three tests are used per student per didactic area since the tracking moment in the final grade happens half-way into the year. The distribution of the three averaged test scores is given in the left panel of Figure 5, and the association between average test scores and assigned track levels is given in the right panel. As can be seen, there is a strong positive relationship between higher test scores and higher assigned track levels, as should be expected.

Bias factors As bias factors of interest, I use a student’s sex and level of parental education. The former is readily available from the PTS, whereas schools are obliged to collect the latter as part of the government’s funding scheme for primary education. So-called ‘student weights’ indicate whether a student might have an expected learning disadvantage when starting education and the Ministry of Education allocates additional funding to schools with more weighted students (Ladd and Fiske, 2011).

10%, 10%-25%, 25%-50%, 50%-75%, and 75%-100% bins (de Wijs et al., 2010; Feenstra et al., 2010; Janssen et al., 2010). Since the dataset used for this paper is many times larger than the sample used by CITO, I further norm the test scores within each bin such that a uniform distribution is obtained (Engzell, Frey, and Verhagen, 2020).

Table 1. Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Max
<i>Outcome</i>					
Assigned Track Level	48,946				
... Vocat. Practical	10,642	21.7%			
... Vocat. Theoretical	14,444	29.5%			
... Vocat. / Applied Science	8,447	17.3%			
... Applied Science / Academic	5358	10.9%			
... Academic	10,055	20.5%			
<i>Ability measures</i>					
Reading	48,946	0.525	0.239	0	1
Maths	48,946	0.52	0.278	0.001	1
Language	48,946	0.521	0.27	0.001	1
<i>Pupil characteristics</i>					
Parental Education: Low	48,946	0.096	0.295	0	1
Sex: Girl	48,946	0.507	0.5	0	1
Cohort: 2017-2018	48,946	0.491	0.5	0	1
<i>School characteristics</i>					
School Denomination	48,946				
... Christian	29,976	61.2%			
... Public	15,359	31.4%			
... Other	3,611	7.4%			
School Urbanity	48,946				
... Lowest	13,207	27%			
... Low	15,869	32.4%			
... Medium	10,598	21.7%			
... High	6,988	14.3%			
... Very High	2,284	4.7%			
School Disadvantage	48,946				
... Low	16,808	34.3%			
... Medium	17,150	35%			
... High	14,988	30.6%			

All other funding is otherwise allocated on a per-student basis and there is no meaningful private / public distinction in The Netherlands. At the student level, there are three weight types: a ‘high’ student weight indicates no expected disadvantage and reflects those students who have at least one parent with a degree above lower secondary education; a ‘low’ student weight indicates neither parent has a degree above lower secondary education but both parents do have a degree above primary education; a ‘lowest’ student weight reflects the case where neither parent has enjoyed lower secondary education. I code this variable into two levels, reflecting high and low parental education.²³ I also add a dummy-coded control variable indicating whether the student graduated in 2018 or in 2019.

²³Robustness checks including the variable as a three level categorical does not lead to meaningful differences in the substantive results.

School characteristics Finally, I include various school level characteristics into the model. First, a school-level version of the weight indicator used for school funding.²⁴ Second, I include the denomination of the school, coded as ‘Public’, ‘Christian’ (consisting of ‘Roman-Catholic’, ‘Protestant’, ‘Reformist’) and ‘Other’ (e.g. Dalton, Montessori, Jena). Finally, I include a measure of the urbanicity of the neighbourhood of the school. Most students go to a primary school in their direct neighbourhood – the average distance between a student’s home and school is 300 meters. Descriptive statistics for the dataset can be found in Table 1.

Representativity The sample is constructed by the Mirror Foundation, a Dutch non-profit research institute and has been used for nationally representative research before (Engzell, Frey, and Verhagen, 2020). As the main goal of the nationally implemented PTS is to allow schools to track the development of their students, there are various tools accessible to schools to analyse student test scores, set goals and make learning plans for individual students. Access to the sample at hand is mediated through schools’ adoption of such a visualization tool. There is thus possibility for non-random selection into the sample. There are various reasons why such selection might be limited.

All schools are publicly funded on a per-student basis with relatively more disadvantaged schools getting additional funding - private funds play no role in the Dutch system. There is thus little reason to expect selectivity in terms of resource differences between schools. Furthermore, the upload and management process of the test data is mandated by law and there is minimal upfront investment on the part of the school to use such tools. Importantly, Figure 6 shows that the sample of schools aligns closely with the total population on most observables. In particular, the distribution of low parental education students is similar to the population, as is the indicator for school-level disadvantage. There is a slight over-representation of smaller schools (101-200 students) and an under-representation of very large schools (500+), but the distribution is similar in terms of urbanicity and denomination. The distribution of track outcomes also closely follows that of the total population.

