Addressing Validity and Generalizability Concerns in Field Experiments*

Gerhard Riener† Sebastian O. Schneider‡ Valentin Wagner§

June 25, 2020

Abstract

In this paper, we systematically analyze the empirical importance of standard conditions for the validity and generalizability of field experiments: the internal and external overlap and unconfoundedness conditions. We experimentally varied the degree of overlap in disjoint sub-samples from a recruitment experiment with more than 3,000 public schools, mimicking small scale field experiments. This was achieved by using different techniques for treatment assignment. We applied standard methods, such as pure randomization, and the novel minMSE treatment assignment method. This new technique should achieve improved overlap by balancing covariate dependencies and variances instead of focusing on individual mean values. We assess the relevance of the overlap condition by linking the estimation precision in the disjoint sub-samples to measures of overlap and balance in general. Unconfoundedness is addressed by using a rich set of administrative data on institution and municipality characteristics to study potential self-selection. We find no evidence for the violation of unconfoundedness and establish that improved overlap, and balancedness, as achieved by the minMSE method, reduce the bias of the treatment effect estimation by more than 35% compared to pure randomization, illustrating the importance of, and suggesting a solution to, addressing overlap also in (field) experiments.

Keywords: External validity, field experiments, generalizability, treatment effect, overlap, balance, precision, treatment assignment, unconfoundedness, self-selection bias, site-selection bias

JEL codes: C9, C90, C93, D04

*We are grateful for comments and advice from Stefania Bortolotti, Daniel Schunk, Matthias Sutter, and 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 Days in Fontainebleau. Schneider gratefully acknowledges the financial support provided by the German Research Foundation (DFG) via RTG 1723 and the Foundation of German Businesses (sdw). The usual disclaimer applies.

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

‡Corresponding author: 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

Academic researchers as well as public policy-makers increasingly employ field experiments in order to gain insights into the effects of policies.[1] The common practice of randomization of subjects in control and treatment groups ensures a causal interpretation of the experimental results as it rules out self-selection into treatments. The condition for causal interpretation of the results, *the unconfoundedness condition*, however, is not sufficient for obtaining robust insights that can be extrapolated to other settings (e.g., Rosenbaum and Rubin, 1983; Hotz, Imbens, and Mortimer, 2005; Czibor, Jimenez-Gomez, and List, 2019). The *overlap condition* requires relevant subgroups to be represented in treatment and control groups (e.g., Rosenbaum and Rubin, 1983), in order to estimate the average treatment effect by comparing an outcome accross treatment groups. For extrapolation of the results across other settings, two further conditions are key: the *external unconfoundedness* and *external overlap* conditions (Hotz, Imbens, and Mortimer, 2005; Allcott, 2015): External unconfoundedness means that in an average sense treatment effects of the population participating in the experiment are not different from those of the target population that is not participating in the experiment. External overlap, on the other hand, requires having all subgroups from the target population represented in the experiment. Thus, not only have field experiments to guarantee (internal) unconfoundedness, they also have to fulfill the external unconfoundedness condition, and have internal and external overlap to yield generalizable, causal insights, making them the "gold standard for drawing inferences about the effect of a policy" (Athey and Imbens, 2017b).

Fulfilling these additional conditions is a major difficulty in many field experiments: Representativeness of the participating sample is the first challenge, as the potential sample usually is small (Athey and Imbens, 2017b). Reasons for this are manifold, and frequent ones are budgetary constraints by the researchers, geographical or institutional preconditions (e.g., if the intervention is implemented at county level), or high attrition rates. The willingness of participants or institutions to participate may aggravate the issue, and in particular result in a violation of external unconfoundedness, causing a site or self-selection bias (Allcott, 2015). Internal overlap or balance in pre-treatment characteristics more generally is also likely to be limited in small samples, thus decreasing external validity and precision of estimates (Bruhn and McKenzie, 2009); this is particularly relevant for experiments in which treatment assignment is implemented at a superordinate level, e.g., at school level instead of student level.

In this paper, we systematically analyze internal and external unconfoundedness and overlap conditions in a field experiment. We provide suggestions how to measure overlap and show how it relates to the precision of treatment effect estimates in a large-scale recruitment experiment with 3,305 public schools. Moreover, we investigate the ability of alternative treatment assignment procedures to pure randomization to generate

[1] See, for example, the rise of behavioral research units installed by governments and inter- or supranational organizations to evaluate policies around the world: https://www.oecd.org/gov/regulatory-policy/behavioural-insights.htm. This is likely to expand in the future as digitization lowers the cost of implementation (Athey and Imbens, 2017a).

overlap and balancedness of treatment groups in this experiment with up to seven treatment groups.[2] We address external unconfoundedness and external overlap by analyzing potential self-selection of public schools in our experiment with rich administrative data.

In order to draw conclusions about the importance of overlap with respect to precision of estimation, we systematically varied the treatment assignment methods in this experiment. We allocated schools – before contacting them – into treatment arms using different treatment assignment methods in several disjointed sub-samples. We compare pure randomization to the new minimum mean squared error (minMSE) treatment assignment method (Schneider and Schlather, 2017). Additionally, we implement two benchmark methods, pair-wise matching and re-randomization based on t-statistics, methods frequently implemented in the evaluation literature and by practitioners (Bruhn and McKenzie, 2009).[3] The basic idea of the minMSE method is to re-randomize treatment status a given number of times, optimizing a theoretically derived statistic of balancedness. This procedure achieves balancedness by maximizing covariate variance in all treatment groups accounting for correlations. We selected the minMSE method as 'treatment method', since (i) it focuses on balance in higher moments of the covariate distribution than just in the mean (Schneider and Schlather, 2017), thus is theoretically suited to achieve overlap, and (ii) for its flexibility to fit the needs of the recruitment experiment.

The recruitment experiment allows us to investigate whether the external unconfoundedness condition is likely to be fulfilled in a large-scale application and to address whether researchers may increase the participation probability in an educational field experiment. The study was conducted in the state of North Rhine-Westphalia (NRW) in Germany, which provides an ideal test bed, as headmasters are by law the gatekeepers for scientific studies. Moreover, the state of NRW provides detailed school- and municipality-specific information. We therefore could contact schools directly, avoiding self-selection at a higher level such as the state school administration or the ministry of education. Additionally, we gathered extensive administrative data from the responsible school authority. In the recruitment e-mail that we sent to headmasters, we varied the content along two dimensions: the main topic of the planned field study and the extrinsic financial incentive to participate. Our recruitment e-mails 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. Schools invited to answer the survey serve as our control group, as it measures the headmasters' willingness to invest a minimum amount of effort by responding to our e-mail. Recruitment e-mails to participate in the field experiment highlighted three different research topics: (i) e-learning, (ii) parental involvement, or (iii) integration of migrant children. In combination with our rich set of administrative data on school and

[2] In this study, we assess different notions of balance in pre-treatment characteristics, where overlap is the notion that we focus on for its theoretical relevance; nevertheless, our experimental variation affects balance in general, and is not limited to overlap. For ease of exposition and when confusion is unlikely, we write overlap and mean overlap and balance in general.

[3] While pure randomization assigned schools to the control group or one of the treatments by pure chance, pair-wise matching finds pairs of units that are comparable in covariates, and then one is randomly assigned to the treatment and one to the control group. Using re-randomization, the researcher performs a random assignment multiple times until a criterion of balancedness – which is sometimes the subjective judgment of the researcher – is met and the process is stopped, or until a certain number of assignments have been created from which the best is selected (see, e.g., Bruhn and McKenzie, 2009). The implemented benchmark method picks the assignment, in which the maximal t-statistic from regressions of considered variables on treatment status is minimal.

municipality characteristics, this allows us to shed light on self-selection into participation and thus on a increased likelihood of a violation of the external unconfoundedness condition.

Our main contribution is the systematic analysis of validity conditions in field experiments. For observational studies, it has been shown that overlap influences the precision of estimates and that limited overlap complicates inference (e.g., Cochran, 1968; Crump et al., 2009; Rothe, 2017). While statistical improvements have been suggested, they remain imperfect as they either rely on a minimal amount of overlap (e.g., Cochran, 1968) or imply a focus on a subsample (e.g., Crump et al., 2009). It remains unclear, however, whether experiments are likely to suffer from limited overlap and what options researchers have to increase overlap already at the design stage, thus completely avoiding any complications due to limited overlap.[4] Appropriate treatment assignment methods have been shown to increase balance in the means of variables in small and moderately sized samples, where otherwise balance can be rather limited (e.g., Bruhn and McKenzie, 2009). However, the abilities of these methods to achieve overlap, which is a stricter requirement with respect to balance targeting the entire covariate distribution, remain unstudied. Theoretically, overlap should be improved whenever the whole distribution of pre-treatment information of the sample is taken into account when assigning treatment, or when the sample size is increased, but if a method economically improves overlap remains unclear. We are the first to empirically document the relevance of overlap, and to show how overlap can be improved, in a (field) experimental setting.

Self-selection into laboratory and artefactual field experiments threatening and violating external unconfoundedness has been documented widely (see, e.g., Lazear, Malmendier, and Weber, 2012; Charness, Gneezy, and Kuhn, 2013; Abeler and Nosenzo, 2015; Schulz et al., 2019). However, whether samples in field experiments deviate substantially from a random selection, and thus potentially violate the external unconfoundedness condition, is a fundamentally different question. In natural field experiments, participants do not know that they are taking part in an experiment, and their participation often depends on the leaders of the entities at which the experiment is conducted. Similar to lab studies, potential site or self-selection implies a non-representative participant pool. However, as the population of interest in field studies is usually more heterogeneous, this is more likely to happen and thus to threaten the scalability of field experiments (Al-Ubaydli, List, and Suskind, 2019; Czibor, Jimenez-Gomez, and List, 2019).[5] As generalizability of field-experiment results is key for policy making, investigating site or self-selection in field experiments that potentially violates external overlap and external unconfoundedness is of utmost importance.[6] As Belot and James (2016) have observed, most field-experiment studies provide little or no information on how partici-

[4]Studying experiments at multiple utilities, Allcott (2015) argues, by use of approximations, that overlap is unlikely to be the cause of rather wrong predictions with respect to the treatment effect from the first treated utilities to the remaining ones; due to lack of the relevant data, however, this remains untested.

[5]Similar to the assumption in Allcott (2015), Czibor, Jimenez-Gomez, and List (2019) list four threats to the generalizability of (field) experimental results: characteristics of experiment, selective noncompliance, non-random selection, and different populations. See also Deaton and Cartwright (2018) for a discussion on randomized controlled trials.

[6]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 to design sample recruitment plans for school studies in the US.

pants were recruited and the experimental sample is rarely compared to the broader population of interest.[7] 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, while Belot and James (2016) analyze the selection of local school authorities in a policy-relevant experiment. However, we still know little about site-selection in a public institutional setting and whether the degree of self-selection in field experiments varies by research topic and extrinsic incentives for participation. We contribute to this literature by investigating site-selection in a field experiment with public schools in Germany, where, contrarily to Belot and James (2016), headmasters of potential partner schools are not pre-selected by school authorities, but addressed directly. Moreover, we are the first to study whether and how participation is affected by the research topic of a study and by provision of extrinsic incentives for participating institutions.

We also contribute to the literature on treatment assignment in field experiments. So far, studies comparing treatment assignment methods have focused on comparing mean values of one or several variables between the experimental groups (Bruhn and McKenzie, 2009; Morgan and Rubin, 2012), but have never considered their ability to balance additional aspects of the covariate distribution, such as dependencies and overlap. For instance, re-randomization based on differences in mean values (e.g., the benchmark method focusing on t-statistics from comparing group means) could result in one group consisting of all middle-aged participants, and another group with all young and all old participants, and thus yield disjunct groups with equal mean age.[8] This illustrates that neither overlap, balanced variances, or balanced dependencies among the covariates are guaranteed by the use of any given treatment assignment method. Moreover, simulations, although powerful and cost effective tools, crucially depend on assumptions that may or may not reflect reality and thus their validity for real-world situations may be limited. Lastly, up to now, different treatment assignment methods have been assessed with respect to their capacity to form balanced experimental groups with binary treatments only, but have not been tested within a real field-experiment setting with several treatment arms (e.g., Greevy et al., 2004; Bruhn and McKenzie, 2009). We compare pure randomization, the minMSE treatment assignment method, re-randomization based on t-statistics and pair-wise matching with respect to their ability to form several comparable treatment groups as measured by overlap and a criteria that accounts for dependencies between the covariates (Hansen and Bowers, 2008). In contrast to previous work, we rely on a real field experiment instead of running simulations.

A very recent literature sheds light on the normative appeal of randomized controlled trials and the underlying treatment assignment procedures. Banerjee et al. (2020) show that an ambiguity averse decision maker who wants to provide robust evidence on potential treatment effects prefers pure randomization or

[7] Belot and James (2016) find that only 3 out of 24 studies compare the experimental sample to the broader population. The authors focus on experiments in the fields of policy evaluation, personnel economics, and development economics in the Top-5 journals and in the American Economic Journal: Applied Economics.

[8] To a certain extent, this could also happen when using stratification/blocking: If, e.g., age is discretized to yield two groups, and randomization is then conducted in both age groups, there is no guarantee that in the resulting experimental groups both children and retired persons are represented; one experimental group might also consist of mostly elderly and those aged close to the cutoff for discretization, while the other experimental group mostly consists of those close to the cutoff and the young ones.

certain forms of re-randomization compared to deterministic treatment assignment. We connect to this literature by empirically comparing several re-randomization techniques and pure randomization.

Our results show that estimations from sub-samples with improved overlap have a significantly lower bias, with a more than 35% reduction in bias, compared to estimations from sub-samples with limited overlap, thus highlighting both the relevance of limited overlap in experiments, and the appropriateness of the minMSE method to achieve overlap and increase precision. This result is comparable to Crump et al. (2009), who empirically show that appropriately censoring data with limited overlap can lead to a reduction of the variance of the average effect by 36%. When addressing overlap already at the design stage, however, censoring can be avoided.

Moreover, overlap and balance decrease with the number of treatment arms, but appropriate treatment assignment can assign up to 2.5 more groups than purely random treatment assignment with the same decrease in balance. These results also hold for an alternative measure of treatment group balance (Hansen and Bowers, 2008). For pair-wise matching, we cannot draw a statistically sound conclusion, but in comparison to re-randomization based on t-statistics, the minMSE method is superior in achieving overlap.

With respect to (external) unconfoundedness and site-selection, we find that schools where headmasters responded positively to our e-mail do not significantly differ from schools that did not respond or actively opted out, with respect to observable characteristics. However, we find that the topic of the experiment increases the number of positive responses. The topics of parental involvement or on the integration of migrant children were more likely to be followed by a positive response than the control group. Interestingly, neither the topic of e-learning nor offering financial extrinsic incentives had a positive effect on responses. These findings show the importance of carefully choosing the characteristics of the study to be highlighted in the recruitment e-mail in order to increase the success of researchers in increasing the institutional gatekeepers' willingness to participate. Taken together, our results suggest pre-testing potential recruitment setups in order to ensure external unconfoundedness and external overlap, show how, by the use of the minMSE method, overlap can be achieved.

2 Theoretical Background: Requirements for Estimation and Extrapolation of Treatment Effects using (Multisite) Experiments

In this section, we lay out the requirements to estimate consistently the average treatment effect in experiments and to extrapolate it to a target population, when the treatment allocation is conducted at a higher level than the individual with only a sample of the higher level entities. Allcott (2015) defines those entities as "sites": a setting in which a program is implemented, characterized by a population of individuals, a treatment, and an economic environment. Examples for a site are schools, hospitals, private companies, or also NGOs. The requirements laid out below guide our empirical investigation, where we build on previous

work by Allcott (2015), closely related to work by Belot and James (2014) and Belot and James (2016). From this, we derive the conditions for the validity of field experiments that we assess empirically.

