Hogrebe and Strietholt *Large-scale Assess Educ (2016) 4:2*
DOI 10.1186/s40536-016-0017-3

Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Does non-participation in preschool affect children's reading achievement? International evidence from propensity score analyses

Nina Hogrebe[1*] and Rolf Strietholt[2,3,4]

*Correspondence: nina.hogrebe@uni-muenster.de
[1] Department of Education, University of Münster, Georgskommende 33, 48143 Münster, Germany
Full list of author information is available at the end of the article

Abstract

While expectations are high for early childhood education to support students' reading literacy, research findings are inconclusive. The purpose of the study is to estimate the effect of preschool non-participation on reading literacy at the end of primary school. That is, what is the average achievement of children who did not attend preschool compared to what it would have been if they had attended preschool? Using PIRLS 2011 data, we employ propensity score matching to approximate a randomized experiment on a large-scale basis to estimate this effect for nine countries. We find that children who did not attend preschool come from disadvantaged backgrounds in all countries. However, with the exception of two countries, our study shows that their reading achievement at the end of primary school is not statistically significantly lower than the performance of matched children from similar backgrounds who attended preschool. Keeping in mind some methodological limitations, we discuss the findings of our study from a policy perspective.

Keywords: Preschool participation, Disadvantaged children, PIRLS, Propensity score matching, Reading achievement

Background

International large-scale assessments (ILSAs) provide representative information about a population's knowledge, skills, or behaviors in certain domains and compare them across countries. In education, such surveys were first conducted in the 1960s and have since then expanded in terms of investigated domains, included countries, and observed populations. However, primary and secondary school students' competencies in core subjects are still the central fields of investigation (Kirsch et al. 2013; Strietholt et al. 2014). ILSAs in early childhood education[1] are scarcer and attract fewer countries to participate. One of the first cross-national studies in early childhood education was the longitudinal Preprimary Project (PPP) (1986–2006), which was conducted by the

[1] Throughout the paper, we refer to early childhood education as defined by the International Standard Classification of Education (ISCED) Level 0. Here, early childhood education targets children below the age of entry into primary school. The respective programs are schoolbased or otherwise institutionalized and characterized by organized and purposeful learning activities outside of the family with an intentional education component. A minimum intensity and duration of at least two hours per day and 100 days a year is required (UNESCO 2012). The term preschool is used synonymously.

International Association for the Evaluation of Educational Achievement (IEA) and related preschool experiences at age four to children's language and cognitive development in school (age seven; Montie 2011). Only seven countries participated in all three study phases.

While young children's competencies have so far been understudied from an international comparative perspective, early childhood education itself is a recurring topic in the discussion of ILSA results, namely in attempts to explain differences in school outcomes. As ILSAs like the Programme for International Student Assessment (PISA) or the Progress in Reading Literacy Study (PIRLS) reveal that some students lack relevant basic skills and that many of them did not attend preschool, expectations are high that early childhood education is a promising approach to support students' literacy. For example, PIRLS 2011 shows a positive relation between preschool attendance and reading achievement at the end of primary school. Based on this finding, Mullis et al. (2012) emphasize the importance of early childhood education for students' later success in school and claim that "an early start [is] crucial in shaping children's reading literacy" (p. 10). A recent report draws similar conclusions based on PISA data (Schleicher 2014). Against this background, early childhood education is often adduced as a policy lever for improving students' future school competencies (e.g. UNESCO 2006).

One is naturally skeptical about inferences based on correlations within cross-sectional data. The difference in achievement scores between children who attended and those who did not attend preschool may not be interpreted in causal terms because of confounding variables that potentially bias the results. The PISA 2012 data, for example, illustrate that the strong relationship between preschool attendance and school performance in mathematics is considerably reduced in nearly all countries when accounting for socio-economic status (OECD 2013). Such selection mechanisms are a well-documented phenomenon. Research on preschool participation shows that disadvantaged children often do not enter preschool programs at all, participate to a lesser extent, or experience a lower quality than their more privileged peers (e.g. Hynes and Habasevich-Brooks 2008; Pianta et al. 2009).

It is the purpose of the current study to make use of the advantages of ILSA data (i.e. especially its representativity and cross-country comparativeness) in order to investigate the effect of preschool non-participation on grade-four-students' reading achievement using the PIRLS 2011 data. More precisely, we want to answer the following policy question: What effects might preschool attendance have had on the reading competencies of children who did not participate in early childhood education? The risk of bias due to selection mechanisms is dealt with by propensity score matching (PSM), which can be interpreted as a simulation of an experiment by conditioning on observed influencing factors.

Theoretical framework

Researchers from various disciplines argue that early childhood education is promising in supporting students' literacy acquisition (e.g. Knudsen et al. 2006). Ecological theories of human development provide a framework to model the factors that potentially influence a child's development. According to Bronfenbrenner (1979, 1990), human development is the result of an interplay between individuals' dispositions and the context in which they grow up. He describes this developmental environment as a system of

nested, interdependent, dynamic structures that range from proximal micro- to the distal macro-levels. Structures at rather distal levels—such as cultural values and societal structures as well as the parents' work life—do not have a direct impact on children's development but are linked to the more proximal environments and thereby exert their influence indirectly. The main effects emanate from the immediate settings. Efforts to support children's literacy should address these early proximal contexts. In this respect, families and early childhood education generally provide the two most important primary learning environments for young children before school.

However, these two contexts are not independent from each other. Family background characteristics and preschool participation correlate, and children form disadvantaged families participate to a lesser extent in early childhood education. An extension of ecological theories of human development can be used to explain these selection processes. According to the Process-Person-Context-Time model, processes vary as a function of persons' characteristics, environmental contexts, and time periods in which they take place (Bronfenbrenner and Morris 2006). Early and Burchinal (2001) applied this model to the context of preschool participation and showed that parental preferences for certain care characteristics (person) and their subsequent choices for or against early childhood education (process) are influenced by their circumstances of life (context). The authors found that higher income is associated with forms of non-parental care including preschools and childcare provided by relatives other than parents while ethnicity is more related to the use of the latter. There is further empirical evidence that shows that parents with less education and a lower occupational status also choose care by a relative over early childhood education as they are more concerned with practical issues such as costs and convenience. On the contrary, highly educated parents are more likely than other parents to send their children to preschools (Grogan 2012; Kim and Fram 2009).

Although we have witnessed an expansion of early childhood education prior to primary school worldwide (UNESCO 2006), research results on the long-term effects of early childhood education programs are somewhat ambiguous (Duncan and Magnusson 2013). One reason for this is that there are different types of preschool programs that are differently influenced by the correlation between family background and preschool participation. A discussion of research on the effectiveness of these programs is therefore linked to respective target groups and the extent to which disadvantaged children are addressed. While some effective small-scale programs target at-risk children, the results of those interventions cannot be generalized to other programs or countries for two reasons. Firstly, they do not entail a random sample of disadvantaged children. Secondly, they are not representative of other targeted programs and less so for universal and voluntary early childhood education systems.

