Self-Selection and Treatment Assignment in Field Experiments*

Gerhard Riener† Sebastian Schneider‡ Valentin Wagner§

December 7, 2019

Click here to get the newest version of this paper.

Abstract

Self-selection and unbalanced treatment assignment are two major concerns in experimental evaluations as they compromise the validity of a study at the external and the internal margin. In this paper we present evidence on the selection of partner institutions into participation and balanced treatment assignment in field experiments. We answer two questions: (1) do stakeholders that choose to participate in a field experiment differ from the population of interest, and (2) does pre-treatment balancedness on observable characteristics translate to lower bias and increased power in a real-world setting, and which treatment assignment method should best be used if that is the case? To this end, we conducted a recruitment experiment inviting institutional gatekeepers to participate in field experiments and varying the salience of the research topic and the stakes of participation. We combine this experimental data with a rich set of administrative data on institution and municipality characteristics to identify a possible self-selection bias. Moreover, we compare pure randomization, pair-wise matching, and re-randomization based on t-statistics to a new treatment assignment method—the minimum mean squared error treatment assignment (min MSE). We find no evidence for a self-selection bias on observable characteristics and establish that balancedness as achieved by the minMSE method reduces bias of treatment effect estimation by 33% compared to pure randomization. The minMSE method performs best in increasing pre-treatment balancedness of observable characteristics compared to pure randomization, matching, and t-statistic based re-randomization.

Keywords: Self-selection bias, field experiments, treatment assignment

JEL codes: C93, I20

*We are grateful for comments and advice from Matthias Sutter, Stefania Bortolotti, seminar participants at the Max Planck Institute for Research on Collective Goods, Bonn, the University of Mainz and at the NTNU Trondheim, and participants of the External Validity, Generalizability and Replicability of Economic Experiments Workshop, IWAEE 2019, NCBEE 2019, and the 8th Field Day in Fontainebleau. The usual disclaimer applies.

†Corresponding author: riener@dice.uni-duesseldorf.de, Düsseldorf Institute for Competition Economics, Universitätsstr. 1, 40225 Düsseldorf.

‡sschneider@coll.mpg.de, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, 53113 Bonn.

§wagnerv@uni-mainz.de, University of Mainz, Jakob-Welder-Weg 4, 55099 Mainz.

1 Introduction

Experimental methods are increasingly used in recent years by academic researchers as well as public policy-makers to enrich the toolkit for evidence based policy making.[1] *"The most credible and influential research designs use random assignment"* (Angrist and Pischke, 2009), as randomization on expectation ensures a causal interpretation of the respective treatment conditions, independently of characteristics of the participating population. The strength of properly conducted randomized controlled trials is that they not only shall allow for causal interpretation within a sample, but they also should provide reliable out of sample predictions for the population of interest. While appropriate treatment assignment is a necessary condition for internal validity of a study, it is not sufficient for achieving external validity and hence the ability to make out of sample predictions (Hotz, G. Imbens, and Mortimer, 2005; Allcott, 2015). However, only if the latter condition is met, randomized controlled trials can be considered as the gold standard for drawing inferences about the effect of a policy (Athey and G. Imbens, 2017b).

Applying randomized field experiments to inform policy-makers about the impact of a given program is likely to increase in the future as digitization lowers the cost of implementation and therefore eases its use (Athey and G. Imbens, 2017a). Nevertheless, in many cases, field experiments remain difficult to implement, for political, or ethical reasons, or because the population of interest is too small (Athey and G. Imbens, 2017b) and therefore often suffer from relatively small samples. This implies the need to choose an appropriate treatment assignment method and to evaluate who self select into participation.

Pure randomization, where the allocation of treatment to individuals is left purely to chance, is often perceived as ex-ante fair and hence politically easier to implement (Burtless, 1995). In large samples, pure randomization achieves balanced treatment and control groups, i.e. groups will have, on average, identical characteristics.[2] However, the probability of differences in characteristics is increasing in small-scale experiments or experiments in which treatment assignment is implemented at a superordinate level, e.g. at school-level instead of student-level. Randomization methods trying to avoid imbalance on characteristics in medium or small-scale experiments are e.g., re-randomization, stratification (also called blocking), or pairwise matching. Recently, building on the criteria of balance by Kasy (2016), Schneider and Schlather (2017) have introduced a new re-randomization method suited for several, possibly continuous covariates which they call min MSE. It improves on earlier re-randomization methods by taking into account more characteristics of the joint distribution of covariates, such as correlation between covariates or their variances instead of just their mean values. The question is whether these methods considerably improve on pure randomization in generating overlap, i.e. in distributing every characteristic in the sample in a way that it appears at least

[1] See for example the rise of behavioral research unity installed by governments and inter- or supranational organizations to evaluate policy around the world https://www.oecd.org/gov/regulatory-policy/behavioural-insights.htm.

[2] While this is theoretically obvious, it is far less obvious, what 'large' means. Bruhn and McKenzie (2009) establish that with samples sizes of 300 units and binary treatment, experimental groups are fairly balanced.

once in every experimental group. Theoretically, overlap is needed to identify, i.e. consistently estimate, the conditional average treatment effect (see e.g. Abadie and G. W. Imbens, 2006) when comparing treatment and control group while controlling for covariates. Moreover, limited overlap might not just lead to biased results but theoretically also cause distorted confidence intervals (Rothe, 2017). This rises the second question related to treatment assignment: Do these methods improve precision of treatment effect estimation?

Survey results indicate that even among experts there is no agreement on which method to use for treatment assignment to achieve balance and precision of estimations (Bruhn and McKenzie, 2009). In simulation studies, Bruhn and McKenzie (2009) show that the method of randomization does not seem to matter in large samples (300 units) but that pair-wise matching outperforms stratification and re-randomization (and pure randomization) in achieving balance in covariates in smaller samples (100 units). Interestingly, while Bruhn and McKenzie (2009) ask which method for treatment assignment to use, Kasy (2016) argues that any randomization is not optimal for treatment assignment in the case of continuous pretreatment information such as e.g., age or income. Instead, he proposes a new assignment method which involves minimizing the expected mean squared error (MSE) of the treatment effect estimator by choice of treatment assignment. Schneider and Schlather (2017) interpret this method as a re-randomization method, introduce it in a simplified theoretical framework, thereby making it applicable for practitioners, and extend it to allow for assigning multiple treatment arms. So far, the different treatment assignment methods have been compared in simulation studies and with binary treatments only but have not been compared with real field experimental data and in settings with more than one treatment arm.

The second key question concerns the external validity of the experiment. A challenging aspect of field experiments is that they often rely on voluntary participation on the side of the subjects of interest or the leaders of the entities at which the experiment is conducted.[3]. This self-selection is likely to lead to a non-representative participant pool which is a major threat to the scalability of field experiments (Al-Ubaydli, List, and Suskind, 2019; Czibor, Jimenez-Gomez, and List, 2019).[4]

Despite its importance, self-selection into field experiments remains a largely understudied topic.[5] Only a few studies assess the implications of self-selection for field experimental research (e.g., Harrison and List, 2004; Ludwig, Kling, and Mullainathan, 2011; Belot and James, 2016) and to the best of our knowledge, no study experimentally varies the characteristics of a field experiment in order to assess their relevance for the probability of selection into field experiments. In many situations, alternatives to experimental interventions

[3]Ethics committees often insist on exit clauses and voluntary participation. This might be a consequence of review boards often being developed and installed for life-sciences that tend to ask different questions than social-scientists (Humphreys, 2019)

[4]According to Czibor, Jimenez-Gomez, and List (2019), there are four threats to the generalizability of (field) experimental results: characteristics of experiment, selective noncompliance, non-random selection, and different populations.

[5]Increased interest in recent years has led to the development of tools to foster the trust in generalizability of education field experiment. For example the website *The Generalizer* https://www.bethtipton.com/generalizations helps designing sample recruitment plans for school studies in the US.

are available outside the experiment which could lead to both positive and negative self-selection at the same time as the possibility of being part of the control group entails opportunity costs (Belot and James, 2014). For example, potential innovative partners that are willing to test new programs might be more open to participate in the experiment, but at the same time, they are also more likely to already have participated in other effective programs lowering their willingness to participate in an additional program (Allcott, 2015). As Belot and James (2016) notice, most (field) experimental studies provide little or no information on how participants were recruited and the experimental sample is never compared to the broader population of interest.[6] So far, there is ample research analyzing non-random participation in laboratory (e.g., Slonim et al., 2013; Krawczyk, 2011; Abeler and Nosenzo, 2015; Charness, Gneezy, and Kuhn, 2013; Cleave, Nikiforakis, and Slonim, 2013; Anderson et al., 2013; Falk, Meier, and Zehnder, 2013; Benndorf, Möllers, and Normann, 2017; Lazear, Malmendier, and Weber, 2012) and artefactual field experiments (e.g., Harrison, Lau, and Rutström, 2009; Frijters, Kong, and Liu, 2015) but whether participants in natural field experiments differ from the general population remains largely unstudied. Moreover, we still know little about site selection bias, this is, whether the probability that a program is adopted or evaluated is correlated with its impacts (Allcott, 2015). Two notable exceptions are the studies by Allcott (2015) and Belot and James (2016). Allcott (2015) tests for site selection bias in the context of the Opower energy conservation programs and Belot and James (2016) analyze the selection of local school authorities into a policy relevant experiment.

This paper contributes to two important questions addressing internal and external validity: (i)does the method of treatment assignment—pure randomization, pair-wise matching, re-randomization based on t-statistics, min MSE method—matter for balancing covariates and increasing the precision and power of detecting treatment effects in a real-world setting?, and (ii) can we detect self-selection of institutional stakeholders into field experiments? We answer these questions by conducting a recruitment experiment in which we ask institutional stakeholder, in our case headmasters, about their willingness to participate in a field study. The recruitment experiment was conducted in Germany's most populous state North Rhine-Westphalia (NRW) with more than 3,000 elementary and secondary schools. We sent personalized emails directly to headmasters who could respond to our email by choosing one of three links: "strong interest" (headmasters are willing to participate), "light interest" (research topic is interesting but headmaster wants to be contacted again later), and "opt-out" (headmasters do not want to be contacted by researchers again). Not clicking on any link was recorded as "no response".

To answer the question which treatment assignment increases balance of covariates and precision of estimates, we randomized schools—before contacting them via email—into treatment arms using pure randomization, matching, re-randomization based on t-statistics, or the min MSE treatment assignment method by Schneider and Schlather (2017). Pure randomization assigned schools to the control group or one of the

[6] Belot and James (2016) focus on experiments in the fields of policy evaluation, personnel and development economics in the Top-5 journals and in the American Economic Journal: Applied Economics.

treatments by pure chance, pair-wise matching—due to its nature—was used in a sub-sample for assigning two treatments only, as it tries to find pairs of units that are comparable, and then one is randomly assigned to the treatment and one to the control group. Re-randomization in general means random assignment until a criterion of balancedness, sometimes subjective judgment of the researcher, is met (see e.g. Bruhn and McKenzie, 2009). The basic idea of the min MSE method as interpreted by Schneider and Schlather (2017) is to re-randomize treatment status a given number of times. With each iteration only some randomly selected units are exchanged, and only if this improves the theoretical criteria of balancedness, namely the mean squared error of the treatment effect as a function of covariates, the new assignment is kept and used as basis for the next iteration. Roughly speaking, the criteria maximizes covariate variances in all treatment groups while accounting for correlations, thereby achieving balancedness.

The response rates also allow us to shed light on self-selection into participation by combining them with a rich set of administrative data on school and municipality characteristics. We then estimate whether schools which positively responded (opting-in) differ along these characteristics from schools which did not respond or actively opted out, thereby identifying a potential self-selection bias. We further investigate how to get headmasters' attention. This is—besides the issue of self-selection—a key question when conducting field experiments. We therefore varied the content of the invitation email to headmasters along two dimensions: the main topic of the planned field study, and the extrinsic incentive to participate. Our invitation emails either invited headmasters to participate in a survey on the collaboration between schools and academia, or invited them to participate in a larger field experiment. Invitation emails to participate in the field experiment highlighted three different research topics: (i) e-learning, (ii) parental involvement, or (iii) integration of migrant children. The survey treatment serves as our control group as it measures headmasters' willingness to invest a minimum amount effort to respond to our email. Varying the topic of the planned field experiment and comparing response rates with those of the control group allows to identify whether we could attract headmasters' attention in this manner. Moreover, we analyze whether financial rewards for participation can attract headmasters' attention by varying whether schools were eligible to receive an extrinsic incentive for participation or not.

We find that the min MSE treatment assignment method is superior to pure randomization, matching, and re-randomization based on simple t-statistics in increasing (i) pre-treatment balancedness of observable characteristics as measured e.g. in overlap, (ii) the precision of treatment effect estimates (reduces bias by 33%), and (iii) power in detecting treatment effects (lower p-values). Furthermore, there is no evidence for non-random selection of schools. Schools which responded positively to our email do not significantly differ with respect to observable characteristics from schools which did not respond or actively opted out. However, we find that the topic of the experiment matters to attract headmasters' attention. Schools which were invited to participate in an experiment on parental involvement or on the integration of migrant children

were more likely to positively respond than schools in the control group. Interestingly, inviting schools to participate in an experiment on e-learning and offering financial extrinsic incentives had no positive effect on headmasters' attention.

2 External validity and site selection

In this section we develop a decision model for the main decision maker—the school principal—of our experiment. The model draws on previous work by Allcott (2015) and Belot and James (2014). However, in contrast to their assumptions, we do not assume that the decision maker knows about the experimental treatment and hence decides upon the perceived effectiveness of the treatment, but rather that the decision maker knows about the topic of the study, and assesses its usefulness on the topic. As Belot and James (2014) recognize, the knowledge of the treatment itself allows for positive selection, i.e. potential participants that expect the treatment intervention to be more effectively implemented outside the experiment and without running the risk of being put in the control group select out of the experiment. In our setup we avoid this problem by not distributing information on the new technology under study. This also allows for a simplification of the decision problem and focus on a different aspect of selection: the selection based on deontological motives.

Following Rubin (1974), we define $T_i \in \{1, 0\}$ as the treatment indicator for unit i with potential outcomes $Y_i(1)$ when treated and $Y_i(0)$ otherwise. The difference in potential outcomes for unit i is $\tau_i = Y_i(1) - Y_i(0)$, and X_i is a vector of covariates and X constitutes the support of the covariates. The target population is the population for which we would like to estimate the *Average Treatment Effect* (ATE). The sample population is the population which was exposed to the experiment. $D_i \in \{1, 0\}$ indicates if unit i is in the sample. The ATE in the target population can be consistently estimated under the following four assumptions (Allcott, 2015):

Assumption 1 Unconfoundedness. $T_i \perp (Y_i(1), Y_i(0))|X_i$

Assumption 2 Overlap. $0 < Pr(T_i = 1|X_i = x) < 1$

Assumption 3 External unconfoundedness. $D_i \perp (Y_i(1) - Y_i(0))|X_i$

Assumption 4 External overlap. $0 < Pr(D_i = 1|X_i = x) < 1$ for all $x \in X$

Suppose the decision maker considers replying to a call and participates in a study. We assume that she will either apply cost benefit calculation of the value of the study for the her institution or use deontological reasoning to reject or accept participation. Let this value depend on the institutional characteristic x and let the cost of participation be fixed at c and some unobserved decision maker type that can either be utilitarian or deontological $w \in \{ut, de\}$. We assume that the value $b(x)$ depends monotonically and continuously increasing on the (univariate) school characteristic x, which, for simplicity, we assume can be mapped into a single dimension, and the cost of participation. Let this value be given by:

$$V(D, x, y) = (b(x) - c)D \quad \text{for decision maker using cost benefit arguments} \tag{1}$$

A decision maker using cost-benefit analysis will decide to participate as long as the cost is smaller than the benefit. Therefore, there exists a unique (intermediate value theorem) cutoff $\bar{x}$ above which the gatekeeper will decide to participate. A decision maker using deontological reasoning—*de*—will participate if the study is legitimate according to her world view. Offering a bounty that reduces costs of participation will not be effective in these cases. We can further refine the notion of legitimacy. Legitimacy can either be denied upon a (i) general rejection of the evaluation (ii) a rejection of the objective of the study. While the latter point can be addressed by framing (as in the experiment), the first point will always lead to non-participation of the decision maker.

Consequences for participation For the *ut*-types of decision maker, we obtain a clear selection pattern. Institutions with characteristics below the cutoff $\bar{x}$ will never participate in the study leading to selection on observables x, violating the external overlap Assumption 4. If the characteristics of interest x are independent of the deontological type of the decision maker, this poses no further selection issues.

3 Experimental setup and treatment assignment

The recruitment experiment was conducted in North Rhine-Westphalia (henceforth NRW) from October 2016 to January 2017. The institutional preconditions in NRW are ideal for our research question on self selection as headmasters are allowed to autonomously decide to participate in scientific studies without the permission of the school authority. This allows us to directly contact important gatekeepers while avoiding potential self selection on a higher level (as might be the case in Belot and James (2016)). Further (practical) reasons for choosing NRW is the high density of schools as NRW is Germany's most populous state and the reducing the costs to implement the proposed experiment being located at the University of Düsseldorf. We decided to We contacted schools that were included in the official school list of the Ministry of Education in NRW as of March 2016 and invited them to participate in our study. To reduce gatekeepers' i.e. headmasters' costs of responding to our inquiry, all contact with the schools was electronically by email asking headmasters on participating in a scientific study. We learned in previous studies that the responsiveness of schools in NRW does not depend on whether we send a posted letter or email (see Panel B of Table 2).[7] Recruitment emails were sent out October 2, 2016 and for those schools who did not respond—neither positively nor negatively—we sent out two reminder emails.[8]

[7] Panel Panel B of Table 2 presents response rates by contact type in the study of Riener and Wagner (2019). The authors varied whether they contacted school by email, posted letter, or a combination of both.

