Andersen, Simon Calmar; Bodilsen, Simon Tranberg; Houmark, Mikkel Aagaard; Nielsen, Helena Skyt

Working Paper

Fade-Out of Educational Interventions: Statistical and Substantive Sources

CESifo Working Paper, No. 10094

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Andersen, Simon Calmar; Bodilsen, Simon Tranberg; Houmark, Mikkel Aagaard; Nielsen, Helena Skyt (2022) : Fade-Out of Educational Interventions: Statistical and Substantive Sources, CESifo Working Paper, No. 10094, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/267326

CESifo WORKING PAPERS

10094
2022

November 2022

Fade-Out of Educational Interventions: Statistical and Substantive Sources

Simon Calmar Andersen, Simon Tranberg Bodilsen, Mikkel Aagaard Houmark, Helena Skyt Nielsen

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10094*

Fade-Out of Educational Interventions: Statistical and Substantive Sources

Abstract

What appears to be ineffectiveness of educational interventions in the long run may actually be caused by statistical artefacts in the equating of tests taken at different time points or by the nature of the skill development in the absence of targeted interventions. We use longitudinal data on the full population of public school students in Denmark to estimate central parameters in the equating of reading test scores and in a skill formation model. We compare the model's predictions to observed fade-out in a randomized controlled trial two and four years after the end of the intervention. Predicted and observed estimates consistently show that about half of the initial effect has faded out after four years. However, because of the concave nature of skill development, the treated students maintain more than 80 % of their time lead.

JEL-Codes: I210, J240.

Keywords: persistence, growth curve, time lead, statistical artefact, test equating, RCT.

Simon Calmar Andersen
Department of Political Science & TrygFondens Centre for Child Research
Aarhus University / Denmark
sca@ps.au.dk

Simon Tranberg Bodilsen
TrygFondens Centre for Child Research & Department of Economics and Business Economics, Aarhus University / Denmark
sibo@econ.au.dk

Mikkel Aagaard Houmark
TrygFondens Centre for Child Research & Department of Economics and Business Economics, Aarhus University / Denmark
mhoumark@econ.au.dk

Helena Skyt Nielsen
TrygFondens Centre for Child Research & Department of Economics and Business Economics, Aarhus University / Denmark
hnielsen@econ.au.dk

November 8, 2022
We acknowledge financial support from the Independent Research Fund Denmark (9038-00045B), Tryg-Fonden and helpful comments from Drew Bailey, Zach Bleemer, Victor Ronda, seminar participants at Aarhus University and participants at the EffEE conference 2022, the 1+1 workshop in Oslo 2022, the Copenhagen Education Network 2022, the 7th WOLFE workshop in York 2022, and the Meeting on Quantitative Educational Research in Denmark 2021. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

1 Introduction

> "*A tighter link between theory, econometric methods and data is essential to compare and reconcile the mixed and sometimes conflicting empirical results across studies, and to understand when and why the impacts of home environment and pre-school interventions fade out.*" (Duncan et al., 2023)

The long-term impacts of educational interventions are crucial for assessing the costs and benefits of different investments in human capital development. An increasing number of rigorously evaluated randomized controlled trials (RCTs) have demonstrated positive effects of early educational interventions (Alan et al., 2019; Burgess et al., 2021; Rege et al., 2021). However, few studies measure effects on skills after the end of the intervention period, and those that do indicate considerable fade-out (Bailey et al., 2020), which might suggest that early investments are ineffective in the long run. Yet, lack of skill measures that apply the same scale across multiple years of child development, lack of data on the typical development of children, and lack of combining theory and data on skill development have hindered proper assessments of fade-out (Duncan et al., 2023).

This study combines theoretical models of test equating (Kolen and Brennan, 2014; Cunha et al., 2021) and skill formation (Cunha and Heckman, 2007) with longitudinal data on more than 900,000 public school students in Denmark to estimate central parameters in the models. We compare the model's prediction to observed fade-out in a randomized controlled trial two and four years after the intervention (Andersen and Nielsen, 2016).

We distinguish three perspectives on fade-out. First, we consider the possibility that apparent fade-out is partly a statistical artefact stemming from the use of different tests with different scales for different age groups. If the skill dispersion among the tested students varies by age, the test scores standardized to have a standard deviation of one may exaggerate or underestimate the true fade-out (Cascio and Staiger, 2012; Agostinelli and Wiswall, 2023). Indeed, without data to equate different test score scales, different assumptions about how to compare the scales of two different tests may lead to conclusions ranging from complete fade-out to no fade-out or even fade-in. The study of Wan et al. (2021) reanalyze data from an RCT targeting mathematics skills. This RCT was characterized by a large degree of fade-out after a two-year period. The authors consider how sensitive this finding is to various order-preserving transformations of the data. They find that fade-out is still present under

most considered transformations, but for some specific scaling of the data they could eliminate or even reverse the fade-out pattern.

Secondly, actual fade-out may be a predictable consequence of the characteristics of the skill formation technology in the population. If an intervention works by boosting some particular skill initially, long-run treatment effects will be present to the extent that the affected skill has a causal impact on future skill development. Thus, limited self-productivity (the effect of skills in one period on the skills in the next period being less than one) may explain fade-out because it means that the comparison group catches up on the treated students, whereas high (above one) self-productivity would predict that treatment effects are increasing over time (Cunha and Heckman, 2007; Bailey et al., 2020). The self-productivity may be limited because, even if an intervention was effective in changing what Bailey et al. (2020) refer to as "trifecta skills,"—i.e. skills that are (i) malleable through intervention, (ii) fundamental for success, and (iii) would not develop eventually in the comparison group—and thereby raising treated students to higher skill growth curves, the intervention would never change pre-treatment factors (such as genes or parental education) that help students who are already on that higher growth curve maintain their advantage. Self-productivity may also be limited if investments in skill development are allocated in such a way as to reduce initial skill differences, for example if educational resources are targeted towards children who are lagging behind. Skill development may also be mediated by the home environment, and an additional source of fade-out may be present if parental investments are compensating (Nicoletti and Tonei, 2020) and if public and parental investments are substitutes (Gensowski et al., 2020).

Thirdly, and perhaps most importantly, even if the nature of the skill formation technology predicts that intervention effects will fade out when defined (as is usual) by the difference in skills between the treatment and control group *at a given point in time* (i.e., on the vertical dimension of the skill growth curve), the same skill formation technology may predict no fade-out if defined as the difference in time at which the control group will catch up to the treatment group, i.e., to *a given level of skills* (the horizontal dimension). If the skill growth curve is concave, i.e., if the marginal return to years of schooling is decreasing (as suggested by Lipsey et al., 2012), the treated students may maintain their time lead, even if the vertical skill-gap is reduced. This third distinction has received less attention in the literature on fade-out. It is important in the educational system, though, since admission to

further education is often based on certain skill levels at a given point in time, and even if students in the comparison group may be close to the required skill level—some of them may never make it in time to advance to further education.

We use data on more than 4 million test scores (from IT-based, adaptive reading tests) from 963,646 pupils, which is the full population of public school students in Denmark. We equate reading tests from grade levels 2, 4, 6, and 8, and thereby establish a common scale. We exploit that a subset of students take two different tests (e.g., the 2nd and the 4th grade test) within thirty days and that another subset of students take the same test twice in a row. This allows us to directly measure the change in dispersion across the grade levels that is due to changes in the variance of the underlying skill level as opposed to changes in test scales or in measurement error. Our equating strategy is robust to different specifications and shows that dispersion in reading skills is reduced as students become older, which actually indicates that using test scales standardized separately at different age-levels would underestimate fade-out.

To estimate the self-productivity parameter, we use data on test-retest reliability and detailed administrative data on the students' socioeconomic status to adjust the raw correlations between test scores at different grade levels. We find robust estimates across different specifications in the range between 0.81 and 0.88.[1] We compare the model's predictions to observed fade-out in, READ, an RCT of a shared book reading intervention implemented in 2nd grade (Andersen and Nielsen, 2016). The results show that the fade-out precisely followed the catch-up pattern predicted by the skill formation model two years after the end of the intervention (in 4th grade). Four years after the intervention, 57% of the treatment effect had faded out, but 77% of the fade-out could be explained by the skill-formation model. When extending our framework to include three possibly interrelated sub-dimensions of reading skills, predicted fade-out was very close to the observed fade-out four year after the intervention for both decoding and language comprehension, whereas observed fade-out in text comprehension followed the prediction after two years, but then dropped more than explained by the theoretical prediction.

Finally, we find that the average speed of skill development is 1.16 standard deviations from 2nd grade to 4th grade, but falling to 0.84 from 4th to 6th grade and only 0.61 from 6th

[1] We note that this is not far from estimates of self-productivity in Cunha and Heckman (2008) and Agostinelli and Wiswall (2023) who also find it to be <1, even though estimates are not directly comparable across these samples and model specifications.

to 8th grade corresponding to a concave skill production function. If we convert the arbitrary scale of test scores (the vertical dimension of the skill growth curve) to a scale of time lead (the horizontal dimension), which is intervals with economically relevant properties (cf. Bond and Lang, 2018), fade-out after four years is about 20% (instead of the 57% when measured using the standardized test score scale). Importantly, when using the time lead scale, the model predictions for the self-productivity of reading skills are consistent with no fade-out between 2nd and 6th grade.