Previous research on the effects of early childhood education on reading literacy at the end of primary school

Literature reviews and meta-analyses that deal with the effects of early childhood education on a student's academic achievement emphasize substantial variation in the effect sizes between the individual studies, which is often associated with the investigated programs and the target-groups they address (Barnett 2011; Burger 2010; Camilli et al. 2010; Chambers et al. 2010; Duncan and Magnusson 2013; Pianta et al. 2009).

Some influential studies on the longer-term effects of targeted preschool programs on children's future academic success started about 50 years ago. Characteristic for these approaches to early childhood education—like the High/Scope Perry Preschool Project (Schweinhart et al. 2005) or the Abecedarian Program (Campbell and Ramey 1994)—is that they focused on a small sample of only about one hundred at-risk children and were based on intense, high quality curricula that included elements of parental support. Research on these interventions employed randomized-controlled trials and provides compelling evidence for the lasting effects such programs can have on a variety of outcomes observed at school age, including reading achievement. However, due to the program's high-quality, constrained regional or, rather, local context, limited and specific target group, and historical and political context the research result have a limited external validity, and it "is difficult to extract policy lessons from these two initiatives for early childhood education programs that states or the federal government might offer today" (Duncan and Magnusson 2013, p. 117).

Current large-scale targeted programs that address disadvantaged children are not able to replicate the results of these small-scale research programs (Barnett 2011). The Head Start impact study, for example, found that children entering Head Start programs at age four tend to show somewhat higher basic reading skills, but the effect was not statistically significant. For younger children (three-year-olds) it was not visible at all (Puma et al. 2012). Similarly, research on Early Head Start (EHS) for children aged two and three did not find any clear impacts on child outcomes when the children were about 10 years old, at least not for the overall sample. Regarding reading competencies, effects could only be identified for children with less pronounced risk profiles (Vogel et al. 2010). Given the results of EHS, one might assume that programs that do not target disadvantaged children but offer universal access for all children at a certain age are also rather ineffective for at-risk children. However, the Effective Provision of Pre-school Education Project (EPPE; 1997–2013)—a seminal European study of universal and voluntary preschool—for example, showed that preschool participation is related to a significant longer-term impact especially for the most disadvantaged children but emphasizes the role that the quality of early education settings within universal preschool systems plays in order to realize positive effects (Sylva et al. 2008).

Research question

The above review of the research on the effects of participating in early childhood education on future reading achievement at the end of primary school reveals inconclusive results. What is needed, then, are representative and international comparative studies with a high degree of generalizability. ILSA provide a valuable data source as they encompass representative samples from multiple countries. Findings from such studies may not only be generalized within countries but also compared across countries as they use the same methods of data collection and measures. ILSAs, however, are typically observational in nature and due to selection mechanisms in preschool participation it is not valid to draw conclusions about the influence of preschool attendance on reading achievement at the end of primary school on the basis of simple correlations. Thus, the purpose of the current study is to investigate the effect that preschool participation might have on non-participating grade-four-students' reading achievement using the

PIRLS 2011 data by employing propensity score matching. Before we elaborate on the setup of the study, we briefly outline some basic principles of causal effects and propensity score methods.

Propensity score methods for causal inferences

Causality and selection bias

Let's say that we are interested in the consequence of withholding a child from preschool for his or her later cognitive achievement. In this case, the treatment condition is not attending preschool. To formalize this problem, we can think about preschool attendance as a binary variable, $A_i = \{0,1\}$. The outcome, student achievement at the end of primary school, is denoted as Y_i. Irrespective of a child's actual preschool attendance, Y_{0i} is the achievement score if this child had attended preschool, while Y_{1i} is the achievement score if the same child had not attended preschool. For any individual there are two potential outcomes: Y_{1i} if $A_i = 1$ and Y_{0i} if $A_i = 0$. The difference of $Y_{1i} - Y_{0i}$ is the causal effect of preschool for an individual. This notation is also referred to as the potential outcome framework or as the Rubin causal model (Rubin 1974; Imbens and Rubin 2015).

Although it does not require more than one child to define the causal effect in theory, inferences about a causal effect require multiple individuals because it is impossible to observe both potential outcomes for the same person. For this reason, we must compare the average achievement of children who were exposed to the treatment $E[Y_{1i}|A_i = 1]$ with the mean results of those who were not, i.e. $E[Y_{0i}|A_i = 0]$. Comparisons between both groups, however, will only permit valid inferences about a causal effect if the two groups do not differ in terms of other predictors of student achievement. With observational data this is rarely the case. With regard to the preschool example, as discussed previously, there is a selection effect into preschool. As the covariates of preschool participation are typically also associated with later student achievement, the simple comparison of the average achievement by treatment status reflects not only the average causal effect of the treatment (i.e., non-participation in preschool) on those who received the treatment but also suffers from selection bias (i.e., family background).

Before we demonstrate how propensity score methods address the issue of selection bias, we need to introduce a conceptual distinction between the average treatment effect (ATE) and the average treatment effect on the treated (ATT; Imbens 2004). The ATE is the average effect of being treated for both treated and untreated children. The ATT is the average effect of being treated for students who actually received the treatment. This distinction is meaningful in observational studies because treated and untreated children may differ. Heckman and Robb (1986) argue that the ATT can be particularly valuable for research questions that concern social policy making. In the same vein, we argue that the ATT is more useful if policy makers consider making preschool compulsory or try to increase participation rates. Such a reform would only have consequences for those who are currently *not* attending preschool. For this reason, we think that the ATT is of greater interest for this particular research context. However, there is no general rule that the ATE or the ATT is of greater utility.

In PIRLS, the target population is students enrolled in grade four in the participating countries (Martin and Mullis 2012). In this regard, DuGoff et al. (2014) distinguish between the sample ATE and ATT (SATE and SATT) and population ATE and ATT

(PATE and PATT). SATE and SATT correspond to the ATE and ATT in an unweighted survey sample, and they may only be generalized to the students in the sample. PATE and PATT correspond to a weighted survey sample which accounts for the sampling design, and they may be generalized to the target population of all fourth graders in the respective countries. Often researchers aim to generalize their findings to the target population so that PATE and PATT are of greater interest.

The central role of the propensity score

In a carefully designed and implemented experimental study, random assignment of treatment overcomes the selection bias. The benefit from randomization is that all covariates will be balanced in treatment and control groups if they are sufficiently large. Propensity score methods replicate a randomized experiment as they aim at balancing covariates. Rosenbaum and Rubin's (1983) key insight is that adjusting for the propensity score removes bias from all observed confounding variables. The propensity score e_i is defined as the probability to receive the treatment conditional on a set of observed covariates:

$$e_i = \Pr(A_i = 1|X_i)$$

To estimate the propensity score for each individual, we can use the predicted probability to receive the treatment from a fitted regression model. Typically, the logistic regression model is used to regress the treatment on the covariates to obtain the propensity scores. Conditional on this score, two groups of treated and control units have the same distribution of observed covariates X_i. That means that the propensity score effectively controls for all observed covariates. Thus, any outcome difference between treated and control units cannot be due to the observed confounding variables. The propensity score translates the problem of selection bias from a multivariate vector of many covariates into a single score e_i for each individual. It is quite obvious that the propensity score approach can only attempt to achieve balance in observed confounding variables whereas randomization automatically also balances unobserved covariates. The key assumption in propensity score analysis is the ignorable treatment assignment, which implies that there is no hidden bias from unobserved confounders (Rosenbaum and Rubin 1983; Imbens 2004). Stuart (2010) points out that this assumption is sometimes more reasonable than it may sound at first. She argues that controlling for observed covariates also controls (at least partly) for unobserved covariates insofar as they are correlated with observed covariates. The Achilles heel of propensity score methods are unobserved covariates that are unrelated to the observed covariates. It should be noted that propensity score techniques exist for observational (nonrandomized) studies with multiple treatments (Imai and van Dyk 2004; McCaffrey et al. 2013), doses of treatment (Imbens 2000; Rosenbaum 2002), and treatments at school level (Stuart 2007).