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 the individual treatment effect, $\tau_i = Y_i(1) - Y_i(0)$, and X_i is a vector of covariates where X constitutes the support of the covariates. The target population is the population for which one would like to estimate the *Average Treatment Effect* (ATE). The sample population is the population that was exposed to the experiment. $D_i \in \{1, 0\}$ indicates if unit i is in the sample population. Sites are numbered, and S_i is the number of the site that individual i belongs to. The ATE at a site s conditional on $X_i = x$ is defined as $\tau_s(x) = \mathbb{E}[\tau_i \mid X_i = x, S_i = s]$ and, following Allcott (2015), we assume that either all individuals belonging to a site are in sample or target, i.e. $D_s \in \{0, 1\}$. Then, 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 (multisite evaluation). $\mathbb{E}[\tau_s(x) \mid D_s = 1] = \mathbb{E}[\tau_s(x) \mid D_s = 0]$

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

The first two conditions ensure that in a given experiment at a given site, the average treatment effect is correctly identified when comparing group means of the outcome variable conditional on X_i, they thus address *internal validity*. The second two conditions are needed to extrapolate the average treatment effects from all sites in a sample to a broader target population and thus address *external validity* of an experiment.

2.1 Empirical Assessment of the Overlap Assumptions: Balance in Pre-Treatment Characteristics

One interpretation of the overlap assumptions is the requirement that no covariate value or combination of covariate values perfectly predicts either receiving the treatment or participating in the experiment (Belot and James, 2016). From the viewpoint of identifying and extrapolating an average treatment effect, it can also be interpreted as a certain notion of balance: All covariate values or combinations of covariate values have to be represented in all experimental groups, or in sample and target sites, respectively. For example, if X were an indicator for being female, both possible values would have to be found in treatment and control groups, or sample and target sites, respectively, to identify the average treatment effect conditional on this value.

We use this interpretation of the overlap condition to assess the importance of its fulfillment for – and more generally the impact of balance of treatment and control groups with respect to pre-treatment covariates

on – the precision of estimating the average treatment effects. This effect serves as a measure of the degree to which the fulfillment of this arguably abstract condition has an impact on the validity of an experiment.

Allcott (2015) argues that the lack of external overlap in the Opower context is unlikely to matter for the prediction error observed when extrapolating treatment effects from the first 10 sites to the later 101 sites. Nevertheless, as the distribution of target sites, $f_{D=0}(x)$, is unknown, a clean test for overlap in the Opower context is not feasible. In the present study, the distribution of the entire population is known, and therefore our study has the potential to fill this gap. Moreover, we can provide a clean test for the importance of the overlap condition for precision of estimation, although strictly speaking we rather address the overlap instead of the external overlap condition. Our results are informative for the external overlap condition as well, as it is technically very similar, and its fulfillment can be ensured with the same mechanism, namely appropriate treatment assignment.

We expect that an increase in overlap leads to an increase in the precision of the estimation. Moreover, treatment assignment methods that aim at balancing the whole covariate distribution – pair-wise matching and the minMSE method – should perform better in creating overlap than pure randomization and mean value-based re-randomization (such as those schemes focusing on t-statistics), hence leading to more precise estimates.

2.2 Empirical Assessment of the Unconfoundedness Assumption in Multisite Evaluations: Site-Selection Bias

External unconfoundedness in single-site evaluations, usually formulated as $D_i \perp (Y_i(1) - Y_i(0))|X_i$, is equivalent to our formulation in Assumption 3 – external unconfoundedness in multisite evaluations when extrapolating from many sample sites to many target sites (Allcott, 2015). For the importance of multisite evaluations in development, health and educational economics and policy evaluation in general, where researchers implement treatments with the help of, e.g., MFIs, hospitals, or schools, we focus on the multisite version of this assumption.

This assumption can be tested by assessing whether any variables moderate both the selection and the treatment effect (Allcott, 2015; Belot and James, 2016). We follow this route and investigate whether self-selection of partner institutions has to be expected in a typical field experiment with multiple sites.[9]

We follow Belot and James (2016) in deriving hypotheses on self-selection of schools. According to their model, the main determinants of selection into an experiment are beliefs about the effectiveness of the treatment, probabilities of being assigned to treatment and control groups, costs of carrying out the intervention and participating in the experiment, and the subsidies provided by the experimenter (e.g., paying for materials and data collection). These determinants can lead to both negative or positive self-selection. Furthermore, to determine the direction of self-selection, Belot and James (2016) exploit the information on observables at the time of selection and pose assumptions regarding the relationship between observables

[9] For a more general model on self-selection in RCTs, see Belot and James (2014).

and the variables determining selection. As our setting is comparable to Belot and James (2016), we follow the authors' idea in formulating hypotheses on self-selection for the three categories of observable factors that could affect the headmasters' willingness to participate: the characteristics of the schools targeted, the degree to which schools are able to implement the intervention, and the degree to which schools care about the outcome of interest.

Characteristics of the target schools For each of the proposed research topics (e-learning, parental involvement, and integration of migrant children), we expect different characteristics of the schools tp play a role for potential self-selection. Integration-Migrant-Children Treatment: We have detailed information at school level about the migration background of children and their parents and expect that a higher share of migrant children is correlated with a higher willingness of headmasters to participate in an experiment on this topic. Parental-Involvement Treatment: We have information on the unemployment rate and the social index of the municipality and expect that schools in municipalities with a higher unemployment rate and lower social index are more interested in participating in an experiment on parental involvement, which might increase the treatment effect. E-Learning Treatment: To proxy the schools' unobserved technological infrastructure, we use information on the land prices in the municipality and the age of the teachers within a school. Both are likely to be correlated with e-learning, as schools in richer municipalities might be technologically better equipped, and schools with younger teachers, on average, younger teachers might have a stronger affinity with technology than schools with older teachers. We therefore expect that schools in richer municipalities and on average younger teachers are more inclined to implement new e-learning programs.

Degree to which schools are able to implement the intervention We have information at school level on the share of teachers employed full-time, compulsory teaching hours, and the number of classes and students. As participating in the study and implementing the intervention incurs costs for the school (e.g., time, rooms, and personnel), we would expect that larger schools have a higher belief about the effectiveness of the treatment and are therefore more likely to participate.

Degree to which schools care about the outcome of interest We do not have direct information on the degree to which schools promote one or the other topic by, e.g., already existing extracurricular programs. Moreover, the degree to which schools care about the outcome of interest is likely to be correlated with headmaster characteristics for which we also have no direct measure. However, characteristics of headmasters and the existence of extracurricular programs are likely to be correlated with the characteristics of the target population described above.

3 Experimental Setup

In this section, we first describe the principles of our recruitment experiment, designed to analyze whether there is evidence for any violation of the (external) unconfoundedness condition in a typical multisite field experiment in an educational setting, and to gain insights on how participation might be increased in these settings. We then describe the experimental design of the integrated experiment on overlap (or balance more generally) and precision of estimation.

3.1 Recruitment Experiment: Testing the Unconfoundedness Condition

We conducted the recruitment experiment in North Rhine-Westphalia (NRW) from October 2016 to January 2017. The institutional preconditions in NRW are ideal for our research question on the external unconfoundedness condition, i.e., site-selection bias, as headmasters are allowed to decide autonomously whether or not they wish their school to participate in scientific studies, without the permission of the school authority. This allows us to contact the relevant gatekeepers directly, while avoiding potential self-selection at a higher administrative level, such as the school authority. Moreover, being Germany's most populous state, NRW has a high number of schools, making it a suitable and relevant test ground. 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 the headmasters' costs of responding to our inquiry, all contact with schools was electronically, and we asked headmasters about participating in a scientific study. Using electronic communication comes at no cost: We learned in previous studies that the responsiveness of schools in NRW does not depend on whether we send a posted letter or an e-mail (see Panel B of Table 2).[10] Recruitment e-mails were sent out on 2 October 2016 and for those schools that did not respond – neither positively nor negatively – we sent out two reminder e-mails.[11] The reminder e-mails were already announced in the first invitation e-mail in order to induce schools to give feedback and in order to achieve a meaningful opting-out measure. In the first reminder, we also announced that they would be contacted again unless they responded by a given deadline. The last short reminder was sent three weeks later.

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 municipalities not associated with a county. We excluded school types (a) and (b), as not all our research topics are relevant 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 likely to be over-researched as they all are home to at least one university, and thus receive many inquiries, e.g., from bachelor and master students, which might introduce noise in the

[10] 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 schools by e-mail, posted letter, or a combination of both.

[11] We sent e-mails in batches of 50 per two-hour interval on mailing day using the internal LimeSurvey procedure to handle invitations to surveys. In total, about 3% of e-mails could not be delivered due to technical reasons. The first reminder e-mail was sent one month after the first contact, on 2 November 2016. The second reminder was sent three weeks after the first.

measurement of willingness to participate in our study. Second, we were concerned about reputation effects and ongoing partnerships in schools in larger cities. We had previously conducted three 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 (we use the data of our previous experiments to shed some light on potential evidence of a violation of the unconfoundedness condition or a site self-selection bias in larger cities in Appendix G). Moreover, schools with already existing partners and ongoing programs might or might not be more likely to participate (Allcott, 2015). In total, 3,305 schools were contacted, which represents 66.29% of all schools in NRW.

Recruitment e-mail We asked headmasters to express their interest in participating in a scientific study. This message introduced the researchers and their expertise in conducting scientific studies in schools, mentioned the respective research question, briefly explained the methodology, and outlined the expected workload for the school (see the online Appendix E for a facsimile of the recruitment e-mail). We kept the information in the e-mail short to increase the likelihood of headmasters reading the message. However, headmasters could access more information on the project – the scientific foundation of the research question, the timeline of the study, the exclusion criteria for participation, and the information about data protection – by clicking on a link provided in the e-mail. Moreover, to measure the schools' responsiveness, headmasters could indicate their interest in participating by clicking one of three links displayed at the bottom of the recruitment e-mail. This decision (plus the idle option of not responding at all) serves as our outcome measure. Clicking the first link, headmasters could express a strong interest in the project and were told that they would 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 saw no capacity to participate at that particular time and wished to be contacted again. 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 directed to a questionnaire asking for further details about the school (e.g., what the position of the respondent is within the school).

3.1.1 Treatments

We implemented a 3×2 plus control treatment structure, 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 e-mail, announcing our plan to conduct an experiment about the respective research topic within schools. The fourth paragraph of the e-mail 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 (see Appendix F for the survey). In this survey, we asked about the headmasters' point of view regarding the collaboration between academia and schools, i.e., how insights gained in academic research can be integrated into the everyday life of schools. Importantly, answering the survey did not involve participation in any experimental study and it required a minimum of the 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 the schools' baseline responsiveness in dealing with inquiries of academic researchers.

E-Learning Treatment In the E-LEARNING TREATMENT, we suggested participating in a study on the use of electronic devices in education. The presented research question was to find out which types of electronic testing formats could be implemented in schools and how they performed 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.[12]

Parental-Involvement Treatment In the PARENTAL-INVOLVEMENT TREATMENT, we asked for participation in a study aiming at analyzing the effect of getting parents involved in their children's education (e.g., the students' behavior in class and their 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-Migrant-Children Treatment In the INTEGRATION-MIGRANT-CHILDREN TREATMENT letter proposing the topic of integrating 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 rapidly into the school environment.

Monetary-Incentive Treatment Beside the research topic, we altered whether schools were offered a financial incentive. Two schools could win a 700 Euro budget in the case of participating. This money could be 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 the

[12] For an overview of all programs initiated within this effort, see, e.g., http://www.bildung-forschung.digital/.

gatekeepers' attention and to increase their willingness to respond to the e-mail. 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 Euro per teacher employed full-time). For more than 70 % of the schools our incentive constitutes a share of 80 to 90 % of the yearly budget.[13]

3.1.2 Implementation

Contrary to most of the fairly similar "audit and correspondence studies", e.g., those summarized in Bertrand and Duflo (2017), we actually intended to implement the suggested experiment in schools after the summer break in 2017. The goal of the planned experiment was randomly to provide schools with a digital class-book and to analyze its impact on the behavior of students in class, as well as their educational achievements, where features of the class-book addressed the treatment topics above. However, as the public demand for data protection measures of sensitive student data is rather high in Germany, the partner institutions were very concerned about this topic. Meeting the highest standards with respect to data security with the digital class-book (e.g., encryption, secure authentification, storage of data on schools' servers instead of on ours, ...) increased the complexity in programming and the support needed by the partner institutions to a degree where it was not feasible to stick to the initial time and budget plan. We therefore decided to postpone the study and implement it in a different format.[14]

3.2 An Experiment on Overlap, Balance in General, and the Relation to the Precision of Estimation

We conducted an experiment within the recruitment experiment outlined above to study empirically the relation between precision and overlap of, or, more generally, balance in observable characteristics in a real-world setting.[15] The key feature of our research design is that we use real treatment effects without the need, but – maybe more importantly – also without the freedom, to make assumptions about the possible nature and magnitude of the treatment effect, as commonly done in simulation studies; in this sense, compared to simulation studies, our hands are tied, and the results are thus more credible and surely reflect relevant real-world conditions.

First, we divided the whole sample of schools into smaller, comparable sub-samples, and experimentally varied the degree of overlap or balance of covariates, more generally, in these sub-samples by use of different

[13] Clearly, in terms of expected value this financial incentive is rather low (11.76 Euro 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 Euro 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. Although we cannot test for headmasters' perceptions, we do not find any evidence either in the survey for the support of expected value calculations.

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

[15] Theoretically, Rothe (2017) shows that limited overlap 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 as measured in standard errors.

treatment assignment methods: Mainly, we use pure randomization, and the minMSE method (Schneider and Schlather, 2017). After our recruitment experiment, we assessed the precision of the estimation in the sub-samples. We then related precision to pre-treatment balance, as measured, e.g., in overlap.

In order to inform researchers about the ability of commonly used treatment assignment methods to achieve overlap or balance, more generally, we implemented two benchmark methods for treatment assignment in two of our twelve sub-samples: a pair-wise matching approach and re-randomization based on t-statistics.

3.2.1 Division of the Sample in Treatment and Control Group (Experiment on Overlap, Balance and Precision)

Our total sample consists of 3,305 schools. From this sample, we randomly draw 12 sub-samples. In order to gain insights into how strongly balance decreases with an increasing number of treatment arms, we draw sub-samples consisting of increasing numbers of schools, so that we can assign between one and six treatment groups with equal numbers of schools; see Table 11.[16] 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.[17] Comparability, or Balance, was checked with the omnibus test of equivalence between groups introduced by Hansen and Bowers (2008). The p-value for the null hypothesis of equivalence ranged from .25 to .99, with a mean of .63, and, most importantly, was never smaller than .10.

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 (i.e., the treatment group'balance'), and the other sub-sample to a comparison method (i.e., the control group). For 10 pairs, pure randomization was the comparison method, and for one pair each, re-randomization based on t-statistics and pair-wise matching were chosen as comparison methods, respectively (see Table 11).

Treatment Assignment for Remaining Schools After having allocated the schools in 12 sub-samples (matching & minMSE sub-sample, rerandomization & minMSE sub-sample, and ten randomization & minMSE sub-samples) to experimental groups, around one third of the sample was not assigned an experimental group. Taking into account the treatment assignments already made, using the minMSE method, we allocated those remaining schools to the control and the treatment groups, with the restriction of having the group sizes as equal as possible and the goal of achieving 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 minimal p-value when testing the null hypothesis of balanced groups between any treatment group and the control group is 0.87.

