Andersen, Simon Calmar; Michel, Bastien; Nielsen, Helena Skyt

**Working Paper**

# Coaching and Implementation: Insights from a Field Experiment in Danish Schools

IZA Discussion Papers, No. 17728

**Provided in Cooperation with:**
IZA – Institute of Labor Economics

*Suggested Citation:* Andersen, Simon Calmar; Michel, Bastien; Nielsen, Helena Skyt (2025) : Coaching and Implementation: Insights from a Field Experiment in Danish Schools, IZA Discussion Papers, No. 17728, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/314625

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17728

# Coaching and Implementation: Insights from a Field Experiment in Danish Schools

Simon Calmar Andersen
Bastien Michel
Helena Skyt Nielsen

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17728

# Coaching and Implementation: Insights from a Field Experiment in Danish Schools

**Simon Calmar Andersen**
*Aarhus University*

**Bastien Michel**
*Aarhus University and Nantes University*

**Helena Skyt Nielsen**
*Aarhus University and IZA*

FEBRUARY 2025



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

# Coaching and Implementation: Insights from a Field Experiment in Danish Schools*

We study the effect of peer coaching separately from the effect of training on teachers' implementation of new teaching techniques. We conducted a preregistered field experiment involving 68 teachers and 1,490 students in Denmark. Teachers in an active control group took part in a teaching program that introduced new teaching techniques. On top of the teaching program, the treatment group received coaching from peers. External observers, blinded to the treatment status, assessed teachers' use of the program techniques in the classroom. While we observe increased transfer to teachers' practices, the overall effects are mixed, calling for caution.

**JEL Classification:** I21, J24

**Keywords:** coaching, knowledge transfer, school teachers, field experiment

**Corresponding author:**
Helena Skyt Nielsen
Department of Economics and Business Economics
and Trygfonden's Centre for Child Research
Aarhus University
Fuglesangs Allé 4
8210 Aarhus V.
Denmark

E-mail: hnielsen@econ.au.dk

* We acknowledge financial support from Egmont Fonden and TrygFonden as well as helpful comments from two referees and colleagues. We thank Sanne Dalgaard Toft for excellent research assistance. The trial was preregistered in the AEA RCT Registry: https://www.socialscienceregistry. org/trials/2419. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

# 1 Introduction

Teachers have a substantial impact on students' short- and long-term outcomes (Chetty, Friedman, and Rockoff 2014) leading school systems to spend tens of billions of dollars annually on professional development and in-service training (Kraft, Blazar, and Hogan 2018). Yet, teacher training seldom translates into observable improvements in student outcomes (Harris and Sass 2011) at least in part due to difficulties in knowing how to translate theory into practice (Frank, Xu, and Penuel 2018; Raaphorst 2018; Cecchini and Harrits 2022; Møller 2022) – alongside the broader enactment challenges (Kennedy 1999, 2016). This has led to the suggestion that teacher training should be combined with peer coaching in the classroom in order to facilitate the implementation of what is taught in teacher training programs.

A large meta-analysis finds that coaching – defined as programs in which peers observe colleagues' professional behavior and provide constructive feedback – has positive effects (Kraft, Blazar, and Hogan 2018). However, this and other meta-analyses also indicate that coaching is rarely implemented in isolation. Instead, it is often combined with group training sessions or courses designed to teach new skills or content knowledge (Kraft, Blazar, and Hogan 2018; see also Kretlow and Bartholomew 2010; Schachter 2015; Kennedy 2016).

This overlap between coaching and content-focused training complicates the task of isolating coaching's specific impact. Since acquiring new skills or knowledge independently contributes to professional development, it becomes difficult to disentangle the effect of coaching from the broader benefits of increased knowledge (Jakobsen, Jacobsen, and Serritzlew 2019; see also Allen et al. 2011).

In this study, we measure the separate effect of coaching by randomly assigning teachers to either an active control group, which received only a teacher training program, or a treatment group, which received both training and coaching. The field experiment involved 68 language–arts teachers instructing a total of 1,490 students. The teaching program consisted of five modules delivered over six months. Teachers in the treatment group participated in the same program, but were also offered coaching from one of their colleagues for 2.5 hours per week throughout the school year. We registered teachers' teaching practices as our primary outcome and students' reading skills and well-being as secondary outcomes.

The results indicate that coaching had a significant impact on our preregistered primary

outcome: the extent to which teachers implemented the teaching techniques introduced during the training sessions in their classrooms. External surveyors, blinded to the teachers' treatment status, evaluated technique usage through classroom observations. These techniques encompassed a range of strategies designed to foster an inclusive learning environment where students with diverse needs could thrive.

However, a closer examination of the overall effect of the coaching intervention warrants caution for two reasons. First, when breaking down the effects by technological factors (*i.e.*, elements designed to enhance the physical and visual organization of the classroom) and behavioral factors (*i.e.*, teachers' actions in the classroom), we find that while coaching led to improvements in the former, it had no statistically significant effect on the latter (although the two point estimates are not statistically different from one another). This contrasts with our preregistered expectation that coaching would have its greatest effect in the latter dimension, behavioral factors being intrinsically harder to change.

Second, we observe a reduction of students' emotional stability in classrooms with coaching, indicating that they experience more negative feelings such as loneliness or insecurity. This suggests that, rather than fostering a more supportive learning environment, the intervention may have inadvertently had a destabilizing effect on students. We find no positive impact on the other preregistered student outcomes: reading skills, conscientiousness, or agreeableness.[1]

A distinctive contribution of our study is the separation of the effect of coaching from the effect of teacher development programs. Our study is related to a recent study on peer observation (without training and coaching per se), which found that having teachers observing, scoring and providing feedback to peers had positive impacts on students' math and English exam results (Burgess, Rawal, and Taylor 2021). Like our intervention, their treatment implied no explicit incentives and supposedly worked by improving teaching skills. Another study closely related to ours tested the effects of two versions of coaching (as opposed to self-reflection) introduced as an add-on to a course using mixed reality simulations, where prospective teacher candidates practiced their classroom management skills (Cohen et al. 2020). The authors find

1. We had also planned to study the potential theoretical mechanisms. Following Hanna, Mullainathan, and Schwartzstein (2014), we asked teachers in a pre-and post-survey about their awareness of and knowledge about the different techniques in order to potentially separate the information, attention, and training effects of the coach. However, the measures used to capture the different mechanisms proved to be irrelevant in the context of this study, as we explain below.

that coaching impacts teaching practice on several dimensions. Relatedly, we observe how coaching affects teachers' implementation of a teaching program in the classroom.

Our results highlight the need for caution in designing coaching programs, as they can lead to unintended negative short-term effects and may not have a net positive impact on overall welfare if they are not sufficiently effective.

We return to a discussion of how our results may be interpreted. First, we present the design of the study, then we lay out the results related to both teacher behavior in the classroom and student outcomes.

## 2 Design of the Coaching Field Experiment

To study the effect of coaching, we implemented a randomized controlled trial in Denmark during the 2017/2018 school year. The study was preregistered and a pre-analysis plan was uploaded to the AEA Registry in the first months of the project (Andersen, Michel, and Nielsen 2024).[2]

### 2.1 The Context

In Denmark, education is compulsory from grade 0 (when children are typically 6 years old) through grade 9 (when they are typically 16 years old). Compulsory school encompasses preschool class (grade 0), primary education (grades 1 to 6), and lower secondary education (grades 7 to 9). Education is free at public schools, which accounted for 79% of the students enrolled in grades 1 to 9 in 2017/2018 (18% were enrolled at a private school and 3% at a special school). The average class size is 22 students per class, which is similar to other OECD countries (for more details, see UVM 2022).

At different points of their primary and lower secondary education, students are requested to take different national tests designed to track their progress mainly in Danish/reading, and mathematics. Every year, students are also requested to take a well-being survey designed to assess their degree of well-being at school.

School years start in August and end in June.

2. The pre-analysis plan is publicly available on the AEA registry: https://www.socialscienceregistry.org/trials/2419

## 2.2 The Teaching Program

All teachers participating in the study were invited to take part in a training program designed to enhance various aspects of their teaching practices, with the goal of creating an inclusive and supportive environment for both students with and without disorders.[3] The course was based on the idea that teaching techniques that are particularly helpful for children with autism spectrum disorders (such as strict organization of assignments, help to regulate emotions, etc.) would generally also be helpful for all students. The course thus aimed to help teachers create an inclusive classroom environment in which both students with developmental disorders and students without special needs would learn and thrive.