[8] The first wave was sent on October 2, 2016. We sent emails in batches of 50 per 2 hour interval on mailing day using the internal LimeSurvey procedure to handle invitations to surveys. In total, about 3% of emails could not be delievered due to technical reasons. The first reminder email was sent one month after the first contact, on November 2, 2016. The second reminder was sent three weeks after the first.

We contacted all (elementary and secondary) schools in NRW that fulfilled our basic requirements. Our three exclusion criteria were: (a) schools with a medical focus, (b) schools that mainly teach adults in second-chance-education or evening schools, and (c) schools in the largest cities of NRW. We excluded school types (a) and (b), as not all our research topics are relevent for them, e.g., the research topic "parental involvement". Schools in larger cities (type (c)) were excluded for two reasons: First, schools in metropolitan areas are over-researched as they all are home to at least one university, and thus receive many inquiries, e.g., by bachelor and master students, which might introduce noise in the measurement of willigness to participate in our study. Second, we were concerned about reputation effects and ongoing partnerships in schools in larger cities. We previously conducted several other experiments in schools in larger cities in NRW (Riener and Wagner, 2019; Fischer and Wagner, 2018; Wagner, 2016) which might cause a positive or negative reputation effect for participation in an additional study. Moreover, schools with already existing partners and ongoing programs might or might not be more likely to participate (Allcott, 2015). In total 3305 schools were contacted which represents 66.29% of all schools in NRW.

The recruitment email: Gatekeepers, i.e. Headmasters, were asked to express their interest in participating in a scientific study using a standardized email message. This message introduced ourselves and our expertise in conducting scientific studies in schools, mentioned the respective research question, briefly explained the methodology we will use, and outlined the workload for the schools. We kept the information in the email intentionally very short to increase the likelihood of headmasters reading the message. However, headmasters could access more information on the project—scientific foundation of the research question, timeline of the study, exclusion criteria for participation, and information about data protection—by clicking on a link provided in the email (see the online Appendix D for a facsimile of the recruitment email). Moreover, to measure schools' responsiveness, headmasters could indicate their interest in participating by clicking one of three links displayed at the bottom of the recruitment email. Clicking the first link, headmasters could express strong interest in the project and were told that they will be contacted again with a detailed plan of the experiment. Choosing the second link, a school could indicate that they were generally interested in the topic but at the time see no capacity to participate. The third link was an opt-out link where schools could opt-out of participating and receiving further reminders. After clicking on one of the three links, schools were forwarded to a questionnaire asking for further details about the school (e.g., what the position of the respondent is within the school).

3.1 Treatments

We implemented a 3×2 design, to study interest in three suggested research topics of the collaborative project, where we varied the provision of incentives (no incentive vs. monetary incentive) independently. All collaborative projects were presented in the same way and were equally long. The treatment variation

was the first and last paragraph of the email, announcing that our plan to conduct an experiment about the respective research topic within schools. The fourth paragraph of the email informed about monetary incentives, if applicable. We administered an additional treatment—the control treatment—where we simply asked headmasters to fill out a questionnaire online. The rationale for this treatment is to lower the bar for participation substantially and construct a benchmark of schools that are willing to contribute to scientific studies, but for whatever reason do not want to participate in an experiment.

Control treatment In the control treatment, we asked schools to participate in an online survey. In this survey, we asked about headmasters' point of view regarding the collaboration between academia and schools, this is, how insight gained in academic research can be integrated in schools' daily life. Importantly, answering the survey did not involve participation in any experimental study and it required a minimum of headmasters' time—approximately five minutes. Due to the low stakes of the survey and the time frame, we interpret the responsiveness in the survey as schools' baseline responsiveness in dealing with inquiries of academic researchers.

E-learning In the E-learning treatment we suggested participating in a study on the use of electronic devices in education. The research question for this topic was to find out which types of electronic testing formats could be implemented in schools and how they perform compared to traditional pen and paper exams. This treatment was motivated by a recent move of the German government to increase spending towards research on digital media in the classroom.[9]

Parental involvement In the parental involvement treatment letter, we asked for participation in a study aiming at analyzing the effect of getting parents involved in their children's education (e.g. students in-class behavior and academic performance). This treatment was motivated by recent academic research using electronic devices (e.g. text messaging) to reduce information frictions between parents and children (e.g. Bergman and Chan, 2019; Kraft and Rogers, 2015). These studies show that active participation of parents in their children's education can lead to favorable educational and behavioral outcomes.

Integration of migrant children In the treatment letter proposing the topic integration of migrant children, we asked schools to participate in a study to analyze how students with a migration background and language difficulties could best be integrated into classroom education. This topic was inspired by the increasing migration to Germany in 2015/16, which was covered widely in the media. It constituted a major challenge for schools to rapidly integrate non-German speaking children into the school environment.

Monetary incentive Besides the research topic, we altered whether or not schools were offered a financial incentive. Two schools could win a 700 Euro budget in the case of participating. This money could be

[9] http://www.bildung-forschung.digital/

used for school internal projects, such as continued education for teachers or study material. We included this aspect in the study to shed light on whether financial incentives have the power to attract gatekeepers' attention and to increase their willingness to respond to the email. In order to evaluate the size of our financial incentive, we show the share of the incentive in terms of the yearly budget of a school for the training of teachers in Table 9 (schools get a yearly budget of 45 Euros per full-time employed teacher). For more than 70 percent of the schools our incentive constitutes a share of 80 to 90 percent of the yearly budget.[10]

We planned to implement the suggested experiment in schools after the summer break in 2017. The goal of the planned experiment in 2017 was to randomly provide schools with a digital class-book and to analyze its impact on students' in-class behavior and educational attainment. Features of the electronic class-book should encompass a communication tool for teachers to send private or public notifications to students and/or parents. Moreover, we planned to incorporate an build-in translator within the communication tool to reduce language barriers and better integrate migrant children and their parents. Another feature we intended to implement within the electronic class-book was an "exercise and testing tool" in which teachers could send students personalized exercises, remind them about assignments, e.g., homework, and to test students online. However, for several reasons—the most important being unexpected complexity in programming leading to an considerable excess of our budget—we were not able to finally conduct the experiment after our initial contact end of 2016, although we had planned it.[11]

3.2 An Experiment on Balance, Precision and the Role of Treatment Assignment Methods

We empirically want to assess the degree to which balance in observable characteristics matters for precision of the estimation of treatment effects in a real-world setting.[12] Directly related, we are interested in the performance of different treatment assignment methods with respect to balance and precision of estimation. To this end, we added another experimental layer to our research design. We conducted an experiment parallel to the experiment outlined above to empirically compare different treatment assignment methods and to study the relation between balance and precision in a real-world setting using real treatment effects.

Taking advantage of our unique setting, we divided the whole sample of more than 3000 schools into smaller, comparable subsamples, and use different treatment assignment mechanisms to vary the degree of balance (of covariates) in these subsamples. We then assess the precision of estimation in the subsamples

[10] Clearly, in terms of expected value this financial incentive is rather low ($\tilde{1}$1,76 Euros if we consider all schools who responded positively). However, headmasters did not know how many schools were contacted or how many schools responded positively and hence could only form a belief about the expected value of the financial incentives. From all our experience, we believe that the absolute value of 700 Euros was the salient figure, and we rather assume that headmasters might have compared the financial incentive to their yearly budget instead of deriving beliefs about participating schools to calculate expected values. However, we cannot proof this, but there is also no evidence from the survey supporting expected value calculations.

[11] Potential (educational) partners need to be contacted at an early stage of the project as activities in a school year are planned well ahead.

[12] Theoretically, Rothe (2017) shows that limited overlap—a certain notion of imbalancedness—may lead to distorted confidence intervals and Greevy et al. (2004) show that balance in observable characteristics indeed leads to a higher precision of estimation.

and relate it to pre-treatment balance. The design of our experiment additionally allows us to compare commonly used treatment assignment methods in particular in settings with one and multiple treatments and multiple (possibly continuous) covariates. We build on results by Bruhn and McKenzie (2009) and compare a new method, the minMSE method, to re-randomization based on individual t-statistics of the different variables considered for treatment assignment as popularized by Bruhn and McKenzie (2009), as well as to pure randomization and where possible—in the case of a binary treatment—to pair-wise matching.

The minMSE method as proposed by Schneider and Schlather (2017) builds on earlier work by Kasy (2016), in particular on his notion of balance and the statistic of balancedness he applies. Building on a simpler theoretical framework, Schneider and Schlather (2017) extend the method by Kasy (2016) to the case of multiple treatment groups. A side effect of their simpler framework is easier implementation, as it works without specifying technical parameters, whereas for the method proposed by Kasy (2016), rather technical parameters have to be specified, such as the R^2 of a regression of considered covariates on *potential* outcomes. Ultimately, this allows for the flexibility to assume a different variance of the outcome of interest in the different experimental groups. Schneider and Schlather (2017) implement this flexibility with optional scaling parameters. Importantly, while Kasy (2016) proposes to optimize balance using the derived statistic of balancedness to finally use the resulting deterministic treatment assignment, Schneider and Schlather (2017) propose to use the statistic for re-randomization using the stochastic simulated annealing algorithm with a finite number of iterations (Kirkpatrick, Gelatt, and Vecchi, 1983): First, a hypothetical treatment assigment is performed by randomly allocating units to treatment groups. Using this hypothetical assignment, the extended statistic of balancedness is computed. For each of a specified number of iterations, a certain amount of units are randomly selected and their hypothetical treatment group assignment is switched. Then, the statistic of balancedness is re-computed. If the balancedness of the hypothetical assignment of the current iteration improves on the balancedness of the hypothetical treatment group assignment of the last iteration, it is used for the next iteration; otherwise, the last hypothetical assignment is used to proceed, or, with a probability that is a decreasing function of the number of iterations, also a worse current iteration is kept. Finally, only the hypothetical treatment group assignment of the last iteration is used for treatment assignment. Thus, for using a finite number of iterations and for using a stochastic algorithm to perform re-randomization, traditional randomization inference can be applied.[13]

For comparison reasons, we implement the version of Schneider and Schlather (2017) also in the case of binary treatment, where we could theoretically also have used the method by Kasy (2016).

The General Setting: Multivariate and Continuous Pre-treatment Characteristics In all considered settings of this added experimental layer we account for the nature of the actually available pre-treatment information. That is: At least some pre-treatment variables are continuous, there are more than three variables available and all of them might be relevant for the outcome measured. Lastly, no exact split is needed,

[13]The implementation of the method in the provided R package `minMSE` is able to automatically provide different alternative test vectors to the actually used assignment vector to perform, e.g., permutation inference for assessment of significance of treatment effects.

such as a sharp 1:1 division of, e.g. both females and males in treatment and control group. Therefore, in the overall setting of our experiment, treatment assignment using stratification is either practically difficult to impossible or not necessary, thus including it would at best yield blurred results.[14] However, we acknowledge that when an exact split is important, pure stratification or stratification in combination with the treatment assignment mechanisms considered here might be the appropriate solution.

The settings that we consider differ in the amount of treatment groups that are to be formed. The simplest setting considers the case of binary treatment, and we increase the number of experimental groups to up to seven.

Binary Treatment: Pair-Wise Matching and the minMSE method Using simulations in a setting of binary treatment, where, as in our study, several continuous pre-treatment characteristics are to be balanced, Bruhn and McKenzie (2009) show that pair-wise matching outperforms stratification and the re-randomization approaches considered.

However, a draw-back of pair-wise matching for assignment of units to treatment groups is its strong dependency on pairs, which is problematic if attrition happens, i.e. if observations that were used for treatment assignment are missing for conducting treatment or for measuring the outcome of interest. Typical examples of settings where attrition is to be expected include repeated measurements at schools where due to illness of participants, 10% of the sample can be expected to be absent at one of the measurement dates or when randomization is performed at the cluster level. In those cases, it is common practice to also drop their counterparts from the sample, to mantain balance and to ensure consistency of estimated treatment effects. This might eventually limit the power of the study below a critical threshold. The minMSE approach does not require this step, and is thus an interesting alternative in these settings.

For the reasons just explained, we compare the pair-wise matching approach with the new minMSE approach in a setting of binary treatment in absence of attrition, where it is unclear which method performs superior. Mimicking a standard use-case for the matching approach to treatment assignment, we use a quite small sample of 30 units that has to be divided into a treatment and a control group. In cases where the units that have to be assigned to experimental groups are clusters, this might already be a big sample.

We implemented an optimal pair-wise matching approach, which improves on the *greedy* approach to pair-wise matching applied in Bruhn and McKenzie (2009), by optimizing the overall generalized distance between observations. We use the R implementation (package 'nbpMatching') that accompanies Lu et al. (2011).

Multiple Treatment Arms: Min-max-t-statistic rerandomization, pure randomization and the minMSE method For settings with more than one treatment group, we are not aware of any simulation results. Moreover, as far as we know, to date, there is no readily implementable, theoretically founded

[14] See also the discussion in Bruhn and McKenzie (2009) on implementing a stratification approach with several and/or continuous variables.

standard approach to allocate units to more than two groups using matching or any alternative treatment allocation method.

However, the min-max-t-statistic-method, popularized by Bruhn and McKenzie, 2009, is extendable to multiple treatment arms relatively easily. For this reason, we compare the minMSE method with this method in case of multiple treatment groups.

A standard use case for re-randomization might be to allocate units to groups of 30 units. We therefore compare the minMSE method with min-max-t-statistic-rerandomization when 7 experimental groups are desired with a group size of 30; yielding a total sample size of 210 units.

We implemented min-max-t-statistic rerandomization by modifying the code from Bruhn and McKenzie, 2009, where, instead of two groups, we assign seven, and when regressing on the treatment indication variable to obtain t-statistics, our indication variable consists of 7 groups instead of 2.

Experimental Variation of Balancedness with Respect to Pretreatment Characteristics To experimentally assess the degree, to which balance affects precision of estimation in a real-world setting with real treatment effects that are unknown at the time of treatment assignment, we vary balancedness of pretreatment characteristics. To maximize the variation in balance, we use purely random treatment assignment for lowest possible balancedness. For the experimental group that we 'treat' with balance, we use the minMSE method. In order to be able to study the decline in balance with an increasing number of experimental groups that are to be formed, we vary the number of experimental groups from two to seven, each group consisting of 20 observations; see Table 11 for details.

Implementation We draw sub-samples of our total sample consisting of 3305 schools.We randomly draw 12 sub-samples consisting of different numbers of schools (see Table 11), and for those sub-samples we draw—without repetition from the whole remaining sample of schools—groups of equal sizes that were comparable to the ones randomly drawn. [15]

In this way, we obtained 24 sub-samples consisting of 12 pairs of pair-wise comparable sub-samples. Of each pair, we randomly allocated one sub-sample to the minMSE approach, and the other sub-sample to a comparison method. Then, within sub-samples, treatment allocation was performed by the assigned method for up to 7 experimental groups, which is the maximum number of groups needed (see Section 3.1 on treatments). This design is illustrated in Table 11.

After having allocated the schools in 12 subgroups (matching vs. minMSE sub-sample, rerandomization vs. minMSE sub-sample and ten randomization vs. minMSE sub-samples) to experimental groups, around one third of the sample was not yet assigned to an experimental group. Taking into account the treatment assignments already made, using the minMSE method, we allocated those remaining schools to the control

[15] Comparability of the groups of schools—or balancedness among the covariates or observables of the groups—was achieved with an algorithm using the same decision criteria as the minMSE approach.

and the treatment groups, with the restriction to have the group sizes as equal as possible and the goal to achieve overall balance across treatments in the whole sample. The resulting assignment to experimental groups is balanced as assessed with the omnibus test by Hansen and Bowers, 2008: the p-value when testing the null hypothesis of balanced groups is above 0.66.

Measures of Balance and Precision We assess balance of pre-treatment characteristics in two ways. The first way is rather a measure for imbalance and has been introduced in Section 2: the Overlap condition. Although overlap is actually needed to identify the conditional average treatment effect, it can be used as a measure of imbalance, by counting the cases in which the overlap condition is not fulfilled for a certain characteristic. This measure thus focuses rather on imbalance than on balance, but as we are interested also in rather adverse cases, we measure balance by assessing the overlap.

The second way to assess balance of multivariate information in treatment groups relies on the test of imbalance developed by Hansen and Bowers (2008). We compare p-values of the test, where the p-value corresponds to the likelihood that the statistic of multivariate differences is due to pure randomness, that is: the lower the p-value, the higher imbalance and the lower is balance.

One measure of precision of estimation is the bias of the estimation. This notion of precision we will label the precision of the experiment. In this sense, an estimation is precise if it is close to the true value that would be obtained by measuring the effect with the whole population or by repeating the experiment sufficiently often with different sub-samples, thereby averaging out any influence that is not due to the treatment. Taking advantage of the fact that the schools in our experiment actually constitute almost the whole population of schools in North-Rhine Westphalia, we can compute the bias of every estimation. We first compute the treatment effects for the whole population. Then we estimate the effects for the sub-samples. Finally, we compute the bias for every sub-sample, i.e. the deviation between the estimated effect for the sub-sample and the effect for the whole population of schools.

The second measure of precision of estimation that we apply is linked to statistical significance.[16] Our measure of power consists of higher or lower p-values of the treatment effect estimations.

4 Results

The result section is organized in the following way. We first describe our data and present descriptive statistics. Second, we compare the minMSE treatment assignment method to alternative assignment techniques with respect to pretreatment blance of covariates (overlap condition and omnibus test of imbalance), precision of the estimator (bias and p-values of treatment effect coefficient), and investigate how balancedness of covariates affects precision. Thereafter, we analyze whether a self-selection bias into participation in experiments exists at the institution level and how to attract gatekeepers' attention. Finally, we present survey

[16] Although we believe that the precision of the experiment should always be considered first, because a wrong estimation that is significantly estimated might even be dangerous, we acknowledge that for many researchers, the statistical interpretation is of great importance as well.

data on gatekeepers perceived usefulness of academic research.

4.1 Data and Descriptive Statistics

We gathered a rich set of official data on observable characteristics at both the municipality level and the school level. School level data were provided by the statistical office of NRW individually for our study, and municipality level data are publicly available from the German statistical offices. These data include—at the school level—inter alia the school type, number of students, average age of teachers, compulsory teaching hours of teachers, and migration background information of students and their parents. Data at municipality level comprise inter alia the number of inhabitants, unemployment rate, election results, land prices, composition of the workforce, and the social index of the municipality. A complete list and detailed description of the background characteristics can be found in Appendix D.