Adjusting for the propensity score

There are two different sets of strategies to adjust for the propensity score. The first approach is propensity score matching. The simplest form of matching is 1:1 nearest neighbor matching, which means to select for each treated unit the control unit with the most similar propensity score. There are several versions of matching including with

or without replacement, increasing numbers of controls per treated unit (k:1 matching), and the restriction that only pairs of treated and untreated units are formed if their propensity scores differ at most by a pre-specified amount (caliper matching). Nearest neighbor matching estimates the (sample or population) ATT because it matches control units to the treated units and discards controls which are not selected as matches. One drawback of matching is that it does not use all available data because control units are disregarded if other controls better match with the propensity scores of the treated units. The second set of strategies uses all data: Stratification (or subclassification) creates groups of individuals with a similar propensity score. For this purpose, the propensity score distribution is divided into, say, five subclasses. Within each subclass outcome differences between treated and control units are then compared separately. To estimate the (sample or population) ATT, we weigh each subclass by the number of treated units in this subclass before pooling them; weighting by the overall number of units in each subclass estimates the (sample or population) ATEs. Another strategy to adjust for the propensity score is weighting. Weighting uses the propensity score to compute weights for each individual. These weights can be computed in different ways to estimate the SATE or the SATT, and they can then be used like sampling weights. They can also be multiplied with the actual sampling weights to estimate PATE or PATT. Austin (2011) and Stuart (2010) elaborate on the different approaches of propensity score techniques and discuss trade-offs.

Balance diagnostics

It is important to consider that treated and untreated units can differ dramatically in observational studies. In such situations propensity score methods may be ineffective. It is, therefore, important to check the balance between the samples after adjusting for the propensity score. Rubin (2001) proposes three balance measures: the standardized bias, the ratio of the variances of the propensity scores, and the ratio of the variances of the residuals of the covariates after regressing them on the propensity scores. The standardized bias is defined as the difference in means divided by the standard deviation in the treatment group. It can be computed for continuous and binary variables. Ideally, standardized biases should be as small as possible, but values less than 0.25 are considered to be acceptable (Harder et al. 2010; Stuart 2010). However, this is rather a rule of thumb than a strict cut-off value and it is advisable to use regression adjustments to remove the remaining bias (see the next paragraph; Austin 2009). The ratio of the variances of the propensity scores and the ratio of the variances for the residuals of the covariates after adjusting for the propensity scores should be close to one (e.g. between 0.5 and 2; Rubin 2001).

Estimating the causal effect

Propensity score methods themselves are not methods for estimating causal effects. The estimation of causal estimates is a separate step after having adjusted for the propensity score through matching, stratification, or weighting. Parametric or nonparametric approaches like Fisher's exact test, Wilcoxon signed-rank test, paired samples *t*-tests, or different regression models may be used. For propensity score matching, for example, we can use the matched samples and regress the outcome of interest on the treatment

variable to estimate the causal effect. An advantage of regression models is that the observed covariates can be used as control variables. Such additional regression adjustments remove the remaining bias if treated and control units are not perfectly balanced after the matching. Weighted least square models may be used to incorporate survey weights (or the weight from propensity score weighting). Finally, is important to consider that ILSAs involve specific design issues, i.e. plausible values as outcomes and complex samples. All analyses have to be repeated for each plausible value and replication techniques have to be used for variance estimation (e.g., Jackknife 2 in PIRLS; see Foy and Kennedy 2008).

Applying propensity score matching to PIRLS data

Sample

In order to answer our research question regarding whether non-participation in preschool affects children's reading achievement, we apply propensity score matching to data from PIRLS 2011 (Mullis et al. 2012). PIRLS provides data on students' reading achievement, preschool attendance, and other factors of home and learning environments in several countries. Our analyses are based on stratified and clustered random samples of students at the end of primary school in nine countries: Chinese Taipei (n = 4293), Germany (n = 4000), New Zealand (n = 5644), Norway (n = 3190), Russia (n = 4461), Singapore (n = 6367), Slovakia (n = 5630), Spain (n = 8580), and Sweden (n = 4622). The countries have well-established early childhood education systems with high enrollment rates and, therefore, are well suited to establish both control and treatment groups for the study. PIRLS provides data from reading literacy tests for fourth graders and background questionnaires for students, parents, teachers, and principals. Here, we rely on data from the reading literacy tests and information from the student and the parental background survey.

Instruments

Treatment

We use a binary treatment variable that is 0 if the student attended preschool for three or more years and 1 if they did not attend preschool. It is important to note that we define non-participation as the treatment condition. Propensity score matching allows us to estimate the ATT; this implies that we estimate the effect of preschool participation for children who did not participate. We think that this effect of is of great interest for policy makers who consider extending preschool participation.

Information on preschool participation is derived from an item of the parental survey (ASDHAPS[2]) where parents are asked if and for how long their child attended early childhood education. Table 1 lists the distribution of how long the students attended early childhood education in the respective countries. For the sake of simplicity, we excluded all students who attended preschool up to 3 years. Of course, preschool participation could also be considered as a continuous treatment. However, for illustrative purposes, we focus on a binary treatment instead of modeling different doses of preschool (see Imai and van Dyk 2004). The further analyses are based on the effective sample sizes

[2] The item names refer to the names in the original data sets.

Table 1 Distribution of preschool attendance

	Attendance of preschool (in years)				Sample size	
	(1) Did not attend	(2) 1 Year or less	(3) Less than 3 years but more than 1 year	(4) 3 Years or more	(5) Original	(6) Present study [(1) and (4)]
Chinese Taipei	53	175	2411	1654	4293	1707
Germany	51	62	952	2935	4000	2986
New Zealand	254	247	3017	2126	5644	2380
Norway	94	70	744	2282	3190	2376
Russia	592	106	615	3148	4461	3740
Singapore	80	98	2215	3973	6367	4053
Slovakia	247	393	1298	3691	5630	3938
Spain	158	222	2351	5848	8580	6006
Sweden	199	114	908	3401	4622	3600
Total	1728	1487	14511	29058	46787	30786

We used multiple imputations to replace missing data, and the reported frequencies are based on the combined results from the five imputed datasets (see the section on missing data)

listed in column (6), which is the sum of the children who did not attend (column 1) and children who attended preschool for at least 3 years (column 4).

Outcome

The outcome variable is the overall reading achievement score. (ASRREA01-ASRREA05; see von Davier et al. 2009; Martin and Mullis 2012, for more technical details). We use all five plausible values as outcomes to estimate the treatment effect and combine the results using Rubin's (1987) rules. Table 2 provides descriptive statistics for the outcome variable.