The teaching program consisted of five one-day modules where teachers would receive training in four types of educational inputs (teaching techniques) that facilitate teaching and learning in an inclusive classroom setting. A first group of teaching practices aimed to improve the *physical environment* of the classroom, which involved creating a structured and friendly classroom setting that supports all students. The design focuses on clear organization and minimizing sensory overload to facilitate learning. The second group of teaching practices aimed to improve the *organization of assignments* given to children, making them structured to be clear and manageable, often broken down into smaller steps. This approach aids in understanding and completion, ensuring that tasks are accessible to all students. The third group of teaching practices aimed to improve the *regulation of senses, attention, and emotions* by incorporating strategies to help students manage sensory input, maintain attention, and regulate emotions. This includes the use of calming areas and sensory tools within the classroom. The fourth group of teaching strategies aimed to improve *pedagogy and teacher mentalization* by training teachers to adopt a reflective and empathetic approach, understanding each student's perspective. This dimension emphasizes the importance of teacher collaboration and consistent application of supportive teaching methods. There is some natural overlap between the different groups of teaching practices. For instance, some elements used as visual support to constitute a suitable physical environment are, for example, used to help regulate

3. The program was adapted from the ASD NEST program, which was implemented in the US and aimed at including students with autism spectrum disorders in ordinary age-appropriate classrooms. The program "employs components of evidence-based models, approaches, and practices" (Koenig et al. 2009). More information on the program can be found on the following webpage: http://steinhardt.nyu.edu/asdnest/ (accessed on August 30, 2023).

emotions. As a consequence, reflecting and practicing each of the tools would often be repeated across modules.

The teaching practices introduced in the training can be further divided into two distinct categories based on their nature: technological and behavioral input factors. *Technological input factors* are input factors designed to enhance the physical and visual organization of the classroom, fostering a structured, calm, and supportive learning environment. They involved organizing the classroom to minimize visual distractions, providing students with a designated area where they can regulate their emotions, recharge, and regain focus, and making visual supports introduced during the training visible in the classroom. *Behavioral input factors* pertain to teachers' actions in the classroom and focus specifically on implementing supportive teaching and learning strategies aimed at improving clarity, fostering engagement, and ensuring accessibility within the classroom. This involved the use of visual supports (5-point voice scales, daily class schedule, timers, etc.), supplementing written instructions with verbal instructions, providing the entire class opportunities for movement, giving students the opportunity to make choices, incorporating students' strengths, interests and learning styles into learning activities, highlighting and reinforcing positive behaviors over addressing negative ones, etc.

Figure 1 shows examples of the techniques taught in the program. Panel 1a shows a technological tool that teachers may use. It is a voice scale that should make it clear and visible to the students how loud or quiet they are supposed to talk at the moment (indicated by the arrow). Panel 1b shows another technological tool designed to encourage students to make independent decisions on what type of exercise they plan to do during the school day, and commit to that decision. Panel 1c is a behavioral tool designed to foster self-regulation in students. The tool instructs the teachers in how they may handle conflicts with the students by asking what happened just before the situation (to learn to prevent), what did the student do himself and what did the other student actually do during the conflict (to learn how to react differently), and what happened afterwards (to learn how to react and resolve the situation). Panel 1d illustrates a behavioral tool that encourages teachers to reflect upon their own and others' thoughts, emotions, needs and intentions. Teachers may speak out loud about what they think or feel as they act, or they may mirror the thoughts or needs they imagine that a student has in a hard situation. The idea is that the student learns to describe thoughts and emotions and realizes that there may be different perspectives in a given situation.

Finally, during each training module, teachers were asked to reflect on their own students and to set an agenda for themselves that details how they would change their teaching practices so as to take what they will learn during the course modules into account. In addition, each course module started with rehearsal and reflection on what they have learned to consolidate knowledge. The course modules were held once a month from August through November, with a fifth one in February.

### 2.3 The Coaching Intervention

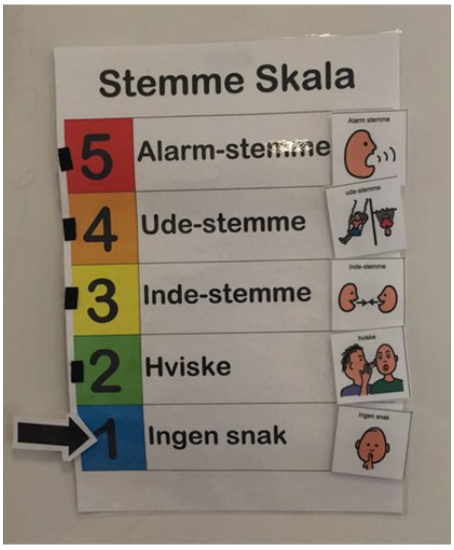
Alongside the teaching program, a coaching intervention was provided to a subset of teachers. The aim was to support them in effectively implementing the teaching practices introduced during the training. As part of this intervention, teachers were paired with a co-teacher, their "coach", and instructed to spend 2.5 hours per week with them for the entire school year. The content of these coaching sessions was to be dedicated to the implementation of what had been covered during the teaching course.

Each participating school appointed one or more individuals to be their school's coach(es), usually another teacher. The appointed person(s) was requested to have received specific training on how to handle students with special needs (learning difficulties; social, emotional, or behavioral problems).

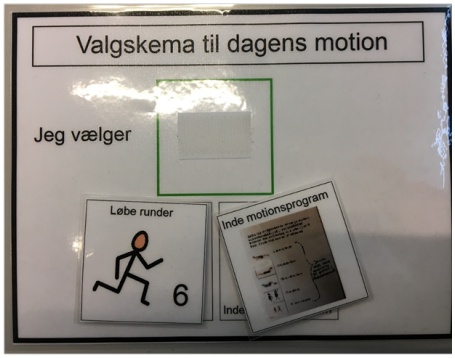
Following the first module of the course, the coaches stayed for an additional two hours to be instructed in their coaching role. They were first introduced to three different co-teaching models: one of the two teachers teaches while the other observes; one of the two teachers teaches while the other assists; both teachers teach on equal terms.[4] Furthermore, they were advised to allocate the 2.5 weekly hours as follows: 30 minutes for preparation, two 45-minute class lessons, and 30 minutes for consolidation. They were taught that their key tasks were listening and guiding rather than advising and instructing. The participant prospective coaches discussed how to establish trust and supportive collaboration, and they practiced appreciative inquiry.

4. This information was given as inspiration and it is unclear whether and how much they used the three models in practice.

**Figure 1:** Examples of techniques taught in the teaching program

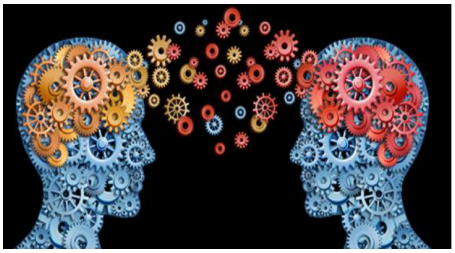



**(a)** Voice scale




**(b)** Choose Exercise


**(c)** Conflict resolution



**(d)** Mentalizing

Translations: **Panel 1a**: "Voice scale." 1 is "No talk," 2 "Whispering," 3 "Indoor voice," 4 "Outdoor voice," 5 "Alarm voice." **Panel 1b** "Scheme to choose exercise of the day." "I choose" "Running rounds," "Indoor exercise program." **Panel 1c** "Before. Preceding (trigger) *What happened just before?* Prevent/ precautionary measures." "During. Observable behavior *What did I do? What did the student do?* Substitute." "After. Consequence *What happened afterwards?* React."

## 2.4 Sampling Strategy and Randomization

### 2.4.1 Sampling Strategy

We recruited teachers from all public schools located in a large municipality in Denmark. More specifically, we invited all language arts teachers teaching grade levels 3–6 to participate. In total, 17 schools and 68 language arts teachers signed up to participate in the experiment, representing a total of 1,490 students.[5]

In four cases, a teacher signed up for more than one of their classes to participate in the experiment. Here, one class was randomly selected to participate.

### 2.4.2 Randomization

As there are strong reasons to expect significant selection with respect to the type of language arts teachers who would enroll in the coaching program, we implemented a randomized controlled trial to measure its impact.

The coaching intervention was randomized among teachers participating in the teaching program. The draw was carried out at the teacher level and stratified by school so as to increase the level of buy-in from participating schools. As a consequence, in each school, we randomly selected half of the language arts teachers to benefit from the coaching intervention. In schools where an uneven number of teachers was enrolled in the experiment, the number of teachers assigned to the coaching intervention was rounded upward or downward based on schools' stated capacities or preferences. In total, out of the 68 teachers (classes) participating in the experiment, 35 were assigned the coaching intervention. Teachers and coaches were informed about the results of the draw between the first and second course (after the completion of the baseline data collection).[6]

Since both control and treated teachers participated in the same teaching program, it can be assumed that they received the same level of information on effective teaching techniques and were equally encouraged to focus on relevant input factors during the training. Consequently, any differences observed between treatment and control classes can be attributed to the effect

5. In parallel to the experiment, grade levels 3–6 mathematics teachers working in the same 17 schools were invited to participate in the teaching program. However, for budgetary reasons, only language arts teachers were eligible for the coaching intervention.