[16] Note that by the design of our recruitment experiment we are limited to six treatment groups (for our 2×3 design, see Section 3.1.1).

[17] Comparability of the groups of schools – or balancedness among the covariates or observables of the groups – was achieved with an algorithm using the same statistic of balance as the minMSE method.

3.2.2 Treatments and the Choice of Treatment Assignment Methods

For this added experimental layer we had to account for the nature and number of the actually available pre-treatment information, in order to not pose any risk to the recruitment experiment. 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, such as a sharp 1:1 division of, e.g., both females and males in the treatment and control groups. Therefore, in the overall setting of our experiment, treatment assignment using stratification is either practically difficult to impossible or not necessary.[18] 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.

As the minMSE method proposed by Schneider and Schlather (2017) is one of the few methods – if not the only one – that is theoretically derived, can handle this setting (multiple, possibly continuous pre-treatment characteristics), in particular also for an increasing number of treatment arms, and can easily be applied via software implementation in Stata and R (package `minMSE` Schneider and Baldini, 2019) without consulting a technical documentation, we opted for this method as our 'treatment' method for the treatment group 'balance'. The flexibility of the method allows us to keep the method constant over the increasing number of treatment arms. In the following, we give a brief explanation of all the methods considered, present their implementation, and discuss the choice of the treatment assignment methods considered as 'control'.

The Minimum Mean Squared Error Treatment Assignment Method 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 one of the implementation options of this notion of balance that he applies. Building on a simpler theoretical framework, Schneider and Schlather (2017) extend the notion of balance 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 while allowing the same degree of flexibility.[19] Importantly, while Kasy (2016) proposes to optimize balance by using the derived statistic of balancedness without randomization, i.e., using a 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

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

[19] For the method resulting from the work 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.

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, a worse current iteration is also 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 inference can be applied: The implementation of the method in the provided R package `minMSE` (Schneider and Baldini, 2019) is even able automatically to provide different alternative test vectors to the actually used assignment vector to perform, e.g., permutation inference for a conservative and non-parametric assessment of significance of treatment effects.

We implemented the method with a preliminary version of the Stata ado-package provided by the authors. We ran 1000 iterations, for being the time equivalent of 500 draws of re-randomization based on t-statistics (see below). Other than that, we used the program's default values, in particular for controlling the optimization process.

Standard Method 'Pure Randomization' Pure randomization is the easiest way of allocating subjects to treatment and control groups. Several means of randomization can be used, e.g., a dice, a coin, birth dates, or also a a software, e.g., R or Stata.

While its advantage is flexible, easy, and credible implementation, in particular also in field settings, the drawback of this method – potentially very different treatment groups leading to wrong estimates and wrong conclusions drawn from the experiment – is consequential and dramatic, and was already discussed almost a century ago by Fisher (1935).

In terms of imbalance, pure randomization can be seen as the reference and should therefore be the method of the control groups. We compare pure randomization to the minMSE method for assignment of two to seven experimental groups, each consisting of 20 schools; see Table 11 for the experimental design.

We implemented pure randomization via the generation of a random variable in Stata that we used for sorting all observations. Then, the row number was divided by the number of experimental groups where the remainder indicates the treatment group.

Benchmark Method 'Pair-Wise Matching' (Binary Treatment Only) Pair-wise matching is a two-step procedure. First, pairs are formed within the sample. The idea is that both units in a pair are as similar as possible; for multivariate pre-treatment characteristics usually a generalized distance, such as the Mahalanobis distance, is used to assess similarity. When pairs are formed, one unit each is randomly assigned to the treatment group, while the other is assigned to the control group.

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. The method is particularly attractive for its ability to consider multivariate, possibly continuous, pre-treatement characteristics and for its theoretical

characteristics: Given a sample to divide into two groups, it maximizes the minimum level of generalized variance within each group (Schneider and Schlather, 2017). Put differently, it makes sure that the resulting groups have representatives of all sub-groups found in the sample, and thus is an ideal candidate for fulfillment of the overlap condition. However, a drawback 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.[20] In those cases, it is common practice also to drop their counterparts from the sample, to maintain balance, and to ensure consistency of estimated treatment effects (e.g., Donner and Klar, 2004; Fiero et al., 2016). This might eventually limit the power of the study below a critical threshold, in particular in cluster randomized trials; see, e.g., the case study in Schneider and Schlather (2017).

We compare the pair-wise matching approach with the new minMSE approach in a setting of binary treatment in absence of attrition. Mimicking a standard use-case for the matching approach to treatment assignment, we use a 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', Beck, Lu, and Greevy, 2016) that accompanies Lu et al. (2011).

Benchmark Method 'Min-max-t-statistic Re-randomization' (Multiple Treatment Arms) Re-randomization generally follows the following principle: Instead of one random assignment of units to treatment and control groups, several assignments are performed. The process is either stopped according to a certain criteria (e.g., a statistic falling below a threshold Morgan and Rubin, 2012) or after a certain number of iterations has been performed. The min-max-t-statistic-method, popularized by Bruhn and McKenzie (2009), consists of selecting the treatment assignment vector with the smallest maximal t-statistic resulting from regressing the pre-treatment characteristics on the group status. For this reason, we compare the minMSE method with this method in case of multiple treatment groups.

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 standard approach to allocate units to more than two groups using matching or any alternative treatment allocation method. The min-max-t-statistic however, can be extended relatively easily to multiple treatment arms. This flexibility is one of its advantages. Moreover, it ensures that the common table in publications of RCT studies showing pre-treatment mean values in covariates across the treatment groups evokes the impression of comparable groups. A disadvantage of all re-randomization mechanisms that we are aware of (e.g., Bruhn and McKenzie,

[20] Typical examples of settings where attrition might be problematic include repeated measurements at schools where, due to illness of participants, 10% of the sample can be expected to be absent on one of the measurement dates or when randomization is performed at the cluster level.

2009; Morgan and Rubin, 2012) – except the minMSE method – is their focus on balancing covariate mean values solely, ignoring higher moments of covariate distributions. This means that it is possible to end up with, e.g., two groups with participants of equal average age, where all elderly and all young participants are in the same group, and all middle-aged participants in the other group. Clearly, such an allocation does not fulfill the overlap condition; moreover, subgroup analysis is not ensured with such an allocation method.

For its flexibility and being a common method, we compare the re-randomization based on t-statistics with the minMSE method in a setting of multiple treatment arms. A common 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-re-randomization when 7 experimental groups are desired with a group size of 30; yielding a total sample size of 210 units.

For implementation of min-max-t-statistic rerandomization, we modified the code from Bruhn and McKenzie (2009) to account for the increased number of treatment groups (seven instead of two) when regressing the covariates on treatment assignment to obtain t-statistics.

3.2.3 Measures of Overlap, Balance, and Precision

Overlap and Balance We assess the balance of pre-treatment characteristics in two ways. The first way 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 on imbalance rather than on balance, but as we are interested in adverse cases, we measure balance by assessing the overlap.

The second way to assess the 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. In other words, the lower the p-value, the higher the imbalance and the lower the balance.

Precision One measure of precision of estimation is the bias of the estimation; it is, for a given experiment, 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 NRW, we interpret the treatment effect in the main outcomes considered (any response, positive response, and participation in our survey) for our treatments (i.e., (no) incentivization combined with the different topics; see Section 3.1.1) using our whole sample as the true value: We first compute the treatment effects for the whole population. Then we estimate the effects for the sub-samples. Finally, we compute the

difference for every sub-sample that we interpret as bias, 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.[21] Our measure of power consists of higher or lower p-values of the treatment effect estimations.

4 Results

This section is organized as follows. We first describe our data and present descriptive statistics. Second, we analyze the effect of overlap on precision of estimation. In a first step, we assess the success of our treatment, i.e., whether using the minMSE method leads to higher levels of overlap (and balance more general) in the 'treated' sub-samples. Then, we compare the degrees of precision (expressed in bias and p-value of treatment effect estimates) in the treated and untreated sub-samples, and relate them to the degrees of balance. Thereafter, we analyze whether we find evidence for a violation of the (external) unconfoundedness condition, that is, we investigate whether in our setting a self-selection bias into participation in experiments exists at the institutional level. Finally, we present results on the treatments in our recruitment experiment, providing insights on how to attract the headmasters' attention.

4.1 Data and Descriptive Statistics

We gathered a rich set of official data on observable characteristics at both the municipality and the school level. School-level data were provided by the statistical office of NRW specifically for our study, and municipality-level data are publicly available from the German statistical offices. These data include – at the school level – the school type, the number of students, the average age of teachers, the compulsory teaching hours of teachers, and information on the migration background of students and their parents. Data at the municipality level comprise the number of inhabitants, unemployment rate, election results, land prices, composition of the workforce, and the social index of the municipality.[22]

Analyzing the response rates of the recruitment experiment, Panel A in Table 1 summarizes whether schools did not respond, actively opted out, showed "light" interest (this means clicking on a link indicating they wish to be contacted later), or responded positively ("strong" interest). We observe that most schools did not respond to our inquiry, ranging from 71.7% in the Parental-Involvement Treatment to 78.2% in the E-Learning Treatment (pooling treatments with and without extrinsic incentive). Active opting-out is highest in our Control Treatment (20.6%) and lowest in the E-Learning Treatment (13.5%). Positive response rates are lowest in the Integration-Migrant-Children Treatment (3.7%) and highest in the Control 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.

[21] 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.

[22] A complete list and detailed description of the background characteristics can be found in Appendix D.

These response rates are comparable to the response rates of other studies with schools in NRW (Riener and Wagner (2019), Fischer and Wagner (2018), and Wagner (2016)). 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 non-response rates in Wagner (2016) of 86.12% and 76.8% in this study. We consider these differences in non-response rates to be small in light of the differences in research topics and the stakes of the experiments, i.e., the effort needed for participation in the study, which vary from very low (this study) to high (in Fischer and Wagner, 2018).

We sent our inquiry to the schools' official e-mail address, and asked for the respondent's position within the school. 840 schools responded to our recruitment e-mail by clicking one of the three links provided and 188 (∼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%). Thus, it is indeed mainly the institutional gatekeeper who handled our inquiries.

Tables 3 and 4 present descriptive statistics on background characteristics of schools and municipalities we will later use in our analysis. Columns (1)-(3) show means of the three treatment groups (E-Learning Treatment, Parental-Involvement Treatment, and Integration-Migrant-Children Treatment), 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 treatments and the control group are small for school and municipality characteristics and moreover insignificant.[23] Hence, the treatment assignment procedures achieved overall balance.

4.2 Assessing the Overlap Condition and its Causal Relationship with Precision of Estimation

In this subsection, we first discuss the results on pre-treatment balancedness of covariates with respect to overlap and balance using a test to detect imbalance as proposed by Hansen and Bowers (2008). Then, we assess the differences in the precision of estimating the treatment effects due to the experimentally induced differences in balance by use of the minMSE method and purely random treatment assignment (and re-randomization based on t-statistics and pair-wise matching, although this is not our focus). As measures of precision, we use the bias of an estimation and the p-value of the estimate 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.

[23] 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; note that our sample is relatively large. The difference between the age of teachers in the E-Learning Treatment and the Control Treatment 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 Treatment and Parental-Involvement Treatment and the Control Treatment would, without controlling for multiple testing, imply significance at the 10% level ($p = 0.061$ and $p = 0.093$, respectively).

4.2.1 Balance

Overlap Figure 1 presents 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. Other school data used in the analysis are not publicly available and were not used for the treatment assignment. Municipality data are constant across municipalities, and the municipality ID is already included in the variables considered. Therefore, all other municipality 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 observe if we included all variables considered for treatment assignment. The difference in balance shown in Figure 1 between the sub-samples where treatment was conducted purely at random and the sub-samples where the minMSE method was used is thus likely 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 sub-sample three in Table 11, there are seven groups to be formed, whereas in sub-sample twelve, the characteristic is to be distributed and thus to be found in only two groups. 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 maximum possible number of groups.

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 sub-samples, variables, and characteristics, the overlap condition is fulfilled in 60% of all cases when assigning treatment purely at 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, $N = 534$, p-value < 0.012); this finding is robust to inclusion of sub-samples where the minMSE method is compared with pair-wise matching or re-randomization based on t-statistics (chi-Squared test, $N = 642$, p-value < 0.01).

Figure 1b compares the average share of fulfillment of the overlap condition. Consider, e.g., 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 0.5 for this variable and this characteristic. For every combination of draw, variable, and characteristic of a variable, we obtain one share of fulfillment.

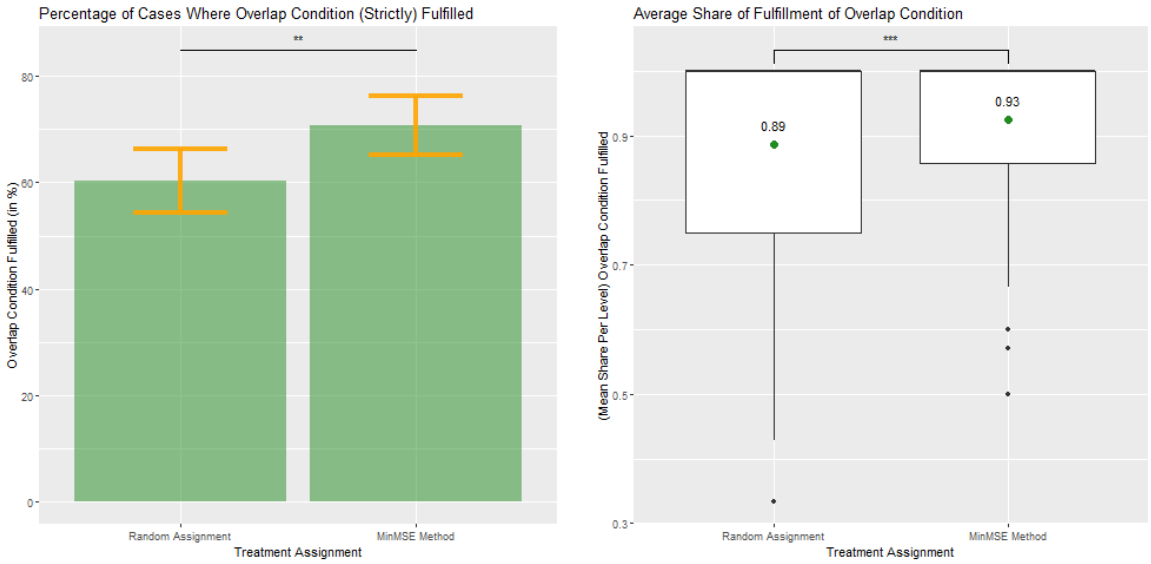
Figure 1b shows that, on average, both treatment assignment perform in a relatively similar fashion. Given that the maximum share is 1, however, the difference is bigger than it seems and it is significant at the 1% level (rank sum test, p-value < 0.01; robust to inclusion of sub-samples where the minMSE method is compared with pair-wise matching or re-randomization based on t-statistics). 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 0.86 (0.5) compared to 0.75 (0.33) 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 Pre-treatment Balance: Overlap Condition

(a) Percentage of Cases Where Overlap Condition is Fulfilled

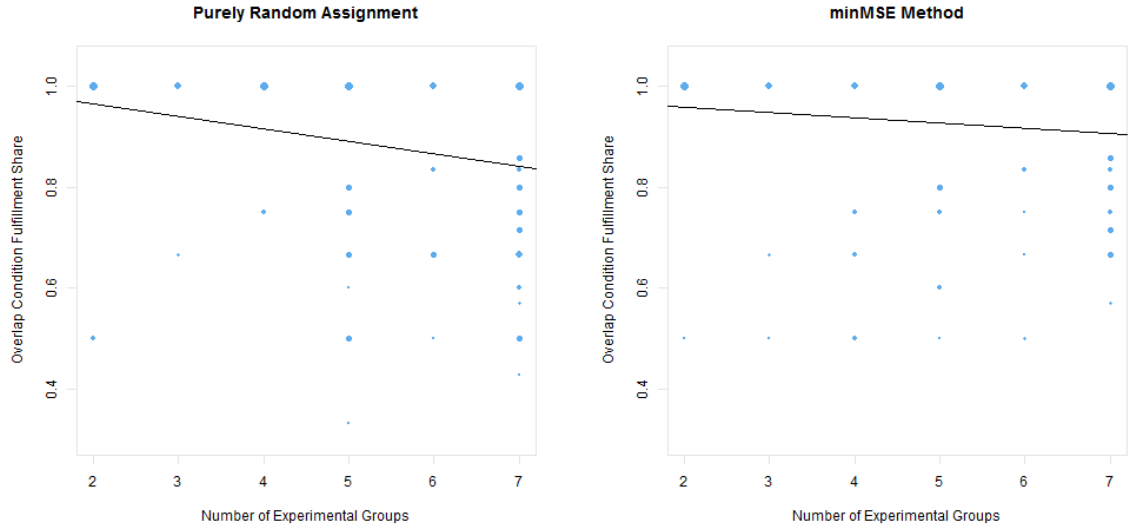
(b) Average Share of Fulfillment of Overlap Condition



*Note: The first graph (Figure 1a) compares the percentage of cases in which the overlap condition is fulfilled. Generally, the overlap condition is considered fulfilled if a characteristic of a variable is found in all treatment groups. ** denotes significance of a chi-squared test at the 5% 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. All results are robust at least at the same levels of significance to inclusion of sub-samples where the minMSE method is compared to matching and re-randomization based on t-statistics.*

The relation between the number of groups to assign and balance is illustrated in Figure 2): The success rate of fulfillment of the overlap condition is higher for the sub-samples with fewer groups to be assigned. However, there is a significant difference (as indicated by an F-test on the coefficients in a model with pooled data, 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. As apparent from Figure 2, the decay in balance as measured by the success rate of fulfillment of 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.