Covariates

The availability of observable covariates that serve as matching variables critically influences the internal validity of the matching. To model the selection into treatment, we refer to the Process-Person-Context-Time model that suggests that participation in early childhood education is associated with socioeconomic status and other background factors of children and their families (see above). International research provides empirical evidence that these factors are key covariates for the selection into treatment across countries (e.g., Grogan 2012; Hirshberg et al. 2005; Kim and Fram 2009; Müller et al. 2014; Zachrisson et al. 2013; Vandenbroeck et al. 2008). Thus, we combine a rich set of students' background and family measures that relate to preschool participation or students' reading competencies from the student and parent surveys (see Table 2 for descriptive statistics)[3]:

- *Parents Attitudes Towards Reading* (ASBHPLR) is a scale based on eight 4-point Likert items (e.g. 'I read only if I have to') from the parental questionnaire.

[3] The questionnaires along with further information on the construction on the scales can be downloaded from the project website (http://timss.bc.edu/pirls2011/).

Table 2 Descriptive statistics and amount of missing data for the outcome and covariates

	Reading Achiev. (ASRREA01–ASRREA05)	Parents Attitudes Towards Reading (ASBHPLR)	Language at Home (ASBG03)	Highest Parental Occupational Status (ASDHOCCP)	Highest Parental Educational Status (ASDHEDUP)	Gender (ITSEX)	Early Home Literacy Activities (ASBHELA)	Books at Home (student survey) (ASBG04)	Books at Home (parental survey) (ASBH14)
Chinese Taipei									
Mean	561.25	9.49	1.51	3.79	2.73	1.54	8.79	2.02	1.96
SD	60.47	1.71	0.52	1.24	0.99	0.50	1.90	1.28	1.32
Missing (%)	0.00	1.44	0.98	6.64	2.80	0.00	1.19	0.84	1.51
Germany									
Mean	544.32	10.14	1.80	3.57	2.50	1.51	10.20	2.18	2.54
SD	62.20	2.21	0.41	1.17	1.26	0.50	1.73	1.15	1.23
Missing (%)	0.00	20.73	10.10	27.18	27.05	0.00	20.08	10.75	20.73
New Zealand									
Mean	530.33	10.73	1.72	4.07	2.89	1.51	10.91	2.17	2.47
SD	87.46	2.19	0.49	1.23	1.13	0.50	2.12	1.16	1.25
Missing (%)	0.00	40.06	1.05	43.04	41.02	0.00	40.52	2.20	40.26
Norway									
Mean	513.05	10.78	1.81	4.44	3.47	1.50	10.08	2.31	2.91
SD	57.44	2.13	0.43	1.01	0.82	0.50	1.72	1.12	1.12
Missing (%)	0.00	9.62	2.76	12.35	12.60	0.00	8.81	3.26	9.44
Russia									
Mean	572.32	9.63	1.84	3.77	3.36	1.51	11.13	1.97	2.29
SD	61.57	1.79	0.41	1.25	0.77	0.50	1.94	1.06	1.15
Missing (%)	0.00	0.94	0.34	4.15	5.65	0.00	1.10	0.29	1.05
Singapore									
Mean	575.46	9.74	1.29	4.18	2.86	1.49	9.53	2.16	1.88
SD	74.71	1.75	0.56	1.14	1.10	0.50	2.11	1.07	1.22
Missing (%)	0.00	2.76	1.55	9.25	5.61	0.00	2.72	1.76	2.87

Table 2 continued

	Reading Achiev. (ASRREA01–ASRREA05)	Parents Attitudes Towards Reading (ASBHPLR)	Language at Home (ASBG03)	Highest Parental Occupational Status (ASDHOCCP)	Highest Parental Educational Status (ASDHEDUP)	Gender (ITSEX)	Early Home Literacy Activities (ASBHELA)	Books at Home (student survey) (ASBG04)	Books at Home (parental survey) (ASBH14)
Slovakia									
Mean	541.51	10.05	1.77	3.59	2.64	1.50	10.55	2.00	2.13
SD	64.12	2.01	0.49	1.37	1.01	0.50	1.85	1.13	1.17
Missing (%)	0.00	2.84	1.14	11.28	4.23	0.00	2.65	1.23	2.95
Spain									
Mean	523.98	10.10	1.73	3.57	2.50	1.50	10.50	2.05	2.35
SD	62.27	2.10	0.55	1.37	1.35	0.50	1.78	1.15	1.21
Missing (%)	0.00	7.32	1.15	13.76	12.21	0.00	7.40	1.39	7.55
Sweden									
Mean	543.58	10.86	1.77	4.20	3.06	1.51	9.95	2.25	2.72
SD	62.66	2.13	0.45	1.17	1.00	0.50	1.78	1.16	1.23
Missing (%)	0.00	13.63	2.42	18.91	20.96	0.00	13.18	3.66	13.95

We used multiple imputations to replace missing data, and the reported means and standard deviations (SD) are based on the combined results from the five imputed datasets (see the section on missing data)

- *Language at Home* (ASBG03) is a single item from the student questionnaire ('How often do you speak <language of test> at home?') with a 3-point response scale: never/sometimes/always or almost always.
- *Highest Parental Occupational Status* (ASDHOCCP) is a variable derived from two items in the parental questionnaire about the father's and mother's main job (e.g., 'Clerk—Includes office clerks; secretaries; typists; data entry operators; customer service clerks').
- *Highest Parental Education* (ASDHEDUP) is a variable derived from two items about the father's and mother's highest educational level according to the ISCED classification.
- *Gender* (ITSEX) was recorded during the sampling.
- *Early Home Literacy Activities* (ASBHELA) is a scale based on nine items from the parental questionnaire about how often parents (or someone else) did a set of activities before the child entered school (e.g. reading books, tell stories) on a 3-point frequency scale (often/sometimes/never or almost never).
- *Books at Home (student survey)* (ASBG04) and *Books at Home (parental survey)* (ASBH14) are items from the student and the parent questionnaire, respectively. There were five response categories (0–10, 11–25, 26–100, 101–200, or more than 200).

A general limitation of all measures is that they were not collected prior to preschool education but at the end of primary school so that they may not only have affected the selection into treatment but also vice versa. This would be a serious threat for our study because we would condition on an outcome of attending preschool. For example, the availability of a publicly funded preschool education is also regarded as a policy to promote female labor force participation. If such covariates are themselves outcomes, they are 'bad controls' and should not be controlled for because that might cause underestimation of the treatment effects (e.g. Angrist and Pischke 2009). However, we think that all the covariates are rather stable measures. For example, the availability of pre-primary education probably affects only one parent's employability (typically the mother). But we do not think that it is reasonable to assume that pre-primary education affects the covariate *Highest Parental Occupational Status* as combined information about the occupational status of both parents. For the sake of simplicity, we treat all covariates as continuous variables except for gender.