6. When schools appointed more than one coach, the teacher-coach pairs were randomly generated.

of coaching.

### 2.4.3 Research Questions

As part of our primary research question, we aimed to assess whether the coaching intervention would have an impact on teacher practices. Additionally, we set out to explore four secondary research questions. First, we planned to measure its impact on student outcomes, particularly reading skills—measured through national standardized reading tests—and socio-emotional skills, assessed via national standardized well-being surveys. Second, we planned to examine whether the intervention's impact on teaching practices was stronger for behavioral or technological practices, hypothesizing a greater effect on the former.[7] Primary and secondary research questions were documented in a pre-analysis plan uploaded to the AEA RCT Registry on September 2017 (between the first and second training session).

We carried out statistical power calculations taking into account the fact that 17 schools and 68 language arts teachers enrolled in the study, representing an average of 4 teachers per school (two in each group). Under the assumptions that baseline covariates would allow us to explain 25% of the variation in our outcome variables, the power analysis suggested that with a power of 80% we would be able to detect a minimum effect size of 0.50 to 0.60 standard deviations in teacher behavior.[8] This aligns closely with the pooled effect size of 0.49 standard

7. The last two secondary research questions related to mechanisms and heterogeneous effects. We had planned to investigate the mechanisms underlying changes in teaching practices as a secondary research question but the preregistered measures to capture mechanisms proved to be inappropriate in this study's context. Specifically, we surveyed teachers at baseline, endline, and at the beginning of each of the four remaining course modules on three key aspects for each teaching practice listed in Appendix A.2: their level of knowledge about the teaching practices (the "information" channel); the extent to which they remember to use the teaching practices taught during the training (the "attention" channel); and their capacity to implement the teaching practices they aim to use (the "training" channel). We preregistered that the strength of these mechanisms would be measured using three proxies: (i) the proportion of teachers who reported being unable to assess the importance of the investigated input factors for students, (ii) the proportion of teachers who reported being unable to rate to assess the extent to which they used these input factors, and (iii) the gap between teachers' reported desire to use an input factor and their actual observed usage, as recorded by our surveyors. However, possibly due to the simultaneous implementation of the training program, the proportion of teachers who reported being unable to assess the importance of the investigated input factors for students or the extent to which they used them is very small—even in the control group. This raises serious concerns about the suitability of these measures for examining the designated mechanisms. We provide more information on our empirical strategy to investigate mechanisms in Appendix A.3. Regarding heterogeneous effects, we had planned to assess whether or not the effect varied with respect to students' cognitive abilities and gender. However, since the main effect point estimates are generally small, and because of our limited statistical power for the student effects, we do not present estimates of heterogeneous effects, as any statistically significant effects could easily be due to magnitude errors and/or Type I errors (false positives).

8. Calculations were performed in Optimal Design (Spybrook et al. 2011).

deviations for the impact of coaching on teacher instruction, as reported in the meta-analysis by Kraft, Blazar, and Hogan (2018).

For student-level outcomes, and under the assumption that baseline covariates would allow us to explain 28% of the variation in our outcome variables, with a power of 80%, we would be able to detect a minimum effect size of 0.23 to 0.30 standard deviations in students' test scores. Parameter values on average class size, expected variance explained by each level's covariates, and inter-class correlations in test score data were estimated from Danish register data (past cohorts). Given that Kraft, Blazar, and Hogan (2018) reported pooled effect sizes of 0.18 standard deviations on student achievement, there is uncertainty about whether our study had sufficient statistical power to detect effects on student outcomes.

## 2.5 Data

In order to measure the impact of the intervention, information was collected at different points in time throughout the experiment. As shown in Table 1, we do not observe any differential attrition rate across groups at follow-up.

### 2.5.1 Outcomes

**Behavior in the Classroom** To measure the impact of the coaching intervention on the extent to which teachers use the teaching techniques taught as part of the teaching program, teachers' practices were assessed through 60-minute classroom observations. Each teacher was observed twice: once at the beginning of the school year and again at its end.

Teachers' techniques were initially assessed using a 34-item observation checklist. Each item fell into one of the four broad types of enabling inputs covered during training: (i) physical environment, (ii) organization of assignments, (iii) regulation of senses, attention and emotions, and (iv) pedagogy and teacher mentalization. Upon completion of the training modules, the observation form was refined and reduced to 24 items at the request of the training organizers, as ten items were ultimately not covered during the training. Our analysis focuses exclusively on the restricted set of 24 items. Each of these items was also categorized as either a behavioral or a technological input factor.[9]

9. Of the 24 items, four were classified as technological input factors—two related to the physical environment and two to the regulation of senses, attention, and emotions.

For each item/educational input, the teachers' technique was assessed on a scale of 1 to 5 (1 indicating that "a technique was not used at all", and 5 indicating that "a technique was used extensively"). These variables were later combined into one single index indicative of teachers' overall practices and calculated as their simple average. We also computed four sub-indices based on the nature of the dimension the questions investigated (each sub-index representing a family of outcomes). All these variables have been standardized to have mean of 0 and a standard deviation of 1 in the control group.

In order to avoid any data collection bias, classroom observations were conducted by research assistants who were specifically recruited and trained for this task and were blinded to the teachers' treatment status. Different research assistants conducted the pre- and post-observations. The assistants were trained in advance to obtain a high level of reliability.

The full observation form is reported in Appendix A.2. It also provides a mapping of the different items into the four types of inputs.

**Student Outcomes** In order to measure the impact of the coaching intervention on students' academic achievements, we use their performance on a standardized national reading test. This test is mandatory for grade 2, 4, 6 and 8 students, and it is implemented to track students' progress throughout school. The test is online, self-scoring and adaptive, and consists of three subdomains: reading comprehension, decoding, and text comprehension. We standardize each subdomain to mean 0 and standard deviation 1. To measure the composite reading skill, we take the mean of the three subdomains and standardize the composite score. For more details on the national test and how to use them, see Beuchert and Nandrup (2018). We observe this outcome for students in grades 4 and 6 only (not for grades 3 and 5).

In order to measure the impact of the coaching intervention on students' socio-emotional skills, we use a national well-being survey. The survey is mandatory for all grade 0 to 9 students and aims to collect information on their well-being at school. As part of this questionnaire, grade 4 to 9 students are requested to answer 40 questions investigating different aspects of their well-being at school. In each dimension, students are asked to rate their well-being on a scale of 1 to 5. Among the 40 items, 3 items have been validated as a measure of students' conscientiousness, 2 for agreeableness, and 3 for emotional stability by Andersen et al. (2020), and we follow their way of measuring the three skills (standardizing each item, averaging

across the relevant items, and standardizing the composite score). Especially conscientiousness, which relates to facet self-control and grit, has been shown to be strongly associated with academic performance in school (Andersen et al. 2020). We observe these measures for grades 4-6 (not for grade 3 where students get a shorter questionnaire) Table A.1 shows the pairwise correlations between the socio-emotional survey variables.

### 2.5.2 Covariates

To increase the precision of the effect estimates, we include various sets of baseline covariates.

First, we include school fixed effects since randomization was stratified at the school level. Second, we include grade fixed effects and baseline values of the primary outcome variables, derived from classroom observations of teacher practices conducted before the randomization results were announced to schools and teachers. Third, we include information on teachers: the number of students who have repeated class (variable standardized to have mean 0 and standard deviation 1), and the experience of the teacher (in years).

## 2.6 Statistical Model

We assess the impact of the intervention on outcome $y_i$ for student $i$ using an Intention-To-Treat (ITT) analysis. Specifically, we estimate the following equation:

$$y_i = aT_i + \Sigma_{j=1}^{17}\mu_j S_j + bX_i + \epsilon_i \tag{1}$$

In this equation, $T$ is a dummy variable indicating whether or not teacher or student $i$ is assigned to the treatment, and $S_j$ are stratum fixed effects (school fixed effects). Parameter $a$ is the parameter of interest and measures the effect of the intervention. While the baseline model is without further covariates, we control for progressively larger sets of baseline covariates, $X$, as a robustness test, as described above. We calculate Huber-White robust standard errors for regressions using teacher-level data, and standard errors clustered at the level of the 68 classrooms included in the sample for regressions using student-level data.