Our findings support a more nuanced view on fade-out of effects of early interventions. Our results have direct implications for the literature on comparing effect sizes across interventions at different age groups (Lipsey et al., 2012; Kraft, 2020). On the one hand, we show that fade-out is a substantive phenomenon that needs to be accounted for in evaluations of interventions. If two interventions implemented at two different grade levels are equally costly and equally effective in terms of their immediate "vertical" effect, our results imply that a later reading intervention will be relatively more valuable because its long run effect at any later age will be larger.[2] For this reason, the presence of fade-out also entails that interventions should be particularly effective immediately before institutional gateways as suggested by Bailey et al. (2017).

On the other hand, if the two interventions implemented at two different grade levels are equal in terms of the time lead they generate (the horizontal dimension), conclusions are different. The fact that we find the skill growth trajectory to be concave suggests that it is generally more difficult to increase skills at later ages. In 2nd grade, students progress about half a standard deviation in one year, whereas in 6th grade, students progress only about a quarter of a standard deviation in one year. So for two interventions that both produce, say, half a year of time lead for the treatment group, the earlier intervention may be more effective in the long run (because it created a larger immediate "vertical" difference in skills). Hence, early investments may generally be more cost-effective, but only to the extent that early skills are sufficiently more malleable and effect sizes sufficiently much larger to outweigh

[2]Of course, depending on the specific intervention, a wider range of measurements may be needed to evaluate the total effect, and for some aspects the relevant metric is not necessarily the long run effect. The most obvious example is an intervention that is effective at raising well-being over some time period. To measure the value of this intervention, ideally we would want to take the integral of the effect over time rather than the effect size at any given point in time.

any subsequently diminishing returns due to fade-out. A simple comparison of immediate effects of interventions relative to their costs is too simplistic for deciding between early and late interventions.

The remainder of the paper is organized as follows: Section 2 describes the conceptual framework. Section 3 presents the institutional setting, the reading RCT, and the data used. Sections 4–7 present estimation results of key parameters in the theoretical model and of treatment effects, and distinguish the three types of fade-out. Section 8 concludes the paper.

2 Theoretical Models of Test Equating and Skill Formation

In this section, we provide a theoretical framework that allows us to study fade-out of educational interventions from three distinct perspectives: Fade-out as a statistical artefact, fade-out caused by the skill production function measured in the vertical dimension, and the same fade-out translated to the horizontal dimension according to the curvature of the skill growth curve.

2.1 Statistical Sources of Fade-Out

One theory in the context of fade-out of educational interventions is that fade-out is a *statistical artefact*. In assessments of interventions it is customary to compare test scores taken at different points in time for an intervention and a control group. As the content and the level of difficulty naturally vary as the participants progress through school, the test scores observed at different stages are transformed such that they share similar statistical properties. Most commonly, test scores at a certain grade are standardized to have a mean of zero and a variance of one, which makes it easy to express effect sizes of interventions in terms of standard deviations. However, if the true skill dispersion among children is increasing over time, an effect of one standard deviation at a lower grade level will translate into a smaller estimated effect in higher grades, even if the true effect on skills was maintained. Thus, in this scenario the typical pattern of diminishing intervention effects over time is partly a statistical artefact.

To formalize this idea, suppose that Z_t represents an observed test score in time period t for a child (to simplify the notation we omit a subscript i to index children). The raw test score is assumed to be an imperfect measure of skills. Following Agostinelli and Wiswall (2023)

we assume that the observed test score provides a noisy signal of the child's stock of skills, θ_t, by specifying that it is a linear function of skills contaminated by an error component, ϵ_t, at the population level:

$$Z_t = \mu_t + \lambda_t \theta_t + \epsilon_t, \qquad t \geq 1. \tag{1}$$

Here μ_t and λ_t are location and scale parameters relating to the specific measure. The error ϵ_t can be interpreted as a measurement error. We assume it is drawn independently of θ_t and that it has a mean of zero and a variance of σ_t^2. A central first assumption is the following:

Assumption 1. *The skill level, θ_t, has a strictly positive impact on test scores, Z_t, in any time period t. That is, $\lambda_t > 0$ for all $t \geq 1$.*

In other words, the test is assumed to be a valid proxy of the underlying skill in the sense that, on average, it assigns a higher test score to an individual with a higher skill level relative to an individual with a lower stock of skills.

In general, for some change in average test scores, $\mathrm{E}[Z_{t+k}] - \mathrm{E}[Z_t]$, we cannot identify whether this change is due to a change in skills or a change in measurements:

$$\mathrm{E}[Z_{t+k}] - \mathrm{E}[Z_t] = (\mu_{t+k} - \mu_t) + \lambda_{t+k}\,\mathrm{E}[\theta_{t+k}] - \lambda_t\,\mathrm{E}[\theta_t].$$

However, for the purpose of evaluating educational interventions, we are mainly interested in the difference of test scores between the treatment and the control group at a specific point in time. To this end, let

$$\Delta_{t+k|t} \equiv \mathrm{E}[Z_{t+k} \mid D_t = 1] - \mathrm{E}[Z_{t+k} \mid D_t = 0],$$

denote the treatment effect on test scores $k \geq 0$ time periods after an intervention carried out a time t between the treatment group ($D_t = 1$) and the control group ($D_t = 0$). We assume the following regarding the impact of an intervention:

Assumption 2. *The only component in (1) affected by an intervention is the skill level, θ_t. That is, $\mathrm{E}[Z_{t+k} \mid D_t = d] = \mu_{t+k} + \lambda_{t+k}\,\mathrm{E}[\theta_{t+k} \mid D_t = d]$ for all $t, k \geq 0$ and $d \in \{0, 1\}$.*

Under Assumption 2, we have that the test score treatment effect reduces to $\Delta_{t+k|t} = \lambda_{t+k}\tau_{t+k|t}$, where we define the treatment effect on actual skills as $\tau_{t+k|t} \equiv \mathrm{E}[\theta_{t+k} \mid D_t =$

$1] - \mathrm{E}[\theta_{t+k} \mid D_t = 0]$. Hence, unless the scale parameters are known, the former expression for $\Delta_{t+k|t}$ shows that we cannot infer the treatment effect on skills based on the treatment effect on test scores for a given intervention.

As mentioned, in evaluations of educational interventions it is customary to standardize test scores to have an unconditional mean of zero and a variance of one, such that effect sizes can be expressed on an easily interpretable scale in terms of standard deviations. The standardized test score in time period t is defined as

$$Y_t = \frac{Z_t - \mathrm{E}[Z_t]}{\sqrt{\mathrm{var}(Z_t)}}.$$

Using the standardized test score, the treatment effect of an intervention can be expressed as

$$\Delta^Y_{t+k|t} \equiv \mathrm{E}[Y_{t+k} \mid D_t = 1] - \mathrm{E}[Y_{t+k} \mid D_t = 0] = \frac{\lambda_{t+k}\tau_{t+k|t}}{\sqrt{\mathrm{var}(Z_{t+k})}} = \frac{\lambda_{t+k}\tau_{t+k|t}}{\sqrt{\lambda^2_{t+k}\,\mathrm{var}(\theta_{t+k}) + \sigma^2_{t+k}}}.$$

If $\Delta^Y_{t+k|t}/\Delta^Y_{t|t} < 1$ for all $k \geq 1$, the effect on test scores fades out over time, since this corresponds to a diminishing effect size k periods after the end of the intervention. The degree of fade-out on standardized test scores is given by

$$\begin{aligned} \frac{\Delta^Y_{t+k|t}}{\Delta^Y_{t|t}} &= \frac{\tau_{t+k|t}}{\tau_{t|t}} \times \frac{\lambda_{t+k}}{\lambda_t} \sqrt{\frac{\lambda^2_t\,\mathrm{var}\,(\theta_t) + \sigma^2_t}{\lambda^2_{t+k}\,\mathrm{var}\,(\theta_{t+k}) + \sigma^2_{t+k}}} \\ &= \text{“Skill Fade-Out”} \times \text{“Statistical Artefact”}. \end{aligned} \tag{2}$$

This shows that, unless we are willing to make the assumption that $\frac{\lambda_t}{\sqrt{\mathrm{var}(Z_t)}} = \frac{\lambda_{t+k}}{\sqrt{\mathrm{var}(Z_{t+k})}}$ for every $k \geq 1$, the degree of fade-out measured in terms of standardized test scores is not identical to the degree of fade-out measured in terms of skills. Thus, for a hypothetical intervention in which the effect on skills fully persists after k periods (i.e., $\tau_{t|t} = \tau_{t+k|t}$), it is still possible to find fade-out in test scores if e.g., the variance of skills increases over time. As the equation also shows, fade-out in test scores can also occur due to an increase in the variance of the measurement error or due to changes in the scale parameter over time. In these cases, fade-out is not substantive but is a statistical artefact. In summary, standardization does not solve the fundamental problem of identifying the treatment effects on skills based on observable test scores.