Missing data

To account for missing data we created five multiply imputed datasets using predictive mean matching (PMM; e.g. Rubin 1987). This was done separately for each country using the treatment and outcome variables and covariates outlined in the previous section. Instead of the five plausible values for reading achievement we used the first plausible value in the imputation models. Missing data range from 0.29 to 10.75 % for the measures from the student survey and from 0.94 to 43.04 % for the measures from the parental survey (see Table 2). We do not observe any missing data for the outcome and gender. We combined each of the imputed datasets with one of the five plausible values of the outcome reading achievement. The propensity score matching and the further

analyses were run on each of the five imputed datasets and the estimates were combined using Rubin's (1987) rules.

Modeling approach

The samples from the respective countries cover 51 (in Germany) to 592 (in Russia) children who did not attend preschool. Each of the students in this group was matched with a student who attended preschool for three or more years but had a similar propensity not to attend preschool using 1:1 nearest neighbor propensity score matching without replacement, i.e. we allowed each child from the control group to be matched only once. Using matching implies estimating the ATT for children who did not participate in preschool, i.e., we estimate the long-term consequences of non-participation on student achievement at the end of primary school. We used logistic regression models to estimate propensity scores for each child by regressing the binary treatment variable on all covariates. The matching was done separately for the five imputed datasets in the respective countries so that the matched samples cover between 102 (in Germany) to 1184 (in Russia) students. In each country, the matched samples were then compared for balance. For this purpose, we used the standardized bias, the ratio of the variances of the propensity scores, and the ratio of the variances of the residuals of the covariates after regressing them on the propensity scores.

In our final analysis model, we use the matched samples to regress the outcome reading achievement on the treatment variable in order to estimate the effect of not attending preschool on later student achievement. As the treatment variable is binary, the estimate is the mean difference in reading achievement between the treatment and the control group. In further analyses, we use regression adjustments to adjust for small imbalances remaining in the matched samples, i.e. we include all matching variables as covariates in the model.

We aim to estimate the PATT to generalize the findings from our study to the PIRLS target population. Thus, we use the PIRLS sampling weight HOUWGT in all regression models. Furthermore, we employ the jackknife repeated replication technique (using the variables JKZONE and JKREP) to estimate sampling errors and account for the complex stratified cluster samples in PIRLS (Lohr 2010; Rutkowski et al. 2010). All analyses are conducted in the R environment using the packages *MICE* (Buuren and Groothuis-Oudshoorn 2011) for the multiple imputation of missing data, *MatchIt* (Ho et al. 2011) for the matching, and *Survey* (Lumley 2014) for the regression analyses with jackknife replications.

Results

Student characteristics

Regarding student characteristics, there were large pre-match imbalances in all countries. The size of these differences varies internationally, and a few covariates are balanced prior to the matching. But in general children who did not attend preschool grew up in home environments where they received less support for learning to read than children who attended preschool for three or more years. Figure 1 illustrates the standardized bias prior to (dots) and after (crosses) the matching for all countries with negative values indicating disadvantages for the children who did not attend preschool. For each covariate, the standardized bias is the mean difference between the treatment and the control

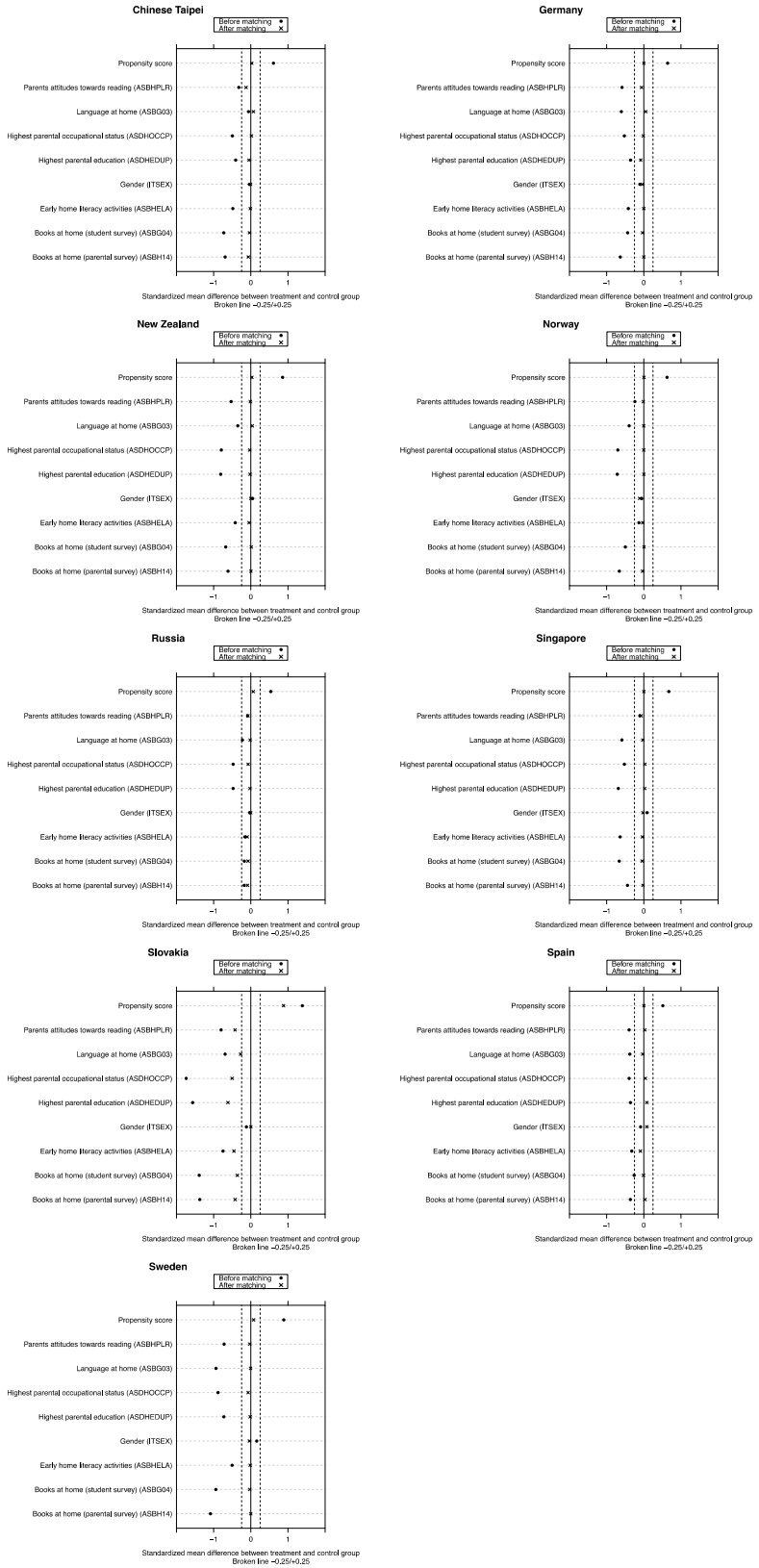

Fig. 1 Standardizes bias pre- and post-match samples

group divided by the standard deviation in the treatment group. The treatment group was used to standardize the mean differences because it is composed of the same students prior and after the matching, while the composition of the control group changes. In Sweden, for example, parents of children who did not attend preschool report more than one standard deviation fewer books at home than parents of children who attended preschool for at least 3 years before the matching. After matching, standardized mean differences for all other covariates were below 0.25 for Sweden and the other countries (i.e., within the broken lines), indicating negligible post-match bias in most countries. The only exception is Slovakia where differences are still well above this threshold for the matched sample indicating non-negligible post-match bias. In this case the matching was not successful because children who did not attend preschool form an extremely disadvantaged group in Slovakia and there were no similarly disadvantaged children who did attend preschool. The comparisons of the variances of the propensity scores and ratio of the variances for the residuals of the covariates after adjusting for the propensity scores provide further support for the comparability of the treatment and the matched control group in all countries but Slovakia. The variance ratios for the propensity score and the residuals fall into the interval from 0.5 to 2 in all countries except for Slovakia.