Analyzing response rates of the recruitment experiment, Panel A in Table 1 summarizes whether schools did not respond, actively opted out, showed "light" interest, this is, clicking on the link indicating to be contacted again later, or responded positively ("strong" interest). We observe that most schools did not respond to our inquiry, ranging from 71.7% in the (pooled) parental involvement treatment to 78.2% in the pooled E-learning treatment. Active opting-out is highest in our baseline treatment (20.6%) and lowest in the E-learning treatment (13.5%). Positive response rates are lowest in the integration of migrant children treatment (3.7%) and highest in the baseline treatment (6.3%), which might be due to the fact that schools simply had to answer a questionnaire without the commitment to participate in an experiment.

These response rates are comparable to the response rates observed in Riener and Wagner (2019) and Fischer and Wagner (2018), and Wagner (2016)—studies conducted with schools in NRW. As apparent from Panel A of Table 2, the non-response rates in secondary schools vary from 67.1% in Riener and Wagner (2019) to 76.9% in Fischer and Wagner (2018), with the non-response rate of this study lying in between (72.4%). Addressing elementary schools, we observe higher non-response rates in Wagner (2016) (86.12%) compared to this study (76.8%). Note, however, that stakes, i.e. the effort needed for participation in the study, of the experiments vary from very low (this study) to high as in Fischer and Wagner (2018). In this light, the response rate observed in this study is comparable to other studies with schools in NRW.

Who responded to the recruitment email? We sent our inquiry to the schools' official email address, and asked for the respondent's position within the school. 840 schools responded to our recruitment email by clicking one of the three links provided and 188 ($\sim$22%) also answered the following questionnaire. As can be seen in Panel B of Table 1, inquiries are answered by headmasters directly most of the time (73.40%), followed by the dean of students (12.23%).

Tables 3 and 4 present descriptive statistics on background characteristics of schools and municipalities

we will later use in our analyzes. Columns (1)-(3) show means of the three treatment groups (e-learning, parental involvement, and integration of migrant children), column (4) describes our control group (scientific contribution), and column (5) shows pooled statistics for the groups in columns (1)-(4). Overall, we observe that differences between treatment groups and the control group are small for school and municipality characteristics and moreover economically insignificant.[17] Hence, it our randomization procedure was successful.

4.2 From Balance to Precision and the Role of Treatment Assignment Methods

In this section, we first discuss the results on pre-treatment balancedness of covariates with respect to two criteria, overlap and a test to detect imbalance due to Hansen and Bowers (2008). Then, we assess the differences in precision of estimation of treatment effects due to the different considered treatment mechanisms. As measures of precision, we use the bias of an estimation and the p-value of the coefficient of the treatment effect resulting from an estimation of the latter. Lastly, we present results on the link between balance and precision on the internal margin of balance, relating the degree of balance with the degree of precision.

4.2.1 Balance

Overlap Figure 1 shows the comparison of purely random treatment assignment and minMSE treatment assignment with respect to balance as measured by overlap (see Assumptions 2 and 4 in Section 2). For this measure, we consider five of the variables used for treatment assignment: all categorical information about schools (type of school, authority type, gender of the headmaster, municipality ID), plus a discretized version of the number of pupils using three equally populated bins. The remaining data are constant across municipalities, and the municipality ID is already included in the variables considered. Therefore, those data are excluded as they would distort the result. Yet, this means that the balance described here is an upper limit of what we would see if we included all variables considered for treatment assignment, and the difference in balance might be interpreted as a lower limit.

We consider the overlap condition as fulfilled for a level of a variable (say "female" of the variable "gender"), if this characteristic is represented in all possible groups. In Draw three, there are seven groups to be formed, whereas in Draw 12, the characteristic is to be distributed and thus to be found in only two groups (see Table 11 for details). In some cases, there are more groups to be formed than a certain characteristic is represented in the respective sample. In these cases, we consider the overlap condition as fulfilled if the characteristic is found in the maximal possible number of groups.

[17] Although we *know* that all differences are actually due to randomness, sometimes, p-values resulting from testing for non-random differences are reported. We find no difference in school or municipality characteristics that would imply significance at the 5% level. The difference between the age of teachers in the E-Learning treatment group and the control group would, without controlling for multiple testing, imply significance at the 10% level (p = 0.064). With respect to municipality characteristics, the difference in election outcome for the christian democratic union (CDU) between the E-Learning and parental involvement treatment groups and the control group would, without controlling for multiple testing, imply significance at the 10% level (p = 0.061 and p = 0.093, respectively).

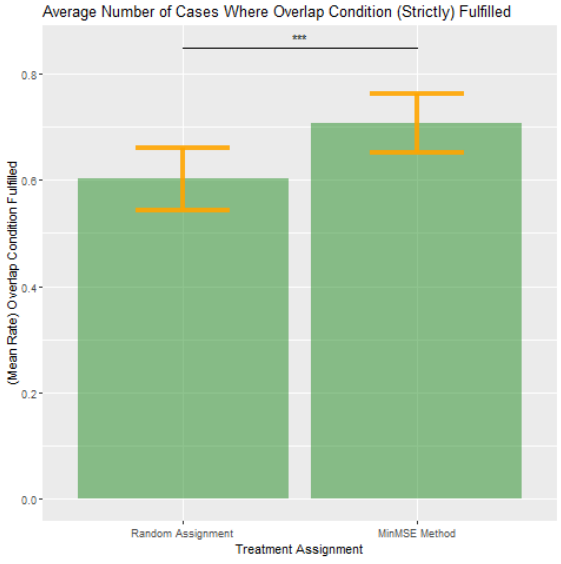
Figure 1a compares how often the overlap condition is fulfilled in the samples where treatment assignment was performed either completely at random or with the minMSE method. Considering all draws, variables and characteristics, the overlap condition is fulfilled in 60% of all cases when assigning treatment purely random, and in 71% of the cases when relying on the minMSE method. The difference in fulfilling the overlap condition between the two treatment assignment methods is significant (Chi-squared test, p-value < 0.001).

Figure 1b compares the average share of fulfillment of the overlap condition. Consider the case of the variable gender. If males were assigned to three of the six groups, but not to the three others, although in the total sample more than six males were present, the share of fulfillment of the overlap condition would be .5 for this variable and this characteristic. For every combination of draw, variable and characteristic of a variable, we obtain one share of fulfillment.

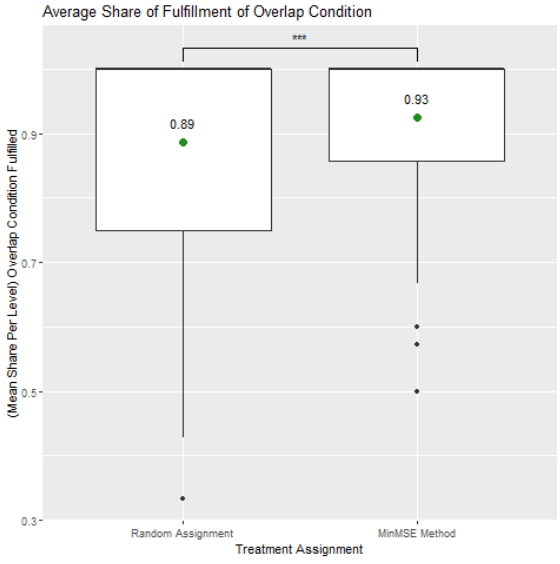
Figure 1b shows that, on average, both treatment assignment approaches perform relatively similar. However, given that the maximum share is 1, the difference is bigger than it seems. Yet, the more striking result is the difference in variance of this share. The 25% quantile (minimum) of the distribution of the share of fulfillment of the overlap condition resulting from the minMSE method is .86 (.5) compared to .75 (.33) when relying on purely random assignment. In that sense, proper treatment assignment may be understood as an "insurance" against adverse "draws" in which balance is really not that good, as indicated by the minimum shares of two methods.

Figure 1: Comparison of Pretreatment Balance: Overlap Condition

(a) Average Number of Cases Where Overlap Condition Fulfilled



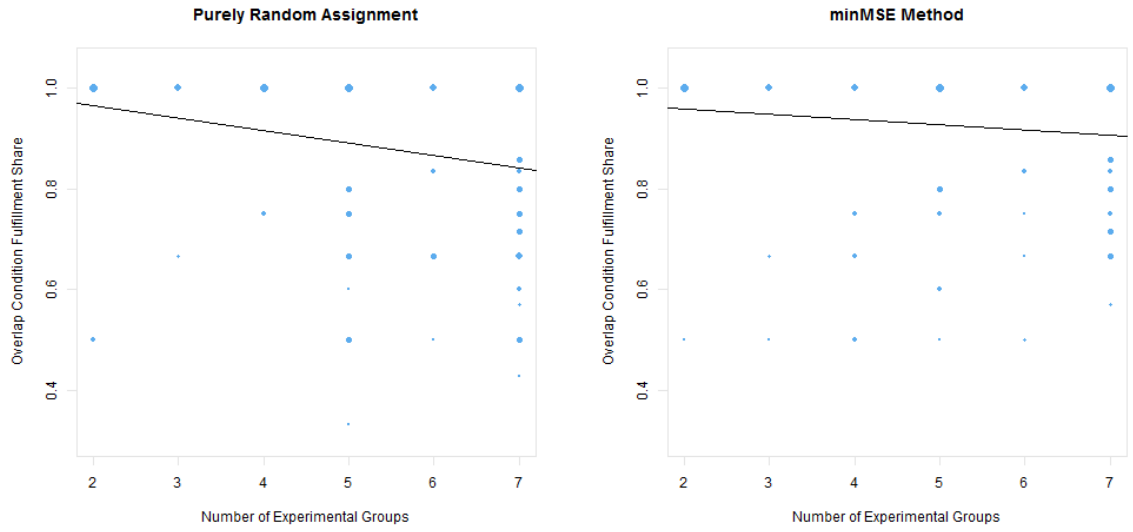
(b) Average Share of Fulfillment of Overlap Condition



*Note: The first graph (Figure 1a) compares the average number of cases, in which the overlap condition is fulfilled. Generally, the overlap condition is considered as fulfilled, if a characteristic of a variable is found in all treatment groups. *** denotes significance of a chi-squared test at the 1% level. The second graph (Figure 1b) compares the average ratio of treatment groups, in which a characteristic is found, to the total number of treatment groups to be assigned in a draw (in the general case). *** denotes significance of a rank sum test at the 1% level.*

As expected, the success rate is higher for the draws with less groups to be assigned. However, there is a significant difference (as indicated by an interaction term, p-value < 0.001) between the treatment assignment methods in the decay of the success rate for the overlap condition with increasing number of groups to be formed. The decay in balance as measured by the overlap condition is 1% per additional treatment group when using the minMSE method, and nearly 2.5 times as much for assignment of groups purely at random; see Figure 2.

Figure 2: Comparison of Pretreatment Balance: Average Share of Fulfillment of Overlap Condition with Increasing Number of Experimental Groups



Note: These graphs present the decay of balance of pre-treatment characteristics as the number of treatment groups to be formed increases for the two treatment assignment approaches considered. Here, balance is measured by the overlap condition (see Assumptions 2 and 4 in Section 2). The difference in slopes (decay) (about -2.5 for purely random assignment vs. -1 for the minMSE Method) is significant at the 1% level.

Omnibus Test of Imbalance (Hansen and Bowers, 2008) Based on the omnibus test of imbalance due to Hansen and Bowers (2008), our second measure of pretreatment balance considers all variables used for treatment assignment. It bases on a statistic that accounts for correlation between the specified variables, thus "corrects" for comparison of multiple variables accross control and treatment group and summarizes all differences in one single statistic that approximately follows a chi-squared distribution.

We run the test for every combination of treatment and control group possible in a draw. Table 12 summarizes these results by reporting the minimal p-value of any comparison between control and treatment group in a draw. Note that low p-values imply low balance, whereas higher p-values indicate better balance. In none of the groups, the null hypothesis of balance is rejected at conventional significance levels. A Wilcoxon rank-sum test statistically confirms the obvious observation that minimal p-values are smaller for the minMSE method – even when including pair-wise matching and rerandomization as comparison methods. The null

hypothesis of equality of balance (as measured by p-values) is rejected with a p-value < 0.01.[18]

The relation between balance based on the omnibus test by Hansen and Bowers (2008) and the number of groups to be formed is the same as when measuring balance with the overlap condition: Balance decreases with increasing number of groups to be assigned. The pearson correlation coefficient between number of groups and the minimal p-value is $\rho = -.91$ (p-value < 0.001) when pooling pure randomization, matching and rerandomization, and it is $\rho = -.49$ and non significant for the minMSE method.

Result 1. ***Balance:*** *Pretreatment balance between control and treatment groups is significantly higher when groups are assigned using the minMSE method compared to alternatives: purely random treatment assignment, pair-wise matching and min-max t-statisic rerandomization. The degree of balance significantly decreases with the number of experimental groups that are to be assigned when using purely random treatment assignment but not when using the minMSE method. The minMSE method allows to assign about 2.5 more groups with the same decrease in balance as compared to pure randomization.*

4.2.2 Precision

One way to assess precision is the bias, i.e. the difference between the true value and the estimated value. We use the bias as one measure to compare the performance of the different treatment assignment mechanisms with respect to precision. To this end, we consider the three main outcomes in this paper: Response (yes/no), positive response (yes/no) and whether the questionnaire was filled (yes/no) and pool the results. Following Bruhn and McKenzie (2009), we also compare precision as indicated by the p-values of the estimations of the treatment effect.

Precision: Bias For every treatment (i.e. every topic with or without incentivization), we can estimate its treatment effect on our main outcomes using the full sample. Assuming this estimation corresponds to the true value, the absolute bias of an estimated treatment effect is simply the absolute difference between this estimated value and an estimation that only uses the observations in a given subsample. We consider all treatment effects resulting from an estimation using the full sample with a statistically significant estimated coefficient at least at the 10% level. Of the estimations using subsamples, we include all estimations, where the p-value of the treatment effect estimation indicates more precision than pure randomness, i.e. where it is below 0.5.

Figure 3a summarizes the result. For comparison, the absolute bias is expressed in standard deviations of the respective treatment effect estimations, i.e. the standard deviations of all estimates of the treatment effect of the respective topic with or without incentivization (e.g. e-learning with incentivization). The (absolute) bias differs significantly between the subsamples in which treatment was assigned purely at random and those in which the minMSE method was used for treatment assignment (Wilcoxon rank sum test, $p < 0.007$). On

[18] This finding is robust at least at the same significance level to not aggregating the p-values over the treatment groups of a draw by using the minimal value and to aggregating by taking the mean p-value of the ombnibus test of imbalance instead of the minimal p-value of any comparison between treatment group and control.

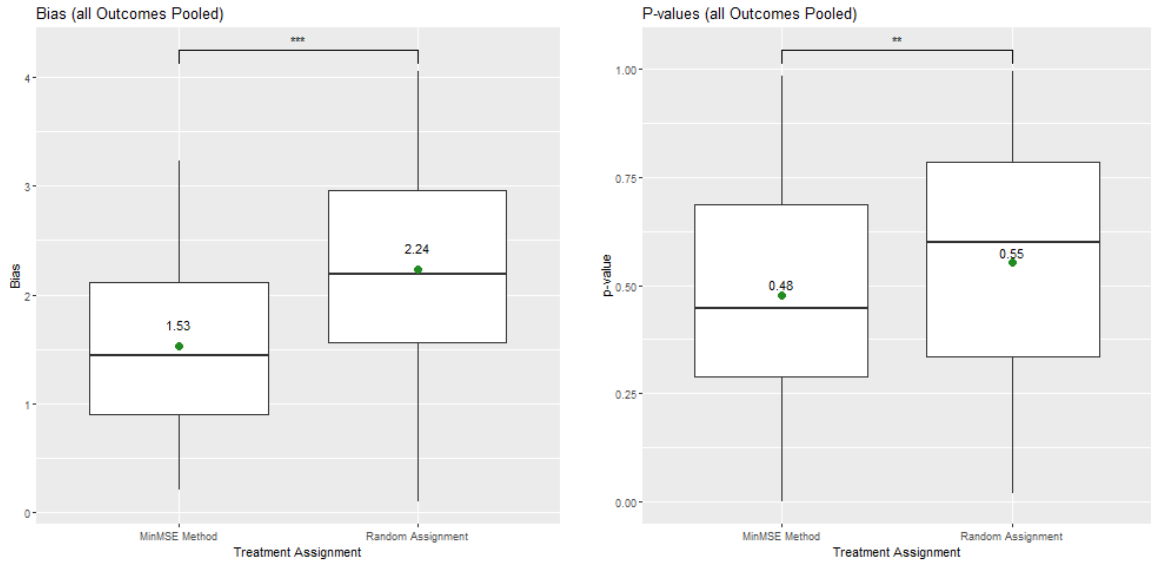
average, the bias is nearly 1.5 times as large when assigning treatments purely at random compared to when using the min MSE method. Moreover, the median bias of estimations in the "purely random assignment" samples (2.2) is larger than the 75% quantile in the "minMSE assignment" samples (2.1; median 1.4).

Precision: P-Value of Treatment Effect Coefficient We can measure precision also by the p-value of an estimation. Again, we consider all the estimations using subsamples that estimate treatment effects that were estimated with a statistically significant coefficient at least at the 10% level with the full sample. Figure 3b summarizes the result. The estimations in the samples, where treatment was assigned using the minMSE method, differ significantly in their p-values when estimating significant treatment effects from the estimations in the samples, where treatment groups were assigned purely at random (Wilcoxon rank sum test, $p < 0.031$). The mean (median) p-value of estimations in the "purely random assignment" samples is .55 (.60), where it is .48 (.45) in the "minMSE assignment" samples.