2.1.1 Identification of the Statistical Artefact

In this subsection, we introduce our identification strategy that allows us to quantify the statistical artefact component of fade-out in (2).

The main difficulty stems from the fact that skills have no natural scale and location. This means that μ_t and λ_t have no absolute values. They are only informative relative to some other skill measure. As suggested by Agostinelli and Wiswall (2023), a possible solution to this problem is to normalize skills relative to the the skill measure available in the first time period:

$$\mathrm{E}[\theta_1] = 0 \quad \text{and} \quad \lambda_1 = 1.$$

This normalization is convenient, as it implies that $\mu_1 = \mathrm{E}[Z_1]$ and that for an intervention in the first time period the contemporaneous treatment effect measured in terms of the raw test score is equal to the treatment effect on skills, i.e., $\Delta_{1|1} = \tau_{1|1}$.

To identify the location and scale parameters in the remaining time periods, we consider an individual who, in the same time period, takes two tests of different difficulty measuring the same underlying skill. Assuming two tests of difficulty levels corresponding to (the individual's grade level at) time period t and $t+1$, respectively, are taken at some time $s \in [t, t+1]$, we then have that the outcomes of the tests are determined by the following set of equations:

$$\begin{aligned} Z_{t,s} &= \mu_t + \lambda_t \theta_s + \epsilon_{t,s} \\ Z_{t+1,s} &= \mu_{t+1} + \lambda_{t+1} \theta_s + \epsilon_{t+1,s}, \end{aligned}$$

where $Z_{t,s}$ denotes a test score of difficulty level t taken at time s. In the case where $s = t$, we simplify notation by writing $Z_{t,t} = Z_t$ and $\epsilon_{t,t} = \epsilon_t$ as in (1).

Since, the outcomes of the two arbitrarily scaled tests both depend on the skill level at time s, we can equate the two equations in terms of the skill level to obtain that

$$\theta_s = \frac{Z_{t,s} - \mu_t - \epsilon_{t,s}}{\lambda_t} = \frac{Z_{t+1,s} - \mu_{t+1} - \epsilon_{t+1,s}}{\lambda_{t+1}},$$

from which the mean and the variance can be equated as follows:

$$\mu_{t+1} = \frac{\lambda_{t+1}}{\lambda_t}(\mu_t - \mathrm{E}[Z_{t,s}]) + \mathrm{E}[Z_{t+1,s}] \tag{3}$$

$$\frac{\lambda_{t+1}}{\lambda_t} = \sqrt{\frac{\mathrm{var}(Z_{t+1,s}) - \mathrm{var}(\epsilon_{t+1,s})}{\mathrm{var}(Z_{t,s}) - \mathrm{var}(\epsilon_{t,s})}}. \tag{4}$$

In principle, if we have knowledge about the variance of the error component over time, we could, based on (3) and (4), estimate the location and scale parameters in all time periods by substituting the theoretical means and variances of the measurements by their sample analogs using observed test scores in combination with the normalization that $\lambda_1 = 1$ and $\mu_1 = \mathrm{E}[Z_1]$.

The final ingredient in our identification scheme comes from exploiting a group of individuals that take two (different) tests of the same difficulty at the same point in time. Importantly, the two tests are assumed to be measuring the same underlying skill attribute. Let $Z_{t,s}^{(1)}$ and $Z_{t,s}^{(2)}$ represent two such measures.[3] A general specification for the outcomes of the two tests are given by the following two equations:

$$\begin{aligned} Z_{t,s}^{(1)} &= \mu_t^{(1)} + \lambda_t^{(1)}\theta_s + \epsilon_{t,s}^{(1)} \\ Z_{t,s}^{(2)} &= \mu_t^{(2)} + \lambda_t^{(2)}\theta_s + \epsilon_{t,s}^{(2)}. \end{aligned} \tag{5}$$

To achieve identification of the scale and location parameters, we need to impose some structure on the relationship between the two tests in (5), which we collect in Assumption 3 below:

Assumption 3. *For all $t, s \geq 1$, the measurement errors $\epsilon_{t,s}^{(1)}$ and $\epsilon_{t,s}^{(2)}$ satisfy $\mathrm{E}[\epsilon_{t,s}^{(j)}] = 0$ and $\mathrm{var}(\epsilon_{t,s}^{(j)}) = \sigma_{t,s}^2$ for $j = 1, 2$, and are statistically independent of each other and with θ_t. Furthermore, we impose equality of the two scale parameters, i.e., $\lambda_t^{(1)} = \lambda_t^{(2)} = \lambda_t$.*

Under Assumption 3 we only allow for systematic differences in the two test scores due to differences between the location parameters. Another consequence of Assumption 3, is that the variance of the two tests coincide meaning that $\mathrm{var}(Z_{t,s}^{(1)}) = \mathrm{var}(Z_{t,s}^{(2)}) = \mathrm{var}(Z_{t,s})$.

We can exploit the existence of two such tests satisfying Assumption 3, by considering the correlation between the two measures, which allows for evaluation of the so-called *test-retest*

[3] In theory, the two tests are taken at the exact same point in time, and the skill level is exactly the same and only the realization of the test measure differs. Empirically, one test must come before the other, and therefore, one may worry that skills are influenced by taking the test. We come back to this in the empirical part and provide multiple robustness checks of the identifying assumptions.

reliability ratio defined by $r_{t,s} \equiv \text{corr}(Z_{t,s}^{(1)}, Z_{t,s}^{(2)})$. Under our modeling assumptions we can express this quantity as

$$r_{t,s} = \frac{\text{var}(Z_{t,s}) - \sigma_{t,s}^2}{\text{var}(Z_{t,s})}. \tag{6}$$

From (6) it follows that (4) equivalently can be expressed as

$$\frac{\lambda_{t+1}}{\lambda_t} = \sqrt{\frac{\text{var}(Z_{t+1,s}) r_{t+1,s}}{\text{var}(Z_{t,s}) r_{t,s}}}. \tag{7}$$

The right-hand side on this equation does only depend upon quantities which can be consistently estimated in the presence of observable data that satisfies the modeling assumptions, by using the sample analogs of the theoretical variances and correlations. Hence, starting from $t = 1$, we can iteratively estimate all the subsequent location and scale parameters using (3) and (7) together with the initial conditions $\lambda_1 = 1$ and $\mu_1 = \text{E}[Z_1]$.

Given we have achieved identification of the location and scale parameters, this also allows us to characterize the first and the second moment of the skill distribution, and hence facilitates an analysis of how skills evolve over time. In particular, we have the following expressions for the expected value and the variance of skills as a function of time:

$$\text{E}[\theta_t] = \frac{\text{E}[Z_t] - \mu_t}{\lambda_t} \quad \text{and} \quad \text{var}(\theta_t) = \frac{\text{var}(Z_t) r_t}{\lambda_t^2}, \tag{8}$$

where $r_t = r_{t,t}$ is the test-retest reliability ratio for two tests taken in the same time period of the same difficulty.

2.2 Skill Formation Technology and Fade-Out in Skills

In the following, we provide a model framework to describe how reading skills evolve over time. Our model of skill formation is closely related to the seminal work by Cunha and Heckman (2007), but specifically tailored to our setting.

Let θ_t denote the true level of reading skills of a child. While the child may possess many different skills, several of which may be related to reading ability, we start by assuming a single skill dimension. We relax this assumption later.

A child's production of reading skills at time $t + 1$ is determined by a time-varying production function, f_t, that takes as inputs the stock of reading skills from the previous time

period, θ_t, investments in reading skills, I_t, a vector of background characteristics, X_t, and a stochastic zero mean production shock, η_t:

$$\theta_{t+1} = f_t(\theta_t, I_t, X_t, \eta_t). \tag{9}$$

Investments are distinguished from the other characteristics in that investments can respond to the child's stock of skills, $I_t = I_t(\theta_t)$. This includes parental investments, which may be compensating or reinforcing, and school investments, which similarly may be tailored more towards the needs of the high or low ability pupils. On the other hand, X_t captures things like genes or the quality of the home environment at birth (typically proxied by socio-economic status) that do not respond directly to the child's skills. Thus, the setup is very general because vectors I_t and X_t together include everything that may affect skill formation. It simply focuses on the determinants of a particular skill, i.e., reading ability.

Different specifications of the production function in (9) have been proposed in extant literature. Cunha and Heckman (2008) consider a skill formation model with a linear technology, Cunha et al. (2010) use a constant elasticity of substitution production function, and the more recent studies of Del Bono et al. (2022) and Agostinelli and Wiswall (2023) employ a parametric translog technology. To capture all the channels that are often thought to be relevant in our setting, we assume the following model which allows for an interaction between skills and investments:

$$\theta_{t+1} = \gamma_{0,t} + \gamma_{1,t}\theta_t + \gamma_{2,t}I_t + \gamma_{3,t}\theta_t I_t + \pi_t' X_t + \eta_t, \tag{10}$$

where the production shock, η_t, is assumed to be independent of both θ_t, I_t, and X_t in any time period.