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



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 0.1% level (robust to including sub-samples in which the minMSE method is compared to matching and re-randomization based on t-statistics.)

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 pre-treatment balance considers all variables used for treatment assignment. It is based on a statistic that accounts for correlation between the specified variables, thus "corrects" for comparison of multiple variables across 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 sub-sample. Table 12 summarizes these results by reporting the minimal p-value of all the comparisons between control and the treatment group(s) in a sub-sample. Note that low p-values imply low balance, whereas higher p-values indicate better balance. In none of the groups is the null hypothesis of balance rejected at conventional significance levels. The obvious observation that minimal p-values are larger for the minMSE method is statistically confirmed: The null hypothesis of equality of balance as measured by minimal (mean) p-values is rejected by a Wilcoxon rank sum test with a p-value < 0.01 (p-value < 0.001) independently of including ($N = 24$) or excluding ($N = 20$) pair-wise matching and re-randomization based on t-statistics as comparison methods.[24]

[24]These findings are also robust (p-values < 0.0001) to comparisons without any aggregation of p-values, i.e., using all p-values of all comparisons between control and the treatment group(s)in a sub-sample ($N = 78$ or $N = 92$, depending on whether or not sub-samples are included where the comparison methods are re-randomization based on t-statistics and matching).

The relation between balance as measured by the omnibus test of 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 the number of groups and the minimal p-value is $\rho = -0.91$ (p-value < 0.001) when pooling pure randomization, matching, and re-randomization based on t-statistics, $rho = -0.87$ (p-value $= 0.001$) for the pure randomization samples alone, and it is $\rho = -0.49$ and non significant for the minMSE method. OLS regressions confirm this picture and yield similar results to measuring balance by the overlap condition.

Result 1. ***Balance:*** *Pre-treatment balance between control and treatment groups (as measured by overlap and the omnibus test of balance by Hansen and Bowers) is significantly higher when groups are assigned using the minMSE method compared to purely random treatment assignment. The degree of balance significantly decreases with the number of experimental groups that are to be assigned when using purely random treatment assignment independently of the measure of balance, but not when using the minMSE method. The minMSE method allows us to assign about 2.5 more groups with the same decrease in balance as compared to pure randomization, and the difference is significant.*

4.2.2 Precision

We compare the performance of the different treatment assignment mechanisms with respect to precision in two ways: bias and p-values. The bias of an estimate in the statistical sense is the difference between the true value and the estimated value. Following Bruhn and McKenzie (2009), we also compare precision as indicated by the p-values of the estimations of the treatment effect. For both measures, we consider the three main outcomes in this paper: response (yes/no), positive response (yes/no), and whether the questionnaire was filled at least partly (yes/no), and we pool the results.

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. As we addressed the whole population of schools in NRW (except those meeting our exclusion criteria), we assume that this estimation corresponds to the true value. Then, 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 sub-sample of a sub-sample. We consider all treatment effects resulting from an estimation by using the full sample with a statistically significant estimated coefficient at least at the 10% level. Of the estimations using sub-sub-samples, 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 the 50% level.

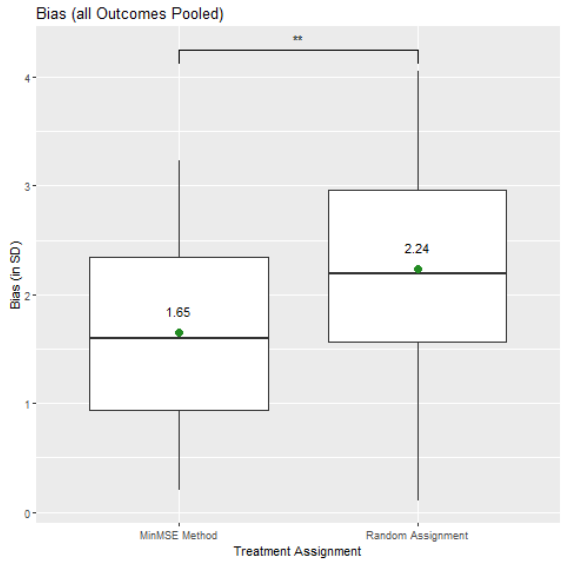
Figure 3a summarizes the result using the bias as a measure of precision. The absolute bias is expressed in standard deviations of the respective treatment effect estimations, i.e., the standard deviations of all estimates of the treatment effect of the respective topic with or without incentivization (e.g., E-Learning Treatment with Monetary-Incentive Treatment). The (absolute) bias differs significantly in the sub-

samples in which treatment was assigned purely at random compared to the minMSE method (Wilcoxon rank sum test, $p < 0.03$; robust to inclusion of sub-samples in which the minMSE method is compared to matching and re-randomization based on t-statistics). On average, the bias is nearly 1.5 times as large when assigning treatments purely at random compared to when using the minMSE method. Moreover, the median bias of estimations in the "purely random assignment" samples is 2.2 and almost as large as the 75% quantile in the comparable "minMSE assignment" samples, where it is 2.3 median 1.4).

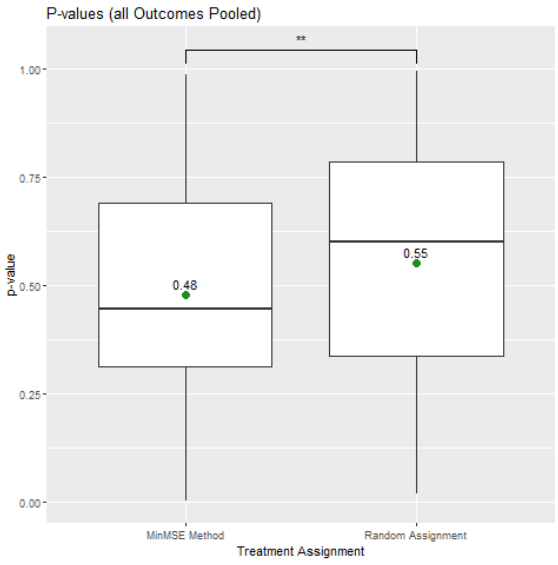
Precision: P-Value of Treatment Effect Estimates We can measure precision also by the p-value of an estimation. Again, we consider all the estimations using sub-samples that estimate treatment effects that were estimated with a statistically significant coefficient at least at the 10% level with the full sample. Figure 3b summarizes the result. The estimations in the sub-samples, where the minMSE method is compared to purely random treatment assignment, differ significantly between the two methods in their p-values when estimating significant treatment effects (Wilcoxon rank-sum test, $p < 0.031$; robust to inclusion of sub-samples in which the minMSE method is compared to matching and re-randomization based on t-statistics). The mean p-value of estimations in the "purely random assignment" samples is 0.55 (median p-value: 0.60), where it is 0.48 (median p-value: 0.45) in the comparable "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: These graphs show precision in estimation of treatment effects considering three outcomes (response, positive response, and whether or not a survey was completed) 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 (at the 10% level). Estimated treatment effects using sub-samples 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 sub-sample is below 0.5 (i.e., more precise than purely random). Figure 3b presents the distribution of p-values of the estimations of the significant treatment effects using the different sub-samples. Stars indicate resuls from Wilcoxon rank sum tests, where ***/** denotes significance at the 1%/5% levels.*

Result 2. ***Precision:*** *Precision is higher when using the minMSE method compared to purely random treatment assignment. Assigning treatment purely at random is associated with an increase in bias of more than 35% compared to the minMSE method.*

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

In Section 4.2.1, we have shown that – independently of the measure – balance is higher when assigning units to treatment groups using the minMSE method as compared to purely random treatment assignment. That is, our treatment or experimental intervention – inducing balancedness by use of proper treatment assignment – was successful. In Section 4.2.2, using two measures of precision, we have shown that precision of treatment effect estimations is also higher when using sub-samples in which treatment was assigned with the minMSE method as compared to those estimations based on sub-samples where treatment was assigned purely at random. Thus, as we have kept everything else as constant as possible between the subgroups in a sub-sample except for the treatment assignment mechanism, i.e., our 'balance' treatment, we have been able to establish causally that imbalance affects precision statistically and economically significantly in our real-world setting – by increasing the bias by 35% on average (with an even stronger effect on the median bias).[25] This finding is supported at the internal margin by a regression analysis (OLS) on the same data:

An increase of 0.1 in the p-value associated with (im)balance due to the test by Hansen and Bowers (2008) (the higher the p-value, the better the balance) results in a decrease in bias of more than 0.25 standard deviations (about a ninth of the average bias in the "purely random" samples, or a sixth in the "minMSE" samples; p-value < 0.001).

Using aggregated values of the bias at the treatment assignment level,[26] the average share of fulfillment of the overlap condition significantly predicts bias: We find that a 10-percent higher fulfillment share of the overlap condition is associated with a more than 0.4 standard deviations smaller bias (p-value < 0.05).

Result 3. ***The Relation between Balance and Precision:*** *(The degree of) balance increases (the degree of) precision. The magnitude of the effect depends on the measures used for balance and precision. Remaining passive with respect to treatment assignment and assigning units purely at random has been shown to increase the bias by more than 33% compared to proper treatment assignment using the minMSE method in our real-world experiment on balance and precision. 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.*

4.2.4 A Note on the Different Treatment Assignment Mechanisms

Our experiment was designed to illustrate how balance affects precision in a real-world setting, that is: using several – possibly continuous – pre-treatment characteristics and several treatment arms. For its theoretically

[25] 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

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

appealing characteristics and its flexibility with respect to treatment arms and the nature and number of covariates to consider, we have therefore used the minMSE method to induce balance and to compare it to completely random treatment assignment.

Our experiment was not designed as a horserace between the minMSE method and pair-wise matching or re-randomization based on t-statistics, but as we have compared each of these methods with the minMSE method in one sub-sample mimicking a typical usecase of these methods (see Table 11), we may nevertheless draw some conclusions.

First, all our results with respect to balance and precision are robust to comparing the minMSE to all alternative methods, including pair-wise matching and re-randomization based on t-statistics: the minMSE method significantly leads to better balance and higher precision in estimation, independently of the measure of balance or precision used.

Second, re-randomization based on t-statistics and the minMSE method are compared in a sub-sample with six subgroups for each method, yielding 12 and 96 outcomes on balance (omnibus test and overlap, respectively) to compare. Wilcoxon rank sum tests confirm that the minMSE method performs better in achieving balance – with respect to the omnibus test by Hansen and Bowers (2008) and with respect to overlap (p-value < 0.03 and p-value < 0.04, respectively). However, although the mean bias and the average p-value of estimates are lower when using the minMSE method compared to re-randomization (bias: .39 vs. .66; p-value: .49 vs. .59), these differences are not statistically significant (p-values are .13 (N = 32) and .27 (N = 36), respectively).

With respect to pair-wise matching, there are only two subgroups for each method in the sub-sample, and so the picture is similar: the minMSE method performs better or just as well in achieving balance (the overlap condition is always fulfilled for both methods), but there are no significant differences with respect to precision.

Result 4. ***Treatment Assignment:*** *With respect to balance and precision, the minMSE method is superior to the pool of alternative treatment assignment methods: random treatment assignment, pair-wise matching, and re-randomization. The minMSE method is also superior in all of the considered measures when it comes to assigning treatments randomly. Re-randomization based on t-statistics yields worse outcomes with respect to balance and precision, but only the differences with respect to balance are significant.*

4.3 Testing the Unconfoundedness Condition: Evidence for Site Selection Bias – Self-selection of Headmasters into Participation?

This section is dedicated to testing for site-selection bias, i.e., testing the external unconfoundedness condition in multisite evaluations. As mentioned in Section 2, following Allcott (2015) and Belot and James (2016), the unconfoundedness condition can be tested by assessing whether any variables moderate both selections into participation and treatment effects. The first step in our context, thus, is to examine whether there is evidence of self-selection of schools into participation in our typical educational field experiment with multiple

sites. We measure participation with two outcomes: First, whether schools respond in any way (negative or positive) to our inquiry, and second, whether schools indicate interest in participating in our study.

Table 5 presents first regression results – marginal effects from probit estimations – on selection into participation: The dependent variable indicates any response of schools to our request, i.e., clicking on one of the three links in the recruitment e-mail, thus indicating an 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 municipality characteristics that determine whether a school is more or less likely to respond to our inquiry. Splitting the sample by treatment groups, we find that schools in the E-Learning Treatment that have a higher share of students with a migration background are less likely to respond to our request and that schools with a higher share of teachers employed full-time show a slightly higher responsiveness in the Control Treatment. These are the only significantly estimated predictors of the 37 investigated predictors. Note that, although we corrected for testing each predictor five times, we would still expect $37 \times 0.05 = 1.85$ 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 presents further results on selection of schools into participation, now focusing on positive responses (instead of any response, as in Table 5), and investigating which school and municipality characteristics predict a positive responsiveness of schools. We observe a similar pattern, i.e., not a single characteristic predicts differences in responding positively in the pooled regressions. The schools' average compulsory teaching hours are slightly positively related with a positive response in the E-Learning Treatment, and in the Control Treatment, schools in municipalities in which smaller parties received a higher vote share are more likely to respond positively. These results provide no evidence either for self-selection into participation – only two predictors of the 37×5 estimated coefficients are significant. Theoretically, the headmasters' response to our inquiry could depend on the headmasters' personal characteristics, e.g., headmaster quality or open-mindedness, which we could not directly observe. However, Altonji, Elder, and Taber (2005) formalize the idea that "selection on the unobservables is the same as selection on the observables". Following their idea, it is unlikely that (headmasters') unobserved characteristics are uncorrelated with all of our observed school and municipality characteristics such as, e.g., the social index of the municipality, the school size, land prices, or the share of migrant pupils. Finding no evidence of self-selection on these observable characteristics indicates that there is indeed no self-selection which is determined by the headmasters' quality. However, we acknowledge that 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 5. ***Violation of unconfoundedness: Self-selection into participation:*** *We find no evidence of selection on observables into participation, and thus no evidence for violation of the (external) unconfoundedness condition.*

4.4 Treatment Effects of the Recruitment Experiment: Attracting the Headmasters' Attention

In this part, we provide insights from the recruitment experiment and analyze the effect of the different treatments on their ability to attract the headmasters' attention, i.e., 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 reward.