Outcomes

Pre-match samples

The estimates from a series of regression models are listed in Table 3. Confidence intervals that do not include zero indicate significant differences at 5 % level. In the upper part we present the result for models using the whole unmatched sample of students who did not attend preschool and all counterparts who attended preschool for 3 years or more (i.e., before the matching). The constant is the mean reading achievement of the students who attended preschool for 3 years or more, and the estimate for the treatment is the mean difference for the group of children who did not attend preschool. The group of children who attended preschool significantly outperform their peers who did not attend preschool in all countries by 19–77 score points on the PIRLS reading scale. These values are similar to those reported in the international PIRLS report (Mullis et al. 2012, p. 128). Minor differences are probably due to the multiple imputation of missing data in our study. However, the observed gap between children who did and did not attend preschool should not be interpreted in causal terms because both groups of children also differ in a set of covariates.

Post-match samples

As a result of the matching, children in the matched sample of students who attended preschool resembled those who did not attend preschool with respect to all covariates. Thus, it is possible to estimate the PATT with the matched samples while using sampling weights. This does not apply to the data from Slovakia where post-match samples showed non-negligible bias. The results from the regression analyses with the matched samples indicate that students who did not attend preschool tend to perform somewhat lower on the PIRLS reading test in comparison to their matches in all countries but Norway. However, the differences are statistically significant (at the 5 % level) in Singapore

Table 3 OLS regression on the effect of preschool attendance on student reading achievement

	Chinese Taipei	Germany	New Zealand	Norway	Russia	Singapore	Slovakia	Spain	Sweden
Original sample									
Constant	560.2	544.7	538.0	510.9	571.8	578.1	544.4	520.0	547.4
	[552.1, 568.3]	[539.5, 549.9]	[531.2, 544.7]	[498, 523.8]	[561.8, 581.8]	[571.8, 584.3]	[538.4, 550.5]	[506.8, 533.2]	[539.7, 555.1]
Treatment	−37.4	−41.7	−60.2	−21.3	−18.7	−76.8	−64.2	−30.0	−41.6
	[−64.5, −10.3]	[−71.6, −11.8]	[−88.5, −31.9]	[−48.2, 5.6]	[−29.6, −7.8]	[−103.2, −50.5]	[−84.5, −43.8]	[−54.1, −5.9]	[−104.4, 21.1]
Covariates	No	No	No	No	No	No	No	No	No
N	1707	2986	2380	2376	3740	4054	3939	6007	3599
Matched sample									
Constant	536.3	524.6	491.0	490.5	561.0	545.2	498.4	502.5	517.6
	[510.7, 561.9]	[497.7, 551.5]	[475, 507.1]	[458.2, 522.7]	[549.9, 572.1]	[519.6, 570.7]	[483.4, 513.3]	[484.4, 520.7]	[506.1, 529.2]
Treatment	−13.5	−21.6	−13.3	−0.8	−7.9	−43.9	−18.1	−12.5	−11.8
	[−48.7, 21.7]	[−57.5, 14.3]	[−37.2, 10.6]	[−36.4, 34.7]	[−22.0, 6.2]	[−78.2, −9.6]	[−39.5, 3.3]	[−35.0, 10.1]	[−33.9, 10.2]
Covariates	No	No	No	No	No	No	No	No	No
N	106	102	508	188	1184	160	494	316	398
Matched sample with covariates as controls									
Constant	518.2	421.8	363.4	409.4	448.4	479.9	495.9	384.3	430.1
	[394.1, 642.3]	[327.9, 515.8]	[297.9, 428.9]	[322.4, 496.4]	[402.0, 494.8]	[398, 561.8]	[420.8, 571.0]	[291.5, 477.0]	[353.4, 506.8]
Treatment	−13.4	−19.2	−17.1	−5.1	−3.5	−39.8	1.2	−14.5	−19.6
	[−44.1, 17.3]	[−50.6, 12.2]	[−37.0, 2.9]	[−35.5, 25.4]	[−16.0, 9.0]	[−64.2, −15.5]	[−17.2, 19.6]	[−33.4, 4.3]	[−36.6, −2.5]
Covariates	Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]
N	106	102	508	188	1184	160	494	316	398

Dependent variable is the overall reading score measured by five plausible values. Pooled results of separate regression models with the five imputed datasets. 95 % confidence intervals in brackets

[a] The eight covariates are Parents Attitudes Towards Reading, Language at Home, Highest Parental Occupational Status, Highest Parental Educational Status, Gender, Early Home Literacy Activities, Books at Home (student survey), Books at Home (parental survey)

only (see the middle part of Table 3). The results for Slovakia should not be interpreted as a causal effect because of the significant post-match imbalance.

In further analyses we included the matching variables as controls in the regression models to adjust for the remaining imbalance in the matched samples. The effect estimate for Sweden becomes significant, but otherwise the estimates in the lower part of Table 3 are quite similar to the results from the regression models without covariates. This finding supports that the observed bias was indeed negligible in the post-match samples in most countries. An exception from this general pattern can be observed for Slovakia where the estimates for the treatment effect decrease from about −18 to 1; both not statistically significant at 5 % level. This finding confirms that the post-match bias is non-negligible in Slovakia.

Mean effect estimate across countries

The fact that PIRLS uses representative random samples allows us to generalize the findings to the population of fourth graders in the respective countries. However, it is obvious that the reduced sample sizes of the matched samples are a natural limitation in terms of statistical power. To take advantage of all data available, we use meta-analytic techniques to combine the results and obtain an overall estimate. To account for the observation that some countries provide more precise effect estimates than others, we weighted each country by the inverse of the variance in their point estimate. In other words, we used the standard error (*SE*) squared to determine the weight *w* for country *j* (see Card 2012):

$$w_j = \frac{1}{SE_j^2}$$

The weighted mean effect estimate ($\overline{EE}$) is then the mean of the weighted effect estimates in the respective countries:

$$\overline{EE} = \frac{\Sigma(w_j EE_j)}{\Sigma w_j}$$

The standard error for this combined estimate is:

$$SE_{\overline{EE}} = \sqrt{\frac{1}{\sum w_j}}$$

Applied to the data from the eight countries with well-balanced covariate distributions (i.e., excluding Slovakia), we receive a weighted mean effect estimate of $\overline{EE} = -14.0$ with the $SE_{\overline{EE}} = 3.6$ (95 % CI −21.1 to −7.1). As the confidence interval does not include zero, the weighted effect estimate is significant at 5 % level. The observation of a significant mean effect estimate across countries may be due to the increased precision of the combined samples. An alternative explanation is that preschool is organized differently across countries, i.e. it is effective in some countries but not in others.