## 2.7 Sample Description, Balance Checks, and Compliance

**Sample Description** Table 1 describes the students, classes, and teachers enrolled in the experiment in the control group and the treatment group. The table also shows the difference between the two groups.

Half of the sample is girls, and 73% of them live with both parents. The average class size is 22, and grades 4–5 dominate compared to grade 3 and grade 6.

Unsurprisingly given the nature of the intervention, about one-third of the classes had 3-4 students with special needs (including children with autism or autism-like features, Asperger's Syndrome, ADHD, behavioral and attention disorder, anxiety, or learning difficulties), and about one-third had 5 or more students with special needs.[10] Moreover, while teachers included in our sample are fairly experienced with 12 years of service on average, only a fourth of them already had benefited from a coaching program prior to the roll-out of the experiment.

**Balance Checks** Coefficients displayed in the balance checks column ("Difference") are obtained by estimating equation (1) (without including covariates $X$) using successively each of the baseline characteristics displayed in the left column of the table as the dependent variable. We do so by using all observations for which endline information is available. The point estimates associated with the treatment variables are not statistically significant at the 10 percent level, except for one coefficient, which reflects a slight overweight of girls in the control group. This suggests that teachers' and students' treatment statuses are uncorrelated with their baseline characteristics.

**Compliance** Table 2 investigates the differential uptake of the interventions across control and treatment teachers. In columns 1 to 8, we analyze the impact of a teacher's treatment status on a range of indicators describing their exposure to the teaching program and to the coaching intervention.[11] First, we find no evidence that treatment teachers attended more professional development training than control respondents: while the point estimate is quite

10. The phrasing of the question specified that children with special needs may be "children with one or more mental diagnoses and children under investigation or with special needs without an actual diagnosis."

11. Covariates and outcomes are almost always observed for the 61 teachers who answer the endline survey. However, in rare cases, slightly fewer or more observations are available. To preserve the anonymity of respondents, we always report the number of observations to be 61 as the gap is less than or equal to three observations.

**Table 1:** Balance checks and student attrition rates

| | Control | | Treatment | | Difference | | N |
|---|---|---|---|---|---|---|---|
| *Student-level information* | | | | | | | |
| Girls | 0.503 | (0.500) | 0.480 | (0.500) | -0.027+ | (0.016) | 1,480 |
| Age | 10.400 | (1.051) | 10.308 | (0.961) | -0.105 | (0.167) | 1,480 |
| Child lives w/ both their parents | 0.724 | (0.447) | 0.730 | (0.444) | 0.013 | (0.021) | 1,490 |
| Repeated at least once | 0.030 | (0.171) | 0.034 | (0.182) | 0.005 | (0.008) | 1,490 |
| *Teacher-level information* | | | | | | | |
| Class size | 22.273 | (3.347) | 21.571 | (4.698) | -0.570 | (0.637) | 68 |
| N students with special needs (ref. $\geq$5) | | | | | | | |
| 0-2 pupils with special needs | 0.333 | (0.479) | 0.229 | (0.426) | -0.090 | (0.112) | 68 |
| 3-4 pupils with special needs | 0.364 | (0.489) | 0.314 | (0.471) | -0.050 | (0.117) | 68 |
| Students' grade (ref. grade 3) | | | | | | | |
| Grade 4 | 0.333 | (0.479) | 0.486 | (0.507) | 0.160 | (0.104) | 68 |
| Grade 5 | 0.333 | (0.479) | 0.314 | (0.471) | -0.020 | (0.110) | 68 |
| Grade 6 | 0.242 | (0.435) | 0.171 | (0.382) | -0.080 | (0.090) | 68 |
| Teacher experience (in years) | 11.485 | (8.910) | 13.371 | (7.975) | 1.540 | (1.957) | 68 |
| Teacher prior trainings | 0.212 | (0.415) | 0.286 | (0.458) | 0.030 | (0.093) | 68 |
| Teacher observations, turnover | 0.121 | (0.331) | 0.171 | (0.382) | 0.040 | (0.079) | 68 |

Notes: Includes only participants that completed the endline questionnaire. To preserve anonymity of respondents, we cannot distinguish attrition in the treatment and control group. To compare the treatment and control group, we regress each of the variables on the treatment indicator, $T$ and school fixed effects. Difference is the coefficient and SE the robust standard error. $p$ is the p-value.
+, *, ** are the 10, 5, and 1 percentage levels.

large (0.452), it is never statistically significant at the 10% level (column 1). This implies that any differences we find in the impact of the coaching intervention across treatment and control teachers can be attributed to the help of the coach, not to any differences in exposure to the teaching program.

Investigating teachers' exposure to coaching intervention, we find little evidence of non-compliance. First, less than 10% of the control teachers benefited from any coaching intervention throughout the year, while an additional 90% of the treatment teachers benefited from one (column 2). Second, we find that the result is entirely driven by differences in the extent to which treatment and control teachers benefit from the coaching intervention evaluated as part of this project (column 3). Alternative coaching programs accessible to all teachers—and, in particular, control teachers—do nothing to reduce this differential exposure rate (column 6).

However, while teachers were supposed to spend 2.5 hours per week with their coach throughout the entire school year, we find that they spent only slightly more than 1.5 hours (67 minutes (column 4) + 36 minutes (column 5)) together.

With a mean of 18.1 years of experience (standard deviation 6.9), the coaches were generally highly experienced teachers.[12]

# 3 Results

Most teachers thought that the NEST elements were not too difficult to use. To get an impression about teachers' overall perception of the program, we asked them about their view of the NEST elements in total (in addition to the survey questions about each individual element as described in Section A.3.1.) Figure 2 shows the distribution on four statements about how easy it was to use the NEST elements on a scale ranging from 1 (fully disagree) through 10 (fully agree). Most responses were in the range from 5 to 10.

## 3.1 Effects on Teacher Behavior

In Table 3, we report the effects of coaching on the overall use of practices taught in the training program and on each of the dimensions and types of inputs measured. As mentioned,

12. Experience ranged from about 5 years to more than 30 years. Exact numbers are not provided in order to preserve anonymity.

**Table 2:** Compliance

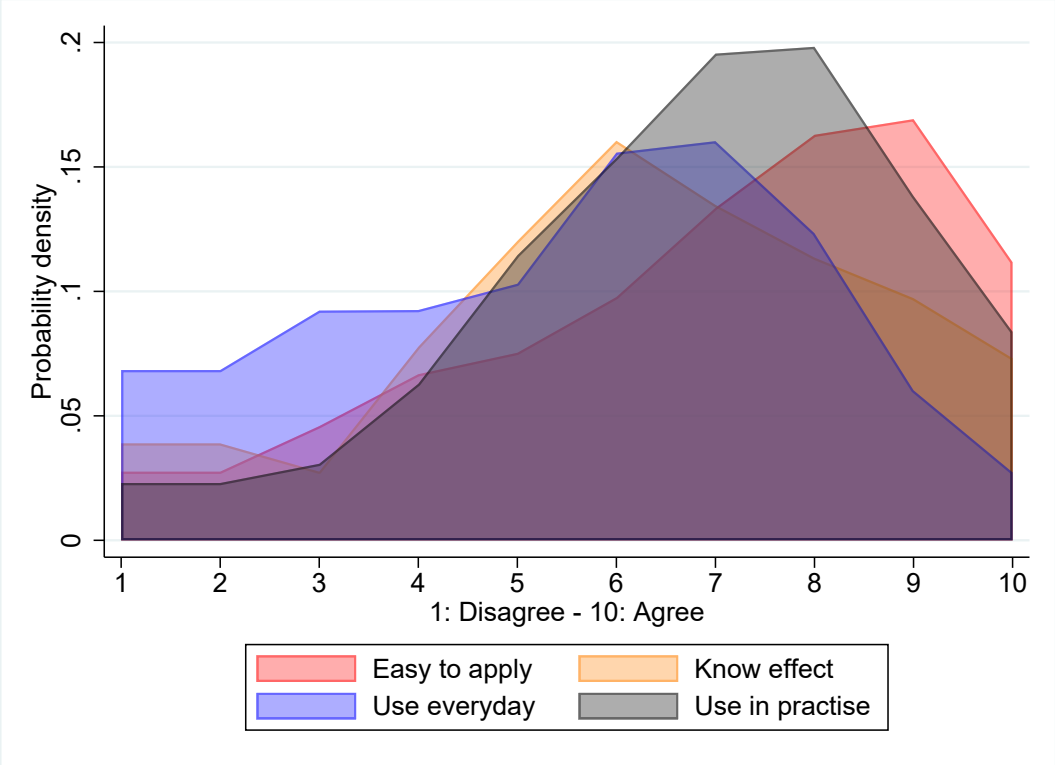
| | | | As part of the project | | | Outside of the project | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | How many NEST element trainings did you attend? | This year, were you ever associated with a co-teacher? | As part of the NEST project, were you associated with a co-teacher? | On average, how much time did the co-teacher spend with you in class per week? (min) | On average, how much time did the co-teacher spend with you in meetings per week? (min) | Outside of the NEST project, were you associated with a co-teacher? | On average, how much time did the co-teacher spend with you in class per week? (min) | On average, how much time did the co-teacher spend with you in meetings per week? (min) |
| Treatment | 0.452 | 0.899** | 0.920** | 67.30** | 35.79** | 0.138 | 4.293 | 3.411 |
| | (0.473) | (0.066) | (0.065) | (8.227) | (4.753) | (0.084) | (3.661) | (3.219) |
| $R^2$ | 0.273 | 0.816 | 0.836 | 0.717 | 0.650 | 0.261 | 0.246 | 0.222 |
| Control group mean | 3.906 | <0.1 | <0.1 | 3 | 3.125 | <0.1 | 1.452 | 1.935 |
| School FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Grade FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Covariates | YES | YES | YES | YES | YES | YES | YES | YES |