Analytical Strategy To study teacher bias in The Netherlands, while addressing possible heterogeneity at the school level, as well as the ordered nature of the outcome, the following functional form is appropriate:

²⁴This measure was introduced in 2019 to replace the student level measure and is a school-level composite of parents’ education, country of origin, and the duration of the mothers’ residence in the country, and whether parents were being assisted by the government for debt restructuring (Posthumus, Bakker, van der Laan, de Mooij, Scholtus, Tepic, van den Tillaart, and de Vette, 2016; Posthumus, Scholtus, and Walhout, 2019)

Figure 6. Sample representativity relative to the population. Each graph compares the sample used (blue) to the total population in The Netherlands (red). Sources: CBS Statline, DUO, and Onderwijsinspectie (CBS (accessed 7th December 2020); Dienst Uitvoering Onderwijs (accessed 10th December 2020); Inspectie van het Onderwijs (accessed 7th December 2020)).

$$y_i^* = \beta_0^{[j]} + \beta_1^{[j]} x_{\text{Maths},i} + \beta_2^{[j]} x_{\text{Reading},i} + \beta_3^{[j]} x_{\text{Language},i} + \beta_4^{[j]} x_{\text{Cohort},i} + \beta_5^{[j]} x_{\text{Bias},i} + \epsilon_i \quad (4)$$

where y_i^* is a latent continuous variable mapped to 5 distinct levels, reflecting the five available track levels. Ability is controlled for through the three ability scores, and cohort differences through a dummy variable. The x_{Bias} variable reflects either the dummy indicator for low parental education or a student's sex, or school-level information. Possible between-school differences are added through the superscript $[j]$. The innovation of (4) relative to existing work is to explicitly address the ordered nature of the outcome, and allow multiple coefficients to differ by school. For illustrative purposes, I will estimate (4) for various levels of complexity, i.e. by assuming the outcome to be continuous as well as ordinal and varying the number of random effects in the model.

Results from conventional analysis

Regression results when estimating model (4) including both sex and parental education are presented in Table 2. I estimate the model for four levels of model complexity: the first column shows coefficient estimates for a standard linear additive model where the outcome is assumed to be an interval variable and without school-level heterogeneity. The results show that performance in maths has the strongest association with a student's assigned track level, followed by reading and then language. All test scores have the expected sign. Additionally, girls get assigned about 0.15 track levels higher than boys, assuming the interval model is correctly specified. Conversely, low parental education students get about 0.17 track levels lower than their peers. The second column estimates the same model but with random intercepts added at the school level. Most coefficients remain unchanged although the effect associated with a student's parental education almost halves. This implies that some of the signal found in the model without random intercepts was due to school-level differences and imply that low parental education students cluster in schools where students tend to get assigned to lower track levels net of observed ability. The proportion of variance explained at the school level is about 16%.

Columns three and four show results when addressing the categorical nature of the outcome through the ordered probit model. Significance, sign and relative magnitude of most coefficients are similar to the interval-model, although the effect of parental education seems considerably larger in the categorical models, even when including random intercepts (column 4). As mentioned, there is added complexity in interpretation, since the cut-off points determining the mapping of latent variable to categorical outcome is of a different magnitude between the two models (cut-offs are also shown in Table 2). For example, the width of the second interval – mapping the outcome to the second track level – is almost twice as large as the next interval, and three times as large as the fourth. This means a fixed increase has a different effect dependent on the distributional location of an observed student.

To get an indication whether the difference in effect sizes are substantial, I normalize coefficient estimates across models. In the left panel of Figure 7 this is illustrated for the models in columns 2 and 4, where each coefficient is normalized by that model's coefficient on maths. As can be seen, most coefficients are of similar size across both the interval- and categorical-outcome models except for the coefficient on parental education which nearly doubles. Such substantial differences reflect potential misspecification in assuming the outcome variable to be continuous, possibly compounded by the fact that low parental education students tend to cluster at the

Table 2. Main results

	Interval	Interval	Categorical	Categorical
Intercept	−0.07*** (0.01)	−0.08*** (0.01)	- -	- -
Cohort: 2018-2019	0.02* (0.01)	0.02*** (0.01)	0.06* (0.03)	0.05*** (0.01)
Ability: Reading	1.74*** (0.03)	1.67*** (0.02)	5.70*** (0.11)	3.61*** (0.04)
Ability: Maths	3.12*** (0.03)	3.08*** (0.01)	9.88*** (0.13)	6.35*** (0.04)
Ability: Language	0.48*** (0.03)	0.57*** (0.02)	1.63*** (0.08)	1.27*** (0.03)
Student: Girl	0.15*** (0.01)	0.16*** (0.01)	0.54*** (0.02)	0.36*** (0.01)
Student: Low Parental Educ.	−0.17*** (0.02)	−0.09*** (0.01)	−0.94*** (0.05)	−0.36*** (0.02)
1 2			5.32*** (0.08)	3.52*** (0.03)
2 3			9.44*** (0.12)	6.22*** (0.04)
3 4			11.73*** (0.14)	7.71*** (0.05)
4 5			13.27*** (0.16)	8.73*** (0.05)
Random intercepts (school)	No	Yes	No	Yes
Clustered standard errors	Yes	-	Yes	-
Adj. R ²	0.78			
Num. obs.	48,946	48,946	48,946	48,946
N Schools	746	746	746	746
AIC		93,779.46	79,117.62	71,616.79
Var: School (Intercept)		0.07		0.39
Var: Residual		0.38		1.00
OOS: Accuracy	57.1%	60.7%	65.1%	68.5%

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Figure 7. Left panel shows coefficient estimates when estimating track levels as an interval variable (red) and as a categorical variable (blue) in a model including random effects on the school level, ability scores and student demographics. All coefficients are normalised per model by the coefficient on maths for comparability. The ratio between the size of the coefficient of the categorical and interval model is depicted on the right. The right panel indicates the out-of-sample predictive performance of various models estimating track levels as an interval outcome (red) or a categorical outcome (blue). Four models are estimated on differing sets of explanatory variables. The first row includes ability scores only. The second includes ability and parental education. The third includes ability and sex. The fourth includes ability and random intercepts. Prediction is done using 250 random splits of the data into a 75% train and 25% test split, stratified at the school level.

lower side of the track distribution. To complement this analysis of the modelling results, I include a predictive routine next.