Figure 3: Comparison of Precision: (Absolute) Bias and P-Values of Estimations

(a) Bias in Estimation

(b) P-Values of Estimations



*Note: This graph shows precision in estimation of treatment effects considering three outcomes (response, positive response and whether a survey was filled in) when assigning treatments purely at random compared to using the minMSE method. Figure 3a presents the distribution of (absolute) bias in estimating significant treatment effects. Estimated treatment effects using subsamples are subtracted from the treatment effect estimated using the full population of schools; the absolute value of the difference is the bias shown here, given that the p-value of the estimation in the subsample is below .5 (i.e. more precise than purely random). Figure 3b presents the distribution of p-values of the estimations of the treatment effects using the different subsamples. Stars indicate resuls from a Wilcoxon rank sum test, where ***/** denotes significance at the 1%/5% level.*

Result 2. ***Precision:*** *Precision is higher when using the minMSE method. Assigning treatment with the minMSE method reduces the bias in estimation of treatment effects by 33% compared to purely random treatment assignment.*

4.2.3 The Relation Between Balance and Precision and the Role of Treatment Assignment

In Section 4.2.1 we have shown that balance is higher when assigning units to treatment groups using the minMSE method as compared to purely random treatment assignment – independently of the measure of balance that we have used. In Section 4.2.2, using two measures of precision, we have shown that also precision of treatment effect estimations is higher when using subsamples in which treatment was assigned with the minMSE method as compared to those estimations based on subsamples where treatment was assigned purely at random. As we have kept everything else as constant as possible between the subgroups in a draw except for the treatment assignment mechanism, which we have shown to affect the balance, this already serves as evidence that balance increases precision considerably in our real world setting – by reducing the bias by 33% on average with the same effect on the median.[19]

This finding is supported by a correlation analysis on the same data: The Pearson correlation coefficient between pre-treatment balance as expressed by p-values corresponding to the omnibus test of (im)balance by Hansen and Bowers (2008) (the higher, the better) and (absolute) bias in treatment effect estimations (the lower, the better) is $\rho = -.42$ (p-value < 0.001).

An increase of .1 in the p-value associated with imbalance due to the test by Hansen and Bowers (2008) results in an increase in bias of .25 standard deviations (a ninth of the average bias in the "purely random" samples, or a sixth in the "minMSE" samples). Using aggregated values of the bias at the draw/treatment assignment level,[20] the average share of fulfillment of the overlap condition significantly predicts bias: We find that a 10 percentage point higher fulfillment share of the overlap condition is associated with a more than .4 standard deviations smaller bias.

Result 3. ***The Relation between Balance and Precision:*** *The degree of balance increases the degree of precision. The exact relationship depends on the measures used for balance and for precision. A 33% lower bias has been attained due to better balance by using a proper treatment assignment method, the minMSE method, in our experiment. Given that the treatment effects in our setting are independent of the covariates used, this result may likely be a lower limit of what can be expected in different settings.*

4.3 Self-selection of gatekeepers into participation?

Table 5 presents the results on selection of institutions, i.e. schools, into participation. We present regression results—marginal effects from probit estimations—where the dependent variable indicates any response to our request, this is, clicking on one of the three links in the recruitment email indicating interest in participating or actively opting out. Our explanatory variables are presented in two groups: (a) school level variables and (b) municipality level variables. We control for multiple testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (2005). Pooling all treatments, there are no school or

[19] Note that this is likely to be a lower limit, since there is no (significant) interaction between covariates and the treatments; see Section 4.4

[20] We use aggregate values on this level, as the fulfillment share can only be meaningfully measured on the draw level; this is the measure used in Section 4.2.1.

municipality characteristics that determine whether a school is more or less likely to respond to our inquiry. When splitting the sample into the different treatment groups, we find that schools in the E-Learning treatment group that have a higher share of students with migration background are less likely to respond to our request and schools with a higher share of full time employed teachers show a slightly higher responsiveness in the general scientific interest treatment. These are the only two significantly estimated predictors of the 37 investigated predictors. Note that, although we corrected for testing each predictor five times, we would still expect $37 \times .05 = 1.85$ predictors to be significant at a 5% significant level. As these coefficients are also economically negligible, we conclude that these results provide no evidence for selection on observable characteristics of the sample that responded.

Table 6 examines school and municipality characteristics that determine positive responses of schools to our request. We find a similar picture, this is, in the pooled regressions not a single characteristic predicts differences in responding positively. Schools' average compulsory teaching hours are slightly positively related with a positive respond in the E-learning treatment group, and in the Scientific Contribution treatment group, schools in municipalities in which smaller parties received a higher vote share are more likely to respond positively. Again, these results provide no evidence for self-selection into participation—only two predictors of the 37×5 estimated coefficients are significant. Administrative school and municipality data regularly are non informative about headmasters' (unobserved) personal characteristics, e.g., headmaster quality or open mindedness which might determine headmasters' reaction to our inquiry. Even if all schools that expressed interest to participate in our study were selected into that sample due to headmasters' personal characteristics, our results show that these are not different from the non-participating schools. Altonji, Elder, and Taber (2005) formalized the idea that "selection on the unobservables is the same as selection on the observables". Following the idea of Altonji, Elder, and Taber (2005) it is unlikely that (headmasters') unobserved characteristics are uncorrelated with all of our observed school and municipality characteristics such as the social index of the municipality, the school size, land prizes or share of migrant pupils. Not finding evidence of self-selection on these observable characteristics indicates that there is no self-selection which is determined by headmasters' quality. However, the idea of selection on observables to construct a proxy for selection on unobservables relies on strong assumptions—e.g., observables and unobservables that are relevant for an outcome are large in number, chosen at random from the full set of factors that determine that outcome, and no single element dominates the outcome.

Result 4. ***Self-selection:*** *We do not find evidence of selection on observables into participation.*

4.4 Attracting gatekeepers' attention?

We now analyze how to attract headmasters' attention, this is, whether their willingness to respond depends (i) on the research topic the proposed scientific study wants to answer, and (ii) the opportunity to receive an extrinsic financial incentive.

Table 7 presents the effect of highlighting and varying the research topic of the proposed study in the initial email on the headmasters' willingness to respond positively, to respond in any way, and to fill out the survey. Columns (1)-(3) report on the pooled effect of being suggested to contribute to a specific research topic compared to being asked to contribute to scientific cooperation in general. Columns (4)-(6) further differentiate by research topic.[21] In general, we find that explicitly stating a research topic in the initial email significantly increases headmasters' willingness to respond positively compared to headmasters who were only asked to participate in a survey on scientific cooperation. However, we do not find a significant effect on whether headmasters responded at all (positive responses and negative responses) and on whether or not they filled out the survey following their response. A positive treatment effect for the number of positive responses and no change in overall responsiveness implies that active opting out decreased. Turning to the proposed research topics, we find that the research question that the study wants to answer does matter for headmasters' responsiveness. Proposing a research topic on getting parents involved or on how to better integrate migrant children increases headmasters' willigness to participate, but there is no statistical significant increase for the topic E-learning. On the contrary, we find that overall responsiveness for the E-learning treatment significantly decreased and that headmasters were less likely to fill out the survey.

Result 5. ***Research topic:*** *The topic of the proposed research question matters to attract headmasters' attention.*

With the incentive treatment, we intended to attract headmasters' attention by offering their schools the possibility to win 700 Euros which could, e.g., spent on teachers' training or teaching material. Table 8 shows whether schools being eligible to receive the financial reward change their responsiveness with respect to positive responses, any response, and filling out the survey. Panel A shows the results pooling all research topics and Panels B, C, and D present results for the respective research topic treatment. Overall, we find no significant effect of the extrinsic incentive on any of the outcome variables neither for the pooled sample nor for each of the respective research topics.[22] As the share of the financial incentives on schools' yearly budget for teacher training varies by school (see Table 9), we analyze if there are heterogeneous responses by the share of the incentive; see Table 10. We find that the size of the share of the yearly budget for teacher training that we offer does not matter to increase headmasters' responsiveness for none of the outcome variables.[23]

Result 6. ***Monetary incentives:*** *Extrinsic financial incentives do not attract headmasters' attention.*

[21] Tables 15 and 16 show that results are robust to calculating bootstrapped standard errors or applying randomization inference.

[22] Tables 17 and 18 show that results are robust to calculating bootstrapped standard errors or applying randomization inference.

[23] Tables 19 and 20 show that results are robust to calculating bootstrapped standard errors or applying randomization inference.

4.5 Survey

Gatekeepers could access further information on the proposed research question by clicking on a link in the recruitment email. Moreover, all headmasters clicking on one of the three links indicating their willingness or non-willingness to participate (opt-out, light interest, opt-in) were guided to a short questionnaire asking about how useful they perceive academic research (see Online Appendix D for the additional information on the linked page and the survey).

Figure 4a shows that only a small fraction (approximately 4%) of schools was actively seeking for additional and more detailed information on the proposed research topic. Moreover, there is no significant difference between the research topics. However, schools in the E-learning treatment group tend to ask more often for further information than schools in the Parental Involvement or the Integration Migration treatment group. Putting this into perspective with the finding that schools in the E-learning treatment group were less likely to respond (positively) compared to schools in the two other treatments, it seems that this was an informed choice; yet, we cannot provide statistical evidence for this conjecture.

Figure 4b shows the share of headmasters that filled in the survey. We find that representatives of schools with a positive response to our inquiry were also significantly more likely to fill in our questionnaire than those from schools indicating light interest and those from schools that actively opted out (Wilcoxon ranksum test, $p < 0.01$)). Overall, roughly 32% of schools whose representative responded in any way also (partly) filled out the questionnaire.

In the survey, we asked inter alia whether schools think that academic research could be integrated in schools' daily life, whether they think academic research is informative for educational policy makers, whether headmasters themselves are generally interested in the results of academic research, whether they find the proposed research topic interesting, whether their school has no personnel capacity to participate in the study, and whether they think that schools receive too many inquiries by researchers. Headmasters were asked to indicate whether they agree or disagree with these statements on a 1 (disagree) to 10 (agree) scale. The answers are shown in Figures 5a and 5b. In line with the regression results in Section 4.4, it seems that headmasters (or their representatives) perceive the research topic they were suggested as interesting. However, they also agree with the statement that there are too many inquiries by researchers to participate in a study and that they lack personnel resources for participation. Moreover, while headmasters seem to be generally interested in the results of academic reseach and think that academic research is useful to inform educational policy makers, they tend to be less optimistic about the possibility to integrate research results into daily school life.

5 Conclusion

The purpose of this paper is twofold, first to assess the relevance of comparable treatment groups for precision and power in treatment effect estimation by use of a new treatment assignment method (min MSE treatment assignment), and second to investigate selection in field experiments. We did so in a particularly relevant environment, the educational system, where the willingness to participate in studies is regularly very low, thus comparability of participating schools to non-participating schools is questionable, and balancedness is hard to achieve through pure randomization—in particular when the unit of randomization is the school level.

We contacted a majority of schools in the largest country of Germany—North-Rhine-Westphalia—to assess the interest of participation in field experiments. Within our main treatment arms we varied the topic and a financial reward for participation. Behind veil of this experiment, we divided the whole sample into smaller, comparable sub-samples, and use different treatment assignment mechanisms to vary the degree of balance (of covariates) in these sub-samples. This allows us to study the influence of balance of covariates on precision of treatment effect estimation.

In many institutional settings, participation in evaluations is voluntary, so an evaluation of the full population of interest—such as in for example Crépon et al., 2013—is hardly feasible due to political constraints. Hence, assessing the severity of selection becomes a crucial stepping stone for the credible evaluation of policies through experimental methods.

Using a broad range of observables at the school and municipality level, we observe no selection into treatment effects. This appears to be good news for researchers trying to reach a broad sample of schools of all backgrounds and geographical locations.

References

Abadie, Alberto and Guido W. Imbens (2006). "Large sample properties of matching estimators for average treatment effects". In: *Econometrica* 74.1, pp. 235–267.

Abeler, Johannes and Daniele Nosenzo (2015). "Self-selection into laboratory experiments: pro-social motives versus monetary incentives". In: *Experimental Economics* 18.2, pp. 195–214.

Allcott, Hunt (2015). "Site selection bias in program evaluation". In: *The Quarterly Journal of Economics* 130.3, pp. 1117–1165.

Altonji, Joseph, Todd Elder, and Christopher Taber (2005). "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools". In: *Journal of Political Economy* 113.1, pp. 151–184.

Anderson, Jon et al. (2013). "Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples". In: *Experimental Economics* 16.2, pp. 170–189.

Angrist, Joshua and Jörn-Steffen Pischke (2009). *Mostly harmless econometrics: An empiricist's companion.* 1st ed. Princeton University Press.

Athey, Susan and Guido Imbens (2017a). "Chapter 3 - The econometrics of randomized experiments". In: *Handbook of Field Experiments.* Ed. by Abhijit Vinayak Banerjee and Esther Duflo. Vol. 1. Handbook of Economic Field Experiments. North-Holland.

— (2017b). "The state of applied econometrics: Causality and policy evaluation". In: *Journal of Economic Perspectives* 31.2, pp. 3–32.

Belot, Michele and Jonathan James (2014). "A new perspective on the issue of selection bias in randomized controlled field experiments". In: *Economics Letters* 124.3, pp. 326–328.

— (2016). "Partner selection into policy relevant field experiments". In: *Journal of Economic Behavior & Organization* 123, pp. 31–56.

Benndorf, Volcker, Claudia Möllers, and Hans-Theo Normann (2017). "Experienced vs. inexperienced participants in the lab: Do they behave differently". In: *Journal of the Economic Science Association* 3.1, pp. 12–25.

Bergman, Peter and Eric Chan (2019). "Leveraging patrents through low-cost echnology: The impact of high-frequency information on student achievement". In: *Journal of Human Resources* forthcoming.

Bruhn, Miriam and David McKenzie (2009). "In pursuit of balance: Randomization in practice in development field experiments". In: *American Economic Journal: Applied Economics* 1.4, pp. 200–232.

Burtless, Gary (1995). "The case for randomized field trials in economic and policy research". In: *The Journal of Economic Perspectives* 9.2, pp. 63–84.

Charness, Gary, Uri Gneezy, and Michael Kuhn (2013). "Experimental methods: Extra-laboratory experiments-extending the reach of experimental economics". In: *Journal of Economic Behavior & Organization* 91, pp. 93–100.

Cleave, Blair, Nikos Nikiforakis, and Robert Slonim (2013). "Is there selection bias in laboratory experiments? The case of social and risk preferences". In: *Experimental Economics* 16.3, pp. 372–382.

Crépon, Bruno et al. (2013). "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment". In: *The Quarterly Journal of Economics* 128.2, pp. 531–580.

Czibor, Eszter, David Jimenez-Gomez, and John List (Jan. 2019). *The dozen things experimental economists should do (more of)*. Working Paper 25451. National Bureau of Economic Research.

Falk, Armin, Stephan Meier, and Christian Zehnder (2013). "Do lab experiments misrepresent social preferences? The case of self-selected student samples". In: *Journal of the European Economic Association* 11.4, pp. 839–852.

Fischer, Mira and Valentin Wagner (2018). *Effects of timing and reference frame of feedback: Evidence from a field experiment*. IZA Discussion Paper Series 11970. IZA - Institute of Labor Economics.

Frijters, Paul, Tao Sherry Kong, and Elaine Liu (2015). "Who is coming to the artefactual field experiment? Participation bias among Chinese rural migrants". In: *Journal of Economic Behavior & Organization* 114, pp. 62–74.

Greevy, Robert et al. (2004). "Optimal multivariate matching before randomization". In: *Biostatistics* 5.2, pp. 263–275.

Hansen, Ben and Jake Bowers (2008). "Covariate balance in simple, stratified and clustered comparative studies". In: *Statistical Science* 23.2, pp. 219–236.

Harrison, Glenn, Morten Lau, and Elisabet Rutström (2009). "Risk attitudes, randomization to treatment, and self-selection into experiments". In: *Journal of Economic Behavior & Organization* 70.3, pp. 498–507.

Harrison, Glenn and John List (2004). "Field experiments". In: *Journal of Economic Literature* 42.4, pp. 1009–1055.

Hotz, Joseph, Guido Imbens, and Julie Mortimer (2005). "Predicting the efficacy of future training programs using past experiences at other locations". In: *Journal of Econometrics* 125.1, pp. 241–270.

Humphreys, Macartan (Apr. 2019). *How to make field experiments more ethical*. The Washington Post, accessed October 16, 2019.

Kasy, Maximilian (2016). "Why experimenters might not always want to randomize, and what they could do instead". In: *Political Analysis* 24.3, pp. 324–338.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (May 1983). "Optimization by simulated annealing". In: *Science* 220.4598, pp. 671–680.

Kraft, Matthew and Todd Rogers (2015). "The underutilized potential of teacher-to-parent communication: Evidence from a field experiment". In: *Economics of Education Review* 47, pp. 49–63.

Krawczyk, Michal (2011). "What brings your subjects to the lab? A field experiment". In: *Experimental Economics* 14.4, pp. 482–489.

Lazear, Edward, Ulrike Malmendier, and Roberto Weber (2012). "Sorting in experiments with application to social preferences". In: *American Economic Journal: Applied Economics* 4.1, pp. 136–163.

Lu, Bo et al. (Feb. 2011). "Optimal nonbipartite matching and its statistical applications". en. In: *The American Statistician* 65.1, pp. 21–30.

Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan (2011). "Mechanism experiments and policy evaluations". In: *Journal of Economic Perspectives* 25.3, pp. 17–38.

Riener, Gerhard and Valentin Wagner (2019). "On the design of non-monetary incentives in schools". In: *Education Economics* 27.3, pp. 223–240.

Romano, Joseph and Michael Wolf (2005). "Stepwise multiple testing as formalized data snooping". In: *Econometrica* 73.4, pp. 1237–1282.

Rothe, Christoph (2017). "Robust confidence intervals for average treatment effects under limited overlap". In: *Econometrica* 85.2, pp. 645–660.

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Schneider, Sebastian and Martin Schlather (2017). *A new approach to treatment assignment for one and multiple treatment groups*. CRC Discussion Papers 228.