Within this framework, we can understand fade-out as the extent to which an exogenous increase in skills in one period does not persist into the next period, i.e., fade-out is given by:

$$1 - \frac{\tau_{t+1|t}}{\tau_{t|t}}.$$

Assumption 2 implies that the only variable that is directly affected by the treatment at time t is θ_t. And since treatment is randomly assigned, anything else that is affected by the

treatment is affected through the initial effect on θ_t. Thus, one can show that (for details, see Appendix A)

$$\tau_{t+1|t} = \tau_{t|t} \, \mathrm{E}\left[\gamma_{1,t} + \gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\Big(I_t + \frac{\partial I_t}{\partial \theta_t}\theta_t\Big)\right],$$

which means that the treatment effect develops as a function of the initial effect and the effect of skills on later skills, i.e.,

$$\tau_{t+1|t} = \tau_{t|t} \, \mathrm{E}\left[\frac{\partial \theta_{t+1}}{\partial \theta_t}\right].$$

Thus, we can also express the degree of fade-out as

$$1 - \frac{\tau_{t+1|t}}{\tau_{t|t}} = 1 - \mathrm{E}\left[\frac{\partial \theta_{t+1}}{\partial \theta_t}\right] \equiv 1 - \delta_t,$$

where we use δ_t to denote the *self-productivity* of skills in period t, which in this model is given by

$$\delta_t = \mathrm{E}\left[\gamma_{1,t} + \gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\Big(I_t + \frac{\partial I_t}{\partial \theta_t}\theta_t\Big)\right].$$

An equivalent expression for the self-productivity (using that $\frac{\partial \theta_{t+1}}{\partial I_t} = \gamma_{2,t} + \gamma_{3,t}\theta_t$) is

$$\delta_t = \mathrm{E}\left[\gamma_{1,t} + \frac{\partial I_t}{\partial \theta_t}\frac{\partial \theta_{t+1}}{\partial I_t} + \gamma_{3,t} I_t\right].$$

We see that the self-productivity has three components: (1) The direct skill persistence, (2) the effect through investments, and (3) the dynamic complementarity or substitutability. Hence, the total degree of fade-out is determined by the combination of these three mechanisms:

1. *The direct skill persistence*: Fade-out may be caused by new skills not building directly on earlier skills, or simply by forgetting. If so, skills are not perfectly persistent (i.e., $\gamma_{1,t} < 1$). This may also be a result of other factors in the production function. For example, if children with a genetic advantage are learning at a faster rate conditional on their current skill level (Houmark et al., 2020), it may appear that skills are highly persistent when this is actually in part due to earlier and later skills being produced by a common factor.

2. *Compensating or reinforcing investments*: Investments may respond to the stock of skills, and if they do so negatively ($\mathrm{E}\left[\frac{\partial I_t}{\partial \theta_t}\right] < 0$), investments are said to be compensating, which will cause fade-out. For example, if parental investments are compensating, it means that if an intervention increases a child's reading skills, parents will respond by lowering their investments in promoting reading skills further (e.g., by shifting to investments in other skills or to investments in siblings' skills). Something similar may happen in schools if teachers tend to focus their efforts on the least skilled children. In the absence of continuing investments in the treatment group, this will lead to a "catch-up" effect by the control group and thereby fade-out of the intervention (Cunha and Heckman, 2007; Bailey et al., 2020). However, the opposite is also possible, in which case the effect of an intervention will be reinforced, possibly even leading to negative fade-out.

3. *Dynamic complementarity or substitutability*: If an increase in skills increases the returns to subsequent investments, this is known as dynamic complementarity ($\gamma_{3,t} > 0$). In this case, the skills learned in one period makes it easier for individuals to obtain new skills in later periods. Conversely, dynamic substitutability entails that investments (e.g., schooling), are most productive for the low-skilled children. If so, an intervention may make some children more skilled at reading, but these children would then learn relatively less in school afterwards. Thus, dynamic substitutability leads to fade-out. Cunha and Heckman (2007) assume dynamic complementarity, whereas Agostinelli and Wiswall (2023) find that skills and investments are substitutable and Rossin-Slater and Wust (2020) find negative interaction effects between preschool and nurse home visiting exposure, which also suggests some substitutability in investments during early childhood.

We can also consider the intervention itself through this framework. Namely, an intervention can be viewed as a particular investment. In this case, an RCT gives identification of both $\gamma_{2,t}$ and $\gamma_{3,t}$ for that particular investment. For example, say that an intervention simply consists of extra teaching, and we find that this is most effective for the low-skilled children (i.e., $\gamma_{3,t} < 0$). This would then also imply that subsequent teaching would have the highest returns for the low-skilled (if we assume that the extra teaching is equivalent to regular teaching). This,

in turn, would entail that part of the treatment effect would fade out over time. Of course, interventions are typically not just an increase in regular teaching, but to the extent that it is, it may be informative about the degree of dynamic complementarity or substitutability more generally.

In practice, estimation of the model in (10) is not straightforward, since reading skills and investments are inherently latent. However, we do propose an estimation scheme in Section 5 that allows us to recover an upper bound on the self-productivity parameter (δ_t). In other words, this allows us to quantify a *lower bound* for the degree of fade-out that should be expected.

2.3 The Skill Growth Curve and Fade-Out in Time Lead

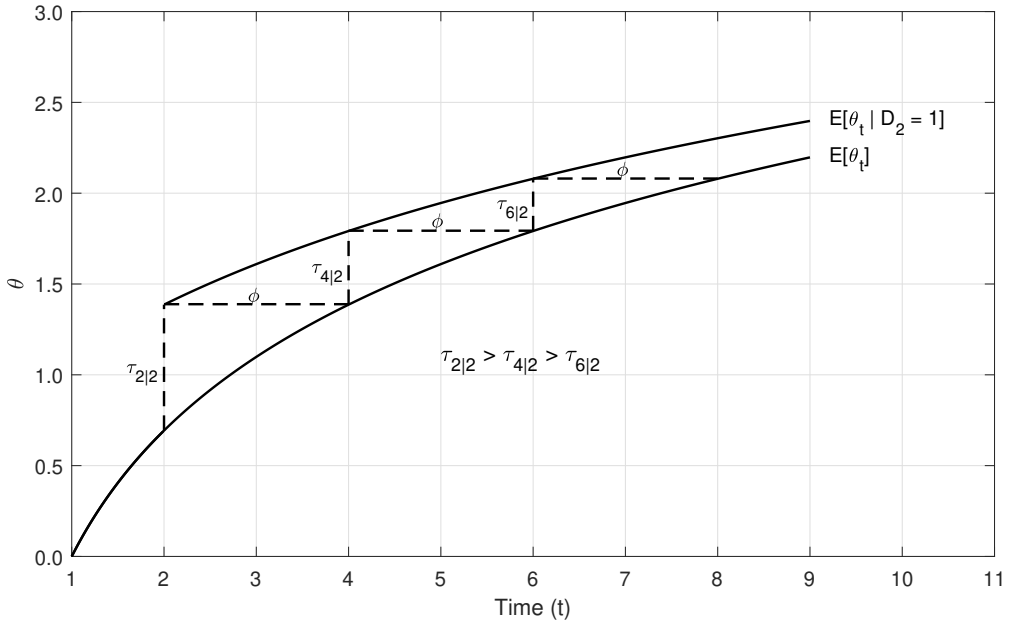
The skill formation technology described in the previous section gives rise to skill growth curves depicting the typical development in skills across time. Studies on fade-out usually focus on the vertical difference in skills at a given point in time. However, the same skill formation technology may also be analyzed in terms of the horizontal dimension of the skill growth curve, i.e., the difference in time it takes to reach a given skill level. In this view, fade-out occurs if the effect sizes from follow-up tests correspond to a reduced time lead in skills compared to the initial assessment of the intervention.

To facilitate such analysis, it is a prerequisite that we can characterize certain features of the distribution of skills over time. For example, assuming the availability of the expected value function $\mathrm{E}[\theta_t]$ (as identified in (8)), it is possible to map an effect size measured at a given point in time into an estimate of a time lead in average skills. More formally, assume an intervention is carried out at time t in which a treatment group is given a certain treatment, whereas the control group does not benefit from this treatment. Moreover, we will assume for simplicity that the control group is large, such that it holds that $\mathrm{E}[\theta_{t+k} \mid D_t = 0] = \mathrm{E}[\theta_{t+k}]$ for all $k \geq 0$.

Suppose that based on follow-up test at times $t, t+1, \ldots, T$, we know the sequence of effect sizes on skills $\{\tau_{t+k|t}\}_{k=0}^{T-t}$. For each follow-up measurement, we can then investigate what time lead in skills a given effect size corresponds to, by finding the value of $\phi_{t+k|t}$ such that

$$\mathrm{E}[\theta_{t+k}] + \tau_{t+k|t} = \mathrm{E}[\theta_{t+k+\phi_{t+k|t}}], \qquad k = 0, 1, \ldots, T-t. \tag{11}$$

Figure 1: Intervention with constant treatment effect measured in time



Notes: This figure illustrates a hypothetical intervention in time period 2, in which the effect on skills diminishes over time, but where the time lead in skills for the treatment group in comparison to the control group remains constant over time. In this example, it is assumed that $\mathrm{E}[\theta_t \mid D_2 = 0] = \mathrm{E}[\theta_t] = \log(t)$ and $\mathrm{E}[\theta_t \mid D_2 = 1] = \log(t + \phi)$ for $t \geq 2$ with $\phi = 2$.