Complementary results from predictive analysis

As a first step, I estimate the out-of-sample predictive performance of both the interval- as well as the categorical-model when including various sets of explanatory variables (Figure 7, right panel). Overall accuracy of the interval models lies between 56-61%, whereas accuracy is about 8% better for the categorical model across specifications implying considerably better fit. For the interval model, continuous outcomes are rounded to the nearest integer to generate concrete track level predictions for each student when using the interval-model.²⁵

The lack of predictive improvement when including student level bias features to the

²⁵Noteworthy is the discrepancy in actual predictive power of the interval-models, compared to that suggested by the R^2 in Table 2 when evaluating actually predicted track levels. This discrepancy is accentuated by the fact that the interval model frequently makes predictions smaller than 0.5, which I assign to the lowest track thus inflating accuracy.

model is striking. Accuracy improves by a mere 0.1% when adding parental education and 0.4% when including sex compared to an ability-only model. Conversely, addressing for differences in tracking between schools improves the predictive power by more than 3.5%. This is noteworthy because much of the existing research on teacher bias focuses on student-level characteristics, even though they add little in terms of making more accurate predictions and thus improving model fit. That the impact of schools has received less attention is likely a consequence of the fact that school variability is usually reported in terms of the estimated variance coefficient of school-level random effects, which are not typically considered in terms of their substantive effect size and contribution to fit.²⁶ This is quite natural in predictive analysis.

A benefit of OOS-predictions is that local model fit, as well as discrepancies between models can be easily diagnosed by comparing predicted outcomes with observed outcomes. This is illustrated in Figure 8 where predictions are visualized for both the interval- and categorical-model using only ability scores as explanatory variables. These figures immediately show why the interval model is misspecified and the clear non-linearity apparent in the data. Forcing an equidistant effect of the explanatory variables onto the outcome leads to poor model fit, in particular for the top track. These graphs further indicate that the lowest track levels tend to be more consistently modelled with students often correctly assigned to the lowest two tracks. The top three tracks have considerably more variability and similar students – in terms of their fixed effects – are often assigned to either of the top three tracks. This holds for both the interval- as well as the categorical-model.

School level heterogeneity One of the key findings from the initial predictive analysis was the considerable improvement in model fit when including random intercepts on the school level. When including school-level characteristics instead, these variables poorly account for the increase in predictive accuracy (Figure 9, left-hand panel) although most coefficients are statistically significant (Table 3). The school-level disadvantage indicator diminishes the student-level association with low parental education (column 2), in line with the earlier finding that part of the parental education effect resides at the school level. In terms of predictive accuracy, school-level disadvantage only improves on the baseline model by 0.7%. The school’s urbanicity and denomination do not improve predictions at all. At the same time, adding

²⁶Typically the Intra-Class Correlation is reported, indicating the relative size of the school-level variance with respect to the residual variance. This metric is dependent on the signal-to-noise ratio in the rest of the independent variables – the measure is also inappropriate in the case of the categorical outcome model (Breen et al., 2018).

Figure 8. Left panel shows predictions of a model estimated when assuming track levels to be an interval variable, including ability scores as explanatory variables. Right panel shows predictions of a model estimated when assuming track levels to be a categorical variable, again using ability scores as explanatory variables. Model predictions are represented by the horizontal axis, while actual track levels are represented by the vertical axis. Vertical dashed lines indicate implied cut-off points mapping numerical predictions to actual track levels. To improve readability, predictions are scattered vertically. Predictions below 0.5 in the interval-model have been assigned the lowest track.

Figure 9. Left panel shows the predictive performance of various categorical outcome models. Bottom two models include student level characteristics (purple). Middle four models include student and selected school level characteristics (red). Top five models include random effects ranging from intercepts only, to random intercepts and selected random coefficients, to a full random effects model. Right panel shows estimated random intercepts (in dark blue) and slopes (in dark red) of a model including all student level demographics and school-level random effects in the sex coefficient (top) and the parental education coefficient (bottom).

Table 3. School characteristics

	Categorical	Categorical	Categorical	Categorical
Cohort: 2018-2019	0.06** (0.02)	0.06** (0.02)	0.06** (0.02)	0.06** (0.02)
Ability: Reading	5.68*** (0.07)	5.68*** (0.07)	5.70*** (0.07)	5.66*** (0.07)
Ability: Maths	9.91*** (0.07)	9.96*** (0.07)	9.90*** (0.07)	10.02*** (0.07)
Ability: Language	1.68*** (0.05)	1.80*** (0.05)	1.64*** (0.05)	1.85*** (0.05)
Student: Girl	0.55*** (0.02)	0.56*** (0.02)	0.54*** (0.02)	0.56*** (0.02)
Student: Low Parental Educ.	-0.95*** (0.04)	-0.64*** (0.04)	-0.99*** (0.04)	-0.68*** (0.04)
School: Other denomination	0.40*** (0.04)			0.24*** (0.04)
School: Public denomination	0.35*** (0.02)			0.28*** (0.02)
School: Medium disadvantage		-0.71*** (0.02)		-0.64*** (0.02)
School: High disadvantage		-1.13*** (0.03)		-1.16*** (0.03)
School: Low urbanity			-0.19*** (0.03)	-0.14*** (0.03)
School: Medium urbanity			-0.18*** (0.03)	-0.28*** (0.03)
School: High urbanity			-0.13*** (0.03)	-0.22*** (0.03)
School: Very high urbanity			-0.56*** (0.05)	-0.55*** (0.05)
1 2	5.48*** (0.04)	4.76*** (0.04)	5.19*** (0.04)	4.74*** (0.05)
2 3	9.62*** (0.06)	9.02*** (0.06)	9.32*** (0.06)	9.03*** (0.06)
3 4	11.91*** (0.07)	11.38*** (0.07)	11.61*** (0.07)	11.40*** (0.07)
4 5	13.47*** (0.08)	12.97*** (0.08)	13.15*** (0.08)	13.01*** (0.08)
AIC	78817.21	77138.01	78977.94	76785.04
Log Likelihood	-39396.60	-38557.01	-39474.97	-38374.52
Num. obs.	48946	48946	48946	48946
OOS: accuracy	65.3%	66.0%	65.3%	66.0%