Slonim, Robert et al. (2013). "Opting-in: Participation bias in economic experiments". In: *Journal of Economic Behavior & Organization* 90, pp. 43–70.

Al-Ubaydli, Omar, John List, and Dana Suskind (May 2019). *The Science of using science: Towards an understanding of the threats to scaling experiments*. Working Paper 25848. National Bureau of Economic Research Working Paper No. 11277.

Wagner, Valentin (2016). *Seeking risk or answering smart? Framing in elementary schools*. DICE Discussion Paper Series 227. Düsseldorf Institute for Competition Economics.

A Tables

Table 1: Descriptive Statistics - Response Rates and Position of Respondent

Panel A: Response Rates by Treatment				
Treatment	*No response*	*Opted out*	*Light interest*	*Strong interest*
E-learning (N=955)	78.22	13.51	4.50	3.77
	(747)	(129)	(43)	(36)
Parental involvement (N=930)	71.72	17.31	5.70	5.27
	(667)	(161)	(53)	(49)
Integration migration (N=930)	74.52	15.27	6.56	3.66
	(693)	(142)	(61)	(34)
Scientific cooperation (N=490)	73.06	20.61	0.00	6.33
	(358)	(101)	(0)	(31)
Panel B: Position of Respondent				
Position (German)	*Position (English)*	*Absolute*	*Share*	*Cumulative*
Oberstudiendirektor	"Headmaster"	138	73.40	73.40
Studiendirektor	"Dean of Students"	23	12.23	85.63
Oberstudienrat	"Senior Teacher"	5	2.66	88.29
Studienrat	"Junior Teacher"	2	1.06	89.35
Referendar	"Trainee Teacher"	8	4.26	93.61
Sekretariat	"Office Staff"	5	2.66	96.27
Sonstige	"Other"	7	3.73	100.00

Note: Panel A summarizes the responses (in %; absolute number in parentheses) of schools depending on the treatment topic. Recipients of the recruitment email could reply by clicking one of three links indicating whether they do not want to participate in the experiment ("Opted out"), are interested but want to be contacted later ("Light interest"), or whether they could imaging participating ("Strong interest"). Schools not responding at all are summarized under "No response". Panel B contains information on the position of the respondent within their school, i.e. the person who filled out the questionnaire. Column (1) of Panel B is the German description of the respondent's position and column (2) is the English translation.

Table 2: Descriptive Statistics - Comparison of Response Rates

Panel A: Response rates

	Secondary Schools			Elementary Schools	
	This study	Riener and Wagner (2019)	Fischer and Wagner (2018)	This study	Wagner (2016)
No Response	72.40	67.06	76.92	76.77	83.13
	(1196)	(114)	(110)	(1269)	(207)
Responded	27.60	32.94	23.08	23.23	16.87
	(456)	(56)	(33)	(384)	(42)
Stakes in Study	very low	low	high	very low	low
Fisher's exact test for difference in response rates		p=0.152	p=0.281		p=0.027

Panel B: Contact type (Riener and Wagner, 2019)

	Letter	*Email*	*Letter+Email*
No Response	66.07	68.42	67.27
	(37)	(39)	(37)
Responded	33.93	31.58	32.73
	(19)	(18)	(18)
Fisher's exact test for difference in response rates		p=0.978	

Note: This table presents descriptive statistics on response rates for studies conducted in NRW by at least one of the authors. In panel A, we compare the response rates in this study to the response rates in Riener and Wagner (2019), Fischer and Wagner (2018), and Wagner (2016). The experiments by Riener and Wagner (2019) and Fischer and Wagner (2018) were conducted in secondary schools and Wagner (2016) conducted his study in elementary schools. The stakes of the studies varied from low-stakes (performance in a test not counting toward the final school grade) in Riener and Wagner (2019) and Wagner (2016) to high-stakes (grade in a high stakes exam) in Fischer and Wagner (2018). A two-sided Fisher's exact test tests for difference in response rates between the studies. Panel B presents response rates by contact type in the study of Riener and Wagner (2019). The authors contact schools either by email only, posted letter only, or both and recorded response rates. A two-sided Fisher's exact test tests for difference in response rates between contact types. In both panels, cell entries represent percentages, and the number of observations in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 3: Descriptive Statistics - School Level Data

	(1) E-Learning	(2) Parental Involvement	(3) Integration Migration	(4) Scientific Contribution	(5) Overall
Gender of headmaster	0.679	0.616	0.634	0.635	0.642
	(0.017)	(0.018)	(0.018)	(0.025)	(0.009)
Av. comp. teaching hours	21.213	21.054	21.108	21.154	21.130
	(0.085)	(0.091)	(0.084)	(0.121)	(0.046)
Students in day care	95.412	95.001	94.769	94.636	95.000
	(0.389)	(0.417)	(0.420)	(0.586)	(0.219)
Age of teachers (full time empl.)	39.858	39.951	40.004	40.555	40.028
	(0.215)	(0.227)	(0.224)	(0.315)	(0.119)
Students migration background	30.373	30.404	28.857	28.719	29.710
	(0.631)	(0.705)	(0.647)	(0.859)	(0.349)
Students migrated	6.385	6.375	6.163	6.604	6.352
	(0.253)	(0.257)	(0.268)	(0.347)	(0.137)
Parents migrated	28.468	28.515	27.349	26.800	27.919
	(0.593)	(0.662)	(0.625)	(0.807)	(0.331)
Number of students	329.547	323.894	331.552	332.293	328.928
	(9.050)	(8.992)	(8.960)	(13.138)	(4.835)
Female students	46.817	47.008	49.558	48.814	47.938
	(0.309)	(0.276)	(2.447)	(1.877)	(0.752)
Non-German students	7.195	7.230	7.194	7.368	7.230
	(0.265)	(0.274)	(0.269)	(0.326)	(0.141)
Non-German female students	3.400	3.412	3.305	3.404	3.377
	(0.131)	(0.133)	(0.121)	(0.152)	(0.067)
Share fullt time empl. teachers	55.915	56.000	55.315	55.675	55.735
	(0.548)	(0.578)	(0.524)	(0.820)	(0.297)
Students speak no German at home	15.987	16.461	15.266	15.070	15.782
	(0.502)	(0.555)	(0.500)	(0.663)	(0.274)
Number of classes	12.497	11.988	12.247	12.120	12.227
	(0.229)	(0.215)	(0.213)	(0.321)	(0.118)
N	955	930	930	490	3305
Proportion	0.289	0.281	0.281	0.148	1.000

Note: This table presents descriptive statistics on observable characteristics of schools by treatment. Cell entries report the group means and standard errors are reported in parentheses. Observable characteristics are described in more detail in Online Appendix D. For none of the differences in group means, the p-value associated with the corresponding t-statistic would imply significance at the 5% level. Without correcting for multiple comparisons only the difference in age of teachers between the E-Learning treatment group and the Scientific Contribution treatment would imply significance at the 10% level (p = 0.064).

Table 4: Descriptive Statistics - Municipality Level Data

	(1) E-Learning	(2) Parental Involvement	(3) Integration Migration	(4) Scientific Contribution	(5) Overall
Inhabitants	371.964	372.022	369.553	368.521	370.791
	(3.854)	(3.893)	(3.873)	(5.471)	(2.069)
Status married	48.457	48.538	48.501	48.479	48.495
	(0.045)	(0.042)	(0.042)	(0.060)	(0.023)
Unemployment rate	2.352	2.347	2.348	2.358	2.350
	(0.025)	(0.026)	(0.025)	(0.035)	(0.013)
Voter turnout 2013	73.238	73.218	73.408	73.205	73.275
	(0.111)	(0.116)	(0.113)	(0.157)	(0.061)
Elections party: CDU	42.911	42.964	43.258	43.587	43.124
	(0.207)	(0.217)	(0.219)	(0.301)	(0.115)
Elections party: SPD	30.091	30.082	29.826	29.646	29.948
	(0.176)	(0.185)	(0.183)	(0.248)	(0.096)
Elections party: FDP	5.189	5.202	5.255	5.174	5.209
	(0.038)	(0.040)	(0.040)	(0.054)	(0.021)
Elections party: Grune	6.932	6.899	6.907	6.854	6.904
	(0.055)	(0.055)	(0.057)	(0.076)	(0.030)
Elections party: DieLinke	5.309	5.275	5.240	5.240	5.270
	(0.034)	(0.035)	(0.035)	(0.050)	(0.019)
Elections party: Other	8.425	8.444	8.379	8.345	8.405
	(0.041)	(0.042)	(0.042)	(0.055)	(0.022)
German citizenship	93.145	93.137	93.121	93.027	93.118
	(0.057)	(0.057)	(0.057)	(0.081)	(0.031)
Education: Uni access	17.489	17.227	17.252	17.311	17.322
	(0.132)	(0.130)	(0.131)	(0.181)	(0.070)
Education: High School	27.099	27.051	27.063	27.042	27.067
	(0.095)	(0.096)	(0.097)	(0.129)	(0.051)
Land prices in 2014	134.365	134.081	133.941	133.947	134.104
	(1.259)	(1.279)	(1.267)	(1.762)	(0.676)
Share people aged 64 or older	20.511	20.534	20.528	20.512	20.522
	(0.026)	(0.027)	(0.027)	(0.036)	(0.014)
Religion: Other	27.400	27.167	27.003	27.218	27.196
	(0.200)	(0.202)	(0.202)	(0.284)	(0.108)
Religion: Protestant	27.951	27.413	27.125	27.545	27.507
	(0.426)	(0.410)	(0.410)	(0.594)	(0.223)
Male Workers	51.595	51.635	51.645	51.632	51.626
	(0.030)	(0.031)	(0.030)	(0.043)	(0.016)
Social index of municipality	30.033	29.750	29.391	30.005	29.769
	(0.508)	(0.517)	(0.517)	(0.723)	(0.274)
N	955	930	930	490	3305
Proportion	0.289	0.281	0.281	0.148	1.000

Note: This table presents descriptive statistics on observable characteristics of municipalities by treatment. Cell entries report the group means and standard errors are reported in parentheses. Observable characteristics are described in more detail in the Online Appendix D. For none of the differences in group means, the p-value associated with the corresponding t-statistic would imply significance at the 5% level. Without correcting for multiple comparisons only the the differences in election outcome for CDU between the E-Learning treatment group and the Scientific Contribution treatment group as well as between the Parental Involvement treatment group and the Scientific Contribution treatment group would imply significance at the 10% level ($p = 0.061$ and $p = 0.093$).

Table 5: Results - Self-selection (Dep. Var: Responded)

	(1) Pooled		(2) E-Learning		(3) Parental Involvement		(4) Migration		(5) Scientific Contribution	
School-level contr.										
Vocational (Gesamtsch.)	0.108	(0.051)	0.065	(0.097)	0.021	(0.090)	0.184	(0.106)	0.132	(0.099)
High School	0.210	(0.094)	0.177	(0.199)	0.186	(0.164)	0.267	(0.161)	0.183	(0.227)
Vocational (Hauptsch.)	0.050	(0.040)	0.009	(0.071)	0.052	(0.053)	0.160	(0.068)	-0.071	(0.100)
Vocational (Realsch.)	-0.010	(0.032)	0.019	(0.051)	-0.021	(0.063)	0.034	(0.063)	-0.171	(0.074)
Other school types	0.054	(0.034)	0.118	(0.068)	0.045	(0.043)	0.035	(0.078)	-0.031	(0.057)
Gender of headmaster	-0.013	(0.020)	-0.031	(0.025)	0.015	(0.028)	-0.021	(0.029)	-0.020	(0.036)
Av. comp. teaching hours	0.007	(0.004)	0.007	(0.004)	0.008	(0.006)	0.012	(0.008)	-0.003	(0.009)
Students in day care	0.005	(0.002)	0.004	(0.006)	0.003	(0.004)	0.006	(0.005)	0.000	(0.006)
Age of teachers (full time empl.)	-0.000	(0.001)	0.001	(0.002)	-0.002	(0.002)	-0.003	(0.002)	0.005	(0.002)
Students migration background	-0.000	(0.002)	-0.012*	(0.004)	-0.001	(0.003)	0.006	(0.004)	0.006	(0.004)
Students migrated	-0.000	(0.001)	0.005	(0.002)	-0.002	(0.003)	-0.002	(0.003)	-0.003	(0.003)
Parents migrated	0.001	(0.001)	0.012	(0.004)	0.002	(0.003)	-0.004	(0.004)	-0.002	(0.003)
Number of students	0.000	(0.000)	0.000	(0.000)	-0.000	(0.000)	0.000	(0.000)	-0.000	(0.000)
Female students	-0.001	(0.001)	-0.001	(0.002)	-0.001	(0.002)	-0.000	(0.000)	-0.001	(0.000)
Non-German students	0.000	(0.003)	-0.007	(0.005)	0.006	(0.005)	-0.002	(0.006)	0.000	(0.008)
Non-German female students	-0.004	(0.005)	0.005	(0.011)	-0.011	(0.011)	0.002	(0.011)	-0.013	(0.018)
Share fullt time empl. teachers	-0.000	(0.001)	-0.000	(0.001)	-0.001	(0.001)	-0.002	(0.002)	0.006***	(0.002)
Students speak no German at home	-0.000	(0.001)	0.002	(0.002)	-0.001	(0.002)	-0.001	(0.002)	-0.002	(0.002)
Number of classes	0.001	(0.003)	-0.007	(0.006)	0.012	(0.006)	-0.001	(0.009)	0.014	(0.008)
Munic.-level contr.										
Inhabitants	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	-0.000	(0.000)	0.000	(0.000)
Status married	0.029	(0.012)	0.026	(0.016)	0.037	(0.030)	0.035	(0.014)	0.024	(0.026)
Unemployment rate	0.003	(0.022)	0.030	(0.029)	0.047	(0.049)	-0.057	(0.030)	-0.027	(0.050)
Voter turnout 2013	-0.003	(0.004)	-0.005	(0.006)	-0.012	(0.008)	0.004	(0.006)	0.013	(0.007)
Elections party: SPD	-0.002	(0.003)	-0.001	(0.004)	-0.007	(0.004)	0.005	(0.005)	-0.001	(0.007)
Elections party: FDP	-0.025	(0.013)	-0.022	(0.018)	-0.033	(0.019)	-0.028	(0.018)	-0.020	(0.030)
Elections party: Grune	0.012	(0.010)	0.007	(0.013)	0.010	(0.014)	0.001	(0.015)	0.023	(0.024)
Elections party: DieLinke	-0.034	(0.017)	-0.025	(0.029)	-0.038	(0.032)	-0.028	(0.028)	-0.036	(0.023)
Elections party: Other	0.010	(0.009)	-0.011	(0.016)	0.013	(0.021)	0.014	(0.021)	0.020	(0.035)
German citizenship	0.006	(0.009)	0.008	(0.007)	0.012	(0.017)	0.007	(0.012)	-0.009	(0.018)
Education: Uni access	0.006	(0.007)	0.006	(0.009)	0.015	(0.012)	0.013	(0.008)	-0.008	(0.014)
Education: High School	0.001	(0.005)	0.003	(0.007)	-0.001	(0.011)	-0.001	(0.007)	-0.001	(0.012)
Land prices in 2014	0.000	(0.001)	0.002	(0.001)	0.001	(0.001)	-0.002	(0.001)	-0.001	(0.001)
Share people aged 64 or older	-0.014	(0.022)	-0.008	(0.029)	0.004	(0.042)	-0.015	(0.037)	-0.060	(0.040)
Religion: Other	-0.006	(0.006)	-0.010	(0.005)	-0.014	(0.009)	0.004	(0.006)	-0.004	(0.012)
Religion: Protestant	0.001	(0.002)	0.003	(0.002)	0.003	(0.003)	-0.003	(0.002)	0.004	(0.004)
Male Workers	0.014	(0.020)	0.067	(0.029)	0.014	(0.042)	0.014	(0.031)	-0.074	(0.040)
Social index of municipality	0.001	(0.001)	0.003	(0.001)	0.003	(0.002)	-0.001	(0.002)	-0.004	(0.002)
N	3305		955		930		930		490	

Note: This table summarizes the determinants of schools' responses to the recruitment email. Dependent variable: Any response = 0 if no response from school in any way; any response = 1 if school's respondent clicked on one of the three links in the recruitment email (opt out, light interest, strong interest). The coefficients are the marginal effects from a probit regression, standard errors in parentheses. Column (1) presents estimates using the full sample pooling all research topics. Columns (2) - (5) present estimates for each research topic separately resulting from corresponding sample splits. P-values are adjusted for multiple hypothesis testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (Romano and Wolf, 2005). * p<0.05, ** p<0.01, *** p<0.001.