Fade-out in time lead is then defined as a situation in which $\{\phi_{t+k|t}\}_{k=0}^{T-t}$ is a decreasing sequence.

Depending on the curvature of the expected value function $\mathrm{E}[\theta_t]$ it is possible to have a situation with fade-out following an intervention in terms of vertical skill differences, but where no or even reverse fade-out occurs in terms of time lead in average skills for the treatment group (horizontal difference). If the skill curve is strictly concave (convex) then holding the time lead in skills constant (i.e., $\phi_{t+k|t} = \phi$) does necessarily imply that the vertical treatment effect on skills must be decreasing (increasing) over time. Conversely, if we hold the treatment effect on skills constant (i.e., $\tau_{t+k|t} = \tau$) the time leads will increase (decrease) over time when the skill curve is strictly concave (convex). The former is illustrated in Figure 1 in a hypothetical setting in which the average skill curve is an increasing strictly concave function of time. Before the intervention at time 2 the treatment and control groups have the same average skill level. After the intervention the treatment group maintains a constant time lead of ϕ in the expected skill level. Due to the concavity of $\mathrm{E}[\theta_t]$, however, this implies that the

effect size in terms of the skill level decreases over time, leading to a conclusion of fade-out in the usual sense.[4]

3 Institutional Setting and Data

To estimate the theoretical parameters in the models of test equating and skill formation, we use administrative register data from all public school students in Denmark. We compare the model's predictions to observed effects from an RCT. Below we describe the Danish education and test system before we describe the data and samples that we use.

3.1 Compulsory Education and Testing in Denmark

For the children in our study, education was compulsory from the calendar year in which they turned six until completing 9^{th} grade. School starts with a one-year kindergarten class and ends with a compulsory school exit exam (around age 16). The exit exam must be passed with grades beyond a certain threshold in order to enroll in upper-secondary education.

Public schools are free of charge and cater for 85% of all students. Testing and grading is not key in the public school, and systematic teacher assessed grades are only used from 8^{th} grade as a preparation for the exit exam. However, nationwide testing was introduced in public schools starting from the school year 2009/10. The emphasis is on reading (four tests) and math (two tests), and the main purpose of the tests is to give feedback to teachers, students and parents about the ability level of the individual child.

The national tests are computerized and adaptive. The tests are adaptive in the sense that the system draws items from an item bank based on the estimated skill level of the student. The tests have a run-in period of three items, with average difficulty level, after which point the estimated skill level is based on the student's responses to all previous items, as well as the difficulty level of the items. The system continuously estimates the uncertainty (standard error of measurement; SEM) of the skill estimate. We use both the final test score (the estimated skill level of the student using a Rasch-scale), and the estimated SEM. For our analysis, we standardize the final test score using the means and the standard deviations from the full set

[4] As a consequence, a moderate fade-out as measured in skills may correspond to a negative fade-out (i.e., fade-in) in the time dimension.

of reading test scores at a certain grade level.

The reading test consists of three skill domains—language comprehension, decoding and text comprehension—each with its own item bank. The system switches between items from the three domains so the three reading domains are tested simultaneously. The system is, of course, blinded to the treatment status of the students.

The national reading tests are mandatory for all public school students in the spring in grades 2, 4, 6, and 8. Teachers can also sign up students to take additional (optional) tests each fall. These optional tests can be taken one grade level before or after the mandatory test, such that, for instance, the 4^{th} grade test can be taken in the fall in grade 3, 4, and 5 (but only twice).

The reading tests at all the grade levels (2–8) test all three domains. However, the tests at each grade level uses different item banks, and therefore, the test score at one grade level is not directly comparable to the test score at another grade level.

3.2 Data Samples

Equating Sample To equate the test scores from different grade levels, we exploit that a subsample of students within a short time span have taken tests from two different grade levels (e.g., some students will during 3^{rd} grade have taken both the 2^{nd} grade test and the 4^{th} grade test). We denote this sample the "Equating Sample". We restrict the sample to students who took two tests of different grade (difficulty) levels no more than 30 days apart. This leaves 1,518 students to link grade 2/4, 2,140 pupils to link grade 4/6 and 1,800 students to link grade 6/8.

Reliability Sample To estimate the test-retest reliability, we use a "Reliability Sample" of students who have taken the same test twice within a short period. We restrict the sample to individuals who took the same test twice no more than five days apart. This leaves 1,943 students in grade 2, 1,354 students in grade 4, 1,594 students in grade 6 and 1,180 students in grade 8.

Gross Sample To estimate the self-productivity parameter in the skill formation model we use the "Gross Sample". The Gross Sample consists of all individuals who complete a reading

test (compulsory or optional) between 2009/10 and 2018/19. In total, we observe 4,064,380 test scores from 963,646 pupils. 2,112,500 (52.0%) of the tests are compulsory, and they are divided between the different grade (difficulty) levels as follows: 1,060,800 (26.1%) of the 2^{nd} grade test, 1,078,930 (26.5%) of the 4^{th} grade test, 1,067,946 (26.3%) of the 6^{th} grade test and 856,704 (21.1%) of the 8^{th} grade test. The lower number of 8^{th} grade tests is mainly because this test is used less for optional testing.

READ Sample Finally, we compare the model predictions of fade-out with observed fade-out of an RCT, which tested the effect of the READ intervention (Andersen and Nielsen, 2016). We denote this the "READ Sample". The READ sample consists of all students who were assigned to either the treatment or the control group in the READ trial, which we describe in more detail below.

3.3 Socio-demographic Data on Students and Their Parents

For all four samples we are able to connect the data on test scores to socio-demographic data on the students and their parents. All citizens in Denmark have a personal ID-number, which is used by all public authorities. We use this number to merge information from different administrative registers including data on their reading test scores, age, gender, and immigrant status as well as information on their parents' education and income. The personal ID-numbers enable us to obtain data on the students many years after the intervention ended.

In subgroup analyses examining potential heterogeneous fade-out, we categorize students as high/low socio-economic status (SES) according to their parents' income and educational level. We do so by first ranking all parents (separately for mothers and fathers) according to their annual income and years of completed education. We then take the average of the child's maternal and paternal rank and rank again to obtain an overall parental education and income rank. Finally, the SES measure is again an average rank of the education and income ranks, and high (low) SES is defined as a rank above (below) the median. Unreported robustness analyses show that we obtain similar results using alternative definitions, e.g. using only parental education or using only maternal SES.

Table 1 shows descriptive statistics on the students and parents in the four samples used for the analyses. The table shows that both the Equating Sample and the Reliability Sample

Table 1: Descriptive statistics

	Gross Sample	Equating Sample	Reliability Sample	READ Sample
Child characteristics				
Female	0.491	0.501	0.482	0.491
	(0.500)	(0.500)	(0.500)	(0.500)
Immigrant	0.103	0.170	0.097	0.207
	(0.304)	(0.375)	(0.295)	(0.405)
Family characteristics				
Years of education (mother)	13.336	13.328	13.129	13.671
	(2.785)	(2.980)	(2.728)	(3.344)
Years of education (father)	12.886	12.973	12.745	13.682
	(2.683)	(2.892)	(2.620)	(3.116)
Yearly income (mother)	292,166	293,043	281,387	278,299
	(215,462)	(256,776)	(200,988)	(229,788)
Yearly income (father)	419,057	418,439	403,681	409,961
	(513,588)	(522,795)	(359,896)	(394,696)
Individuals	963,646	5,458	6,047	1,572
Observations	4,064,380	10,916	12,094	4,414

Notes: This table reports empirical means and standard deviations (in parentheses) of a set of background variables for the different samples used in the analysis.

are very similar to the Gross Sample, although relatively more immigrants are included in the Equating Sample (17% as opposed to 10%). This is even more so for the READ sample (21%), because the intervention took place in the municipality of Aarhus, the second-largest city in Denmark. But on average the READ sample is otherwise very similar socioeconomically to both the Gross Sample and the Equating Sample.

3.4 READ: A Shared Book Reading Intervention

To compare the theoretical model prediction to observed fade-out of an intervention, we use a shared book reading program called READ. The program is well-suited for this purpose for two reasons. First, the intervention was directly targeted at reading skills, which makes assumption 2 plausible. Second, previous research has shown that the intervention effect was positive and suggested lasting effects on reading skills.

READ took place in the school year 2013/14 and it was implemented in 2nd grade. Families received books and instructions on how the parents could support their children's reading skill development by asking open-ended questions about the books that the children read aloud. The program applied a growth mindset (Dweck, 2006) approach by explaining to parents that

reading abilities are malleable, and that parents should reward their child's effort rather than performance. Details about the program, the randomized controlled trial used to evaluate it, and the short-term effects are reported by Andersen and Nielsen (2016), so here we just summarize the main features of the trial.