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

All coefficients are normalised by the coefficient for maths. Reference group in the full model are schools of Christian denomination in neighbourhoods of very low urbanity with low disadvantage.

even more heterogeneity to the model – by allowing all coefficients in the model to vary at the school-level – does not appreciably improve predictive performance either. This implies that most school heterogeneity comes from structural over- or under-evaluation of students rather than differential treatment along specific student characteristics. The right hand panel of Figure 9 further confirms that bias towards girls and students of low parental education are unimodal, whereas there is considerable dispersion of the random intercepts which seem bimodal.

Group Biases Given the complexities required to model teacher bias appropriately, prediction can also be used to generate understanding into substantive effects. When studying e.g. sex bias, the coefficient for sex is difficult to interpret using the categorical outcome model, although it is clear that this model fits the data better.

More so, bias might not manifest itself through a conditional mean difference but instead throughout the mechanism – for example if girls are differentially evaluated for their math scores compared to boys. By applying prediction to different evaluation and estimation sets, as described above, both a level difference as well as differences throughout the functional form can be assessed. I start with assessing the substantive effect of the latter since it most closely resembles the question of bias, i.e. how differences in assessment practices manifest themselves between groups (bottom schema, Figure 1).

For the question of sex bias, two models are estimated by partitioning the estimation set along the grouping variable of interest – sex in this case. One model is estimated using only data on girls and another using data on boys. A similar partition is made for the evaluation set: one only including boys and one only girls. By making predictions for either evaluation set using both estimated models, we effectively obtain estimates of the question what would happen if boys were evaluated as if they were girls and vice-versa. Results of this approach are presented in Figure 10 for both student sex as well as parental education.

In the top left panel, track levels are predicted for boys using a model estimated on boys (dark blue) and using a model estimated on girls (dark red). The average difference between the two approaches is about 0.15, implying that if boys get tracked as if they were girls they would on average be assigned 0.15 track levels higher. The top right panel shows a similar reversed result. Note that the implied bias of 0.15 is about the same as was found in the interval model, which might be expected since the distribution of girls and boys is very similar across observables which means that the coefficient is less affected by non-linearities in the process of mapping the latent variable to the outcome space.

The bottom two panels of Figure 10 show the same approach for parental education. When students of high parental education get tracked as if they were of low parental education (bottom left), they get assigned about 0.23 track levels lower than they would otherwise. This is considerably more than the interval model suggested. Conversely, low parental education students would gain about 0.13 track levels when treated as high parental education students (bottom right). The difference in effects stem from the fact that most low parental education students are clustered in the lower two track levels where less movement across tracks is possible. When the set of students of low parental education is resampled such that it resembles the set of high parental education students in terms of their observed ability, the positive effect for low parental education becomes of a similar size to the reverse effect described earlier.

Figure 10. Top two panels show predicted track levels for boys (top left) and girls (top right) when using a model fit to boys (dark red) and a model fit to girls (dark blue). Bottom two panels show predicted track levels for students with high parental education (bottom left) and low parental education (bottom right) when using a model fit to high parental students (red) or low parental students (blue). Horizontal dashed lines indicate average track level amongst the full group. Prediction is done using 250 random splits of the data into a 75% train and 25% test split, stratified at the school level.

For both sources of bias, a more simple routine can be estimated by simply intervening on the observed bias feature and comparing predictions – i.e., changing the observed sex for males to female and vice-versa (top schema, Figure 1). For sex bias, this yields identical results with boys obtaining 0.16 track levels higher when intervening on observed sex, whereas for girls the reverse effect is found. For parental education, intervening on high parental education students leads to a reduction of 0.17 track levels. Intervening on low parental education students leads to an increase of 0.13 track levels. In other words, there seems to be more evidence for different evaluation practices between students of varying parental education than varying sex. Across both routines, bias towards low-SES students is considerably larger than found in the interval model.

Figure 11. Figure shows the difference between the average predicted track levels for a school when using a full random effects model and a population model without any variation on the school-level. Horizontal line indicates no difference between school- and population-model. Prediction is done using 250 random splits of the data into a 75% train and 25% test split, stratified at the school level. Blue dots indicate schools where at least 95% of estimates are above or below zero (30% of schools). Green dots indicate school where at least 95% of estimates are above / below or end at zero (21% of schools). Red dots indicate schools where less than 95% of estimate are above or below zero (49.0% of schools).