Table 6: Results - Self-selection (Dep. Var: Positive Response)

	(1) Pooled		(2) E-Learning		(3) Parental Involvement		(4) Migration		(5) Scientific Contribution	
School-level contr.										
Vocational (Gesamtsch.)	0.054	(0.036)	0.077	(0.060)	0.020	(0.069)	0.070	(0.081)	0.004	(0.069)
High School	0.148	(0.061)	0.206	(0.108)	0.124	(0.109)	0.122	(0.139)	0.113	(0.125)
Vocational (Hauptsch.)	-0.057	(0.027)	-0.046	(0.049)	-0.094	(0.050)	-0.013	(0.055)	-0.062	(0.057)
Vocational (Realsch.)	-0.020	(0.023)	0.019	(0.037)	-0.021	(0.040)	-0.043	(0.045)	-0.063	(0.045)
Other school types	0.009	(0.020)	0.047	(0.032)	0.033	(0.034)	-0.032	(0.056)	-0.042	(0.051)
Gender of headmaster	-0.003	(0.012)	-0.008	(0.018)	0.013	(0.022)	0.004	(0.019)	-0.021	(0.022)
Av. comp. teaching hours	0.005	(0.002)	0.008*	(0.003)	0.001	(0.004)	0.009	(0.005)	0.007	(0.004)
Students in day care	0.004	(0.002)	0.005	(0.003)	0.004	(0.003)	0.003	(0.004)	0.003	(0.004)
Age of teachers (full time empl.)	-0.000	(0.001)	0.000	(0.001)	-0.003	(0.001)	-0.000	(0.002)	0.003	(0.002)
Students migration background	0.001	(0.001)	-0.004	(0.003)	-0.001	(0.002)	0.004	(0.003)	0.003	(0.002)
Students migrated	-0.000	(0.001)	0.003	(0.001)	-0.003	(0.002)	-0.003	(0.001)	0.001	(0.002)
Parents migrated	-0.000	(0.001)	0.005	(0.003)	0.001	(0.002)	-0.002	(0.003)	-0.002	(0.002)
Number of students	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	-0.000	(0.000)	0.000	(0.000)
Female students	-0.001	(0.001)	-0.000	(0.001)	-0.002	(0.001)	-0.001	(0.002)	-0.002	(0.002)
Non-German students	-0.001	(0.002)	-0.006	(0.003)	0.006	(0.004)	-0.006	(0.005)	-0.003	(0.005)
Non-German female students	0.003	(0.004)	0.004	(0.007)	-0.004	(0.009)	0.015	(0.009)	-0.003	(0.011)
Share fullt time empl. teachers	-0.000	(0.000)	0.000	(0.001)	-0.001	(0.001)	-0.001	(0.001)	0.001	(0.001)
Students speak no German at home	-0.001	(0.001)	-0.001	(0.001)	-0.000	(0.001)	-0.002	(0.001)	-0.000	(0.001)
Number of classes	0.001	(0.002)	-0.002	(0.003)	0.003	(0.004)	0.006	(0.007)	0.003	(0.005)
Munic.-level contr.										
Inhabitants	0.000	(0.000)	-0.000	(0.000)	0.000	(0.000)	-0.000	(0.000)	-0.000	(0.000)
Status married	0.008	(0.005)	0.007	(0.010)	0.022	(0.016)	0.009	(0.012)	0.022	(0.012)
Unemployment rate	-0.002	(0.010)	0.018	(0.015)	0.052	(0.023)	-0.039	(0.029)	-0.069	(0.028)
Voter turnout 2013	0.001	(0.002)	0.007	(0.004)	-0.003	(0.004)	0.003	(0.004)	-0.004	(0.004)
Elections party: SPD	-0.001	(0.001)	0.001	(0.002)	-0.004	(0.003)	-0.002	(0.003)	-0.002	(0.003)
Elections party: FDP	-0.005	(0.005)	0.010	(0.010)	-0.008	(0.011)	-0.030	(0.011)	0.008	(0.013)
Elections party: Grune	0.003	(0.005)	-0.008	(0.009)	-0.001	(0.010)	0.015	(0.009)	0.013	(0.008)
Elections party: DieLinke	-0.006	(0.011)	0.019	(0.021)	-0.006	(0.019)	-0.023	(0.022)	-0.004	(0.015)
Elections party: Other	0.015	(0.007)	0.010	(0.013)	-0.010	(0.013)	0.036*	(0.014)	0.013	(0.015)
German citizenship	-0.001	(0.003)	0.003	(0.005)	0.004	(0.006)	-0.015	(0.009)	-0.005	(0.007)
Education: Uni access	-0.001	(0.002)	0.002	(0.005)	0.008	(0.007)	-0.010	(0.005)	0.000	(0.007)
Education: High School	0.003	(0.002)	-0.006	(0.004)	-0.003	(0.006)	0.018*	(0.005)	0.001	(0.005)
Land prices in 2014	0.000	(0.000)	0.001	(0.000)	0.001	(0.001)	-0.001	(0.001)	-0.000	(0.001)
Share people aged 64 or older	-0.012	(0.010)	-0.038	(0.020)	0.009	(0.025)	-0.034	(0.018)	0.029	(0.020)
Religion: Other	-0.006	(0.003)	-0.006	(0.003)	-0.006	(0.006)	-0.006	(0.005)	-0.006	(0.006)
Religion: Protestant	0.002	(0.001)	0.002	(0.001)	0.003	(0.002)	0.001	(0.002)	0.000	(0.002)
Male Workers	0.001	(0.009)	0.011	(0.012)	0.048	(0.023)	-0.044	(0.019)	-0.029	(0.021)
Social index of municipality	0.001	(0.000)	0.000	(0.001)	0.002	(0.001)	-0.000	(0.001)	0.001	(0.001)
N	3305		955		930		930		490	

Note: This table summarizes the determinants of schools' response to the recruitment email. Dependent variable: Positive response = 0 if no response from school in any way or active opt out; positive response = 1 if school's respondent indicated light or strong interest in participation. The coefficients are the marginal effects from a probit regression, standard errors in parentheses. Column (1) presents estimates using the full sample pooling all research topics. Columns (2) - (5) present estimates for each research topic separately resulting from corresponding sample splits. P-values are adjusted for multiple hypothesis testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (Romano and Wolf, 2005). * p<0.05, ** p<0.01, *** p<0.001.

Table 7: Results - Role of Treatment Topic

	(1) Positive Response	(2) Responded	(3) Filled Out Survey	(4) Positive Response	(5) Responded	(6) Filled Out Survey
Treated	0.036**	-0.029	-0.027			
	(0.012)	(0.020)	(0.015)			
Incentive	0.005	0.021	0.010	0.005	0.021	0.009
	(0.011)	(0.020)	(0.008)	(0.011)	(0.020)	(0.008)
E-Learning				0.019	-0.064*	-0.052**
				(0.016)	(0.025)	(0.018)
Parental Inv.				0.047***	0.001	-0.012
				(0.013)	(0.025)	(0.017)
Integration Migr.				0.040**	-0.024	-0.019
				(0.014)	(0.019)	(0.016)
School-level contr.	Yes	Yes	Yes	Yes	Yes	Yes
County-level contr.	Yes	Yes	Yes	Yes	Yes	Yes
N	3305	3305	3305	3305	3305	3305

Note: This table presents coefficients (marginal effects) of probit regressions. The dependent variable in columns (1) and (4) is a binary variable indicating whether schools positively responded to the recruitment email. The dependent variable in columns (2) and (5) is a binary variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. The dependent variable in columns (3) and (6) is a binary variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Treated* is a binary variable that takes the value of 0 if schools were in the Scientific Contribution Treatment and takes the value of 1 if schools were in the E-Learning, Parental Involvement, or Integration Migration Treatment. *Incentive* is a binary variable that takes the value of 0 if schools were not offered any financial incentive for participation and takes the value of 1 if schools were offered a financial incentive. Standard errors in parentheses. P-values are adjusted for multiple hypothesis testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (Romano and Wolf, 2005). * p<0.05, ** p<0.01, *** p<0.001.

Table 8: Incentive Treatment

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Filled Out Survey	(6) Filled Out Survey
Panel A: Pooled (N=3305)						
Incentive	0.013	0.013	0.012	0.014	0.001	0.002
	(0.011)	(0.011)	(0.019)	(0.019)	(0.008)	(0.008)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel B: E-Learning (N=955)						
Incentive	-0.002	0.001	0.007	0.009	0.005	0.009
	(0.020)	(0.019)	(0.032)	(0.031)	(0.014)	(0.013)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel C: Parental Involvement (N=930)						
Incentive	0.030	0.036	0.041	0.042	0.009	0.014
	(0.025)	(0.023)	(0.028)	(0.028)	(0.017)	(0.016)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel D: Integration Migration (N=930)						
Incentive	-0.015	-0.016	0.006	0.002	0.006	0.005
	(0.017)	(0.016)	(0.027)	(0.027)	(0.018)	(0.016)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Panel A reports on the pooled sample and panels B to D on each research topic separately resulting from corresponding sample splits. The dependent variable in columns (1) and (4) is a binary variable indicating whether schools positively responded to the recruitment email. The dependent variable in columns (2) and (5) is a binary variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. The dependent variable in columns (3) and (6) is a binary variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Incentive* is a binary variable that takes the value of 0 if schools were not offered any financial incentive for participation and takes the value of 1 if schools were offered a financial incentive. Standard errors in parentheses. As none of the coefficients turns out to be significant, we did not adjust p-values for multiple hypothesis testing. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 9: Share of Incentive on School's Yearly Budget for Training of Teachers

(1) Share of Incentive	(2) Absolute	(3) Percent	(4) Cumulative
80< x ≤90%	1,003	71.90	71.90
70< x ≤80%	56	4.01	75.91
60< x ≤70%	50	3.58	79.49
50< x ≤60%	73	5.23	84.72
40< x ≤50%	68	4.87	89.59
30< x ≤40%	87	6.24	95.83
20< x ≤30%	51	3.66	99.49
10< x ≤20%	7	0.50	99.99
Total	1,395	100.00	

Note: This table summarizes the size of the financial incentive (700 Euros) relative to the school's yearly budget for teacher training. $\textit{Share of Incentive} = \frac{700\ \textit{Euro}}{\textit{School's yearly budget}}$.

Table 10: Incentive Treatment - Share Budget

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Filled Out Survey	(6) Filled Out Survey
Panel A: Pooled (N=3305)						
Share Budget	0.005	0.013	-0.000	0.013	-0.004	0.004
	(0.013)	(0.014)	(0.024)	(0.026)	(0.011)	(0.012)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel B: E-Learning (N=955)						
Share Budget	-0.024	-0.014	-0.017	-0.008	0.007	0.015
	(0.023)	(0.023)	(0.040)	(0.040)	(0.018)	(0.017)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel C: Parental Involvement (N=930)						
Share Budget	0.013	0.027	0.020	0.037	-0.009	0.014
	(0.030)	(0.030)	(0.036)	(0.040)	(0.020)	(0.021)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel D: Integration Migration (N=930)						
Share Budget	-0.010	-0.009	0.012	0.007	0.009	0.007
	(0.021)	(0.019)	(0.035)	(0.036)	(0.023)	(0.020)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Panel A reports on the pooled sample and panels B to D on each research topic separately resulting from corresponding sample splits. The dependent variable in columns (1) and (4) is a binary variable indicating whether schools positively responded to the recruitment email. The dependent variable in columns (2) and (5) is a binary variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. The dependent variable in columns (3) and (6) is a binary variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Share Budget* measures the share of the 700 Euro budget on schools' yearly budget for teacher training. Standard errors in parentheses. As none of the coefficients turns out to be significant, we did not adjust p-values for multiple hypothesis testing. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 11: An Experiment on Balance and Precision: Design

Sub-sample	Method	Control	Treatment Group						
		0	1	2	3	4	5	6	$\sum$
1	Matching	15	15						30
	min MSE	15	15						30
2	Re-randomization (t-statistics)	30	30	30	30	30	30	30	210
	min MSE	30	30	30	30	30	30	30	210
3	Randomization	20	20	20	20	20	20	20	140
	min MSE	20	20	20	20	20	20	20	140
4	Randomization	20	20	20	20	20	20	20	140
	min MSE	20	20	20	20	20	20	20	140
5	Randomization	20	20	20	20	20	20		120
	min MSE	20	20	20	20	20	20		120
6	Randomization	20	20	20	20	20	20		120
	min MSE	20	20	20	20	20	20		120
7	Randomization	20	20	20	20	20			100
	min MSE	20	20	20	20	20			100
8	Randomization	20	20	20	20	20			100
	min MSE	20	20	20	20	20			100
9	Randomization	20	20	20	20				80
	min MSE	20	20	20	20				80
10	Randomization	20	20	20	20				80
	min MSE	20	20	20	20				80
11	Randomization	20	20	20					60
	min MSE	20	20	20					60
12	Randomization	20	20						40
	min MSE	20	20						40
	Total	490	490	420	380	300	220	140	2440

Note: This Table illustrates the experimental design of the treatment assignment experiment, see Section 3 for details. It shows, for each randomly drawn subsample, its sample size, which method of treatment assignment was used in the sample and its comparable subgroup, how many experimental groups were assigned, and how many units were assigned to each experimental group. For example, the units in Draw 1 were assigned to one treatment group or the control group, using either pair-wise matching or the min MSE approach, with 15 units in each experimental group, i.e. each method had to allocate 30 units to two experimental groups for this draw.

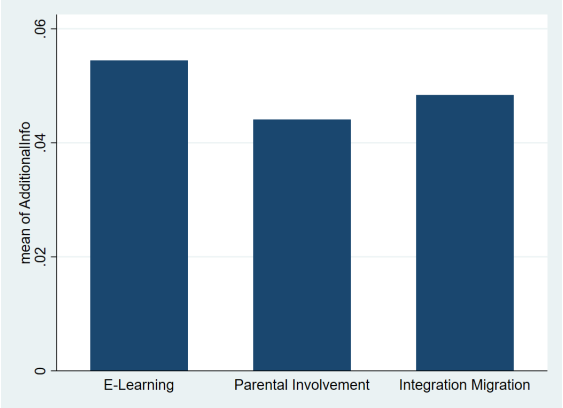
Table 12: Results - Treatment Assignment: Balance

Comparison Method	Draw	Groups	p(minMSE)	p(ComparisonMethod)
Matching	1	1	0.55	0.47
Rerandomization	2	6	0.40	0.25
Randomization	3	6	0.38	0.22
Randomization	4	6	0.45	0.22
Randomization	5	5	0.57	0.30
Randomization	6	5	0.53	0.35
Randomization	7	4	0.57	0.31
Randomization	8	4	0.38	0.31
Randomization	9	3	0.52	0.28
Randomization	10	3	0.48	0.39
Randomization	11	2	0.38	0.42
Randomization	12	1	0.72	0.50

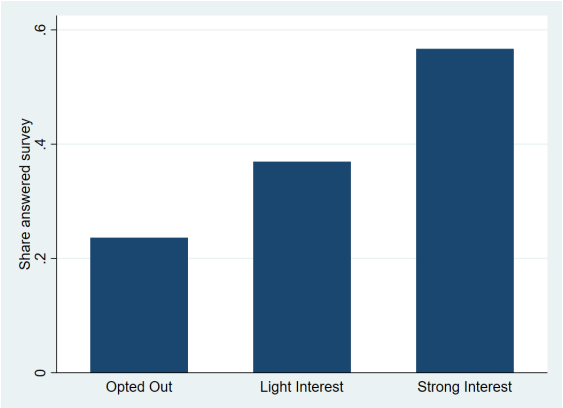
Note: This Table shows the p-values resulting from the test of imbalance due to Hansen and Bowers (2008) when testing for imbalance between the treatment groups in a draw that were allocated with the same treatment assignment method. Lower p-values are associated with a higher chance of imbalance. If several groups are to be compared, the minimal p-value is reported. For example, as Table 11 shows, in Draw 2, six treatment groups were assigned, thus the test was applied to test for imbalance of each of these six groups and the control group; the lowest of these six p-values is 0.4 when assigning units with the min MSE method (fourth column) and 0.25 when assigning units with the comparison method (last column), which in the case of Draw 2 is re-randomization based on t-statistics.

B Graphs

Figure 4: Effort Invested in Dealing with Inquiry



(a) Requested Additional Information (by Treatment)

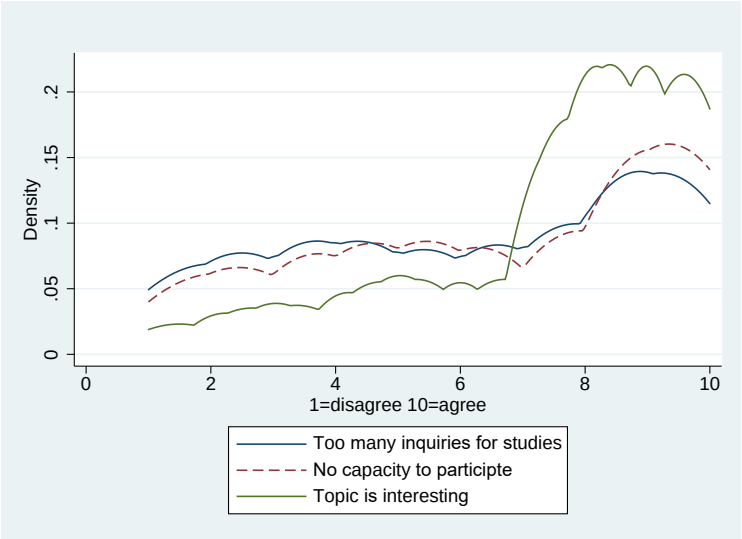


(b) Answered Survey (by Response)

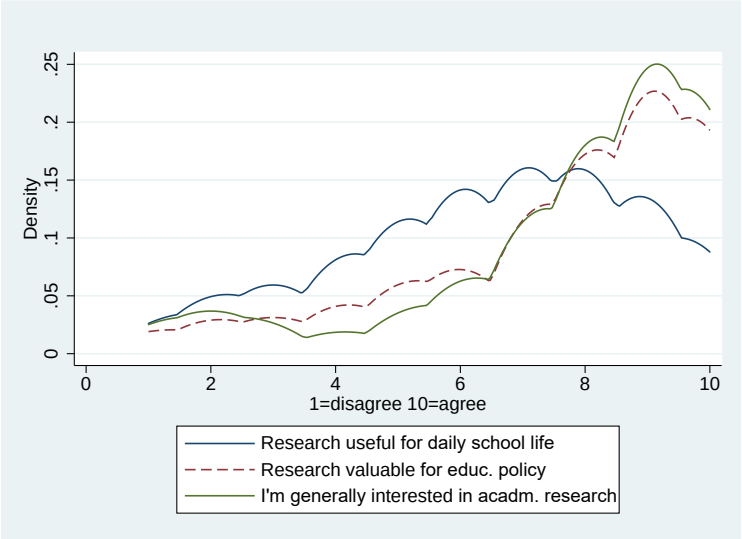
Note: This figure presents two measures on headmasters' effort spent on dealing with our inquiry. Figure 4a shows the share of headmasters who clicked on the link to access more detailed information about the planned experiment by research topic. In total, 138 headmasters were interested in receiving more information. Figure 4b presents the share of headmasters who answered the survey after clicking on one of the three links in the recruitment email (share = number of surveys answered/number of headmasters who responded to inquiry) In total, 269 headmasters filled in the survey.