The study was a cluster-randomized trial run in collaboration with the local government in the municipality of Aarhus. 28 schools signed up 74 classrooms. Classrooms were stratified in groups of four based on language proficiency of the children. Within each stratum, two classrooms were randomly assigned to treatment and two to business-as-usual control group condition. Figure B.1 in the Appendix illustrates the design of the trial and the flow of the 1,587 students who were enrolled.[5] The duration of the program was 16 weeks, and the average costs per child approximately DKK 500 (USD 76).

3.4.1 Empirical Strategy for Evaluating the Effect of READ

We use the following regression model to test the effect of the intervention at various time points after the beginning of the intervention:

$$Y_{i,j,t} = \alpha_t + \beta_t D_i + s_{j,t} + \epsilon_{i,j,t}, \tag{12}$$

where $Y_{i,j,t}$ denotes the standardized reading test score for child i in stratum j at time t. The variable D_i is an indicator variable that equals one if student i is in the treatment group and zero otherwise, and $\epsilon_{i,j,t}$ is a zero mean error term. The parameter β_t represents the intention to treat effect in period t and is the parameter of interest. Finally, to increase precision of the estimates, we include indicators of 19 strata used for randomization captured by $s_{j,t}$. In an unreported robustness analysis we also include a vector of student, parent and school covariates. The pattern of fade-out, however, is similar with and without these covariates included.

[5]The present analysis data set includes all students for whom we observe at least one test score, which is a few more students compared to the previous analyses, where students with some missing covariates were excluded. Furthermore, we use outcomes 7 months after the trial onwards in order to make use of the mandatory national tests, which are available for the entire population.

4 Fade-Out as a Statistical Artefact

As shown in Section 2.1, an intervention causing a skill increase of some percentage of a standard deviation initially will be observed as a smaller percentage increase if measured by a later test score, if the true reading skill variance is increasing over time. This is so even if the skill gap remains the same in an absolute sense. To account for the potential change in skill dispersion over time, we use the Equating Sample and the Reliability Sample to equate reading tests designed for grade 2, 4, 6, and 8.

Ideally, to follow the theoretical framework outlined in Section 2.1.1, we would want students to take each pair of tests *at the same time*. Of course, our empirical application can only approximate this, as we need to allow for some time to pass between the two tests in order to observe any students. Deciding on this threshold entails a bias/variance trade-off. We discuss this in more detail in Appendix C, where we also perform a series of robustness checks and sensitivity analyses. In our preferred specification, we allow up to 30 days between tests for the Equating Sample and up to five days between tests for the Reliability Sample.

The empirical estimates derived from each of these samples that we need for the transformation are listed in Table 2. For the Equating Sample (students who have taken two tests of different difficulty within 30 days), the table reports raw means and standard deviations. Because the test scores used in this sample are standardized using the means and the standard deviations from the full set of reading test scores at a certain grade level, sample means and standard deviations of the test scores in this sample will generally differ from zero and one, respectively.

For the Reliability Sample (students who have completed two tests of the same difficulty less than five days apart), the table reports the empirical correlation between two test scores for the set of students included in the sample at each grade level. The empirical correlations provide us with estimates of the test-retest reliability ratios as discussed in Section 2.1.

From the means, standard deviations and reliability ratios, we can derive the μ and λ parameters (which we set to 0 and 1 in the first period where they can be chosen arbitrarily).[6]

Inspecting the means and standard deviations from the Equating Sample, two things are evident. First, we see that the mean test score gap is largest between 2^{nd} and 4^{th} grade and

[6] Appendix C shows that the results are not sensitive to the exact time lap between the first and the second test in the two samples.

Table 2: Adjusting for differences in location and scale

	Equating					
	2^{nd} grade	4^{th} grade	4^{th} grade	6^{th} grade	6^{th} grade	8^{th} grade
Mean	0.315	-1.003	0.307	-0.728	0.293	-0.490
SD	0.977	1.163	0.932	0.962	1.014	1.041
Obs.	1518		2140		1800	

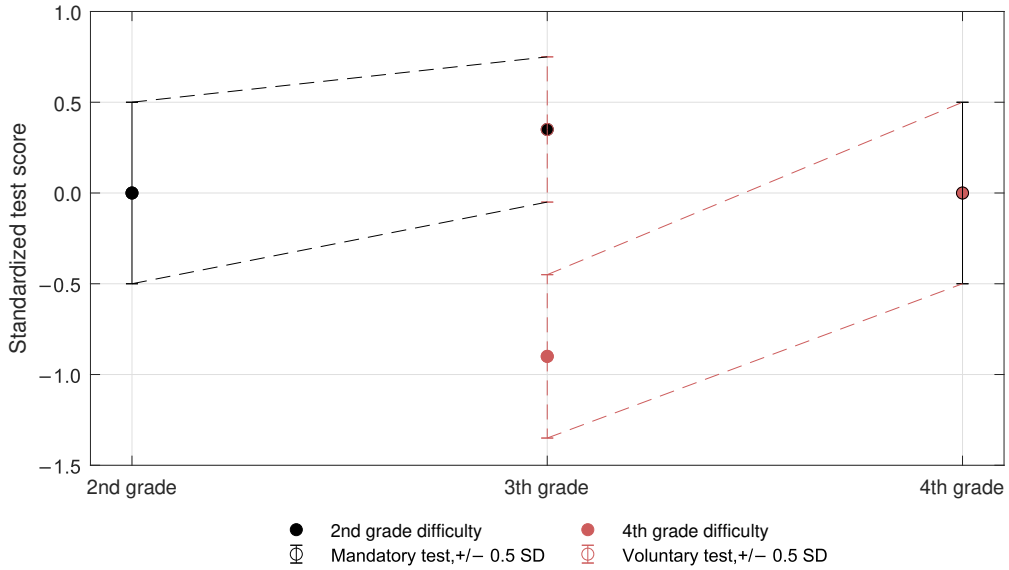
	Parameters			
	2^{nd} grade	4^{th} grade	6^{th} grade	8^{th} grade
μ_t	0	-1.390	-2.439	-3.388
λ_t	1	1.229	1.239	1.314
r_t	0.789	0.840	0.802	0.856
Obs.	3886	2708	3188	2360

Notes: This table reports means and standard deviations for the Equating Sample, and the reliability ratios for the Reliability Sample. From these, the location and scale parameters are derived in each period (normalized to 0 and 1 in 2^{nd} grade).

smallest between 6^{th} and 8^{th} grade. This is indicative of the growth in absolute skill level declining over time. Second, we see that the standard deviations are always larger for the test designed for the higher grade level within each comparison. This is indicative of the variance in the true skill distribution *narrowing* over time.

To illustrate this intuitively, Figure 2 plots the mean and the width of a one standard deviation band around the mean for the 2^{nd} and 4^{th} grade tests, when taken in the respective grades by the full population and in 3^{rd} grade by the equating subsample. In 2^{nd} and 4^{th} grade, the corresponding tests are constructed to have zero mean and a standard deviation of one. In 3^{rd} grade, the subsample is exposed to the tests with both a 2^{nd} grade and a 4^{th} grade level difficulty. As one would expect, when the 2^{nd} grade test is taken in 3^{rd} grade, the average student scores higher, while the converse is true when the 4^{th} grade test is taken in 3^{rd} grade. Second, we see that the variation is also different. In particular, when the 2^{nd} grade test is taken in 3^{rd} grade, the scores on average deviate less from the mean. This is not true to the same extent for the 4^{th} grade test taken in 3^{rd} grade. Hence, when the two tests are taken at the same point in time, it is evident that the easier test also has the smallest standard deviation. But as the standardization is based on the corresponding 2^{nd} and 4^{th} grade results, this implies that the skill variance is decreasing over this period. This can also be directly

Figure 2: Equating test scores



Notes: This figure illustrates how the 2^{nd} and 4^{th} grade tests are equated using the subsample of students taking both tests in 3^{rd} grade. In the population, both tests are standardized to have mean zero and a standard deviation of one in their respective grade level. The scores in 3^{rd} grade in the equating subsample are then used to infer the average skill growth and the change in the variance between the two tests.