Differential evaluation processes across schools The same predictive approach can be applied to school effects as well. By estimating two models on the estimation set – one with and one without school effects – and making predictions on the evaluation set using both models, the impact of school heterogeneity can be estimated. In other words: track predictions are made using a model estimated without any school heterogeneity – the ‘population’ model – and compared to predictions by a model estimated with school heterogeneity – the ‘school-varying’ model. As always, predictions are made for many splits of the data into estimation and evaluation sets.

The results are shown in Figure 11. Comparing predictions using the population model and the school-varying model, results show the impact of school differences on average track levels. Some schools assign students an entire track lower or higher than the average school – as proxied for by the model without school effects – would. Based on frequent resampling, confidence intervals for these effects can be visualised. 30% of the schools consistently deviate from the population model (in blue) with a further 21% of school’s confidence intervals ending at zero (in green). In the most extreme case, the most over-biasing and under-biasing schools differ by almost two entire track levels, implying considerable differences in evaluation practices across different schools.

Discussion & Concluding Remarks

This paper set out to illustrate the use of prediction as an analytical tool for the social sciences. Predictions provide a natural descriptive tool to diagnose model fit, identify misspecification, and compare different models. Additionally, predictions can generate insights into complicated functional forms which might yield model coefficients that are difficult to interpret. Usually, such complexity in interpretation leads researchers to opt for more simplistic models which retain a level of direct interpretability. Unfortunately, this ease of interpretation is meaningless if the functional form is misspecified and can lead to biased results. A case in point is the downward bias found in the estimated effect of low parental education found in this paper's empirical application of prediction when ignoring the categorical nature of track levels. Predictions also illustrated why previously estimated models were misspecified, a richness of detail and understanding into model fit which is not usually accessible in traditional metrics of model fit, but comes natural from a predictive perspective.

Substantively, this paper investigated whether teachers in The Netherlands are biased in their track assignment process towards students based on their sex or parental education. This paper is the first to answer these questions using data after a 2015 law change that made teachers more fundamental to the tracking process. Although broadly in line with work prior to 2015, I find that the size of bias towards low-SES students might have been structurally under-estimated in the past by a lack of appreciation for the categorical nature of the outcome variable. Equally important, this paper's predictive approach shows that even though bias towards student-level observables is present, it is less consequential than the variation in tracking processes between schools, corroborating calls to better study institutional differences. More broadly, this paper once again lays bare the risks of the Dutch educational system's early tracking process to educational equity.

This paper then has two main substantive takeaways. First, that teacher bias towards pupil characteristics that tend to be irregularly associated with track outcomes can suffer from bias when ignoring the ordinality of the outcome variable. As a consequence, past work should be re-analysed to ensure this bias will not have affected substantive conclusions and categorical models should be used in future analyses of tracking. Second, institutional differences in tracking are considerable and confirm fears that a mixture of freedom for schools to pursue didactic philosophies of their choice, teacher-based tracking, and a lack of feedback mechanisms have led to deviations or drift in tracking practices amongst schools. More guidance in calibrating tracking decisions across schools seems appropriate, in particular stimulating inter-

700 actions between teachers at different schools could be a fruitful avenue to ensure
701 tracking is consistently applied throughout the system.

702 There are also limitations to this study. Previous work on The Netherlands has shown
703 that migrant background is another potential source of bias, although findings have
704 been mixed in the past. This source of bias could not be evaluated in this paper
705 due to a lack of data on student's migrant background. However, these students
706 have historically tended to score lower on standardized test scores and it can thus be
707 expected that similar bias in estimation might be present, as was found for students of
708 low parental education. Again, a re-analysis of past work addressing the categorical
709 nature of the outcome would be appropriate. Furthermore, due to the 2015 changes
710 to the tracking mechanism, this paper had to use a different metric of 'ability' than
711 most prior work on The Netherlands: I used a student's grades in the final two years,
712 whereas most prior work uses the school leaver's test. This means that a definitive
713 verdict on whether teacher bias has increased is difficult to give.

714 More generally, this paper champions prediction as a method that can be used to
715 broadly complement existing empirical techniques often used by sociologists. Predic-
716 tion provides new vantage points into model fit on a natural and disaggregated level,
717 which might be overlooked with the in-sample model diagnostics usually applied in
718 the field. Predictive analysis is also amenable to interventional reasoning and can be
719 used to generate intuitive insights into complex models. Unfortunately, prediction
720 is still severely under-utilised within the social sciences, often for the wrong reasons.
721 Prediction is too often conflated with the use of convoluted models and the intense
722 philosophical debate whether prediction and explanation are conceptually the same
723 has generated an unnecessary dichotomy between the two. When viewing prediction
724 as a simple tool to evaluate a model's performance, a host of applications become
725 available, none of which require a dogmatic choice between prediction or explanation.
726 We can and should simply apply both, especially when models become increasingly
727 complicated – as is often the case when studying the complexities of social life. Hope-
728 fully, this paper provides a clear example how prediction can be straightforwardly
729 applied to empirical work, and some of the many benefits it would offer to sociolo-
730 gists' empirical analyses.