Table 7 presents the effect of explicitly mentioning and varying the research topic of the proposed study in the initial e-mail on the headmasters' willingness to respond positively, to respond in any way, and to complete 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.[27] In general, we find that explicitly stating a research topic in the initial e-mail significantly increases the headmasters' willingness to respond positively compared to only asking for participation in a survey on scientific cooperation. However, we do not find a significant effect on whether the headmasters responded at all (positive responses and negative responses) and on whether or not they completed 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 in terms of headmaster responsiveness. Proposing a research topic as in Parental-Involvement Treatment or Integration-Migrant-Children Treatment increases the headmasters' willingness to respond positively, but there is no statistically significant increase for the E-Learning Treatment. On the contrary, we find that the overall responsiveness for the E-Learning Treatment significantly decreased and that the headmasters were less likely to complete the survey. Personal conversations with school staff indicated that potential reasons for this result might be the lack (or poor equipment) of digital infrastructure of schools, technologically untrained teachers, lack of personnel capacity for additional tasks like installing and maintaining devices and associated infrastructure, and the perceived low support from the government in effectively implementing new technologies in teaching practices.

Result 6. ***Research topic:*** *The topic of the proposed research question matters to attract the headmasters' attention.*

With the Monetary-Incentive Treatment, we tested whether offering the possibility to win 700 Euros that could, e.g., be spent on teacher training or teaching material can attract the headmasters' attention.

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

Table 8 shows whether schools that are eligible to receive the financial reward change their responsiveness with respect to positive responses, any response, and filling out the survey. Panel A presents results pooling all research topics and Panels B, C, and D present results for the respective research topic. 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.[28] As the share of the financial incentives on the schools' yearly budget for teacher training varies by school (see Table 9), we analyze whether there are heterogeneous responses depending on the relative size of the offered financial incentive; see Table 10. We find that the size of the share of the yearly budget for teacher training we offer does not matter to increase the headmasters' responsiveness for none of the outcome variables.[29]

Result 7. ***Monetary incentives:*** *Financial incentives do not attract the headmasters' attention.*

4.5 Survey

The headmasters could access further information on the proposed research question by clicking on a link in the recruitment e-mail. Moreover, all headmasters who clicked on one of the three links indicating their willingness or refusal to participate (opt-out, light interest, opt-in) were guided to a short questionnaire asking about how useful they perceived academic research to be in the educational domain (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 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 tend to ask more often for further information than schools in the Parental-Involvement Treatment or Integration-Migrant-Children Treatment. Putting this into perspective with the finding that schools in the E-Learning Treatment were less likely to respond (positively) compared to schools in the two other treatments, it seems that not responding was an informed choice; yet, we cannot provide statistical evidence for this conjecture.

Figure 4b shows the share of headmasters who completed the survey. We find that the 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 rank-sum test, $p < 0.01$). Overall, roughly 32% of schools whose representative responded in any way also (partly) filled out the questionnaire.

The survey included questions assessing the possibility of integrating academic research in the schools' daily life and on the belief whether 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

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

[29] Tables 21 and 22 show that results are robust to calculating bootstrapped standard errors or applying randomization inference.

study, and whether they think that schools receive too many inquiries from researchers. Headmasters were asked to indicate whether they agree or disagree with these statements on a 1 (totally disagree) to 10 (totally agree) Likert 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 suggested research topics as interesting. However, they also agree with the statement that there are too many inquiries from researchers to participate in a study and that they lack personnel resources for participation. Moreover, while the headmasters seem to be generally interested in the results of academic research and think that academic research is useful to inform education policy-makers, they tend to be less optimistic about the possibility of integrating the research results into daily school life.

5 Summary and conclusion

The present study systematically analyzes the relevance of internal and external overlap and unconfoundedness for the validity and generalizability of field-experimental results from an empirical perspective. For this purpose, we designed an experiment on overlap and precision together with a recruitment experiment within the educational system in the most populous state of Germany – North Rhine-Westphalia –, where we contacted schools and assessed their interest in participating in field experiments. We varied the degree of overlap in disjoint sub-samples of our sample population, and related their pre-treatment balance or overlap with the precision of estimating the treatment effects in these sub-samples. For this purpose, we compare standard treatment assignment methods – pure randomization, pair-wise matching, and a re-randomization method based on t-statistics – to the minimum mean squared error (minMSE) treatment assignment method. To address unconfoundedness, we use a rich set of administrative school and municipality characteristics obtained from the statistics department of the state of North Rhine-Westphalia and study whether those schools that express interest in participating in our study differ with respect to these characteristics from those that express no interest or did not answer at all to our request to begin with. In this context, we also varied the topic of the proposed research topic, and the incentivization for schools.

Our results demonstrate that overlap is statistically and economically important in a typical field-experiment setting: Limited overlap increases the bias in estimating treatment effects by more than 33% compared to improved overlap. Appropriate treatment assignment methods, such as the minMSE method, which aims at balancing the whole covariate distribution instead of focusing on mean values, has been demonstrated to achieve improved overlap. Moreover, in our setup, we find no evidence for systematic self-selection of schools into participation. Thus, we present an important context in which there is no indication of violation of the (external) unconfoundedness condition. Finally, exploiting the different treatment arms of the recruitment experiment suggests that the proposed research topic matters in terms of attracting the attention of headmasters, but that offering an extrinsic financial incentive does not.

Our findings are relevant for all researchers who are interested in a correct causal evaluation of interventions. In the spirit of Cochran (1968) and Crump et al. (2009), we not only demonstrate the relevance of overlap, but also show how to address it in practice by presenting a remedy for limited overlap – in our case in experimental settings, where the data yet is to be generated. We show that the minMSE method increases overlap considerably compared to pure randomization, but also compared to re-randomization based on t-statistics (that is: focusing on balancing group means). While it is theoretically less surprising that a treatment assignment method considering the whole covariate distribution should improve overlap, in practice, the tool at hand might still only marginally improve the situation. Here, our results may serve as a guide for practitioners when planning and implementing their intervention. Together with our findings on unconfoundedness, our research may also serve those who are interested in the generalization or extrapolation of their results to other (institutional) settings where the evaluation of the full population of interest — such as in Crépon et al. (2013) – is hardly feasible due to political or other constraints. We however suggest that for large scale field experiments that usually are costly, prior testing of possible recruitment strategies should be taken into consideration and evaluated against their ability to create representative study samples. This not only increases confidence in the results and their interpretation of the study at hand, but contributes to a wider understanding of the mechanisms of site selection in different contexts.

References

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.

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: *The Journal of Economic Perspectives* 31.2, pp. 3–32.

Banerjee, Abhijit V., Sylvain Chassang, Sergio Montero, and Erik Snowberg (Apr. 2020). "A Theory of Experimenters: Robustness, Randomization, and Balance". In: *American Economic Review* 110.4, pp. 1206–1230.

Beck, Cole, Bo Lu, and Robert Greevy (2016). *nbpMatching: Functions for otimal non-bipartite matching*. R package version 1.5.1.

Belot, Michèle 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.

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.

Bertrand, Marianne and Esther Duflo (2017). "Chapter 8 – Field experiments on discrimination". In: *Handbook of Field Experiments*. Ed. by Abhijit Vinayak Banerjee and Esther Duflo. Vol. 1. Handbook of Economic Field Experiments. North-Holland, pp. 309–393.

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.

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.

Cochran, William G. (June 1968). "The effectiveness of adjustment by subclassification in removing bias in observational studies". In: *Biometrics* 24.2, pp. 295–313.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora (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.

Crump, Richard K., V. Joseph Hotz, Guido Imbens, and Oscar A. Mitnik (2009). "Dealing with limited overlap in estimation of average treatment effects". In: *Biometrika* 96.1, pp. 187–199.

Czibor, Eszter, David Jimenez-Gomez, and John List (2019). "The dozen things experimental economists should do (more of)". In: *Southern Economic Journal* 86.2, pp. 371–432.

Deaton, Angus and Nancy Cartwright (2018). "Understanding and misunderstanding randomized controlled trials". In: *Social Science & Medicine* 210, pp. 2–21.

Donner, Allan and Neil Klar (2004). "Pitfalls of and controversies in cluster randomization trials". In: *American Journal of Public Health* 94.3, pp. 416–422.

Fiero, Mallorie, Shuang Huang, Eyal Oren, and Melanie Bell (2016). "Statistical analysis and handling of missing data in cluster randomized trials: A systematic review". In: *Trials* 17.1.

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.

Fisher, Ronald A (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum (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.

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.

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, Scott, Daniel Gelatt, and Mario Vecchi (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.

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, Robert Greevy, Xinyi Xu, and Cole Beck (2011). "Optimal nonbipartite matching and its statistical applications". In: *The American Statistician* 65.1, pp. 21–30.

Morgan, Kari Lock and Donald B. Rubin (2012). "Rerandomization to improve covariate balance in experiments". en. In: *The Annals of Statistics* 40.2, pp. 1263–1282.

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.

Rosenbaum, Paul R. and Donald B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

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 O. and Giulia Baldini (2019). *minMSE: Implementation of the minMSE treatment assignment method for one or multiple treatment groups*. R package version 0.1.1.

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

Schulz, Jonathan, Uwe Sunde, Petra Thiemann, and Christian Thöni (2019). *Selection into experiments: Evidence from a population of students*. IZA Discussion Paper Series 12807. IZA - Institute of Labor Economics.

Al-Ubaydli, Omar, John List, and Dana Suskind (2019). *The Science of using science: Towards an understanding of the threats to scaling experiments*. Working Paper 25848. National Bureau of Economic Research.

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Unconditional				*Conditional*	
Treatment	*No response*	*Opted out*	*Light interest*	*Strong interest*	*Opted out*	*Light interest*	*Strong interest*
E-Learning (N=955)	78.22	13.51	4.50	3.77	62.02	20.67	17.31
	(747)	(129)	(43)	(36)	(129)	(43)	(36)
Parental-Inv. (N=930)	71.72	17.31	5.70	5.27	61.22	20.15	18.63
	(667)	(161)	(53)	(49)	(161)	(53)	(49)
Integr.-Migr.-Children (N=930)	74.52	15.27	6.56	3.66	59.92	25.74	14.34
	(693)	(142)	(61)	(34)	(142)	(61)	(34)
Control (N=490)	73.06	20.61	0.00	6.33	76.52	0.00	23.48
	(358)	(101)	(0)	(31)	(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. Columns (1)-(4) are the unconditional response rates and columns (5)-(7) are the response rates, conditional on having answered the recruitment email. Recipients of the recruitment email could reply by clicking one of three links indicating that they did not want to participate in the experiment ("Opted out"); were interested, but wanted to be contacted later ("Light interest"), or could imagine participating ("Strong interest"). Schools that did not respond 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	*E-mail*	*Letter + E-mail*
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 explores differences in the 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 explores differences 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-Inv.	(3) Integr.-Migr.-Children	(4) Control	(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)
Average 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)	39.858	39.951	40.004	40.555	40.028
	(0.215)	(0.227)	(0.224)	(0.315)	(0.119)
Students with migration background	30.373	30.404	28.857	28.719	29.710
	(0.631)	(0.705)	(0.647)	(0.859)	(0.349)
Students who migrated	6.385	6.375	6.163	6.604	6.352
	(0.253)	(0.257)	(0.268)	(0.347)	(0.137)
Parents who 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 of teachers employed full-time	55.915	56.000	55.315	55.675	55.735
	(0.548)	(0.578)	(0.524)	(0.820)	(0.297)
Students who 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 the age of teachers between the E-Learning Treatment and the Control Treatment would imply significance at the 10% level (p = 0.064).

Table 4: Descriptive Statistics – Municipality-Level Data

	(1) E-Learning	(2) Parental-Inv.	(3) Integr.-Migr.-Children	(4) Control	(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: CDU	42.911	42.964	43.258	43.587	43.124
	(0.207)	(0.217)	(0.219)	(0.301)	(0.115)
Elections: SPD	30.091	30.082	29.826	29.646	29.948
	(0.176)	(0.185)	(0.183)	(0.248)	(0.096)
Elections: FDP	5.189	5.202	5.255	5.174	5.209
	(0.038)	(0.040)	(0.040)	(0.054)	(0.021)
Elections: Grüne	6.932	6.899	6.907	6.854	6.904
	(0.055)	(0.055)	(0.057)	(0.076)	(0.030)
Elections: Die Linke	5.309	5.275	5.240	5.240	5.270
	(0.034)	(0.035)	(0.035)	(0.050)	(0.019)
Elections: 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 of 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 differences in election outcome for CDU between the E-Learning Treatment and the Control Treatment as well as between the Parental-Involvement Treatment and the Control Treatment 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-Inv.		(4) Integr.-Migr.-Children		(5) Control	
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)
Average 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)	-0.000	(0.001)	0.001	(0.002)	-0.002	(0.002)	-0.003	(0.002)	0.005	(0.002)
Students with migration background	-0.000	(0.002)	-0.012*	(0.004)	-0.001	(0.003)	0.006	(0.004)	0.006	(0.004)
Students who migrated	-0.000	(0.001)	0.005	(0.002)	-0.002	(0.003)	-0.002	(0.003)	-0.003	(0.003)
Parents who 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 of teachers employed full-time	-0.000	(0.001)	-0.000	(0.001)	-0.001	(0.001)	-0.002	(0.002)	0.006***	(0.002)
Students who 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: SPD	-0.002	(0.003)	-0.001	(0.004)	-0.007	(0.004)	0.005	(0.005)	-0.001	(0.007)
Elections: FDP	-0.025	(0.013)	-0.022	(0.018)	-0.033	(0.019)	-0.028	(0.018)	-0.020	(0.030)
Elections: Grune	0.012	(0.010)	0.007	(0.013)	0.010	(0.014)	0.001	(0.015)	0.023	(0.024)
Elections: DieLinke	-0.034	(0.017)	-0.025	(0.029)	-0.038	(0.032)	-0.028	(0.028)	-0.036	(0.023)
Elections: 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 of 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 (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-Inv.		(4) Integr.-Migr.-Children		(5) Control	
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)
Average 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)	-0.000	(0.001)	0.000	(0.001)	-0.003	(0.001)	-0.000	(0.002)	0.003	(0.002)
Students with migration background	0.001	(0.001)	-0.004	(0.003)	-0.001	(0.002)	0.004	(0.003)	0.003	(0.002)
Students who migrated	-0.000	(0.001)	0.003	(0.001)	-0.003	(0.002)	-0.003	(0.001)	0.001	(0.002)
Parents who 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 of teachers employed full-time	-0.000	(0.000)	0.000	(0.001)	-0.001	(0.001)	-0.001	(0.001)	0.001	(0.001)
Students who 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: SPD	-0.001	(0.001)	0.001	(0.002)	-0.004	(0.003)	-0.002	(0.003)	-0.002	(0.003)
Elections: FDP	-0.005	(0.005)	0.010	(0.010)	-0.008	(0.011)	-0.030	(0.011)	0.008	(0.013)
Elections: Grune	0.003	(0.005)	-0.008	(0.009)	-0.001	(0.010)	0.015	(0.009)	0.013	(0.008)
Elections: DieLinke	-0.006	(0.011)	0.019	(0.021)	-0.006	(0.019)	-0.023	(0.022)	-0.004	(0.015)
Elections: 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 of 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 (2005). * p<0.05, ** p<0.01, *** p<0.001.