Notes: The table shows the effect of assignment to the coaching intervention on teachers' actual level of exposure to the intervention. We regressed each of the outcome variables displayed in columns 1–8 on a dummy variable indicative of respondents' treatment assignment, school and grade fixed effects, teachers' number of years of experience (standardized), and a categorical variable indicating the number of children with special needs in the class. We report the coefficient and standard error associated with the treatment variable. Robust standard errors are computed. N= 61. In rare cases, N < 61. However, to preserve the anonymity of respondents, we do not report the small deviations.

"NEST" is the name used for the overall project including the teaching program modules.

+, *, ** are the 10, 5, and 1 percentage levels.

**Figure 2:** Distribution of responses to questions about the use of the NEST elements




Notes: Teachers were asked to what extent they agreed or disagreed with four statements: [Easy to apply] "NEST elements are easy to apply." [Know effect] "It is easy to know whether the NEST elements actually benefit the students." [Use everyday] "It is easy to remember to use the NEST elements in everyday work." [Use in practice] "It is easy to know how to use the NEST elements in practice when you teach a class." Response categories range from 1 (fully disagree) to 10 (fully agree). The figure is based on kernel density estimates. Response categories 1 and 2 are averaged to preserve anonymity.

the use of tools is measured on a 1 to 5 scale, with 1 indicating that "a technique was not used at all" and 5 indicating that "a technique was widely used."

The coaching intervention increased the overall teaching score by 0.44 standard deviations (statistically significant at the 10% level), suggesting a quite substantial impact on the extent to which teachers adopted the teaching techniques introduced during the training. We observe that this effect is primarily driven by teaching practices focused on the regulation of senses, attention and emotions, which increased by 0.5 standard deviations (statistically significant at the 5% level). In contrast, the impact was less pronounced for teaching practices related to pedagogy and teacher mentalization (+0.278 standard deviations) and improvements to the physical environment (+0.167 standard deviations).

We further examine the effects of the intervention on teachers' practices by distinguishing between behavioral input factors (*e.g.*, using direction instead of correction, providing support during transitions, maintaining a calm voice) and technological input factors (*e.g.*, making 5-point voice scales or charts visible in the classroom, setting up a break area). While we initially expected that coaching would have a greater impact on behavioral input factors—given their intrinsic difficulty to change—we find that the point estimates for technological input factors are larger (though not statistically different). They are also statistically significant when point estimates for behavioral input factors are not. This suggests that the coaching intervention may not have had the anticipated impact and was potentially less effective in shifting the harder-to-change practices it was specifically designed to influence.

The coefficients are stable across specifications as we progressively control for a larger set of baseline covariates.

### 3.2 Effects on Student Outcomes

We find no statistically significant effect suggesting that the coaching intervention improved student outcomes. Table 4 shows that the effects on the composite reading score, as well as on two subdomains—language comprehension and text comprehension—are generally small and consistently fail to be statistically significant at the 10 percent level.

However, we find some evidence suggesting that the coaching intervention may have had a negative effect on students' well-being. Indeed, while the effects on two of the measures

**Table 3:** ITT estimates of the coaching intervention's impact on observed teacher behavior

| | (1) | (2) | (3) |
|---|---|---|---|
| Teacher practices, overall | $0.389^{+}$ | $0.424^{+}$ | $0.440^{+}$ |
| | (0.231) | (0.223) | (0.227) |
| Teacher practices, Physical Environment | 0.194 | 0.160 | 0.167 |
| | (0.221) | (0.217) | (0.225) |
| Teacher practices, Organization of Assignments | 0.260 | 0.335 | 0.321 |
| | (0.225) | (0.224) | (0.230) |
| Teacher practices, Regulation of Senses, Attention and Emotions | 0.396 | $0.481^{+}$ | $0.496^{*}$ |
| | (0.268) | (0.240) | (0.245) |
| Teacher practices, Pedagogy and Teacher Mentalization | 0.198 | 0.215 | 0.278 |
| | (0.248) | (0.248) | (0.242) |
| Teaching practices, tec. factors | $0.473^{*}$ | $0.410^{*}$ | $0.422^{*}$ |
| | (0.196) | (0.181) | (0.192) |
| Teaching practices, beh. factors | 0.328 | 0.364 | 0.376 |
| | (0.236) | (0.235) | (0.239) |
| School FE | YES | YES | YES |
| Grade FE | | YES | YES |
| Baseline value | | YES | YES |
| Repeat class | | | YES |
| Teacher experience | | | YES |

Notes: Data for this table was collected by student assistants who directly observed the teachers in their classroom and graded their practices using a standardized questionnaire. Entries are coefficients (robust standard errors in parentheses). All outcome variables have been standardized to have mean of 0 and a standard deviation of 1 in the control group.
+, *, ** are the 10, 5, and 1 percentage levels. N=61.

**Table 4:** ITT estimates of the coaching intervention's impact on student outcomes

| | (1) | (2) | (3) |
|---|---|---|---|
| Reading test score | 0.00530 | -0.0279 | -0.0345 |
| | (0.070) | (0.067) | (0.068) |
| Language comprehension | -0.0280 | -0.0242 | -0.0270 |
| | (0.065) | (0.055) | (0.061) |
| Decoding | 0.0702 | 0.0703 | 0.0601 |
| | (0.072) | (0.049) | (0.048) |
| Text comprehension | -0.0290 | -0.100 | -0.105 |
| | (0.076) | (0.083) | (0.083) |
| Conscientiousness index | -0.0719 | -0.0332 | 0.0102 |
| | (0.067) | (0.067) | (0.059) |
| Agreeableness index | -0.0947 | -0.0634 | -0.0620 |
| | (0.070) | (0.067) | (0.067) |
| Emotional stability index | -0.194** | -0.163* | -0.178** |
| | (0.066) | (0.064) | (0.059) |
| School FE | YES | YES | YES |
| Grade FE | | YES | YES |
| Baseline value | | YES | YES |
| Student covariates | | | YES |
| Teacher experience | | | YES |

Notes: The national test in reading is mandatory for grade 4 and 6 students. The wellbeing survey is mandatory for grades 4-6. Entries are coefficients (classroom-clustered standard errors in parentheses). N= [899;1,222]. All outcome variables have been standardized to have mean of 0 and a standard deviation of 1 in the population.
+, *, ** are the 10, 5, and 1 percentage levels.

of socio-emotional skills (conscientiousness and agreeableness) are close to zero and not statistically significant at the 10 percent level, the effect on emotional stability is negative and statistically significant at the one percent level, suggesting that students in classrooms where the coaching intervention was implemented experienced more negative feelings such as loneliness and insecurity. The size of this effect (0.16 to 0.19 standard deviations) is substantial compared to effect sizes reported in many other educational interventions evaluated through field experiments (Kraft 2020). [13]

13. While emotional stability in students is concerned with feeling secure and not feeling lonely, the teaching practices denoted Regulation of Senses, Attention, and Emotions are associated with self-regulation using break areas, calming material, motion breaks and learning to put problems into perspective, see A.2. Therefore, the significant effects on those two variables are not necessarily related even though the wording overlaps.