Figure 5: Survey Answers

(a) Questionnaire: Capacity for Academic Research



(b) Questionnaire: Interest in Academic Research



Note: This figure presents results (kernel density estimates) on the questions asked in the survey. Figure 5a summarizes answers regarding schools' capacity to participate in a field study. In particular, headmasters were asked to agree or disagree on a 1 to 10 scale to the statements that (i) there are too many inquiries for studies, (ii) schools do not have the personnel capacity to participate, and (iii) the proposed research topic is relevant for the school. Figure 5b shows answers on questions that asked more generally about headmasters' opinion about the usefulness of academic research, this is, whether they agree (on a 1 to 10 scale) that (i) insights from academic research can be transferred to everyday school life, (ii) academic research is valuable and informative for educational policy makers, and (iii) that they are in general interested in the findings of academic research.

C Online Appendix - Additional Tables

C.1 Randomization Check With Bootstrapped Standard Errors

Table 13: Descriptive Statistics - School-level data

	(1) E-Learning	(2) Parental Involvement	(3) Integration Migration	(4) Scientific Contribution	(5) Overall
Gender of headmaster	0.679	0.616	0.634	0.635	0.642
	(0.017)	(0.017)	(0.016)	(0.026)	(0.009)
Av. comp. teaching hours	21.213	21.054	21.108	21.154	21.130
	(0.080)	(0.084)	(0.084)	(0.114)	(0.049)
Students in day care	95.412	95.001	94.769	94.636	95.000
	(0.401)	(0.392)	(0.418)	(0.548)	(0.202)
Age of teachers (full time empl.)	39.858	39.951	40.004	40.555	40.028
	(0.197)	(0.232)	(0.206)	(0.295)	(0.124)
Students migration background	30.373	30.404	28.857	28.719	29.710
	(0.618)	(0.684)	(0.671)	(0.838)	(0.355)
Students migrated	6.385	6.375	6.163	6.604	6.352
	(0.247)	(0.274)	(0.280)	(0.364)	(0.132)
Parents migrated	28.468	28.515	27.349	26.800	27.919
	(0.576)	(0.650)	(0.591)	(0.771)	(0.352)
Number of students	329.547	323.894	331.552	332.293	328.928
	(9.169)	(9.251)	(8.708)	(13.286)	(5.232)
Female students	46.817	47.008	49.558	48.814	47.938
	(0.313)	(0.261)	(2.291)	(1.830)	(0.685)
Non-German students	7.195	7.230	7.194	7.368	7.230
	(0.269)	(0.279)	(0.280)	(0.348)	(0.136)
Non-German female students	3.400	3.412	3.305	3.404	3.377
	(0.129)	(0.138)	(0.129)	(0.151)	(0.067)
Share fullt time empl. teachers	55.915	56.000	55.315	55.675	55.735
	(0.583)	(0.591)	(0.512)	(0.846)	(0.273)
Students speak no German at home	15.987	16.461	15.266	15.070	15.782
	(0.474)	(0.532)	(0.483)	(0.657)	(0.277)
Number of classes	12.497	11.988	12.247	12.120	12.227
	(0.228)	(0.209)	(0.210)	(0.305)	(0.115)
N	955	930	930	490	3305
Proportion	0.289	0.281	0.281	0.148	1.000

Note: This table presents descriptive statistics on observable characteristics of schools by treatment. Cell entries report the group means and bootstraped standard errors (200 repetitions) are reported in parentheses. Observable characteristics are described in more detail in the online appendix.

Table 14: Descriptive Statistics - Municipality-level data

	(1) E-Learning	(2) Parental Involvement	(3) Integration Migration	(4) Scientific Contribution	(5) Overall
Inhabitants	371.964	372.022	369.553	368.521	370.791
	(3.827)	(3.617)	(3.609)	(5.314)	(2.053)
Status married	48.457	48.538	48.501	48.479	48.495
	(0.049)	(0.043)	(0.042)	(0.061)	(0.023)
Unemployment rate	2.352	2.347	2.348	2.358	2.350
	(0.024)	(0.027)	(0.025)	(0.035)	(0.013)
Voter turnout 2013	73.238	73.218	73.408	73.205	73.275
	(0.113)	(0.115)	(0.117)	(0.157)	(0.058)
Elections party: CDU	42.911	42.964	43.258	43.587	43.124
	(0.224)	(0.220)	(0.228)	(0.316)	(0.119)
Elections party: SPD	30.091	30.082	29.826	29.646	29.948
	(0.180)	(0.190)	(0.172)	(0.257)	(0.089)
Elections party: FDP	5.189	5.202	5.255	5.174	5.209
	(0.036)	(0.037)	(0.038)	(0.050)	(0.020)
Elections party: Grune	6.932	6.899	6.907	6.854	6.904
	(0.055)	(0.054)	(0.061)	(0.078)	(0.029)
Elections party: DieLinke	5.309	5.275	5.240	5.240	5.270
	(0.036)	(0.037)	(0.032)	(0.050)	(0.020)
Elections party: Other	8.425	8.444	8.379	8.345	8.405
	(0.042)	(0.041)	(0.041)	(0.059)	(0.022)
German citizenship	93.145	93.137	93.121	93.027	93.118
	(0.060)	(0.060)	(0.058)	(0.079)	(0.030)
Education: Uni access	17.489	17.227	17.252	17.311	17.322
	(0.119)	(0.143)	(0.116)	(0.181)	(0.067)
Education: High School	27.099	27.051	27.063	27.042	27.067
	(0.089)	(0.099)	(0.095)	(0.136)	(0.048)
Land prices in 2014	134.365	134.081	133.941	133.947	134.104
	(1.292)	(1.149)	(1.252)	(1.858)	(0.738)
Share people aged 64 or older	27.400	27.167	27.003	27.218	27.196
	(0.195)	(0.190)	(0.213)	(0.292)	(0.103)
Religion: Other	27.400	27.167	27.003	27.218	27.196
	(0.200)	(0.202)	(0.202)	(0.284)	(0.108)
Religion: Protestant	27.951	27.413	27.125	27.545	27.507
	(0.429)	(0.400)	(0.413)	(0.615)	(0.209)
Male Workers	51.595	51.635	51.645	51.632	51.626
	(0.028)	(0.031)	(0.033)	(0.045)	(0.017)
Social index of municipality	30.033	29.750	29.391	30.005	29.769
	(0.536)	(0.564)	(0.531)	(0.702)	(0.288)
N	955	930	930	490	3305
Proportion	0.289	0.281	0.281	0.148	1.000

Note: This table presents descriptive statistics on observable characteristics of municipalities by treatment. Cell entries report the group means and bootstrapped standard errors (200 repetitions) are reported in parentheses. Observable characteristics are described in more detail in the online appendix.

C.2 Treatment Effects With Randomization Inference and Bootstrapped Standard Errors

Table 15: Results - Role of Treatment Topic

	(1) Positive Response	(2) Responded	(3) Filled Out Survey	(4) Positive Response	(5) Responded	(6) Filled Out Survey
Treated	0.036*	-0.029	-0.027			
	(0.015)	(0.174)	(0.139)			
Incentive	0.005	0.021	0.010	0.005	0.021	0.009
	(0.629)	(0.265)	(0.230)	(0.647)	(0.323)	(0.217)
E-Learning				0.019	-0.064*	-0.052*
				(0.279)	(0.012)	(0.016)
Parental Inv.				0.047**	0.001	-0.012
				(0.005)	(0.958)	(0.501)
Integration Migr.				0.040*	-0.024	-0.019
				(0.016)	(0.220)	(0.281)
School-level contr.	Yes	Yes	Yes	Yes	Yes	Yes
Munic.-level contr.	Yes	Yes	Yes	Yes	Yes	Yes
N	3305	3305	3305	3305	3305	3305

Note: This table presents coefficients (marginal effects) of probit regressions. Dependent variable in columns (1) and (4) is a dummy variable indicating whether schools positively responded to the recruitment email. Dependent variable in columns (2) and (5) is a dummy variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. Dependent variable in columns (3) and (6) is a dummy variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Treated* is a dummy variable that takes the value of 0 if schools were in the Scientific Contribution Treatment and takes the value of 1 if schools were in the E-Learning, Parental Involvement, or Integration Migration Treatment. *Incentive* is a dummy variable that takes the value of 0 if schools were not offered any financial incentive for participation and takes the value of 1 if schools were offered a financial incentive. Bootstrapped standard errors (200 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table 16: Results - Role of Treatment Topic

	(1) Positive Response	(2) Responded	(3) Filled Out Survey	(4) Positive Response	(5) Responded	(6) Filled Out Survey
Treated	0.036*	-0.029	-0.027*			
	(0.020)	(0.180)	(0.030)			
Incentive	0.005	0.021	0.010	0.005	0.021	0.009
	(0.640)	(0.160)	(0.270)	(0.640)	(0.160)	(0.270)
E-Learning				0.019	-0.064***	-0.052***
				(0.060)	(0.000)	(0.000)
Parental Inv.				0.047***	0.001	-0.012
				(0.000)	(0.940)	(0.310)
Integration Migr.				0.040***	-0.024	-0.019
				(0.000)	(0.190)	(0.080)
School-level contr.	Yes	Yes	Yes	Yes	Yes	Yes
County-level contr.	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Dependent variable in columns (1) and (4) is a dummy variable indicating whether schools positively responded to the recruitment email. Dependent variable in columns (2) and (5) is a dummy variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. Dependent variable in columns (3) and (6) is a dummy variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Treated* is a dummy variable that takes the value of 0 if schools were in the Scientific Contribution Treatment and takes the value of 1 if schools were in the E-Learning, Parental Involvement, or Integration Migration Treatment. *Incentive* is a dummy variable that takes the value of 0 if schools were not offered any financial incentive for participation and takes the value of 1 if schools were offered a financial incentive. Standard errors of randomization inference (100 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table 17: Incentive Treatment

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Filled Out Survey	(6) Filled Out Survey
Panel A: Pooled (N=3305)						
Incentive	0.013	0.013	0.012	0.014	0.001	0.002
	(0.214)	(0.219)	(0.505)	(0.491)	(0.943)	(0.764)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel B: E-Learning (N=955)						
Incentive	-0.002	0.001	0.007	0.009	0.005	0.009
	(0.923)	(0.968)	(0.814)	(0.794)	(0.738)	(0.597)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel C: Parental Involvement (N=930)						
Incentive	0.030	0.036	0.041	0.042	0.009	0.014
	(0.254)	(0.192)	(0.149)	(0.136)	(0.622)	(0.443)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel D: Integration Migration (N=930)						
Incentive	-0.015	-0.016	0.006	0.002	0.006	0.005
	(0.381)	(0.357)	(0.811)	(0.937)	(0.744)	(0.776)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Panel A is the pooled sample and in panels B to D the sample is split by research topic. Dependent variable in columns (1) and (4) is a dummy variable indicating whether schools positively responded to the recruitment email. Dependent variable in columns (2) and (5) is a dummy variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. Dependent variable in columns (3) and (6) is a dummy variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Incentive* is a dummy variable that takes the value of 0 if schools were not offered any financial incentive for participation and takes the value of 1 if schools were offered a financial incentive. Bootstrapped standard errors (200 repetitions) in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 18: Incentive Treatment

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Filled Out Survey	(6) Filled Out Survey
Panel A: Pooled (N=3305)						
Incentive	0.013	0.013	0.012	0.014	0.001	0.002
	(0.360)	(0.330)	(0.480)	(0.370)	(1.000)	(0.670)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel B: E-Learning (N=955)						
Incentive	-0.002	0.001	0.007	0.009	0.005	0.009
	(0.910)	(0.980)	(0.760)	(0.760)	(0.770)	(0.590)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel C: Parental Involvement (N=930)						
Incentive	0.030	0.036	0.041	0.042	0.009	0.014
	(0.210)	(0.150)	(0.150)	(0.130)	(0.620)	(0.430)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel D: Integration Migration (N=930)						
Incentive	-0.015	-0.016	0.006	0.002	0.006	0.005
	(0.470)	(0.440)	(0.870)	(0.930)	(0.640)	(0.750)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Panel A is the pooled sample and in panels B to D the sample is split by research topic. Dependent variable in columns (1) and (4) is a dummy variable indicating whether schools positively responded to the recruitment email. Dependent variable in columns (2) and (5) is a dummy variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. Dependent variable in columns (3) and (6) is a dummy variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Incentive* is a dummy variable that takes the value of 0 if schools were not offered any financial incentive for participation and takes the value of 1 if schools were offered a financial incentive. Standard errors of randomization inference (100 repetitions) in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 19: Incentive Treatment - Share Budget

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Filled Out Survey	(6) Filled Out Survey
Panel A: Pooled (N=3305)						
Share Budget	0.005	0.013	-0.000	0.013	-0.004	0.004
	(0.734)	(0.396)	(0.991)	(0.619)	(0.727)	(0.724)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel B: E-Learning (N=955)						
Share Budget	-0.024	-0.014	-0.017	-0.008	0.007	0.015
	(0.318)	(0.597)	(0.628)	(0.845)	(0.699)	(0.515)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel C: Parental Involvement (N=930)						
Share Budget	0.013	0.027	0.020	0.037	-0.009	0.014
	(0.641)	(0.421)	(0.610)	(0.417)	(0.676)	(0.517)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel D: Integration Migration (N=930)						
Share Budget	-0.010	-0.009	0.012	0.007	0.009	0.007
	(0.602)	(0.661)	(0.757)	(0.859)	(0.709)	(0.761)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Panel A is the pooled sample and in panels B to D the sample is split by research topic. Dependent variable in columns (1) and (4) is a dummy variable indicating whether schools positively responded to the recruitment email. Dependent variable in columns (2) and (5) is a dummy variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. Dependent variable in columns (3) and (6) is a dummy variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Share Budget* measures the share of the 700 Euro budget on schools' yearly budget for teacher training. Bootstrapped standard errors (200 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table 20: Incentive Treatment - Share Budget

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Filled Out Survey	(6) Filled Out Survey
Panel A: Pooled (N=3305)						
Share Budget	0.005	0.013	-0.000	0.013	-0.004	0.004
	(0.710)	(0.570)	(0.330)	(0.520)	(0.290)	(0.700)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel B: E-Learning (N=955)						
Share Budget	-0.024	-0.014	-0.017	-0.008	0.007	0.015
	(0.080)	(0.210)	(0.240)	(0.420)	(0.690)	(0.290)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel C: Parental Involvement (N=930)						
Share Budget	0.013	0.027	0.020	0.037	-0.009	0.014
	(0.810)	(0.520)	(0.860)	(0.330)	(0.150)	(0.700)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes
Panel D: Integration Migration (N=930)						
Share Budget	-0.010	-0.009	0.012	0.007	0.009	0.007
	(0.860)	(0.940)	(0.560)	(0.610)	(0.520)	(0.640)
School-level contr.	No	Yes	No	Yes	No	Yes
Munic.-level contr.	No	Yes	No	Yes	No	Yes

Note: This table presents coefficients (marginal effects) of probit regressions. Panel A is the pooled sample and in panels B to D the sample is split by research topic. Dependent variable in columns (1) and (4) is a dummy variable indicating whether schools positively responded to the recruitment email. Dependent variable in columns (2) and (5) is a dummy variable indicating whether schools responded in any way (active opt out, light interest, strong interest) to the recruitment email. Dependent variable in columns (3) and (6) is a dummy variable indicating whether schools (partly) filled out the survey attached to the recruitment email. *Share Budget* measures the share of the 700 Euro budget on schools' yearly budget for teacher training. Standard errors of randomization inference (100 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

D Online Appendix - Description of background data, Recruitment Email, and Survey

Description of Variables - School level

Type of school: There are 12 different school types in NRW. Among them elementary school, high school and three types of vocational school (Hauptschule, Realschule, Gesamtschule) are the most prominent school types representing appox. 82% of all schools. The remaining 7 school types are subsummed as "Other school types". Elementary school in Germany runs from age 6 to 10 and thereafter students are tracked into secondary exucation. *Hauptschule* (grades 5 to 9 or 10) provides pupils with a basic general education that prepares them for a vocational job, *Realschule* (grades 5 to 10) also prepares students for a vocational job, but also offers the possibility to attend the advanced level of the high school if grades are good enough, *Gesamtschule* (grades 5 to 10 or 12) offers a longer period of common learning and the possibility to obtain all degrees of secondary education, and *High School - Gymnasium* (grades 5 to 12) is the most academic school type preparing students to apply to university.

Gender of headmaster: Gender of the headmaster was obtained from schools' websites.

Average compulsory teaching hours: For each school, we know the sum of how many compulsory hours teachers have to teach. The average compulsory teaching hour is the sum of compulsory teaching hours divided by the sum of all teachers (full time employed, part time employed, trainee teachers).

Age of teachers (full time employed): Average age of all full time employed teachers within a school.

Share full time employed teachers: Share of teachers who are full time employed.

Students in day care: Share of students attending afternoon childcare.

Students migrated: Share of students not born in Germany (migrated to Germany with or without family manmbers).

Parents migrated: Share of students with at least one parent not born in Germany (includes students born in Germany if at least one parent not born in Germany).

Students migration background: Share of students with some migration background in the family (one parent or both parents not born in Germany and/or child not born in Germany). Note that this variable is

not the sum of students migrated and parents migrated. The sum of students migrated and parents migrated would double count students that migrated together with their parents.

Number of students: The total number of students attending the school.

Female students: Number of female students attending the school.

Non-German students: Share of students not having a German passport

Non-German female students: Share of female students not having a German passport.

Students speak no German at home: Share of students who do not speak German with their parents.

Number of classes: Total number of classes (all grade levels) within a school.

Municipality level data

Inhabitants: Number of inhabitants of the municipality.

Status married: Share of inhabitants being married.

Unemployment rate: Statistic for the current period on the share of unemployed workers.

Voter turnout: Share of eligible voters who have voted.

Elections party [name of party]: Share of votes for the respective political party.

German citizenship: Share of people with German citizenship.

Education: High School: Share of people with high school degree.

Education: Uni access: Share of people with university degree.

Land prices in 2014: Land prices in corresponding cities in 2014.