Table 7: Results – Role of Treatment Topic

	(1) Positive Response	(2) Responded	(3) Completed Survey	(4) Positive Response	(5) Responded	(6) Completed 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 Involvement				0.047***	0.001	-0.012
				(0.013)	(0.025)	(0.017)
Integration Migrant Children				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) completed the survey attached to the recruitment email. *Treated* is a binary variable that takes the value of 0 if schools were in the Control Treatment and takes the value of 1 if schools were in the E-Learning Treatment, Parental-Involvement Treatment or Integration-Migrant-Children 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 (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) Completed Survey	(6) Completed 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 Migrant Children (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) completed 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. *Share of Incentive* $= \frac{700\ Euro}{School's\ yearly\ budget}$.

Table 10: Incentive Treatment – Share Budget

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Completed Survey	(6) Completed 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 Migrant Children (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) completed 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
	minMSE	15	15						30
2	Re-randomization (t-statistics)	30	30	30	30	30	30	30	210
	minMSE	30	30	30	30	30	30	30	210
3	Randomization	20	20	20	20	20	20	20	140
	minMSE	20	20	20	20	20	20	20	140
4	Randomization	20	20	20	20	20	20	20	140
	minMSE	20	20	20	20	20	20	20	140
5	Randomization	20	20	20	20	20	20		120
	minMSE	20	20	20	20	20	20		120
6	Randomization	20	20	20	20	20	20		120
	minMSE	20	20	20	20	20	20		120
7	Randomization	20	20	20	20	20			100
	minMSE	20	20	20	20	20			100
8	Randomization	20	20	20	20	20			100
	minMSE	20	20	20	20	20			100
9	Randomization	20	20	20	20				80
	minMSE	20	20	20	20				80
10	Randomization	20	20	20	20				80
	minMSE	20	20	20	20				80
11	Randomization	20	20	20					60
	minMSE	20	20	20					60
12	Randomization	20	20						40
	minMSE	20	20						40
	Total	490	490	420	380	300	220	140	2440

Note: This table illustrates the experimental design of the experiment on balance and precision, 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 sub-sample 1 were assigned to one treatment group or the control group, using either pair-wise matching or the minMSE approach, with 15 units in each experimental group, i.e., each method had to allocate 30 units to two experimental groups for this sub-sample.

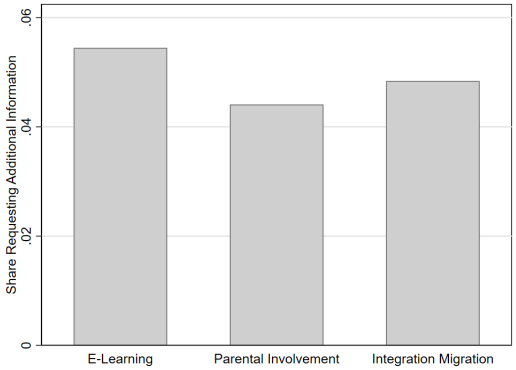
Table 12: Results – Treatment Assignment: Balance in Comparison to min-MSE method

Comparison Method	Sub-sample	Groups	p(minMSE)	p(ComparisonMethod)
Matching	1	1	0.55	0.47
Rerandomization (t-statistics)	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

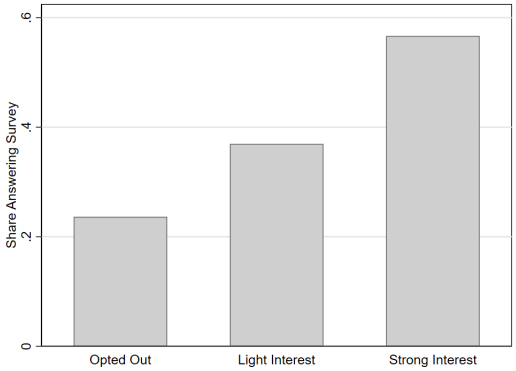
Note: This table shows the p-values resulting from the test of imbalance due to Hansen and Bowers (2008) when testing for imbalance between the treatment groups in a sub-sample that were allocated with the same treatment assignment method. Lower p-values are associated with a higher chance of imbalance.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

B Graphs

Figure 4: Effort Invested in Dealing with Inquiry



(a) Requested Additional Information (by Treatment)

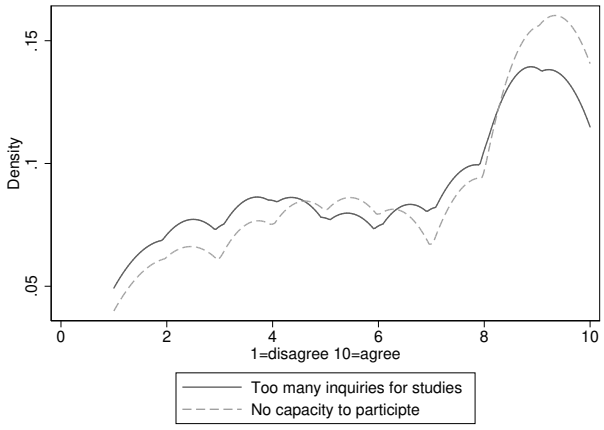


(b) Answered Survey (by Response)

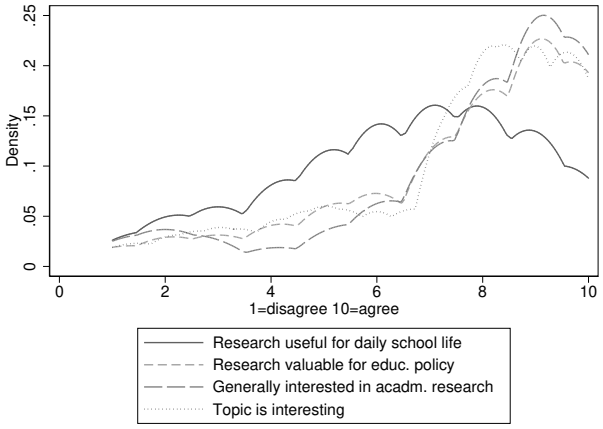
Note: This figure presents two measures on the headmasters' effort spent on dealing with our inquiry. Figure 4a shows the share of headmasters who clicked on the link to access more detailed information about the planned experiment by research topic. In total, 138 headmasters were interested in receiving more information. Figure 4b presents the share of headmasters who answered the survey after clicking on one of the three links in the recruitment e-mail (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 the schools' capacity to participate in a field study. In particular, headmasters were asked to agree or disagree, on a scale from 1 to 10, with the statements that (i) there are too many inquiries for studies, and (ii) schools do not have the personnel capacity to participate. Figure 5b shows answers to questions that asked more generally about the headmasters' opinion about the usefulness of academic research, i.e., 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, (iii) the proposed research topic is relevant for the school, and (iv) whether headmasters 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-Inv.	(3) Integr.-Migr.-Children	(4) Control	(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)
Average 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)	39.858	39.951	40.004	40.555	40.028
	(0.197)	(0.232)	(0.206)	(0.295)	(0.124)
Students with migration background	30.373	30.404	28.857	28.719	29.710
	(0.618)	(0.684)	(0.671)	(0.838)	(0.355)
Students who migrated	6.385	6.375	6.163	6.604	6.352
	(0.247)	(0.274)	(0.280)	(0.364)	(0.132)
Parents who 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 of teachers employed full-time	55.915	56.000	55.315	55.675	55.735
	(0.583)	(0.591)	(0.512)	(0.846)	(0.273)
Students who 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-Inv.	(3) Integr.-Migr.-Children	(4) Control	(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: CDU	42.911	42.964	43.258	43.587	43.124
	(0.224)	(0.220)	(0.228)	(0.316)	(0.119)
Elections: SPD	30.091	30.082	29.826	29.646	29.948
	(0.180)	(0.190)	(0.172)	(0.257)	(0.089)
Elections: FDP	5.189	5.202	5.255	5.174	5.209
	(0.036)	(0.037)	(0.038)	(0.050)	(0.020)
Elections: Grune	6.932	6.899	6.907	6.854	6.904
	(0.055)	(0.054)	(0.061)	(0.078)	(0.029)
Elections: DieLinke	5.309	5.275	5.240	5.240	5.270
	(0.036)	(0.037)	(0.032)	(0.050)	(0.020)
Elections: 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 of 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.

Table 15: 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*	(0.045)	0.065	(0.520)	0.021	(0.827)	0.184	(0.118)	0.132	(0.265)
High School	0.210*	(0.035)	0.177	(0.376)	0.186	(0.293)	0.267	(0.151)	0.183	(0.509)
Vocational (Hauptsch.)	0.050	(0.209)	0.009	(0.903)	0.052	(0.347)	0.160*	(0.047)	-0.071	(0.523)
Vocational (Realsch.)	-0.010	(0.774)	0.019	(0.710)	-0.021	(0.739)	0.034	(0.627)	-0.171	(0.077)
Other school types	0.054	(0.132)	0.118	(0.067)	0.045	(0.343)	0.035	(0.673)	-0.031	(0.677)
Gender of headmaster	-0.013	(0.557)	-0.031	(0.258)	0.015	(0.627)	-0.021	(0.494)	-0.020	(0.650)
Average teaching hours	0.007	(0.072)	0.007	(0.144)	0.008	(0.280)	0.012	(0.156)	-0.003	(0.791)
Students in day care	0.005	(0.075)	0.004	(0.444)	0.003	(0.514)	0.006	(0.198)	0.000	(0.958)
Age of teachers (full-time)	-0.000	(0.765)	0.001	(0.567)	-0.002	(0.517)	-0.003	(0.209)	0.005	(0.077)
Students with migration background	-0.000	(0.936)	-0.012*	(0.033)	-0.001	(0.874)	0.006	(0.171)	0.006	(0.268)
Students who migrated	-0.000	(0.980)	0.005*	(0.040)	-0.002	(0.449)	-0.002	(0.489)	-0.003	(0.458)
Parents who migrated	0.001	(0.377)	0.012*	(0.037)	0.002	(0.603)	-0.004	(0.337)	-0.002	(0.706)
Number of students	0.000	(0.819)	0.000	(0.322)	-0.000	(0.462)	0.000	(0.979)	-0.000	(0.308)
Female students	-0.001	(0.479)	-0.001	(0.476)	-0.001	(0.783)	-0.000	(0.935)	-0.001	(0.845)
Non-German students	0.000	(0.880)	-0.007	(0.244)	0.006	(0.323)	-0.002	(0.782)	0.000	(0.964)
Non-German female students	-0.004	(0.539)	0.005	(0.653)	-0.011	(0.423)	0.002	(0.867)	-0.013	(0.555)
Share of teachers employed full-time	-0.000	(0.557)	-0.000	(0.768)	-0.001	(0.253)	-0.002	(0.149)	0.006**	(0.002)
Students who speak no German at home	-0.000	(0.885)	0.002	(0.267)	-0.001	(0.732)	-0.001	(0.763)	-0.002	(0.582)
Number of classes	0.001	(0.728)	-0.007	(0.278)	0.012	(0.102)	-0.001	(0.958)	0.014	(0.241)
Munic.-level contr.										
Inhabitants	0.000	(0.905)	0.000	(0.787)	0.000	(0.875)	-0.000	(0.917)	0.000	(0.863)
Status married	0.029	(0.561)	0.026	(0.613)	0.037	(0.626)	0.035	(0.569)	0.024	(0.794)
Unemployment rate	0.003	(0.969)	0.030	(0.705)	0.047	(0.704)	-0.057	(0.695)	-0.027	(0.816)
Voter turnout 2013	-0.003	(0.400)	-0.005	(0.462)	-0.012	(0.218)	0.004	(0.549)	0.013	(0.196)
Elections: SPD	-0.002	(0.696)	-0.001	(0.759)	-0.007	(0.217)	0.005	(0.425)	-0.001	(0.948)
Elections: FDP	-0.025	(0.083)	-0.022	(0.297)	-0.033	(0.159)	-0.028	(0.184)	-0.020	(0.545)
Elections: Grune	0.012	(0.314)	0.007	(0.640)	0.010	(0.595)	0.001	(0.944)	0.023	(0.485)
Elections: DieLinke	-0.034	(0.104)	-0.025	(0.457)	-0.038	(0.367)	-0.028	(0.435)	-0.036	(0.295)
Elections: Other	0.010	(0.433)	-0.011	(0.594)	0.013	(0.591)	0.014	(0.569)	0.020	(0.666)
German citizenship	0.006	(0.791)	0.008	(0.783)	0.012	(0.777)	0.007	(0.848)	-0.009	(0.847)
Education: Uni access	0.006	(0.820)	0.006	(0.861)	0.015	(0.695)	0.013	(0.767)	-0.008	(0.912)
Education: High school	0.001	(0.960)	0.003	(0.895)	-0.001	(0.959)	-0.001	(0.974)	-0.001	(0.988)
Land prices in 2014	0.000	(0.885)	0.002	(0.259)	0.001	(0.780)	-0.002	(0.581)	-0.001	(0.914)
Share of people aged 64 or older	-0.014	(0.890)	-0.008	(0.906)	0.004	(0.975)	-0.015	(0.911)	-0.060	(0.743)
Religion: Other	-0.006	(0.745)	-0.010	(0.503)	-0.014	(0.479)	0.004	(0.853)	-0.004	(0.894)
Religion: Protestant	0.001	(0.816)	0.003	(0.576)	0.003	(0.711)	-0.003	(0.713)	0.004	(0.796)
Male Workers	0.014	(0.863)	0.067	(0.377)	0.014	(0.904)	0.014	(0.907)	-0.074	(0.719)
Social index of county	0.001	(0.701)	0.003	(0.234)	0.003	(0.525)	-0.001	(0.798)	-0.004	(0.468)
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. Bootstrapped standard errors (200 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table 16: 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	(0.155)	0.077	(0.310)	0.020	(0.834)	0.070	(0.456)	0.004	(0.991)
High School	0.148*	(0.028)	0.206	(0.161)	0.124	(0.407)	0.122	(0.439)	0.113	(0.884)
Vocational (Hauptsch.)	-0.057	(0.061)	-0.046	(0.415)	-0.094	(0.220)	-0.013	(0.849)	-0.062	(0.592)
Vocational (Realsch.)	-0.020	(0.371)	0.019	(0.669)	-0.021	(0.656)	-0.043	(0.413)	-0.063	(0.598)
Other school types	0.009	(0.706)	0.047	(0.192)	0.033	(0.451)	-0.032	(0.623)	-0.042	(0.824)
Gender of headmaster	-0.003	(0.773)	-0.008	(0.693)	0.013	(0.594)	0.004	(0.838)	-0.021	(0.717)
Average teaching hours	0.005	(0.066)	0.008	(0.068)	0.001	(0.902)	0.009	(0.150)	0.007	(0.678)
Students in day care	0.004*	(0.029)	0.005	(0.228)	0.004	(0.358)	0.003	(0.475)	0.003	(0.890)
Age of teachers (full-time)	-0.000	(0.548)	0.000	(0.845)	-0.003	(0.118)	-0.000	(0.945)	0.003	(0.449)
Students with migration background	0.001	(0.322)	-0.004	(0.389)	-0.001	(0.769)	0.004	(0.330)	0.003	(0.632)
Students who migrated	-0.000	(0.854)	0.003*	(0.045)	-0.003	(0.254)	-0.003	(0.058)	0.001	(0.859)
Parents who migrated	-0.000	(0.822)	0.005	(0.268)	0.001	(0.676)	-0.002	(0.528)	-0.002	(0.678)
Number of students	0.000	(0.533)	0.000	(0.513)	0.000	(0.683)	-0.000	(0.669)	0.000	(0.958)
Female students	-0.001	(0.098)	-0.000	(0.851)	-0.002	(0.238)	-0.001	(0.443)	-0.002	(0.645)
Non-German students	-0.001	(0.650)	-0.006	(0.144)	0.006	(0.287)	-0.006	(0.326)	-0.003	(0.726)
Non-German female students	0.003	(0.465)	0.004	(0.610)	-0.004	(0.696)	0.015	(0.204)	-0.003	(0.877)
Share of teachers employed full-tome	-0.000	(0.515)	0.000	(0.599)	-0.001	(0.346)	-0.001	(0.455)	0.001	(0.542)
Students who speak no German at home	-0.001	(0.071)	-0.001	(0.716)	-0.000	(0.779)	-0.002	(0.331)	-0.000	(0.881)
Number of classes	0.001	(0.715)	-0.002	(0.558)	0.003	(0.528)	0.006	(0.454)	0.003	(0.829)
Munic.-level contr.										
Inhabitants	0.000	(0.963)	-0.000	(0.974)	0.000	(0.881)	-0.000	(0.920)	-0.000	(0.969)
Status married	0.008	(0.762)	0.007	(0.827)	0.022	(0.710)	0.009	(0.868)	0.022	(0.942)
Unemployment rate	-0.002	(0.962)	0.018	(0.731)	0.052	(0.852)	-0.039	(0.670)	-0.069	(0.935)
Voter turnout 2013	0.001	(0.554)	0.007	(0.170)	-0.003	(0.650)	0.003	(0.626)	-0.004	(0.638)
Elections: SPD	-0.001	(0.535)	0.001	(0.642)	-0.004	(0.330)	-0.002	(0.680)	-0.002	(0.860)
Elections: FDP	-0.005	(0.470)	0.010	(0.430)	-0.008	(0.620)	-0.030	(0.071)	0.008	(0.813)
Elections: Grune	0.003	(0.625)	-0.008	(0.415)	-0.001	(0.927)	0.015	(0.311)	0.013	(0.661)
Elections: DieLinke	-0.006	(0.694)	0.019	(0.431)	-0.006	(0.823)	-0.023	(0.435)	-0.004	(0.951)
Elections: Other	0.015	(0.161)	0.010	(0.574)	-0.010	(0.595)	0.036	(0.105)	0.013	(0.793)
German citizenship	-0.001	(0.929)	0.003	(0.852)	0.004	(0.909)	-0.015	(0.560)	-0.005	(0.973)
Education: Uni access	-0.001	(0.959)	0.002	(0.939)	0.008	(0.792)	-0.010	(0.714)	0.000	(0.999)
Education: High school	0.003	(0.781)	-0.006	(0.689)	-0.003	(0.955)	0.018	(0.303)	0.001	(0.990)
Land prices in 2014	0.000	(0.783)	0.001	(0.437)	0.001	(0.835)	-0.001	(0.787)	-0.000	(0.940)
Share of people aged 64 or older	-0.012	(0.763)	-0.038	(0.532)	0.009	(0.947)	-0.034	(0.670)	0.029	(0.933)
Religion: Other	-0.006	(0.599)	-0.006	(0.532)	-0.006	(0.870)	-0.006	(0.732)	-0.006	(0.959)
Religion: Protestant	0.002	(0.628)	0.002	(0.564)	0.003	(0.884)	0.001	(0.891)	0.000	(0.992)
Male Workers	0.001	(0.989)	0.011	(0.853)	0.048	(0.889)	-0.044	(0.576)	-0.029	(0.974)
Social index of county	0.001	(0.814)	0.000	(0.879)	0.002	(0.753)	-0.000	(0.996)	0.001	(0.969)
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. Bootstrapped standard errors (200 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