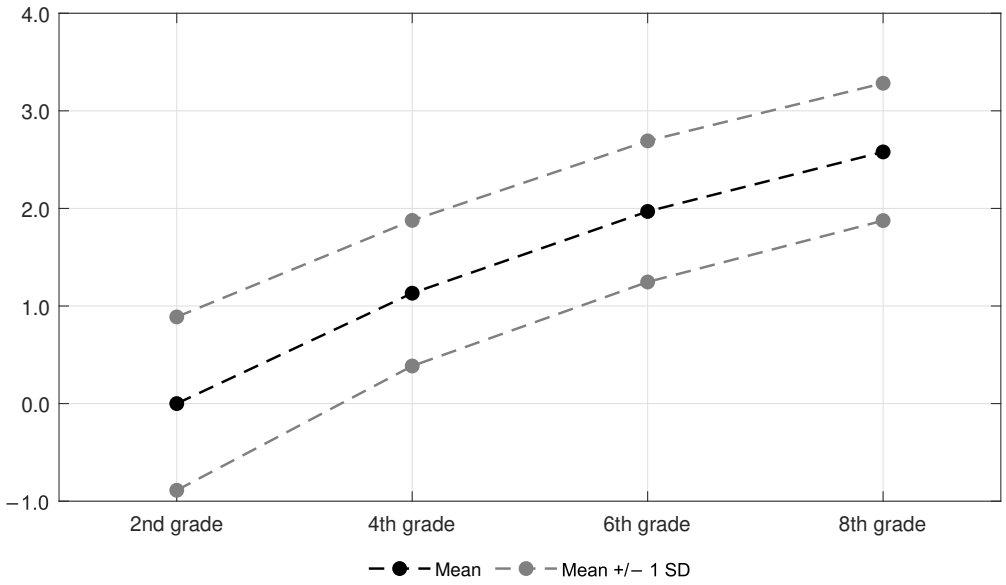
observed from Figure 2; the skill distribution appears to be narrowing from 2^{nd} to 3^{rd} grade, and this is not followed by a corresponding widening from 3^{rd} to 4^{th} grade; this implies that overall, from 2^{nd} to 4^{th} grade, the skill distribution is narrowing.

This method is known as linear equating. It relies on the assumption that the tests in different periods are equally reliable measures of the underlying skill level.[7] To relax this assumption, we have to factor in the reliability ratios, which are test-retest correlations from the Reliability Sample of students who take the same test twice in close succession. Only after adjusting for any change in reliability do we arrive at the final location and scale parameters in Table 2. The fact that λ_t is increasing in t and that $\mu_{t+1} - \mu_t$ is decreasing in t implies that the initial indications hold, and the underlying skill trajectory is characterized by declining growth and decreasing variance.

These two characteristics—declining growth curves and decreasing skill variance—are illustrated in Figure 3, which plots the expected skill curve (see Equation (8)) implied by

[7] Another assumption is that the distribution of the two equated test scores have the same shape, e.g., both being normally distributed, otherwise the linear approximation will be a poor fit. We test this assumption by instead applying the non-parametric equipercentile equating method, which reveals that the linear fit is a good approximation.

Figure 3: Reading skill curve



Notes: This figure illustrates the expected reading skill trajectory from 2^{nd} to 8^{th} grade based on our equating procedure. The dots represent the equated scores corresponding to the mean and one standard deviation above/below the mean for the respective grade levels. The means and the standard deviations are computed using the formulas in (8) using the values of μ_t, λ_t, and r_t provided in Table 2.

the parameters in Table 2. As described earlier, if the skill distribution is widening, this could partly explain fade-out as a statistical artefact. Instead, we observe a narrowing skill distribution. This is in contrast to Cascio and Staiger (2012), who find that a widening skill distribution explains a minor fraction of the fade-out they observe. In our case, the implication is, conversely, that the true fade-out is more substantial than what we might conclude at first.

4.1 Implications for Fade-Out in READ

We first report the effect of the READ intervention and its pattern of fade-out over time using the test scores standardized at each grade level separately (as is common practise). We then compare this to fade-out that is adjusted for the statistical artefact caused by narrowing of the skill distribution.

The upper panel in Table 3 displays the effect sizes for the READ intervention measured in 2^{nd}, 4^{th} and 6^{th} grade, corresponding to 7, 31 and 55 months post-intervention. This shows that the effect is statistically significant in both 2^{nd} and 4^{th} grade, with no apparent fade-out between the two periods. On the other hand, when measured in 6^{th} grade, we see that the

Table 3: Effect size estimates of the READ intervention

Grade:	2^{nd}	4^{th}	6^{th}
	7 months	31 months	55 months
Treatment	0.178**	0.192**	0.105
	(0.083)	(0.082)	(0.070)
Treatment	0.264***	0.212***	0.142**
(low SES)	(0.095)	(0.079)	(0.069)
Treatment	0.007	0.018	-0.028
(high SES)	(0.064)	(0.081)	(0.073)
Obs.	1,513	1,480	1,421

Notes: This table reports parameter estimates from regressions used to link the READ intervention to test scores in reading over time. We regress the test outcome on the treatment indicator, controlling for indicators of the 19 strata used for randomization. Lower panel report results by low and high SES subgroups. Standard errors are reported in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

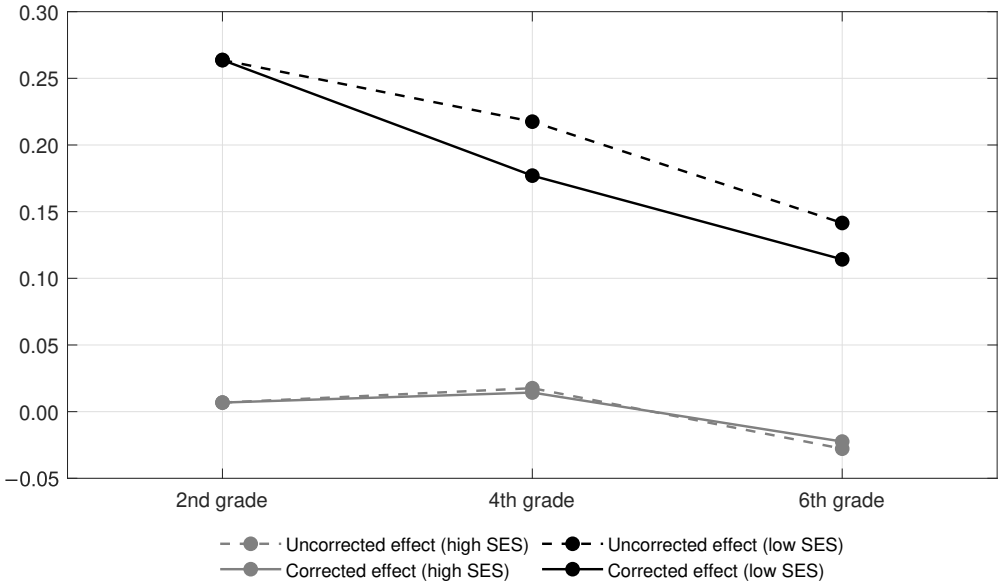
effect estimate has shrunk by almost 50% and is no longer statistically significant. Thus, it appears that a large fraction of the effect has faded out at this point.

The lower panel in Table 3 shows that point estimates for the high SES children are close to zero in all periods. For the low SES children on the other hand, we see large effects in all periods, though the effect also appears to fade out gradually.

Figure 4 compares the effect of READ on the commonly used test scores (standardized at each grade level separately) to the effect on test scores corrected for the statistical artefact. Results are reported for low- and high-SES students separately. We correct the point estimate in grade t for the statistical artefact, by dividing the estimate with the corresponding scale parameter, λ_t (in accordance with the theory presented in Section 2.1). We see that for the low-SES children, fade-out is slightly stronger when we correct for the statistical artefact due to the narrowing of the skill distribution in the population. The correction is more pronounced for the 4^{th} grade measure, because the equating procedure showed that the skill distribution was narrowing the most from 2^{nd} to 4^{th} grade. The corrected effects show a pattern of gradual fade-out of between 30% and 40% each period.

Because of the strong heterogeneity in treatment effects between low- and high-SES students and the apparent lack of an effect (and hence fade-out) for high-SES students, we focus on the

Figure 4: Treatment effect of READ by SES



Notes: This figure compares the apparent fade-out of the READ intervention when evaluated based on the observed standardized test scores to the fade-out once we correct the test scores for the statistical artefact due to changes in skill variance and measurement error (separately for low and high SES students).

low-SES students in the remaining part of the paper.

5 Vertical Fade-out Caused by the Skill Formation Technology

Now that we have corrected for the part of the observed fade-out that was due to the lack of a common scale, a significant degree of fade-out remains present. Does this mean that it is inevitable that an intervention will be followed by some fade-out (at least in our setting)? Answering this question is equivalent to saying something about the self-productivity of reading skills, δ_t (specifically, whether $\delta_t < 1$). As detailed in Section 2.2, the one-period fade-out of an intervention is given by $\frac{\tau_{t+1|t}}{\tau_{t|t}} = \delta_t$. More generally, we have that $\frac{\tau_{t+k|t}}{\tau_{t|t}} = \prod_{j=0}^{k-1} \delta_{t+j}$, for the k-period fade-out of an intervention, where $\delta_{t+j} = \mathrm{E}\left[\frac{\partial \theta_{t+j+1}}{\partial \theta_{t+j}}\right]$.