# 4 Discussion

Overall, the field experiment confirmed the main hypothesis that coaching can act as a facilitator or catalyst for the implementation of theory into practice. However, contrary to our initial expectation, the intervention's impact was not stronger on input factors that are inherently more difficult to alter, such as teachers' actions in the classroom, than on those that are relatively easier to modify, such as elements aimed at improving the classroom's physical and visual organization. In fact, point estimates for behavioral input factors were lower and less statistically significant than those for technological input factors. Even though the difference in effects was not statistically significant, one interpretation of this result may be that it is easier for coaches to observe and provide feedback on the use of such more tangible techniques than the behavioral techniques. This raises questions about the effectiveness of the studied intervention in changing teachers' pedagogical practices in the classroom.

Besides the main effects on teacher behavior, we found no indications that student learning was improved by the treatment. On the contrary, we found that emotional stability was reduced. If this effect is not a false positive, two possible explanations emerge. First, any positive impact of increased teaching practices in this dimension due to the coaching intervention may have been offset by unintended disruptive effects. For instance, the periodic presence of an additional adult in the classroom may have inadvertently caused disturbances, or the newly introduced teaching practices may require more time and practice to be implemented effectively, initially feeling awkward or unnatural. Second, the content of the training program itself (adopted from the ASD NEST program in New York City) may not have been well-suited to the context of this study.

We find no evidence that the increased use of teaching practices promoted in the teacher training program may have induced any negative impact on students, suggesting that the observed negative effect would more likely be driven by potential disruptive effects of the coaching intervention itself. In Table 5, we present the results of a regression where teachers' observed use of NEST elements at follow-up (overall score) is regressed on student outcomes. We find no indication that greater adoption of these teaching practices is negatively correlated with student outcomes—if anything, the relationship appears to be positive. Specifically, the correlation with emotional stability is close to zero, while the correlations with conscientiousness

**Table 5:** Correlations between teachers' observed use of NEST elements (overall score) and student outcomes

| | (1) | (2) | (3) |
|---|---|---|---|
| Reading test score | -0.00351 | -0.0219 | -0.0204 |
| | (0.042) | (0.043) | (0.045) |
| Language comprehension | 0.0246 | 0.00109 | 0.00251 |
| | (0.037) | (0.038) | (0.040) |
| Decoding | -0.0637 | -0.0418 | -0.0403 |
| | (0.038) | (0.032) | (0.032) |
| Text comprehension | 0.0304 | -0.00781 | -0.00716 |
| | (0.044) | (0.049) | (0.050) |
| Conscientiousness index | 0.0747$^{+}$ | 0.113$^{**}$ | 0.0987$^{**}$ |
| | (0.039) | (0.036) | (0.030) |
| Agreeableness index | 0.0531 | 0.0790$^{*}$ | 0.0701$^{+}$ |
| | (0.039) | (0.038) | (0.037) |
| Emotional stability index | 0.00874 | 0.0412 | 0.0378 |
| | (0.033) | (0.032) | (0.034) |
| School FE | YES | YES | YES |
| Grade FE | | YES | YES |
| Baseline value | | YES | YES |
| Student covariates | | | YES |
| Teacher experience | | | YES |

Notes: The national test in reading is mandatory for grade 4 and 6 students. Entries are coefficients (classroom-clustered standard errors in parentheses). N= [800;1,098]. All outcome variables have been standardized to have mean of 0 and a standard deviation of 1 in the population.
+, *, ** are the 10, 5, and 1 percentage levels.

and agreeableness are positive and statistically significant when controlling for grade fixed effects, teachers' observed use of NEST elements at baseline, and baseline covariates (student covariates and teacher experience).

The lack of stronger effects on the students could have many explanations. Despite the observed effect on teachers' use of the techniques, the behavioral change may have been too weak to translate into more substantial student improvements. While the coaching intervention provided no explicit incentives to change effort, non-monetary incentives in terms of, for example, peer pressure or image concerns may arise. However, such effects are likely not lasting or strong enough to shift students' outcomes. Even though our intervention is much more intensive than the peer observations studied by e.g. Burgess, Rawal, and Taylor (2021),

the increase in effort driven by non-monetary incentives may still be limited.

It may also be that the techniques taught at the teaching program were not effective enough or that the teachers' translation of the theoretical knowledge into practice was not done adequately. Alternatively, it may take more time before the change in teacher behavior reflects positively on the students' learning and socio-emotional development. Yet, the negative effect on emotional stability provides some indication that there is a risk that the effects of the coaching intervention come at a cost.

It is also worth noting that the effects that we observe are isolated results of the coaching intervention on top of the teaching program, which was also received by the active control group. This means that the Hawthorne effects that stem from being part of an intervention or from the awareness of being observed and measured by a group of researchers may affect the treatment and control group equally and therefore cannot contribute to the differences in outcomes. As a result, our design captures the marginal impact of adding a coaching program on top of a teacher training program. While previous studies on coaching have generally reported positive effects, they often did not isolate the impact of coaching from that of concurrently implemented training or courses. In contrast, our conclusion is more nuanced and suggests that the impact of the coaching component alone may, in some cases, be quite limited.

# 5 Conclusion

Teacher training and other types of professional development may be a way of increasing teacher skills. However, it has proven difficult to translate new knowledge from training programs into practice, and therefore, further guidance in terms of coaching may be needed. Yet, evidence on the standalone effect of coaching, without being paired with training programs, remains scarce.

The results presented here indicate that coaching may act as a catalyst for this process. However, they also highlight the need for caution regarding the extent of its impact and the expectations surrounding this type of intervention, while warning of potential unintended consequences. Teachers receiving the coach intervention became better at transferring what they learned at the teaching program into changes in teaching practices in the classroom.

However, we did not find indications that this translated into improved student outcomes. In fact, we actually found evidence suggesting that coaching intervention may have had short-term negative impacts on students. Whether this is due to the content of the training program or to the coaching intervention itself remains to be seen. We see the potential for more research on different types of coaching combined with different types of teacher training.

## References


Allen, Joseph P., Robert C. Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun. 2011. "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement." *Science* 333 (6045): 1034–1037.

Andersen, Simon, Bastien Michel, and Helena Nielsen. 2024. "Impact and Mechanisms: Why Consulting Matters in Human Capital Intensive Organizations. Evidence from a Field Experiment on Teacher Coaching." *AEA RCT Registry.*

Andersen, Simon Calmar, Miriam Gensowski, Steven G. Ludeke, and Oliver P. John. 2020. "A Stable Relationship between Personality and Academic Performance from Childhood through Adolescence. An Original Study and Replication in Hundred-Thousand-Person Samples." *Journal of Personality* 88 (5): 925–939.

Beuchert, Louise Voldby, and Anne Brink Nandrup. 2018. "The Danish National Tests at a Glance." *Nationaloekonomisk Tidsskrift,* 1–37.

Burgess, Simon, Shenila Rawal, and Eric S Taylor. 2021. "Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools." *Journal of Labor Economics* 39 (4): 1155–1186.

Cecchini, Mathilde, and Gitte Sommer Harrits. 2022. "The Professional Agency Narrative—Conceptualizing the Role of Professional Knowledge in Frontline Work." *Journal of Public Administration Research and Theory* 32 (1): 41–57.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–2679.

Cohen, Julie, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. "Teacher coaching in a simulated environment." *Educational evaluation and policy analysis* 42 (2): 208–231.

Frank, Kenneth A., Ran Xu, and William R. Penuel. 2018. "Implementation of Evidence-Based Practice in Human Service Organizations: Implications from Agent-Based Models." *Journal of Policy Analysis and Management* 37 (4): 867–895.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning Through Noticing: Theory and Evidence from a Field Experiment." *The Quarterly Journal of Economics* 129 (3): 1311–1353.

Harris, Douglas N., and Tim R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95 (7): 798–812.

Jakobsen, Morten, Christian Bøtcher Jacobsen, and Søren Serritzlew. 2019. "Managing the Behavior of Public Frontline Employees through Change-Oriented Training: Evidence from a Randomized Field Experiment." *Journal of Public Administration Research and Theory* 29 (4): 556–571.

Kennedy, Mary M. 1999. "The role of preservice teacher education." *Teaching as the learning profession: Handbook of policy and practice,* 54–85.

———. 2016. "How does professional development improve teaching?" *Review of educational research* 86 (4): 945–980.

Koenig, Kristie P., Jamie Bleiweiss, Susan Brennan, Shirley Cohen, and Dorothy E. Siegel. 2009. "The ASD Nest Program: A Model for Inclusive Public Education for Individuals with Autism Spectrum Disorders." *Teaching Exceptional Children* 42 (1): 6–13.

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher* 49 (4): 241–253.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research* 88 (4): 547–588.

Kretlow, Allison Graves, and Christina C. Bartholomew. 2010. "Using Coaching to Improve the Fidelity of Evidence-Based Practices: A Review of Studies." *Teacher Education and Special Education* 33 (4): 279–299.