C.3 Treatment Effects With Randomization Inference and Bootstrapped Standard Errors

Table 17: Results – Role of Treatment Topic

	(1) Positive Response	(2) Responded	(3) Completed Survey	(4) Positive Response	(5) Responded	(6) Completed 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 Involvement				0.047**	0.001	-0.012
				(0.005)	(0.958)	(0.501)
Integration Migrant Children				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) completed the survey attached to the recruitment email. *Treated* is a dummy variable that takes the value of 0 if schools were in the Control Treatment and takes the value of 1 if schools were in the E-Learning Treatment, Parental-Involvement Treatment, or Integration-Migrant-Children 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 18: Results – Role of Treatment Topic

	(1) Positive Response	(2) Responded	(3) Completed Survey	(4) Positive Response	(5) Responded	(6) Completed 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 Involvement				0.047***	0.001	-0.012
				(0.000)	(0.940)	(0.310)
Integration Migrant Children				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) completed the survey attached to the recruitment email. *Treated* is a dummy variable that takes the value of 0 if schools were in the Control Treatment and takes the value of 1 if schools were in the E-Learning Treatment, Parental-Involvement Treatment, or Integration-Migrant-Children 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 19: Incentive Treatment

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Completed Survey	(6) Completed 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 Migrant Children (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) completed 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 20: Incentive Treatment

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Completed Survey	(6) Completed 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 Migrant Children (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) completed 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 21: Incentive Treatment – Share Budget

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Completed Survey	(6) Completed 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 Migrant Children (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) completed 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 22: Incentive Treatment – Share Budget

	(1) Pos. Response	(2) Pos. Response	(3) Responded	(4) Responded	(5) Completed Survey	(6) Completed 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 Migrant Children (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) completed 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 E-mail, and Survey

Description of Variables – School level

- Type of school: There are 12 different school types in NRW. Among these, the elementary school, the 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 subsumed as "Other school types". Elementary school in Germany runs from age 6 to 10 and thereafter students are tracked into secondary education. *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 of attending 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 of obtaining all degrees of secondary education, and *High School - Gymnasium* (grades 5 to 12) is the most academic school type, preparing students for university.

- Gender of headmaster: The gender of the headmaster was obtained from the 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 (employed full-time, employed part-time, trainee teachers).

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

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

- 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 members).

- Parents migrated: Share of students with at least one parent was 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 were not born in Germany and/or the child was not born in Germany). Note that this variable is not the sum of the students who migrated and the parents who migrated. The sum of the students who migrated and the parents who migrated would double count the students who migrated jointly 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 who do not have a German passport

- Non-German female students: Share of female students who do not have 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.

Description of variables – Municipality level

- Inhabitants: Number of inhabitants of the municipality.

- Status married: Share of inhabitants who are married.

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

- Voter turnout: Share of eligible voters who have voted.

- Election results for (name of party): Share of votes for the respective political party.

- German citizenship: Share of people who have German citizenship.

- Education: High School: Share of people who have a high-school degree.

- Education: Uni access: Share of people who have a 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 Catholic: Share of catholic people.

- Religion Other: Share of people who are neither protestant nor catholic.

- Male Workers: Share of male workers.

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

E Online Appendix – E-mail communication

E.1 Initial contact e-mail [Translated from German]

Research project on the topic of [TREATMENT]

Dear Sir or Madam,

The Universities of Düsseldorf, Mannheim and Göttingen are currently planning a joint research project in the field of the economics of education and in particular on the topic *"E-learning in schools: Opportunities and Risks". In this research project we would like to to examine which types of electronic exams are operable in schools and how students perform in these exams compared to written exams.*

[The text in italics was replaced for the specific treatments with the text in italics below.]

"Integration of children with a migration background". In this research project, we want to investigate how migrant students and students with language disadvantages can be successfully integrated in the classroom.

"Parental participation in the school development of their children". In this research project, we want to investigate whether childrens' school related behavior (e.g., disturbing in class, lack of concentration) is malleable by getting parents involved.

Methodology We use methods of experimental economic research, i.e., randomized field trials, to be able to answer our research question causally. We have already gained valuable experience in conducting field experiments in schools, e.g., in 2014 we conducted a study on motivation in mathematics education with 25 secondary schools with in total 2,113 pupils, and in a study in 2015 on a similar topic, 20 primary schools participated with in total 1,377 pupils.

Requirements Our aim is to minimize the workload for teachers and pupils. Therefore, the research project is planned to take place in only **one** regularly scheduled lesson and we will provide all the necessary materials. Each school is eligible to participate, i.e., there is no need for a digital infrastructure (computers, etc.). You only have to give your consent for participation and coordinate the exact timing to conduct the research project with us. We will then be responsible for all further steps. All school grades can participate; however, participation is restricted to a maximum of three classes per grade.

The following text was displayed only for the Monetary-Incentive Treatments. As a "gesture of appreciation", we will randomly choose two participating schools to receive a "funding budget" of EURO 700 each. This budget can be used for internal and external teacher training, online training or teaching materials, and school trips or excursions. You are free to choose the content of the teacher training and the teaching materials as well as the provider of the teacher training. The only requirement is that they are compatible with the educational mission of the school.

For more information about the research project, please click on the following link:

Content of link in subsection E.3

Please give us your feedback We would be very happy if your school participated in our research project *[Treatment]*. To implement the project successfully, we need a minimum number of schools. Therefore, we would appreciate if you could briefly indicate whether you are interested in the research topic by clicking on one of the following links below. Clicking on the link is, of course, not yet a binding commitment to participate.

The research topic is interesing and participation is conveivable.

Please contact me later.

The research topic is not interesting to us. Please do not contact us anymore.

Sincerely,

E.2 Initial contact e-mail – Baseline [Translated from German]

Invitation to participate in a survey the "integration of scientific studies into everyday school life"

Dear Sir or Madam,

The Universities of Düsseldorf, Mannheim and Göttingen are currently conducting university-based fundamental research in the field of the economics of education. In recent years, the number of research projects in the economics of education has grown rapidly; so far, however, we know very little about the transfer of research findings into schools' everyday life. In this research project, we want to learn more about the schools' perspective, in particular which topics are currently relevant for schools and how academic researchers can successfully cooperate with schools. For this purpose, we would be very happy if you could answer our short survey which should not take more than 5 minutes.

Link to questionnaire

Sincerely,

E.3 Content of further information links in initial contact e-mail

E-Learning Treatment

Further information on the research project "E-learning in schools: Opportunities and risks"
Digital learning platforms offer the possibility of innovative forms of teacher-pupil interaction. Furthermore, these platforms offer the possibility for individual adaptation of the learning content and immediate feedback of the learning progress. However, we still know very little about the mechanisms that affect the quality of digital learning, particularly whether students are able to remember in the long term the material they have learned. With this research project, we want to investigate in particular what types of digital forms of learning can be implemented in school, and how they perform compared to traditional written exams. Another open question is whether there are spillover effects of using digital platforms on, e.g., learning a programming language or mathematics. Scientific studies have so far shown that teachers were able individually to adjust the level of difficulty and the speed of learning in mathematics to the needs of the pupils by using "smart boards", which in the classroom increased pupils' performance (Cabus et al., 2015). However, other studies conclude that individualized learning through digital learning platforms does not enhance educational attainment (Cornelisz et al., 2017).

Why should you participate? Your participation is decisive for the success of educational research. Beyond contributing to academic research, you will also gain new insights which you might use for your everyday school life, as we share our research findings and send you a summary of the most interesting findings.

Time schedule of the research project The study is scheduled for the second half of 2017 (right after the summer holidays).

Who can participate? All regular school types can participate in the school year 2017/2018. Pupils of all grades can participate; however, there is a maximum of 3 classes per grade.

Publication and data protection We will use the results of this study only for scientific publications. In any case, we will upload a discussion paper to our homepage. We aim to publish the study in an international specialist journal. We will anonymize all student data, so that no conclusions can be drawn about the respective students. The school's identity will also be anonymized. Only the participating researchers can access the data. These data will not be shared with other scientists or third parties. Furthermore, the data will not be used for Bachelor or Master theses (students won't have access to the data).

References Cabus, Sofie, Carla Haelermanns, and Sonha Franken (2017). "SMART in mathematics? Exploring the effects of in-class-level differentiation using SMARTboard on math proficiency". In: *British Journal of Educational Technology* 48, pp. 145-161. Cornelisz, Ilja, Chris van Klaveren, and Sebastiaan Vonk (2015). ''The effect of adaptive versus static practicing on student learning - Evidence from a randomized field experiment". In: TIER Working Paper Series, WP 15/06.

Parental-Involvement Treatment

Further information on the research project "Parental participation in the school development of their children" Social relationships and emotions are decisive factors in teaching and learning processes. Besides the important role of a positive student-teacher relationship as well as student-student relationship, a positive student-parent relationship is crucial as well. In this research project, we want to investigate whether getting parents involved in school causes a change in their children's classroom behavior (disturbance in class, lack of concentration, etc.). In particular, we want to know what forms of involvement are effective and practicable. Scientific studies in France have shown that afternoon programs for parents from socially disadvantaged families have a positive impact on children's behavior in class (Avvisasti et al., 2013). Moreover, parental involvement in everyday school life seems particularly promising for lower-performing pupils, since these children tend to have a higher preference of signalling their academic achievements to their parents (Wagner and Riener, 2015).

Why should you participate? Your participation is decisive for the success of educational research. Beyond contributing to academic research, you will also gain new insights which you might use for your everyday school life, as we share our research findings and send you a summary of the most interesting findings.

Time schedule of the research project The study is scheduled for the second half of 2017 (right after the summer holidays).

Who can participate? All regular school types can participate in the school year 2017/2018. Pupils of all grades can participate; however, there is a maximum of 3 classes per grade.

Publication and data protection We will use the results of this study only for scientific publications. In any case, we will upload a discussion paper to our homepage. We aim to publish the study in an international specialist journal. We will anonymize all student data, so that no conclusions can be drawn about the respective students. The school's identity will also be anonymized. Only the participating researchers can access the data. These data will not be shared with other scientists or third parties. Furthermore, the data will not be used for Bachelor or Master theses (students won't have access to the data).

References Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin (2014). "Getting parents involved: A field experiment in deprived schools". In: *The Review of Economic Studies* 81, pp. 57-83. Wagner, Valentin and Gerhard Riener (2015). *"Peers or parents? On non-monetary incentives in schools "*. In: DICE Discussion Papers 203, Heinrich Heine University Düsseldorf, Düsseldorf Institute for Competition Economics (DICE).

Integration-Migrant-Children Treatment

Further information on the research project "Integration of children with a migration background" Teaching styles and the constellation of pupils within a class are changing due to the increasing number of children with a migration background. Teachers have to integrate these new pupils into the class and respond to their needs. In this research project, we want to investigate how schools can successfully integrate students with a migration background and language disadvantages. Moreover, we want to know the effect of a changing class composition on the incumbent pupils. Do incumbent children improve their school performance because they help migrant children and thus consolidate the content they have already learned, or is there a detrimental effect on incumbent pupils due to a change in teaching styles (e.g., higher focus on the needs of migrant children). Scientific studies have produced mixed results so far; e.g., Ohinata and van Ours (2013) show that learning conditions decreased with an increasing number of migrant children, but this did not result in a deterioration incumbents' school performance. In contrast, Jensen and Rasmussen (2011) find that a higher proportion of migrant children has a negative impact on math and reading scores. This negative effect seems to be stronger for incumbent children without a migration background than for incumbent children with a migration background.