Sensitivity

While matching analyses can adjust for the observed confounding variables, unmeasured confounding is the Achilles heel of this approach. Unmeasured confounding variables result in biased effect estimates in observational studies. Sensitivity analyses can be helpful in assessing the degree of this bias. VanderWeele and Arah (2011) discuss a flexible approach for both continuous and categorical outcomes and unobserved confounders. The basic idea is to hypothesize an unmeasured covariate U. The magnitude of bias due to U depends on two sensitivity parameters: δ is the relationship between U and the outcome, and γ is the relationship between U and the treatment variable. The magnitude of bias d_{a+} is the product of the two sensitivity parameters:

$$d_{a+} = \delta\gamma$$

We can use d_{a+} to adjust the effect estimates from the matched samples. For the sake of simplicity it is useful to assume that U is uncorrelated with the other observed covariates. The standard errors and confidence intervals of the bias-corrected estimates are then precisely the same as the original estimates. This is a useful feature when we think about the sensitivity of the study. For example, the results from the regression analyses with the matched samples (with covariates as controls) suggest a significant treatment effect for Singapore and Sweden. The effect estimate for Sweden is −19.6 points with a 95 % CI from −36.6 to −2.5 (see Table 3). There is a statistically significant treatment effect at the 5 % level because the 95 % CI does not contain 0. We then use theorem $d_{a+} = \delta\gamma$ to conduct a simple sensitivity analysis: the 95 % CI would contain 0 if the observed treatment effect estimate is biased by $d_{a+} = -2.5$ points or more. We can hypothesize a dichotomous confounding variable U, say, poverty, which was not measured in PIRLS. If we assume that facing poverty results in a 40 points decrease in reading achievement,[4] a difference in poverty of $\gamma = -2.5/-40 = 0.0625$ between children who did not attend and those who did attend preschool for at least 3 years would eliminate the significant effect in Sweden. Given that we observed large differences for many covariates (see Fig. 1), a 6.25 % difference in poverty between treatment and control group is not implausible. In other words, we think that the analyses for Sweden are sensitive to unobserved covariates.

In the same vein, we conducted sensitivity analyses for Singapore and the combined effect estimator across countries. The effect estimate in Singapore is −39.8 points with a 95 % CI from −64.2 to −15.5. It would require a $\gamma = -15.5/-40 = 0.3875$ difference to eliminate the significant effect in Singapore. The combined estimate across countries is −14.0 points with a 95 % CI from −21.1 to −7.1. Here, it would require a $\gamma = -7.1/-40 = 0.1775$ difference to eliminate the significant effect. These sensitivity analyses indicate that the original results for Singapore and the combined estimate are somewhat more robust to an unobserved covariate.

The results from the sensitivity analyses depend on how we specify the sensitivity parameters. However, we observed that children who did not attend preschool were disadvantaged on basically all observed covariates. Although not entirely impossible, it seems unlikely that there are other covariates where children who did not attend

[4] The home questionnaire was extended to measure poverty among the PIRLS students in Germany. Children who grow up in poverty performed 40 points below children who did not face poverty (Bos et al. 2012).

preschool are privileged. For this reason, we find it reasonable to rule out the possibility that our research design overestimates the treatment effect. Note that there are other sensitivity-analysis techniques (see Liu et al. 2013 for an overview).

Discussion and conclusion

Our study provides only limited empirical evidence for the hypothesis that preschool attendance of children who did not participate does affect their reading achievement at the end of primary school as measured by the PIRLS 2011 data. In all countries preschool non-attendance is particularly high among children from disadvantaged backgrounds. To estimate the causal effect of preschool non-attendance on later reading achievement, we successively matched children who did and did not attend preschool in all countries but Slovakia. In six out of eight countries preschool participation is not statistically significantly associated with later student achievement. This finding is in line with some of the early childhood education effectiveness research that found effects for less disadvantaged children only (Vogel et al. 2010). As program quality was identified as a crucial requirement for realizing effects, a possible interpretation of our results is that the preschool systems established in those six countries are not designed in a beneficial way for disadvantaged children (see also Duncan and Magnusson 2013). Therefore, it might be interesting to take a closer look at Singapore and Sweden's approach as they are the only countries where preschool attendance positively impacts disadvantaged children's reading scores. However, the results for these two countries must be interpreted with caution because they are sensitive to unobserved covariates, particularly for Sweden.

The results of our study should also be interpreted in light of its limitations. Propensity score methods can be used to approximate a randomized experiment using observational data from international studies. However, a key assumption of this approach is that of strongly ignorable treatment assignment (Rosenbaum and Rubin 1983). This means that conditional on the observed covariates the treatment assignment was independent of unobserved covariates that are correlated with the outcome. Although PIRLS provides a variety of background data about students and their families, the data may not provide perfect measures of family background. The reliability of the information about early home literacy activities, for example, is limited because it was collected several years later at the end of primary school. Furthermore, the available measures may not capture all relevant facets of children's backgrounds. Matching adjusts for observed confounders but there may still be bias due to unobserved covariates. Although it is impossible to precisely foresee how such unobserved covariates may affect our estimates with the data at hand, it seems worth considering potential confounders. If children who did not attend preschool would also be disadvantaged on unobserved predictors of student achievement, we would overestimate the group differences that are already non-significant in all but two countries. Although our analyses provide more robust evidence than simple correlations, the lack of randomization limits our ability to make definite statements about causal effects.

In our study we estimated the effect of not attending preschool for children who actually did not attend preschool (PATT). These children come from disadvantaged family backgrounds, and our findings should by no means be generalized to all children. One

should also bear in mind that ineffectiveness regarding reading achievement in grade four does not mean that there might not be effects earlier or later on in children's school career (see the discussion on fading-out effects in Barnett 2011; Duncan and Magnusson 2013). Additionally, early childhood education might affect other outcomes than basic achievement and cognitive test scores.

Nonetheless, our findings are relevant for policy makers as they provide information on the effect of preschool attendance on later school achievement for disadvantaged children. Our study suggests that their low performance is not due to not attending preschool. From a policy perspective, this finding calls for alternative interventions—or a different approach to early childhood education—to support these children.

Author details
[1] Department of Education, University of Münster, Georgskommende 33, 48143 Münster, Germany. [2] Institute for School Development Research, Technische Universität Dortmund, Dortmund, Germany. [3] Centre for Educational Measurement, University of Oslo, Oslo, Norway. [4] Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden.

Received: 9 January 2016 Accepted: 20 January 2016
Published online: 03 February 2016

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083–3107. doi:10.1002/sim.3697.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. doi:10.1080/00273171.2011.568786.

Barnett, W. S. (2011). Effectiveness of early education interventions. *Science, 333*(6045), 975–978. doi:10.1126/science.1204534.

Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (Eds.). (2012). *IGLU 2011—Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich (Reading competencies of primary school children in German from an international comparative perspective)*. Münster: Waxmann.

Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

Bronfenbrenner, U. (1990). The ecology of cognitive development. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 10*(2), 101–114.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* ((Series Ed.) 6 ed., Vol. 1, pp. 793–828). New York, NY: Wiley.

Burger, K. (2010). How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. *Early Childhood Research Quarterly, 25*(2), 140–165. doi:10.1016/j.ecresq.2009.11.001.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3). Retrieved from http://www.jstatsoft.org/v45/i03.