Møller, Anne Mette. 2022. "Mobilizing Knowledge in Frontline Work: A Conceptual Framework and Empirical Exploration." *Perspectives on Public Management and Governance* 5 (1): 50–62.

Raaphorst, Nadine. 2018. "How to Prove, How to Interpret and What to Do? Uncertainty Experiences of Street-Level Tax Officials." *Public Management Review* 20 (4): 485–502.

Schachter, Rachel E. 2015. "An Analytic Study of the Professional Development Research in Early Childhood Education." *Early Education and Development* 26 (8): 1057–1085.

Spybrook, J., H. Bloom, R. Congdon, C. Hill, A. Martinez, Stephen W. Raudenbush, and A. TO. 2011. *Optimal Design plus Empirical Evidence: Documentation for the "Optimal Design" Software.* Technical report. William T. Grant Foundation.

UVM. 2022. *The Danish education system.* The Ministry of Children and Education https://ufm.dk/en/publications/2022/files/the-danish-education-system.pdf.

# A Appendix

## A.1 Supplementary Information

**Table A.1:** Pairwise Correlation Socio-emotional Variables

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| | **Agreeableness** | | | | | | | | |
| 1 | I try to understand my friends when they are sad or in a bad temper | 1 | | | | | | | |
| 2 | I am good at working together with others in a group | 0.18 | 1 | | | | | | |
| | **Conscientiousness** | | | | | | | | |
| 3 | How often can you manage the things you set your mind to? | 0.27 | 0.26 | 1 | | | | | |
| 4 | Can you concentrate during lessons? | 0.18 | 0.27 | 0.35 | 1 | | | | |
| 5 | If I am interrupted during class, I can quicly concentrate again | 0.20 | 0.30 | 0.38 | 0.57 | 1 | | | |
| | **Emotional stability** | | | | | | | | |
| 6 | Do you feel lonely? | 0.02 | 0.13 | 0.11 | 0.17 | 0.12 | 1 | | |
| 7 | Other students accept me as I am | 0.14 | 0.25 | 0.23 | 0.26 | 0.25 | 0.38 | 1 | |
| 8 | How often do you feel secure at school? | 0.08 | 0.23 | 0.23 | 0.27 | 0.26 | 0.37 | 0.42 | 1 |

## A.2 Observation Form

The form on the next pages shows the items used by blinded observers of teacher practices in the classroom. The first part of the form (not presented) included names of the school, class, and observer. The four sections presented here are divided into (i) physical environment, (ii) organization of assignments, (iii) regulation of senses, attention, and emotions, and (iv) pedagogy and teacher mentalization (corresponding to the themes of the course modules). Items were further divided into technological factors (*i.e.*, elements designed to enhance the physical and visual organization of the classroom) and behavioral factors (*i.e.*, teachers' actions in the classroom). Items in the first category are labeled with the 'Tech' tag, which is displayed immediately after the item ID in the first column of the tables. All other items fall under the category of behavioral factors.

Upon completion of the training modules, the observation form was refined and reduced to 24 items at the request of the training organizers, as ten items were ultimately not covered during the training. These 10 items are highlighted in yellow, with their IDs removed. Our analysis focuses exclusively on the restricted set of 24 items.

## Section C: Physical Environment

| Question | | Response options | |
|---|---|---|---|
| | | **To what extent does the classroom live up to this?** (Write numbers from 1 to 5, where 1=not at all and 5=exceedingly high | **Notes** (Reasons for your assessment). |
| **C1** | 1. Classroom environment accommodates sensory sensitivity and prevents sensory overload (no visual distractions, but e.g. noise, smell and glaring lights are minimized with tools and practices such as "socks on the bottom of the chair legs" or chair silencers). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **C2** | 2. The classroom is organized to minimize visual distractions (such as mess, bright colors and too much furniture close to each other). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **C** | 3. General visual support is used to clarify expectations and academic concepts during the lessons and during individual work / group work (e.g., a sign indicating what the children should do when they want to answer a question, or when the lesson is over). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **C3 Tech** | 4. 5-point voice scales are visible in the classroom. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **C4** | 5. 5-point voice scales are used to specify the voice volume the pupils are allowed to use (for example, "no talk" as the lowest and "outside-voice" as the loudest). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **C5 Tech** | 6. Daily class schedule is visible in the classroom. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **C6** | 7. Daily class schedule is included as a support in transitions / refocusing. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **C7** | 8. Visual timers are used in classroom transitions and activities to show that time passes and how much time there is left. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

**Section D: Organization of Assignments**

| **Questions** | | **Response options** | |
|---|---|---|---|
| | | **To what extent does the framework live up to this?** (Write numbers from 1 to 5, where 1=not at all and 5=exceedingly high | **Notes** (Reasons for your assessment). |
| **D1** | 1. Information about future activities / transitions / expectations is given in advance. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **D** | 2. New, challenging material and / or content will be displayed to the pupils before instruction in the material / content. Pupils are hereby prepared in advance on potential challenges. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **D** | 3. Roleplaying is done with whole classes / in small groups to prepare the pupils for a new / difficult situation (e.g., fire drills, teamwork, playing a math game). Typically, the teacher plays the part first, after which the role-play is performed with the pupils. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **D2** | 4. Extra time is given to the pupils to process and respond to oral communication before, e.g., an answer is expected. I.e. That pupils are not expected to respond right away, but are given some time, e.g. to understand a question. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **D3** | 5. Written instructions are used as a supplement to verbal instructions (e.g. regarding problem solving, where it can be written on the whiteboard, what the pupils should do). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **D4** | 6. Visual charts that breaks down an assignment into a sequence of steps are used in routines and activities (e.g. in the form of checklists, where you can tick off each step during problem solving. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **D** | 7. Complex academic activities are broken down to clarify and dissolve steps and sequences in an assignment. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **D** | 8. It is clarified with the children which roles they should have during the instruction (e.g. when they should listen, when they should answer and when to they should give input). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

**Section E: Regulation of Senses, Attention and Emotions**

| **Questions** | | **Response options** | |
|---|---|---|---|
| | | **To what extent does the framework live up to this?** (Write numbers from 1 to 5, where 1=not at all and 5=exceedingly high | **Notes** (Reasons for your assessment). |
| **E1 Tech** | 1. In the classroom, there is a break area, which is, is inviting, accessible to the pupils and offers calming materials. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **E2 Tech** | 2. 5-point scales are visible in the classroom (e.g. a scale where pupils can indicate their experience of the size of a problem to learn how to put problems into perspective). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **E3** | 3. 5-point scales are referred to, to clarify abstract concepts (e.g. a scale where pupils can indicate their experience of the size of a problem to learn how to put problems into perspective). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **E4** | 4. The opportunity to move around is offered for the whole class (e.g. motion break or movement from one activity to another). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **E5** | 5. The teacher creates and introduces routines regarding breaks during the lesson (in other words, it is clear to the pupils what to do after completing tasks - and breaks thus do not generate a lot of noise). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **E6** | 6. The pupils are given the opportunity to make choices (e.g. whether they want to sit in the classroom or in the library when solving a given assignment, whether they want to write by hand or use a computer and which extra assignment they want to do). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **E7** | 7. The pupils' strengths, interests and learning styles are incorporated into learning activities. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **E8** | 8. Class reward system - with clear, specific behavioral expectations. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **E9** | 9. Motivation and reward are used in ways other than using a class reward system. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

**Section F: Pedagogy and Teacher Mentalization**

| Questions | | Response options | |
|---|---|---|---|
| | | **To what extent does the framework live up to this?** (Write numbers from 1 to 5, where 1=not at all and 5=exceedingly high | **Notes** (Reasons for your assessment). |
| **F1** | 1. Pupils are told what *to do* rather than what *not to do*. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **F2** | 2. The teacher "catches the pupils in being good" by giving behavior-specific praise rather than reproving. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **F** | 3. The teacher is aware of the pupils in the class and quickly detects who needs social and academic support. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **F** | 4. There is a positive atmosphere in the class between teachers and pupils, and they show that they like each other (e.g. shows physical presence, jokes and laughs with each other). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **F3** | 5. Through mirroring the teacher shows that the pupils' reasons to behave as they do are understandable. In other words that the teacher, e.g. interprets and explains what the cause of the inappropriate behavior may be. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **F** | 6. Self-talk is used to facilitate Social Thinking® and problem solving. I.e. that the teacher verbally puts words on his own thoughts and considerations, i.e. almost thinking aloud. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **F4** | 7. The teacher emphasizes when the pupils show flexibility and praise them for it. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

| | | | |
|---|---|---|---|
| **F** | 8. The teacher creates an atmosphere where it is okay for pupils to embark on a new academic area where they do not fully control all the details. | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |
| **F** | 9. The teacher uses self-regulation techniques during the instruction (e.g. ask for a break, take 3 deep breaths, count to 10 slowly and yoga exercises). | Numerical (1-5)<br>❑ 1<br>❑ 2<br>❑ 3<br>❑ 4<br>❑ 5 | *(Text field)* |

## A.3 Mechanisms

### A.3.1 Registered measures

To get some leverage on the potential mechanisms leading from coaching to behavioral changes in the classroom, we surveyed teachers to measure their information about the program, their attention, and their training in using the techniques taught during the program. Teachers were surveyed at baseline, at endline, and at the beginning of each of the four remaining course modules.