References

- Karl L Alexander, Doris R Entwisle, and Maxine S Thompson. School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review*, pages 665–682, 1987.
- George Ansalone. Tracking: Educational differentiation or defective strategy. *Educational Research Quarterly*, 34(2):3, 2010.
- Valerio Baćak and Edward H Kennedy. Principled machine learning using the super learner: an application to predicting prison violence. *Sociological Methods & Research*, 48(3):698–721, 2019.
- Dominik Becker. The impact of teachers’ expectations on students’ educational opportunities in the life course: An empirical test of a subjective expected utility explanation. *Rationality and society*, 25(4):422–469, 2013.
- Julian R Betts. The economics of tracking in education. In *Handbook of the Economics of Education*, volume 3, pages 341–381. Elsevier, 2011.
- Ines Böhmer, Cornelia Gräsel, Sabine Krolak-Schwerdt, Thomas Hörstermann, and Sabine Glock. Teachers’ school tracking decisions. In *Competence Assessment in Education*, pages 131–147. Springer, 2017.
- Thijs Bol and Herman G Van de Werfhorst. Educational systems and the trade-off between labor market allocation and equality of educational opportunity. *Comparative Education Review*, 57(2):285–308, 2013.
- Thijs Bol, Jacqueline Witschge, Herman G. van de Werfhorst, and Jaap Dronkers. Curricular tracking and central examinations: Counterbalancing the impact of social background on student achievement in 36 countries. *Social Forces*, 92(4):1545–1572, 2014.
- Simon Boone, Sarah Thys, Piet Van Avermaet, and Mieke Van Houtte. Class composition as a frame of reference for teachers? the influence of class context on teacher recommendations. *British Educational Research Journal*, 44(2):274–293, 2018.
- Hennie Brandsma and Simone Doolaard. Differences in effectiveness between primary schools and their impact on secondary school recommendations. *School effectiveness and school improvement*, 10(4):430–450, 1999.
- Richard Breen, Kristian Bernt Karlson, and Anders Holm. Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, 44:39–54, 2018.
- Jere E Brophy and Thomas L Good. Teachers’ communication of differential expectations for children’s classroom performance: Some behavioral data. *Journal of educational psychology*, 61(5):365, 1970.
- David T Burkam, Laura LoGerfo, Doug Ready, and Valerie E Lee. The differential

- effects of repeating kindergarten. *Journal of Education for Students Placed at Risk*, 12(2):103–136, 2007.
- A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.
- CBS (accessed 7th December 2020). Statline Data. <https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurtten-2004-2019>. URL <https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurtten-2004-2019>.
- Daniele Checchi and Luca Flabbi. Intergenerational mobility and schooling decisions in italy and germany. In *IZA Discussion Paper 2876, Bonn, Germany*. 2006.
- Anna K. Chmielewski. An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education*, 120(3):293–324, 2014.
- Rune Haubo Bojesen Christensen. ordinal—regression models for ordinal data. *R package version*, 28:2015, 2015.
- Johan Coenen and Chris Van Klaveren. Better test scores with a same-gender teacher? *European Sociological Review*, 32(3):452–464, 2016.
- Hester de Boer, Roel J Bosker, and Margaretha P. C van Der Werf. Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1):168, 2010.
- A de Wijs, F Kamphuis, F Kleintjes, and M Tomesen. Wetenschappelijke verantwoording: Spelling voor groep 3 tot en met 8. Cito, 2010.
- Dienst Uitvoering Onderwijs (accessed 10th December 2020). Open onderwijsdata. URL <https://duo.nl/open-onderwijsdata/databestanden/po/leerlingen-po/po-totaal/bo-gewicht-leeftijd.jsp>.
- Ben Domingue. Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, 79(1):1–19, 2014.
- Ben Domingue, Klint Kanopka, Sam Trejo, and Jeremy Freese. An ordinal model for analysis of years of education. 2021.
- G Driessen, E Smeets, L Mulder, and H Vierke. De relatie tussen prestaties en advies: onder-of overadvisering bij de overgang van basis-naar voortgezet onderwijs. *Inspectie van het Onderwijs, Onderadvisering in beeld*, pages 59–81, 2007.
- Geert Driessen, Peter Slegers, and Frederik Smit. The transition from primary to secondary education: Meritocracy and ethnicity. *European sociological review*, 24(4):527–542, 2008.
- J Dronkers, M Van Erp, M Robijn, and J Roeleveld. Krijgen leerlingen in de grote

- steden en met name in amsterdam te hoge adviezen? *Tijdschrift voor Onderwijs-research*, 23:17–30, 1998.
- Per Engzell, Arun Frey, and Mark D Verhagen. Learning inequality during the covid-19 pandemic. 2020.
- H Feenstra, F Kamphuis, F Kleintjes, and R Krom. Wetenschappelijke verantwoording: Begrijpend lezen voor groep 3 tot en met 6. Cito, 2010.
- Ronald F Ferguson. Teachers’ perceptions and expectations and the black-white test score gap. *Urban education*, 38(4):460–507, 2003.
- Simon Field, Malgorzata Kuczera, and Beatriz Pont. No more failures: Ten steps to equity in education. Technical report, OECD, 2007.
- Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. In *Handbook of labor economics*, volume 4, pages 1–102. Elsevier, 2011.
- Laura Froehlich, Sarah E Martiny, Kay Deaux, and Sog Yee Mok. It’s their responsibility, not ours. *Social Psychology*, 2015.
- Andrew Gelman and Jennifer Hill. *Data analysis using regression and multi-level/hierarchical models*. Cambridge university press, 2006.
- Sara Geven, Anatolia Batruch, Herman van de Werfhorst, et al. Inequality in teacher judgements, expectations and track recommendations: A review study, 2018.
- Sabine Glock and Sabine Krolak-Schwerdt. Does nationality matter? the impact of stereotypical expectations on student teachers’ judgments. *Social Psychology of Education*, 16(1):111–127, 2013.
- Sabine Glock, Sabine Krolak-Schwerdt, Florian Klapproth, and Matthias Böhmer. Beyond judgment bias: How students’ ethnicity and academic profile consistency influence teachers’ tracking judgments. *Social Psychology of Education*, 16(4):555–573, 2013.
- Sabine Glock, Sabine Krolak-Schwerdt, and Ineke M Pit-ten Cate. Are school placement recommendations accurate? the effect of students’ ethnicity on teachers’ judgments and recognition memory. *European Journal of Psychology of Education*, 30(2):169–188, 2015.
- James J Heckman, John Eric Humphries, and Gregory Veramendi. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126(S1):S197–S246, 2018.
- Steffen Hillmert and Marita Jacob. Selections and social selectivity on the academic track: A life-course analysis of educational attainment in germany. *Research in Social Stratification and Mobility*, 28(1):59–76, 2010.
- Matthew Hindman. Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):48–62, 2015.