Camilli, G., Vargas, S., Ryan, S., & Barnett, B. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3), 579–620. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=15440.

Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65*(2), 684–698. doi:10.1111/j.1467-8624.1994.tb00777.x.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.

Chambers, B., Cheung, A., Slavin, R. E., Smith, D., & Laurenzano, M. (2010). *Effective early childhood education programs: A systematic review*. Retrieved from the Best Evidence Encyclopedia website: http://www.bestevidence.org/word/early_child_ed_Sep_22_2010.pdf.

Davier, M. v., Gonzales, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In: *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 2, pp. 9-36). Hamburg/Princeton NJ: IEA-ETS Research Institute.

DuGoff, E. E., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research, 49*(1), 284–303. doi:10.1111/1475-6773.12090.

Duncan, G. J., & Magnusson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*(2), 109–132. doi:10.1257/jep.27.2.109.

Early, D. M., & Burchinal, M. R. (2001). Early childhood care: Relations with family characteristics and preferred care characteristics. *Early Childhood Research Quarterly, 16*(4), 475–497. doi:10.1016/S0885-2006(01)00120-X.

Foy, P., & Kennedy, A. M. (Eds.). (2008). *PIRLS 2006 user guide for the international database*. Chestnut Hill, MA: Boston College.
Grogan, K. E. (2012). Parents' choice of pre-kindergarten: the interaction of parent, child and contextual factors. *Early Child Development and Care, 182*(10), 1265–1287. doi:10.1080/03004430.2011.608127.
Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234–249. doi:10.1037/a0019623.
Heckman, J. J., & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 63–107). New York, NY: Springer.
Hirshberg, D., Huang, D. S.-C., & Fuller, B. (2005). Which low-income parents select child-care? Family demand and neighborhood organizations. *Children and Youth Services Review, 27*(10), 1119–1148. doi:10.1016/j.childyouth.2004.12.029.
Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*. Retrieved from http://gking.harvard.edu/matchit.
Hynes, K., & Habasevich-Brooks, T. (2008). The ups and downs of child care: Variations in child care quality and exposure across the early years. *Early Childhood Research Quaterly, 23*(4), 59–574. doi:10.1016/j.ecresq.2008.09.001.
Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statististical Association, 99*(467), 854–866. doi:10.1198/016214504000001187.
Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710. doi:10.1093/biomet/87.3.706.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*(1), 4–29. doi:10.3386/t0294.
Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences. An introduction*. New York, NY: Cambridge University Press.
Kim, J., & Fram, M. S. (2009). Profiles of choice: Parents' patterns of priority in child care decision-making. *Early Childhood Research Quarterly, 24*(1), 77–91. doi:10.1016/j.ecresq.2008.10.001.
Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). Dordrecht, The Netherlands: Springer.
Knudsen, E. I., Heckman, J. J., Cameron, J. L., & Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences, 27*, 10155–10162. doi:10.1073/pnas.0600888103.
Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science, 14*(6), 570–580. doi:10.1007/s11121-012-0339-5.
Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Brooke/Cole.
Lumley, T. (2014). Survey: analysis of complex survey samples. *R package version*, 3.30.
Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32*(19), 3388–3414. doi:10.1002/sim.5753.
Montie, J. E. (2011). Thr growth and developmet of the preprimary project. In C. Papanastasiou, T. Plomp, & E. C. Papanastasiou (Eds.), *IEA 1958–2008: 50 years experiences and memories* (Vol. 1, pp. 165–185). Nicosia: Cultural Center of the Kykkos Monastery.
Müller, N., Strietholt, R., & Hogrebe, N. (2014). Ungleiche Zugänge zum Kindergarten (Unequal access to preschool). In K. Drossel, R. Strietholt, & W. Bos (Eds.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* (pp. 33–46). Münster: Waxmann.
Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011. International results in reading.* Retrieved from http://timssandpirls.bc.edu/pirls2011/international-results-pirls.html.
Organization for Economic Development and Cooperation (OECD). (2013). *PISA 2012 Results: Excellence through equity: Giving every student the chance to succeed* (Vol. II). Paris: OECD Publishing.
Pianta, R., Barnett, W., Burchinal, M., & Thornburg, K. (2009). The effects of preschool education: What we know, how public policy is or is not aligned with the evidence base, and what we need to know. *Psychological Science in the Public Interest, 10*(2), 49–88. doi:10.1177/1529100610381908.
Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., & Downer, J., (2012). *Third grade follow-up to the Head Start impact study. Final report* (OPRE report no. 2012-45). Retrieved from the Administration for Children and Families, US Department of Health and Human Services website: http://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf.
Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. doi:10.1093/biomet/70.1.41.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. doi:10.1037/h0037350.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3–4), 169–188. doi:10.1023/A:1020363010465.
Rutkowski, L., Gonzales, E. J., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. doi:10.3102/0013189X10363170.
Schleicher, A. (2014). *Equity, excellence and inclusiveness in education: Policy lessons from around the world*. Paris: OECD Publishing.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., & Nores, M. (2005). *Lifetime effects: The high/scope perry preschool study through age 40*. Ypsilanti, MO: High/Scope Press.

Strietholt, R., Gustafsson, J.-E., Rosèn, M., & Bos, W. (2014). Outcomes and causal inference in international comparative assessments. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 10–18). Münster: Waxmann.

Stuart, E. A. (2007). Estimating causal effects using school-level datasets. *Educational Researcher, 36*(4), 187–198. doi:10.3102/0013189X07303396.

Stuart, E. A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science, 25*(1), 1-21. doi:10.1214/09-STS313.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2008). *The effective provision of pre-school education (EPPE) project: Final report from the primary phase: Pre-school, school and family influences on children's development during key stage 2 (Age 7–11).* Retrieved from the Institute of Education, University of London website: http://www.ioe.ac.uk/End_of_primary_school_phase_report.pdf.

UNESCO. (2006). *Strong foundations: Early childhood care and education*. Paris: UNESCO.

UNESCO. (2012). International standard classification of education: ISCED 2011. Retrieved from the UNESCO Institute for Statistics website: http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx.

Vandenbroeck, M., De Visscher, S., Van Nuffel, K., & Ferla, J. (2008). Mothers' search for infant child care: The dynamic relationship between availability and desirability in a continental European welfare state. *Early Childhood Research Quarterly, 23*(2), 245–258. doi:10.1016/j.ecresq.2007.09.002.

VanderWeele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology, 22*(1), 42–52. doi:10.1097/EDE.0b013e3181f74493.

Vogel, C. A., Xue, Y., Moiduddin, E. M., Carlson, B. L., & Kisker, E. E. (2010). *Early Head Start children in grade 5: Long-term follow-up of the Early Head Start Research and Evaluation Study sample* (OPRE report no. 2011-8). Retrieved from the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services website: http://www.acf.hhs.gov/programs/opre/resource/early-head-start-children-in-grade-5-long-term-followup-of-the-early-head.

Zachrisson, H. D., Janson, H., & Nærde, A. (2013). Predicting early center care utilization in a context of universal access. *Early Childhood Research Quarterly, 28*(1), 74–82. doi:10.1016/j.ecresq.2012.06.004.