Why should you participate? Your participation is decisive for the success of educational research. Beyond contributing to academic research, you will also gain new insights which you might use for your everyday school life, as we share our research findings and send you a summary of the most interesting findings.

Time schedule of the research project The study is scheduled for the second half of 2017 (right after the summer holidays).

Who can participate? All regular school types can participate in the school year 2017/2018. Pupils of all grades can participate; however, there is a maximum of 3 classes per grade.

Publication and data protection We will use the results of this study only for scientific publications. In any case, we will upload a discussion paper to our homepage. We aim to publish the study in an international specialist journal. We will anonymize all student data, so that no conclusions can be drawn about the respective students. The school's identity will also be anonymized. Only the participating researchers can access the data. These data will not be shared with other scientists or third parties. Furthermore, the data will not be used for Bachelor or Master theses (students won't have access to the data).

References Jensen, Peter and Astrid Rasmussen (2011). "The effect of immigrant concentration in schools on native and immigrant children's reading and math skills". In: *Economics of Education Review* 30, pp. 1503-1515. Ohinata, Asako, and Jan Van Ours (2013). "How immigrant children affect the academic achievement of native dutch children". In: *The Economic Journal* 123, F308-F331.

E.4 Reminder e-mail

Dear Sir of Madam,

We recently invited you to express your interest in a research project on TREATMENT. We noticed that you have not yet responded to our e-mail and would like you to do so by 22 November 2016 at the latest if you still wish to do so.

We would be happy if your school participated in the project on TREATMENT. If you are not interested, please click the appropriate link and we will stop sending reminder e-mails.

Interested in participating

Stop contacting me

If the topic is in general interesting for you, but you don't see any possibility for participation at the moment, please click the following link.

Generally interested

Further information [LINK]

Sincerely,

F Online Appendix – Initial questionnaire

First page

Your position in school

- ☐ Headmaster
- ☐ Dean of students
- ☐ Senior Teacher
- ☐ Junior Teacher
- ☐ Trainee Teacher
- ☐ Office Staff
- ☐ Other (please specify)

Can we contact you by phone for more information? Please enter a suggested date:

Date

How many teachers are employed full-time at your school?

Number of teachers

Last page Please rate the following statements on the following scale from (1) fully disagree to (5) fully agree.

1. The topic is interesting.
2. There are already too many requests for studies at schools.
3. The school has no time capacity for this type of study.
4. The school has no human resources for studies of this kind.
5. I find education economics studies valuable for the development of education policy.
6. I am interested in the results of scientific studies.
7. The results of scientific studies can be integrated into everyday school life.

Could you please describe briefly how academic researchers should ideally cooperate with schools:

Cooperation between researchers and school

G Online Appendix – Self-selection in Independent Cities

We use data of three other experiments of the authors (Riener and Wagner, 2019; Fischer and Wagner, 2018; Wagner, 2016) to shed light on a potential self-selection bias in independent cities. These studies contacted schools located in the 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 the students' test performance. Fischer and Wagner (2018) also conducted their experiment in secondary schools (contacted schools = 143) to analyze the role of the timing and the 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. However, schools in the three cities might not be representative for schools in other independent cities in NRW, as they are the largest cities. Düsseldorf, Cologne and Bonn, represent the largest cities out of the 22 independent cities in NRW. Düsseldorf and Cologne are the largest and second-largest cities in NRW, respectively, 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%).

Descriptive statistics We check whether background characteristics of schools contacted in the studies of Riener and Wagner (2019), Fischer and Wagner (2018), and Wagner (2016) differ from the average school in the independent cities in NRW. Table 23 shows that schools in Bonn, Cologne, and Düsseldorf (the cities in which the other experiments were conducted) have on average more students, more classes, , fewer students in daycare, a lower share of teachers employed full-time, and teachers who have to teach, on average, fewer lessons. 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). These differences are most likely due to the fact that Cologne and Düsseldorf are the two most populous cities in NRW. Looking at the entire population of schools in NRW reveals that schools in the independent cities differ on average substantially from schools in municipalities (Table 24). Schools in 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, the schools do not differ in the students' gender composition. These differences exist for both elementary and secondary schools (see Tables 26 and 25).

Table 23: Descriptive Statistics: NRW Cities vs Previous experiments by authors (in independent cities)

	(1) NRW Cities - Secondary Schools	(2) Riener and Wagner, 2019	(3) Fischer and Wagner, 2018	(4) NRW Cities - Elementary Schools	(5) Wagner, 2016	(6) (1) vs. (2), p-value	(7) (1) vs. (3), p-value	(8) (4) vs. (5), p-value
Average 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 with 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 who 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 who 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 of teachers employed full-time	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 who 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 presents descriptive statistics on observable characteristics of schools in independent cities (full sample and schools contacted in independent cities in previous studies (Riener and Wagner, 2019; Fischer and Wagner, 2018; Wagner, 2016). Cell entries in columns (1)-(5) report the group means, and standard errors are reported in parentheses; columns (6)-(8) report p-values of testing the difference between schools in this studies and the other experiments. Columns (1)-(3) represent secondary schools and columns (4)-(5) elementary schools. Observable characteristics are described in more detail in the Online Appendix D. We do not test for differences in observables at municipality level, as we do not include these observables in our analysis due to the small number of cities (N=3) in the other experiments.

Table 24: Descriptive Statistics – NRW Sample: Municipalities vs Independent Cities

	(1) Municipality	(2) Independent City	(3) Overall	(4) (1) vs. (2), p-value
Average teaching hours	22.968	23.718	23.216	0.000
	(0.049)	(0.060)	(0.038)	
Teachers employed full-time	17.367	20.287	18.332	0.000
	(0.260)	(0.420)	(0.223)	
Teachers employed part-time	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 with migration background	0.301	0.436	0.345	0.000
	(0.003)	(0.005)	(0.003)	
Students who migrated	0.063	0.096	0.074	0.000
	(0.001)	(0.003)	(0.001)	
Parents who migrated	0.283	0.394	0.320	0.000
	(0.003)	(0.005)	(0.003)	
Students who 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 presents descriptive statistics on observable characteristics of the full sample of schools in NRW. Cell entries report the group means, and standard errors are reported in parentheses. Column (1) is the NRW sample of schools in municipalities, column (2) is the NRW sample of schools in independent cities, (3) is the full sample of NRW schools (municipalities + cities), and column (4) presents p-values testing the difference between columns (1) and (2). Observable characteristics are described in more detail in the Online Appendix D. We do not test for differences in observables at municipality level, as we do not include these observables in our analysis due to the small number of cities (N=3) in the other experiments.

Table 25: Descriptive Statistics – NRW Sample: Secondary Schools - Municipalities vs. Cities

	(1) Municipality	(2) Independent City	(3) Overall	(4) (1) vs. (2), p-value
Average teaching hours	23.006	23.661	23.211	0.000
	(0.077)	(0.098)	(0.061)	
Teachers employed full-time	27.124	32.955	28.952	0.000
	(0.407)	(0.680)	(0.355)	
Teachers employed part-time	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 with migration background	0.278	0.405	0.318	0.000
	(0.004)	(0.008)	(0.004)	
Students who migrated	0.067	0.099	0.077	0.000
	(0.002)	(0.004)	(0.002)	
Parents who migrated	0.260	0.363	0.293	0.000
	(0.004)	(0.007)	(0.004)	
Students who 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 presents descriptive statistics on observable characteristics of the secondary schools in NRW. Cell entries report the group means, and standard errors are reported in parentheses. Column (1) is the NRW sample of secondary schools in municipalities, column (2) is the NRW sample of secondary schools in independent cities, (3) is the full sample of secondary schools in NRW (municipalities + cities), and column (4) presents p-values testing the difference between columns (1) and (2). Observable characteristics are described in more detail in the Online Appendix D. We do not test for differences in observables at municipality level, as we do not include these observables in our analysis due to the small number of cities (N=3) in the other experiments.

Table 26: Descriptive Statistics – NRW Sample: Elementary Schools - Municipalities vs Cities

	(1) Municipality	(2) Independent City	(3) Overall	(4) (1) vs. (2), p-value
Average teaching hours	22.931	23.766	23.220	0.000
	(0.060)	(0.075)	(0.048)	
Teachers employed full-time	7.882	9.653	8.495	0.000
	(0.090)	(0.139)	(0.078)	
Teachers employed part-time	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 with migration background	0.323	0.461	0.371	0.000
	(0.005)	(0.008)	(0.004)	
Students who migrated	0.059	0.094	0.071	0.000
	(0.002)	(0.003)	(0.002)	
Parents who migrated	0.305	0.420	0.345	0.000
	(0.004)	(0.007)	(0.004)	
Students who 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	984	2845	
Proportion	0.654	0.346	1.000	

Note: This table presents descriptive statistics on observable characteristics of the elementary schools in NRW. Cell entries report the group means, and standard errors are reported in parentheses. Column (1) is the NRW sample of elementary schools in municipalities, column (2) is the NRW sample of elementary schools in independent cities, (3) is the full sample of elementary schools in NRW (municipalities + cities), and column (4) presents p-values testing the difference between columns (1) and (2). Observable characteristics are described in more detail in the Online Appendix D. We do not test for differences in observables at municipality level, as we do not include these observables in our analysis due to the small number of cities (N=3) in the other experiments.

Results - Self-selection Table 27 presents the results on selection of schools into participation. We present regression results – marginal effects from probit estimations – where the dependent variable indicates any response to the inquiry to participate in an experiment. We include only municipality-level covariates due to the small number of municipalities in which the three experiments were conducted (N=2). We control for multiple testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (2005). We find mild but not systematic evidence for self-selection of schools in responding to our inquiry. In Riener and Wagner (2019), schools with a higher share of migrated students and a higher share of female students are more likely to respond to our inquiry. In Wagner (2016), schools with on average older teachers are more likely to respond, but covariates which are significant in Wagner (2016) do not show up as significant in this study. In the study of Fischer and Wagner (2018), schools with on average older teachers and a higher number of students are more likely to respond. Table 28 examines school characteristics that determine positive responses of schools to our request. In Riener and Wagner (2019), vocational schools and schools with a higher share of migrated students are more likely to respond positively to our inquiry. In Wagner (2016), schools with a higher share of migrated girls and a lower share of students speaking German at home are more likely to respond positively, and in Fischer and Wagner (2018), no covariate ends up significantly different from zero. Tables 29 and 30 present robustness checks using bootstrapping with 200 repetitions. Overall, these results suggests that there is evidence for mild but unsystematic self-selection of the independent cities. However, a caveat of using data of Riener and Wagner (2019), Fischer and Wagner (2018), and Wagner (2016) is that the cities where they conducted their experiment are not representative for of cities in NRW, as shown in Table 23.

Table 27: Results – Independent Cities: Self-selection (Dep. Var: Responded)

	(1) Riener and Wagner (2019)		(2) Wagner (2016)		(3) Fischer and Wagner (2018)	
Average teaching hours	0.018	(0.039)	0.028	(0.026)	0.037	(0.033)
Age of teachers (full-time)	-0.003	(0.005)	0.008**	(0.003)	0.011**	(0.003)
Students with migration background	-0.637	(0.826)	-0.083	(0.671)	0.633	(0.613)
Students who migrated	0.882***	(0.024)	-0.216	(0.228)	0.290	(0.533)
Parents who 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 of teachers employed full-time	-0.439	(0.787)	-0.281	(0.394)	-0.969	(0.490)
Students who 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 the determinants of schools' response to an inquiry to participate in a scientific study in independent cities. 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 for schools in Riener and Wagner (2019), column (2) are schools in Wagner (2016), and column (3) are schools in Fischer and Wagner (2018). Observable characteristics are described in more detail in the Online Appendix D. We do not include observables at municipality-level due to the small number if cities (N=3) in the experiments. P-values are adjusted for multiple hypothesis testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (2005). * p<0.05, ** p<0.01, *** p<0.001.

Table 28: Results – Independent Cities: Self-selection (Dep. Var: Positive Response)

	(1) Riener and Wagner (2019)		(2) Wagner (2016)		(3) Fischer and Wagner (2018)	
Average teaching hours	0.000	(0.016)	0.033	(0.015)	-0.015	(0.019)
Age of teachers (full-time)	0.005	(0.009)	0.006	(0.003)	-0.002	(0.002)
Students with migration background	-0.156	(0.168)	0.100	(0.357)	-0.441	(0.187)
Students who migrated	0.625**	(0.052)	0.032	(0.147)	-0.021	(0.192)
Parents who 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 of teachers employed full-time	-0.561	(0.233)	-0.272	(0.170)	-0.105	(0.234)
Students who 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 the determinants of schools' response to an inquiry to participate in a scientific study in independent cities. Dependent variable is a binary variable indicating whether schools positively responded to the inquiry. The coefficients are the marginal effects from a probit regression, standard errors in parentheses. Column (1) presents estimates for schools in Riener and Wagner (2019), column (2) are schools in Wagner (2016), and column (3) are schools in Fischer and Wagner (2018). Observable characteristics are described in more detail in the Online Appendix D. We do not include observables at municipality-level due to the small number if cities (N=3) in the experiments. P-values are adjusted for multiple hypothesis testing using the stepdown procedure (with 100 repetitions) proposed by Romano and Wolf (2005). * p<0.05, ** p<0.01, *** p<0.001.

Bootstrapping

Table 29: Results – Independent Cities: Self-selection (Dep. Var: Responded)

	(1) Riener and Wagner (2019)		(2) Wagner (2016)		(3) Fischer and Wagner (2018)	
Average teaching hours	0.018	(0.982)	0.028	(0.504)	0.037	(0.954)
Age of teachers (full-time)	-0.003	(0.791)	0.008**	(0.005)	0.011	(0.961)
Students with migration background	-0.637	(0.817)	-0.083	(0.976)	0.633	(0.969)
Students who migrated	0.882	(0.710)	-0.216	(0.614)	0.290	(0.995)
Parents who 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 of teachers employed full-time	-0.439	(0.967)	-0.281	(0.598)	-0.969	(0.910)
Students who 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 the determinants of schools' response to an inquiry to participate in a scientific study in independent cities. 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 for schools in Riener and Wagner (2019), column (2) are schools in Wagner (2016), and column (3) are schools in Fischer and Wagner (2018). Observable characteristics are described in more detail in the Online Appendix D. We do not include observables at municipality-level due to the small number if cities (N=3) in the experiments.Bootstrapped standard errors (200 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table 30: Results – Independent Cities: Self-selection (Dep. Var: Positive Response)

	(1) Riener and Wagner (2019)		(2) Wagner (2016)		(3) Fischer and Wagner (2018)	
Average teaching hours	0.000	(1.000)	0.033	(0.971)	-0.015	(0.995)
Age of teachers (full-time)	0.005	(0.996)	0.006	(0.961)	-0.002	(0.994)
Students with migration background	-0.156	(0.998)	0.100	(0.992)	-0.441	(0.987)
Students who migrated	0.625	(0.985)	0.032	(0.995)	-0.021	(0.999)
Parents who 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 of teachers employed full-time	-0.561	(0.997)	-0.272	(0.967)	-0.105	(0.996)
Students who 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 the determinants of schools' response to an inquiry to participate in a scientific study in independent cities. Dependent variable is a binary variable indicating whether schools positively responded to the inquiry. The coefficients are the marginal effects from a probit regression, standard errors in parentheses. Column (1) presents estimates for schools in Riener and Wagner (2019), column (2) are schools in Wagner (2016), and column (3) are schools in Fischer and Wagner (2018). Observable characteristics are described in more detail in the Online Appendix D. We do not include observables at municipality-level due to the small number if cities (N=3) in the experiments.Bootstrapped standard errors (200 repetitions) in parentheses. * p<0.05, ** p<0.01, *** p<0.001.