- J Benjamin Hinnant, Marion O'Brien, and Sharon R Ghazarian. The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of educational psychology*, 101(3):662, 2009.
- William C Horrace and Ronald L Oaxaca. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327, 2006.
- Mieke Van Houtte and Peter A. J. Stevens. Study involvement of academic and vocational students: Does between-school tracking sharpen the difference? *American Educational Research Journal*, 46(4):943–973, 2009.
- Inspectie van het Onderwijs (accessed 7th December 2020). Onderwijsdata. <https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata>.
URL <https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata>.
- Judith Janssen, Norman Verhelst, Ronald Engelen, and Floor Scheltens. Wetenschappelijke verantwoording: Rekenen-wiskunde voor groep 3 tot en met 8. Cito, 2010.
- Paul Jungbluth. *De ongelijke basisschool. Etniciteit, sociaal milieu, sekse, verborgen differentiatie, segregatie, onderwijskansen en schooleffectiviteit*. [Sl]: Nijmegen: ITS, 2003.
- Lee Jussim and Kent D Harber. Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and social psychology review*, 9(2):131–155, 2005.
- Timo M Kauppinen. Schools as mediators of neighbourhood effects on choice between vocational and academic tracks of secondary education in helsinki. *European Sociological Review*, 24(3):379–391, 2008.
- Florian Klapproth. Biased predictions of students' future achievement: An experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Studies in Educational Evaluation*, 59:67–75, 2018.
- Joeke Kuyvenhoven and Willem R Boterman. Neighbourhood and school effects on educational inequalities in the transition from primary to secondary education in amsterdam. *Urban Studies*, 2020.
- Helen F Ladd and Edward B Fiske. Weighted student funding in the Netherlands: A model for the US? *Journal of Policy Analysis and Management*, 30(3):470–498, 2011.
- Valerie E Lee and David T Burkam. *Inequality at the starting gate: Social background differences in achievement as children begin school*. ERIC, 2002.
- Kimberley Martine Lek. *Teacher knows best? On the (dis)advantages of teacher*

- judgments and test results, and how to optimally combine them.* PhD thesis, Utrecht University, 2020.
- J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*, volume 7. Stata press, 2006.
- Samuel R. Lucas. Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology*, 106(6): 1642–1690, 2001.
- Elke Lüdemann and Guido Schwerdt. Migration background and educational tracking. *Journal of Population Economics*, 26(2):455–481, 2013.
- H Luyten and RJ Bosker. Hoe meritocratisch zijn schooladviezen? *Pedagogische Studiën*, 2004.
- Gangadharrao S Maddala. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press, 1986.
- Oscar Marcenaro-Gutierrez and Anna Vignoles. A comparison of teacher and test-based assessment for spanish primary and secondary students. *Educational Research*, 57(1):1–21, 2015.
- Robert D Mare. Social background and school continuation decisions. *Journal of the american statistical association*, 75(370):295–305, 1980.
- Sönke Hendrik Matthewes. Better together? heterogeneous effects of tracking on student achievement. *The Economic Journal*, 131(635):1269–1307, 2021.
- Clark McKown and Rhona S Weinstein. Teacher expectations, classroom context, and the achievement gap. *Journal of school psychology*, 46(3):235–261, 2008.
- Gerhard Meisenberg. Verbal ability as a predictor of political preferences in the united states, 1974–2012. *Intelligence*, 50:135–143, 2015.
- Carina Mood. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, 26(1):67–82, 2010.
- Martin Neugebauer, Marcel Helbig, and Andreas Landmann. Unmasking the myth of the same-sex teacher advantage. *European sociological review*, 27(5):669–689, 2011.
- Onderwijsinspectie. *Onderwijsverslag ‘De Staat van het Onderwijs’*. Ministerie van Onderwijs, Cultuur en Wetenschap, 2016. Available at <https://www.onderwijsinspectie.nl/documenten/publicaties/2016/04/13/staat-van-het-onderwijs-2014-2015>.
- Onderwijsinspectie. *Onderwijsverslag ‘De staat van het Onderwijs’*. Ministerie van Onderwijs, Cultuur en Wetenschap, 2019. Available at <https://www.onderwijsinspectie.nl/documenten/rapporten/2019/04/10/rapport-de-staat-van-het-onderwijs-2019>.
- Harry Anthony Patrinos. School choice in the Netherlands. CESifo DICE Report 2/2011, ifo Institute for Economic Research, Munich, 2011.

- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Hanneke Posthumus, Bart Bakker, Jan van der Laan, Martine de Mooij, Sander Scholtus, Mersiha Tepic, Jeroen van den Tillaart, and Nander de Vette. Herziening gewichtenregeling primair onderwijs: Fase I. Statistics Netherlands, 2016.
- Hanneke Posthumus, Sander Scholtus, and Jaap Walhout. De nieuwe onderwijs achterstandenindicator primair onderwijs: Samenvattend rapport. Statistics Netherlands, 2019.
- Tijana Prokic-Breuer and Jaap Dronkers. The high performance of dutch and flemish 15-year-old native pupils: Explaining country differences in math scores between highly stratified educational systems. *Educational Research and Evaluation*, 18(8): 749–777, 2012.
- Douglas D Ready and David L Wright. Accuracy and inaccuracy in teachers’ perceptions of young children’s cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, 48(2):335–360, 2011.
- Jozef MM Ritzen, Jan Van Dommelen, and Frans J De Vijlder. School finance and school choice in the netherlands. *Economics of Education Review*, 16(3):329–335, 1997.
- JAAP Roeleveld, GEERT Driessen, GUUSKE Ledoux, JORIS Cuppen, and JOOST Meijer. Doelgroepleerlingen in het basisonderwijs. *Historische ontwikkeling en actuele situatie*. Amsterdam/Nijmegen: Kohnstamm Instituut/ITS, 2011.
- Christine Rubie-Davies, John Hattie, and Richard Hamilton. Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, 76(3):429–444, 2006.
- Christine M Rubie-Davies. *Teacher expectations in education*. Routledge, 2017.
- Frida Rudolphi and Robrt Erikson. Social selection in formal and informal tracking in sweden. In *Models of Secondary Education and Social Inequality*. Edward Elgar Publishing, 2016.
- Tobias Rüttenauer and Volker Ludwig. Fixed effects individual slopes: Accounting and testing for heterogeneous effects in panel data or other multilevel models. *Sociological Methods & Research*, 2020.
- Jaap Scheerens, Melanie Ehren, Peter Slegers, and Renske de Leeuw. Country background report for the Netherlands. OECD Review on evaluation and assessment frameworks for improving school outcomes, 2012.
- Philip A Schrodtt. Seven deadly sins of contemporary quantitative political analysis. *Journal of peace research*, 51(2):287–300, 2014.
- Galit Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.
- Ed Smeets, Jos Kuijk, and Geert Driessen. *Handreiking bij het opstellen van het basisschooladvies*. 12 2014.

- Nicole S Sorhagen. Early teacher expectations disproportionately affect poor children's high school performance. *Journal of Educational Psychology*, 105(2):465, 2013.
- Anna Südkamp, Johanna Kaiser, and Jens Möller. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of educational psychology*, 104(3):743, 2012.
- Laura Marie Tach and George Farkas. Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research*, 35(4):1048–1079, 2006.
- Nicole Tieben and Maarten HJ Wolbers. Transitions to post-secondary and tertiary education in the netherlands: a trend analysis of unconditional and conditional socio-economic background effects. *Higher education*, 60(1):85–100, 2010.
- Anneke Timmermans, Hans Kuyper, and Greetje van der Werf. Schooladviezen en onderwijsloopbanen. *Voorkomen, risicofactoren en gevolgen van onder-en overadvisering. Groningen: GION*, 2013.
- Anneke C. Timmermans and Christine M. Rubie-Davies. Do teachers differ in the level of expectations or in the extent to which they differentiate in expectations? relations between teacher-level expectations, teacher background and beliefs, and subsequent student performance. *Educational Research and Evaluation*, 24(3-5):241–263, 2018. doi: 10.1080/13803611.2018.1550837. URL <https://doi.org/10.1080/13803611.2018.1550837>.
- Anneke C Timmermans, Hans Kuyper, and Greetje van der Werf. Accurate, inaccurate, or biased teacher expectations: Do dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology*, 85(4):459–478, 2015.
- Jochem Tolsma and Maarten HJ Wolbers. Onderwijs als nieuwe sociale scheidslijn? de gevolgen van onderwijsexpansie voor sociale mobiliteit, de waarde van diploma's en het relatieve belang van opleiding in nederland. 2010.
- Herman G van de Werfhorst. Inequality in learning is a major concern after school closures. *Proceedings of the National Academy of Sciences*, 118(20), 2021.
- Herman G Van de Werfhorst and Jonathan JB Mijs. Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual review of sociology*, 36:407–428, 2010.
- Linda van den Bergh, Eddie Denessen, Lisette Hornstra, Marinus Voeten, and Rob W. Holland. The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, 47(2):497–527, 2010.

- Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.
- Herman G. van der Werfhorst, Louise Elffers, Sjoerd Karsten, et al. *Onderwijsstelsels vergeleken: leren, werken en burgerschap*. Didactief onderzoek, 2015.
- Mark Verhagen. A pragmatist’s guide to using prediction in the social sciences. *Not published yet*, 2021.
- Penelope W. St. J. Watson, Christine M. Rubie-Davies, Kane Meissel, Elizabeth R. Peterson, Annaline Flint, Lynda Garrett, and Lyn McDonald. Teacher gender, and expectation of reading achievement in new zealand elementary school students: essentially a barrier? *Gender and Education*, 31(8):1000–1019, 2019.
- Duncan J Watts. Common sense and sociological explanations. *American Journal of Sociology*, 120(2):313–351, 2014.
- Cristobal Young and Katherine Holsteen. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1):3–40, 2017.