Share people aged 64 or older: Share of people aged 64 years or older.

Religion Protestant: Share of protestant people.

Religion Other: Share of people not being protestant or catholic.

Male Workers: Share of workers being male.

Social index of municipality: Index incorporating information on the unemployment rate, social assistance rate, migrant quota and quota of apartments in single-family homes.

E Online Appendix - Email communication

E.1 Initial contact email

E.2 Reminder email

F Online Appendix - Initial questionnaire

G Online Appendix - Self-selection in administrative independent cities

Using data of three previously conducted experiment in NRW by the authors Riener and Wagner, 2019; Fischer and Wagner, 2018; Wagner, 2016, we can take a look into potential self selection of schools in administratively independent cities. All three studies contacted schools in the administrative independent cities Bonn, Cologne and Düsseldorf. Riener and Wagner, 2019 contacted 168 secondary schools and investigate how the type and design of non-monetary incentives affect students' performance in a low stakes test. (Fischer and Wagner, 2018) also conducted their experiment in secondary schools (contacted schools = 143) to analyze the role of the timing and reference frame of feedback in a high stakes test. Wagner, 2016 contacted 245 elementary schools and manipulated the grading scheme of a low stakes math test. Limitations of these data are....there is some degree of selection by researchers (reachable by public transport), small N, Three cities might not be representative for all administrative independent cities (they are the largest cities in NRW). Düsseldorf, Cologne and Bonn represent the largest cities out of the 22 administrative independent cities in NRW. Düsseldorf and Cologne are the largest and second largest cities in NRW and Bonn is placed 10th. However, within the three cities, the authors contacted almost all schools (Riener and Wagner, 2019 93.58%, Fischer and Wagner, 2018 79,89%, and Wagner, 2016 85.66%).

First, we check whether schools' background characteristics differ between administrative independent cities and municipalities. Table...

Table 22 shows that schools in administrative independent cities differ on average substantially from schools in municipalities. Schools in administrative independent cities are larger (more students, classes, and teachers), have a higher share of migrant children, teachers are younger and have more compulsory teaching hours. However, schools do not differ in students' gender composition. These differences exist for both elementary and secondary schools (see Tables 24 and 23).

In Table 21, we compare the characteristics of an average school in the cities Bonn, Cologne and Düsseldorf with the average school in administrative independent cities in NRW. Schools in Bonn, Cologne and Düsseldorf have on average more students, more classes, less students in day care, a lower share of full time employed teachers, and teachers have a less compulsory teaching hours. Moreover, the share of children with a migration background is higher in the study of Riener and Wagner, 2019, but not in the studies by Fischer and Wagner, 2018 and Wagner, 2016.

Table 21: Descriptive Statistics: NRW-Data vs Experiments

	(1) NRW Secondary Schools	(2) Riener and Wagner, 2019	(3) Fischer and Wagner, 2018	(4) NRW Elementary Schools	(5) Wagner, 2016	(6) (1) vs. (2), p-value	(7) (1) vs. (3), p-value	(8) (4) vs. (5), p-value
Av. comp. teaching hours	21.100	21.619	21.156	21.160	22.106	0.031	0.892	0.000
	(0.074)	(0.186)	(0.181)	(0.055)	(0.126)			
Students in day care	90.033	81.945	77.628	100.000	100.000	0.000	0.000	
	(0.401)	(1.430)	(1.528)	(0.000)	(0.000)			
Age of teachers	41.125	40.400	39.934	38.925	37.627	0.112	0.015	0.000
	(0.141)	(0.315)	(0.312)	(0.187)	(0.321)			
Students migration background	27.797	47.290	43.649	31.635	46.398	0.000	0.000	0.000
	(0.491)	(1.776)	(1.943)	(0.492)	(1.555)			
Students migrated	6.816	12.959	9.723	5.885	9.387	0.000	0.000	0.000
	(0.218)	(0.989)	(0.850)	(0.167)	(0.538)			
Parents migrated	25.997	39.756	37.064	29.854	42.890	0.000	0.000	0.000
	(0.466)	(1.629)	(1.788)	(0.465)	(1.464)			
Number of students	447.099	662.125	758.510	209.967	242.249	0.000	0.000	0.000
	(8.503)	(24.713)	(25.601)	(1.933)	(5.065)			
Female students	46.851	48.838	50.053	49.033	49.040	0.000	0.000	0.000
	(1.495)	(0.831)	(0.940)	(0.101)	(0.231)			
Non-German students	7.853	16.697	12.002	6.604	11.372	0.000	0.000	0.000
	(0.219)	(1.047)	(0.767)	(0.175)	(0.555)			
Non-German female students	3.535	50.834	52.382	3.218	50.542	0.000	0.000	0.000
	(0.100)	(1.004)	(1.133)	(0.089)	(0.944)			
Share full time empl. teachers	62.881	58.162	56.081	48.541	55.538	0.000	0.000	0.000
	(0.357)	(0.975)	(0.981)	(0.403)	(0.947)			
Students speak no German at home	14.114	28.530	24.534	17.460	30.693	0.000	0.000	0.000
	(0.359)	(1.547)	(1.559)	(0.409)	(1.493)			
Number of classes	15.310	18.881	20.182	9.124	10.131	0.000	0.000	0.000
	(0.194)	(0.537)	(0.602)	(0.079)	(0.201)			

Note: This table summarizes . Standard Deviation in parentheses.

Table 22: Descriptive Statistics: NRW-Data - Municipalities vs Cities

	(1) Municipality	(2) Independent City	(3) Overall	(4) (1) vs. (2), p-value
Av. comp. teaching hours	22.968	23.718	23.216	0.000
	(0.049)	(0.060)	(0.038)	
Teachers fullt time empl. teachers	17.367	20.287	18.332	0.000
	(0.260)	(0.420)	(0.223)	
Teachers part time empl. teachers	9.814	11.056	10.224	0.000
	(0.127)	(0.208)	(0.110)	
Age of teachers	46.613	45.709	46.314	0.000
	(0.067)	(0.084)	(0.053)	
Number of classes	12.304	13.873	12.823	0.000
	(0.114)	(0.178)	(0.097)	
Number of students	332.277	379.267	347.797	0.000
	(4.667)	(7.300)	(3.959)	
Female students	0.469	0.472	0.470	0.262
	(0.001)	(0.002)	(0.001)	
Non-German students	0.074	0.138	0.095	0.000
	(0.001)	(0.003)	(0.001)	
Non-German female students	0.035	0.065	0.045	0.000
	(0.001)	(0.001)	(0.001)	
Students migration background	0.301	0.436	0.345	0.000
	(0.003)	(0.005)	(0.003)	
Students migrated	0.063	0.096	0.074	0.000
	(0.001)	(0.003)	(0.001)	
Parents migrated	0.283	0.394	0.320	0.000
	(0.003)	(0.005)	(0.003)	
Students speak no German at home	0.162	0.288	0.204	0.000
	(0.003)	(0.005)	(0.003)	
Students in day care	0.951	0.940	0.947	0.004
	(0.002)	(0.003)	(0.002)	
Female students in day care	0.443	0.440	0.442	0.343
	(0.002)	(0.002)	(0.001)	
N	3670	1810	5480	
Proportion	0.670	0.330	1.000	

Note: This table summarizes . Standard Deviation in parentheses.

Table 23: Descriptive Statistics: NRW-Data Sec. Schools - Rural Areas vs Cities

	(1) Municipality	(2) Independent City	(3) Overall	(4) (1) vs. (2), p-value
Av. comp. teaching hours	23.006	23.661	23.211	0.000
	(0.077)	(0.098)	(0.061)	
Teachers fullt time empl. teachers	27.124	32.955	28.952	0.000
	(0.407)	(0.680)	(0.355)	
Teachers part time empl. teachers	12.608	16.096	13.701	0.000
	(0.224)	(0.363)	(0.194)	
Age of teachers	47.731	47.162	47.552	0.001
	(0.098)	(0.114)	(0.077)	
Number of classes	15.554	18.719	16.546	0.000
	(0.191)	(0.294)	(0.163)	
Number of students	457.361	555.575	488.148	0.000
	(8.313)	(13.360)	(7.133)	
Female students	0.447	0.448	0.448	0.894
	(0.003)	(0.004)	(0.002)	
Non-German students	0.081	0.148	0.102	0.000
	(0.002)	(0.005)	(0.002)	
Non-German female students	0.037	0.067	0.046	0.000
	(0.001)	(0.002)	(0.001)	
Students migration background	0.278	0.405	0.318	0.000
	(0.004)	(0.008)	(0.004)	
Students migrated	0.067	0.099	0.077	0.000
	(0.002)	(0.004)	(0.002)	
Parents migrated	0.260	0.363	0.293	0.000
	(0.004)	(0.007)	(0.004)	
Students speak no German at home	0.143	0.251	0.177	0.000
	(0.003)	(0.007)	(0.003)	
Students in day care	0.900	0.869	0.890	0.000
	(0.004)	(0.006)	(0.003)	
Female students in day care	0.394	0.379	0.389	0.000
	(0.002)	(0.003)	(0.002)	
N	1809	826	2635	
Proportion	0.687	0.313	1.000	

Note: This table summarizes . Standard Deviation in parentheses.

Table 24: Descriptive Statistics: NRW-Data Elemen. Schools - Rural Areas vs Cities

	(1) Municipality	(2) Independent City	(3) Overall	(4) (1) vs. (2), p-value
Av. comp. teaching hours	22.931	23.766	23.220	0.000
	(0.060)	(0.075)	(0.048)	
Teachers fullt time empl. teachers	7.882	9.653	8.495	0.000
	(0.090)	(0.139)	(0.078)	
Teachers part time empl. teachers	7.098	6.826	7.004	0.058
	(0.085)	(0.115)	(0.068)	
Age of teachers	45.525	44.490	45.167	0.000
	(0.085)	(0.108)	(0.067)	
Number of classes	9.145	9.806	9.374	0.000
	(0.075)	(0.099)	(0.060)	
Number of students	210.688	231.269	217.806	0.000
	(1.820)	(2.417)	(1.466)	
Female students	0.490	0.492	0.491	0.308
	(0.001)	(0.001)	(0.001)	
Non-German students	0.068	0.129	0.089	0.000
	(0.002)	(0.004)	(0.002)	
Non-German female students	0.033	0.064	0.044	0.000
	(0.001)	(0.002)	(0.001)	
Students migration background	0.323	0.461	0.371	0.000
	(0.005)	(0.008)	(0.004)	
Students migrated	0.059	0.094	0.071	0.000
	(0.002)	(0.003)	(0.002)	
Parents migrated	0.305	0.420	0.345	0.000
	(0.004)	(0.007)	(0.004)	
Students speak no German at home	0.181	0.319	0.229	0.000
	(0.004)	(0.007)	(0.004)	
Students in day care	1.000	1.000	1.000	
	(0.000)	(0.000)	(0.000)	
Female students in day care	0.490	0.492	0.491	0.308
	(0.001)	(0.001)	(0.001)	
N	1861	5984	2845	
Proportion	0.654	0.346	1.000	

Note: This table summarizes . Standard Deviation in parentheses.

Table 25: Results - Other Experiments: Responded

	(1) WagnerRiener2015		(2) Wagner2016		(3) FischerWagner2018	
Av. comp. teaching hours	0.018	(0.039)	0.028	(0.026)	0.037	(0.033)
Age of teachers (full time empl.)	-0.003	(0.005)	0.008**	(0.003)	0.011**	(0.003)
Students migration background	-0.637	(0.826)	-0.083	(0.671)	0.633	(0.613)
Students migrated	0.882***	(0.024)	-0.216	(0.228)	0.290	(0.533)
Parents migrated	0.732*	(0.369)	0.412	(0.694)	-0.203	(0.266)
Number of students	-0.000	(0.000)	0.000	(0.001)	0.001*	(0.000)
Female students	-0.858*	(0.420)	0.118	(0.313)	-0.732	(0.619)
Non-German students	-0.956	(0.522)	0.388	(0.483)	-0.513	(0.473)
Non-German female students	0.378	(0.233)	0.104	(0.086)	0.422	(0.533)
Share fullt time empl. teachers	-0.439	(0.787)	-0.281	(0.394)	-0.969	(0.490)
Students speak no German at home	-0.226	(0.821)	-0.471	(0.190)	-0.243	(0.500)
Number of classes	-0.003	(0.007)	0.004	(0.032)	-0.022	(0.007)
Students in day care	-0.337	(0.521)			1.075	(0.325)
Vocational (Gesamtsch.)	0.128	(0.234)				
Vocational (Hauptsch.)	0.209	(0.180)				
Vocational (Realsch.)	0.068	(0.212)			-0.350*	(0.162)
N	166		243		141	

Note: This table summarizes . p-vlaues are adjusted for MHT using RomanoWolf with 100 reps. * smaller 0.05, ** smaller 0.01, *** smaller 0.001.

Table 26: Results - Other Experiments: Responded

	(1) WagnerRiener2015		(2) Wagner2016		(3) FischerWagner2018	
Av. comp. teaching hours	0.018	(0.982)	0.028	(0.504)	0.037	(0.954)
Age of teachers (full time empl.)	-0.003	(0.791)	0.008**	(0.005)	0.011	(0.961)
Students migration background	-0.637	(0.817)	-0.083	(0.976)	0.633	(0.969)
Students migrated	0.882	(0.710)	-0.216	(0.614)	0.290	(0.995)
Parents migrated	0.732	(0.613)	0.412	(0.871)	-0.203	(0.963)
Number of students	-0.000	(0.975)	0.000	(0.958)	0.001	(0.971)
Female students	-0.858	(0.882)	0.118	(0.796)	-0.732	(0.899)
Non-German students	-0.956	(0.767)	0.388	(0.768)	-0.513	(0.901)
Non-German female students	0.378	(0.194)	0.104	(0.717)	0.422	(0.671)
Share fullt time empl. teachers	-0.439	(0.967)	-0.281	(0.598)	-0.969	(0.910)
Students speak no German at home	-0.226	(0.918)	-0.471	(0.507)	-0.243	(0.987)
Number of classes	-0.003	(0.956)	0.004	(0.950)	-0.022	(0.981)
Students in day care	-0.337	(0.964)			1.075	(0.897)
Vocational (Gesamtsch.)	0.128	(0.901)				
Vocational (Hauptsch.)	0.209	(0.957)				
Vocational (Realsch.)	0.068	(0.989)			-0.350	(0.835)
N	166		243		141	

Note: This table summarizes .BOOTSTRAPPED. P values * smaller 0.05, ** smaller 0.01, *** smaller 0.001.

Table 27: Results - Other Experiments: Pos. Resp.

	(1) Riener and Wagner, 2019		(2) Wagner, 2016		(3) Fischer and Wagner, 2018	
Av. comp. teaching hours	0.000	(0.016)	0.033	(0.015)	-0.015	(0.019)
Age of teachers (full time empl.)	0.005	(0.009)	0.006	(0.003)	-0.002	(0.002)
Students migration background	-0.156	(0.168)	0.100	(0.357)	-0.441	(0.187)
Students migrated	0.625**	(0.052)	0.032	(0.147)	-0.021	(0.192)
Parents migrated	0.127	(0.056)	0.059	(0.392)	0.201	(0.109)
Number of students	-0.000	(0.000)	0.001	(0.001)	-0.000	(0.000)
Female students	-0.754	(0.500)	0.286	(0.551)	-0.335	(0.221)
Non-German students	-1.280	(0.500)	0.602	(0.417)	0.175	(0.060)
Non-German female students	0.112	(0.350)	0.173***	(0.064)	0.273	(0.261)
Share fullt time empl. teachers	-0.561	(0.233)	-0.272	(0.170)	-0.105	(0.234)
Students speak no German at home	0.344	(0.296)	-0.426**	(0.141)	0.212	(0.126)
Number of classes	0.015	(0.008)	-0.006	(0.020)	0.006	(0.003)
Students in day care	-0.488	(0.360)			0.027	(0.136)
Vocational (Realsch.)	0.132	(0.177)			0.048	(0.071)
Vocational (Gesamtsch.)	0.069	(0.109)				
Vocational (Hauptsch.)	0.367**	(0.170)				
N	166		243		141	

Note: This table summarizes . p-vlaues are adjusted for MHT using RomanoWolf with 100 reps. * smaller 0.05, ** smaller 0.01, *** smaller 0.001.

Table 28: Results - Other Experiments: Pos. Resp.

	(1) Riener and Wagner, 2019		(2) Wagner, 2016		(3) Fischer and Wagner, 2018	
Av. comp. teaching hours	0.000	(1.000)	0.033	(0.971)	-0.015	(0.995)
Age of teachers (full time empl.)	0.005	(0.996)	0.006	(0.961)	-0.002	(0.994)
Students migration background	-0.156	(0.998)	0.100	(0.992)	-0.441	(0.987)
Students migrated	0.625	(0.985)	0.032	(0.995)	-0.021	(0.999)
Parents migrated	0.127	(0.998)	0.059	(0.996)	0.201	(0.991)
Number of students	-0.000	(0.994)	0.001	(0.982)	-0.000	(0.997)
Female students	-0.754	(0.992)	0.286	(0.982)	-0.335	(0.972)
Non-German students	-1.280	(0.974)	0.602	(0.969)	0.175	(0.992)
Non-German female students	0.112	(0.992)	0.173	(0.964)	0.273	(0.988)
Share fullt time empl. teachers	-0.561	(0.997)	-0.272	(0.967)	-0.105	(0.996)
Students speak no German at home	0.344	(0.991)	-0.426	(0.971)	0.212	(0.994)
Number of classes	0.015	(0.908)	-0.006	(0.993)	0.006	(0.995)
Students in day care	-0.488	(0.996)			0.027	(0.999)
Vocational (Gesamtsch.)	0.069	(0.998)				
Vocational (Hauptsch.)	0.367	(0.995)				
Vocational (Realsch.)	0.132	(0.999)			0.048	(0.998)
N	166		243		141	

Note: This table summarizes . BOOTSTRAPPED. P values * smaller 0.05, ** smaller 0.01, *** smaller 0.001.