For each of the teaching techniques listed in Appendix A.2, the teacher questionnaire included three questions at the endline. We use their response to those three questions to measure information, attention, and training.

**Information** In order to investigate the effect of coaching on teachers' perceived information from the program, we use perceived importance of various input factors. Following Hanna, Mullainathan, and Schwartzstein (2014), we proxy whether or not the teachers are knowledgeable about a given input factor by whether or not they are capable of assessing what its optimal level is. More precisely, a teacher is considered knowledgeable about a given input factor if they do not answer "I don't know" to a question investigating how important teachers perceive the input factor to be: "How important is this element for the students' learning? (Write numbers from 1 to 5, where 1 = very low importance and 5 = very important)." We label this variable "Unknown information."

Given that treatment and control groups will benefit from the exact same course modules, information may rise in similar ways in both groups immediately after the sessions. However, in the case of information frictions—in this case when teachers believe they know about the effect of an instrument taught at the teaching program—coaches may help update their information about the technique.

Table A.2 shows the mean number of responses in each response category. Table A.3 shows the number of 'Don't know' responses for each item. It is clear that 'Don't know' was not a frequent choice among the teachers. The limited variation in our measures using 'Don't know' means that we risk null results due to limitations in measurement. We return to this issue after presenting the results.

**Attention** Again following Hanna, Mullainathan, and Schwartzstein (2014), we proxy whether or not a teacher pays attention to a given input factor by whether or not they are capable of assessing the extent to which they use that input factor when teaching. More precisely, a teacher is considered as paying attention to an input factor if they do not answer "I don't know" to a question investigating the extent to which teachers use that input factor: "To what extent do you think you use this element in your teaching? (Write numbers from 1 to 5, where 1 = to a very low extent and 5 = to a very high extent)." We label this "Unknown attention.")

**Training.** Teachers may know the optimal input of a technique, they may pay attention to their own use of that technique, but still be unable to use it. In order to investigate the strength of this mechanism, we combine their own desired use of the input with the surveyors' observation. The teachers are asked about the extent to which they wish they used—or more often used—the input factors when teaching. Based on their response, we construct an additional measure of teachers' lack of training, which is proxied by the difference between the extent to which teachers report they wish they used a specific technique and the extent to which they actually used it (as observed by our surveyors). The greater the difference, the more problematic the lack of training is. We label this "Training."

### A.3.2 Suitability of measures to study mechanisms

**Table A.2:** Mean number of responses by category

| Category | Very low | Low | Neither | High | Very high | Dont_know |
|---|---|---|---|---|---|---|
| Emotions | 8 | 7 | 14 | 16 | 10 | 3 |
| Environment | 6 | 4 | 11 | 14 | 20 | 1 |
| Organization | 3 | 10 | 17 | 16 | 10 | 1 |
| Pedagogy | 3 | 10 | 17 | 17 | 9 | <1 |
| Total | 6 | 7 | 14 | 15 | 13 | 1 |

**Table A.3:** Number of Don't Know responses

| **Tool** | **Qeustion type** | Frequency |
|---|---|---|
| C1. Classroom is environment organized to prevent sensory overload | attention | <5 |
| | information | <5 |
| | wish | <5 |
| C2. The classroom is organized to minimize visual distractions | attention | <5 |
| | information | <5 |
| | wish | <5 |
| C3. 5-point voice scale is visible in the classroom | attention | <5 |
| | information | <5 |
| | wish | <5 |
| C4. 5-point voice scale is used in the classroom | attention | <5 |
| | information | <5 |
| | wish | <5 |
| C5. Daily class schedule is visible in the classroom | attention | <5 |
| | information | <5 |
| | wish | <5 |
| C6. Daily class schedule is used to support transitions | attention | <5 |
| | information | <5 |
| | wish | <5 |
| C7. Visual timers are used during transitions and activities | attention | <5 |
| | information | <5 |
| | wish | <5 |
| D1. Information about future activities are given beforehand | attention | <5 |
| | information | <5 |
| | wish | <5 |
| D2. Extra time is given to pupils to process and respond to oral communication | attention | <5 |
| | information | <5 |
| | wish | <5 |
| D3. Written instructions are used to supplement verbal instructions | attention | <5 |
| | information | <5 |
| | wish | <5 |
| D4. Visual charts are used to break down assignments into steps | attention | <5 |
| | information | <5 |
| | wish | <5 |
| E1. An inviting area exists where students can calm down | attention | 6 |
| | information | 5 |
| | wish | <5 |
| E2. 5-point scales are visible in the classroom | attention | 9 |
| | information | <5 |
| | wish | <5 |
| E3. 5-point scales are referred to to clarify abstract concepts | attention | 10 |
| | information | <5 |
| | wish | <5 |
| E4. The opportunity to move around is offered to the whole class | attention | <5 |
| | information | <5 |
| | wish | <5 |
| E5. The teacher creates routines during the breaks | attention | <5 |
| | information | <5 |
| | wish | <5 |
| E6. Pupils are given the opportunity to make choices | attention | <5 |
| | information | <5 |
| | wish | <5 |
| E7. Pupils' strengths, interests and capabilities are used in learning activitie | attention | <5 |
| | information | <5 |
| | wish | <5 |
| E8. A class reward system is used with clear specific behavioral expectations | attention | 7 |
| | information | 7 |
| | wish | <5 |
| E9. Motivations and rewards (other than the class reward system) are used | attention | 7 |
| | information | 7 |
| | wish | <5 |
| F1. Pupils are told what to do rather than what not to do | attention | <5 |
| | information | <5 |
| | wish | <5 |
| F2. The teacher praises rather than reproves pupils' behaviors | attention | <5 |
| | information | <5 |
| | wish | <5 |
| F3. The teacher helps pupils understand how they feel and behave | attention | <5 |
| | information | <5 |
| | wish | <5 |
| F4. The teacher emphasizes when pupils show flexibility and praises them for it | attention | <5 |
| | information | <5 |
| | wish | <5 |

**Table A.4:** ITT estimates of the coaching intervention's impact on potential mechanisms

| | (1) | (2) | (3) |
|---|---|---|---|
| Unknown information | 0.0173 | 0.00957 | 0.00513 |
| | (0.018) | (0.018) | (0.017) |
| Unknown attention | -0.0171 | -0.0316 | -0.0391 |
| | (0.039) | (0.046) | (0.048) |
| Training | -0.178 | -0.109 | -0.102 |
| | (0.262) | (0.254) | (0.268) |
| School FE | YES | YES | YES |
| Grade FE | | YES | YES |
| Baseline value | | YES | YES |
| Repeat class | | | YES |
| Teacher experience | | | YES |

Notes: Entries are coefficients (robust standard errors in parentheses).
+, *, ** are the 10, 5, and 1 percentage levels. N=53.

### A.3.3 Effect of the coaching intervention on registered measures of mechanisms

Table A.4 shows estimates of the effect of the coaching intervention on potential mechanisms: information, attention, and training. None of the results are statistically significant. An insignificant result is not the same as showing there is no effect, so we cannot conclude that the coaching intervention did not work through any of these potential mechanisms—only that the results are inconclusive.

## A.4 Data Availability Statement

The analyses of this paper are based on administrative registers maintained by Statistics Denmark. Analysis of these data can only be conducted on servers hosted by Statistics Denmark. Statistics Denmark guidelines as well as current legislation entail that these data cannot be made publicly available. The authors of the paper will ensure that the analysis data sets as well as any programs needed to replicate the results of this paper are archived for at least five years following the date of publication. In the interest of scientific validation and replication of the analyses of this paper, the Department of Economics and Business Economics, Aarhus University, will assist researchers who are interested in validating the results of the paper. Statistics Denmark must approve any researcher who is to have access to

the data and data access can be obtained from Aarhus only. Replications requests should be directed to the ECONAU data management team at datamanager@econ.au.dk.