In the following, we attempt to determine the expected degree of fade-out of an intervention by estimating δ_t directly. Exploiting additional data allows us to credibly put an upper bound on this parameter by correcting for measurement error, as well as to approach the causal effect by controlling for various confounding variables. Our estimation procedure relies on the

following simple regression model:

$$\theta_{t+1} = c_t + \delta_t \theta_t + \pi_t' X_t + \eta_t, \tag{13}$$

where X_t and η_t are the same quantities as defined in Section 2.2. By comparing the model in (13) with the true skill formation model in (10), we see that the former model does not explicitly account for investments as a determinant of skills. The reason why we may abstract from investments in (13) is that the investment channels are included in the self-productivity, δ_t, which, as explained in Section 2.2 consists the direct skill persistence as well as any compensating or reinforcing investments and any dynamic complementarity or substitutability (see Appendix A for additional details).

However, if we attempt to estimate (13) using only the corrected test scores, our estimate of the self-productivity parameter will be biased for two reasons: measurement error in the explanatory variable and omitted variables in X_t. First, because the tests measure the pupils' true skills with some error, the estimates of the relationship between skills and subsequent skills will be biased *downwards*. Second, if background characteristics such as parental resources, as explained, predict both current and future skills, they will cause omitted variable bias if not included in the regression model. Intuitively, when predicting expected fade-out, we need to adjust for the fact that the intervention did not affect all factors that influence the association between reading skills in one period and the next. The intervention changed the students' reading skills, but not their parents' education, let alone their genetic endowments. Although ambiguous in theory, such variables will in practice overwhelmingly affect both current and future skills in the same direction.[8] In their absence, the relationship between skills and subsequent skills will then be biased *upwards*.

We have already dealt with the measurement error issue when correcting for the statistical artefact. To adjust for this error component, we simply need to utilize the reliability ratios defined in Section 2.1 once more. To see this, note that our re-scaled measure of reading skills ($\bar{Z}_t$), though now an unbiased measure of latent skills, will still be measured with some error,

[8]For example, parental cognitive and financial resources, child genetic endowments and non-cognitive skills, neighbourhood and school quality etc. will surely have a positive effect on both initial skills, and future skill development conditional on initial skills. It is difficult to think of characteristics for which the effects would work in opposite directions.

i.e.,

$$\bar{Z}_t \equiv \frac{Z_t - \mu_t}{\lambda_t} = \theta_t + \bar{\epsilon}_t, \tag{14}$$

where we have defined a new zero mean error term as $\bar{\epsilon}_t \equiv \epsilon_t/\lambda_t$. If we estimate (13) using $\bar{Z}_t$ (instead of θ_t), our estimate of δ_t will be biased towards zero due to the presence of measurement errors in the independent variable. This attenuation bias, however, can in our setting be expressed as the ratio between the variance of θ_t and the variance of $\bar{Z}_t$, which can readily be shown to equal the reliability ratio given by $r_t = \frac{\text{var}(Z_t)-\sigma_t^2}{\text{var}(Z_t)}$. Hence, since we are able to estimate r_t, we can run an errors-in-variables (EIV) regression, which in practice simply up-scales the estimate of δ_t by $\frac{1}{r_t}$.[9]

In Table 4, we first run a naive regression of the corrected measure of reading skills on the reading skills in the previous period. These associations are shown in column 1. Taken at face-value, they suggest that we should expect fade-out approaching 30-45% per two-year time period. However, the associations adjusted for attenuation bias are then displayed in column 2 of Table 4. After being up-scaled to account for the uncertainty in each measure, the estimates are much closer to unity, suggesting more limited fade-out of 15–30% per period. We can interpret these estimates as an upper bound on the self-productivity of reading skills. As explained, they may still be biased upwards because of omitted variables (e.g., other skills, genes, or other family characteristics) affecting both current and future reading skills. This means that the estimates are conservative when meaning to explain fade-out, because overestimating the self-productivity is the same as underestimating the degree of fade-out. Although we cannot eliminate this bias completely, we can add various control variables to X_t to reduce the upper bound and approach the true effects.

Thus, in columns 3–5, we add a range of control variables. First, in column 3, we add only maternal and paternal educational attainment, which we expect to be the best proxies for the influence of the family on reading skill formation. Indeed, we see that this has some effect on the coefficients, which shrink by 2–3 percentage points. Next, in column 4, we add a wider range of parental controls, including income, wealth, employment history and age at the child's birth. We see that the coefficients on the self-productivity of reading skills are

[9]For more complex relationships, i.e., when there are several miss-measured skills in each period, the bias will depend on the relative distributions of all the true skills and the measurement errors (Lockwood and McCaffrey, 2020)

Table 4: Estimates of the self-productivity of reading skills

	(1)	(2)	(3)	(4)	(5)
	4^{th} grade reading score				
δ_2	0.566	0.717	0.687	0.686	0.685
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	817,241				
	6^{th} grade reading score				
δ_4	0.714	0.849	0.828	0.827	0.826
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	806,933				
ME correction	( )	(X)	(X)	(X)	(X)
Parental education	( )	( )	(X)	(X)	(X)
Full parental controls	( )	( )	( )	(X)	(X)
Child birth controls	( )	( )	( )	( )	(X)

Notes: This table reports parameter estimates from regressions of test scores in reading in 4^{th} and 6^{th} grade to test scores in reading two years prior. In columns 2–5, the estimate is adjusted for measurement error. In columns 3–5, we additionally add a set of control variables. Parental education refers to the years of education of each parent. Full parental controls include, for each parent, their income and wealth ranks, their employment status, and their age at the child's birth. Child birth controls include immigrant status, birth order, birth weight, gestational age, Apgar score, and whether the mother smoked during the pregnancy.

virtually unaffected by this, suggesting that parental education is indeed capturing almost all observable variation in family background that matters for the self-productivity of reading skills. Finally, in column 5, we also add a range of other characteristics capturing the child's birth conditions including immigrant status, birth order, birth weight, gestational age, Apgar score, and whether the mother smoked during the pregnancy. Again, the coefficients practically remain the same.

The reason for the minor additional difference caused by adding these controls is not necessarily that they have no influence on skill formation. For example, parental income significantly predicts both current and later test scores even after controlling for parental education. But compared to the correlation between earlier and later test scores, the correlations between income and test scores are small, meaning that the magnitude of the omitted variable bias is very limited.[10] Although we would ideally add an even broader range of controls, it

[10] If the correlation between the omitted variable and the test scores are one tenth of that between earlier and later test scores, the bias is approximately $\frac{1}{10} \times \frac{1}{10}$, i.e., one percent. After controlling for parental education, the correlation between any of the other controls and test scores are much less than one tenth of the correlation between earlier and later test scores.

seems reasonable to believe, based on Table 4, that this would only cause a slight further reduction in the point estimates. Still, we consider the estimates an upper bound on the self-productivity of reading skills.

Comparing the upper and lower panel in Table 4, it is interesting to note that reading skills appear to be more persistent between 4^{th} and 6^{th} grade (lower panel) than between 2^{nd} and 4^{th} grade (upper panel). This also suggests that there is more re-shuffling happening early in terms of students moving up or down in the skill distribution. This could happen for various reasons. One explanation could be that other factors, such as the quality of the home environment, is more important in the earlier grades. Another explanation could be that investments become more reinforcing later on.

5.1 Observed and Predicted Fade-Out of READ

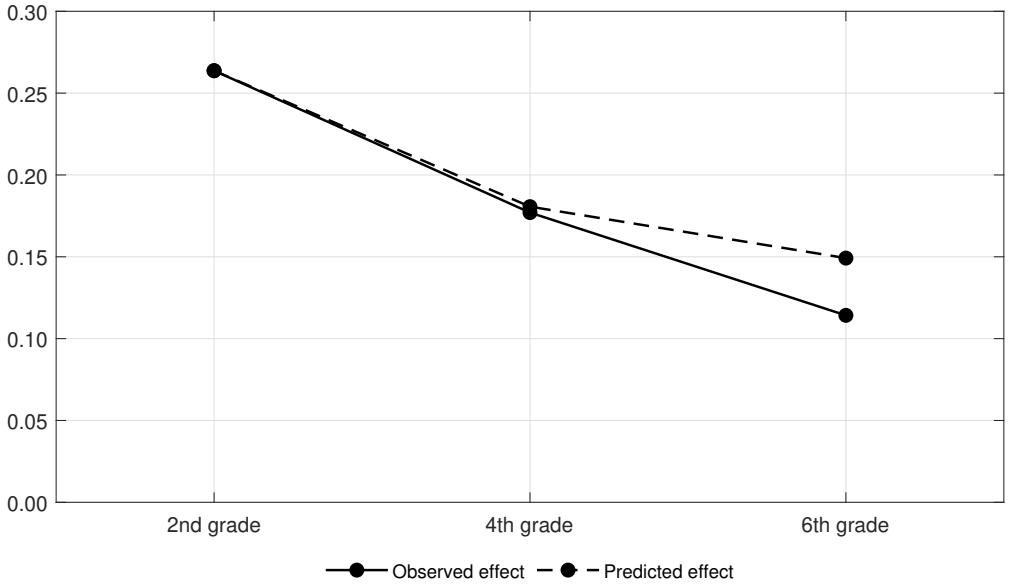
In Figure 5, we plot the observed fade-out (corrected for the statistical artefact) next to the predicted fade-out for the low SES children. We see that the observed fade-out of the READ intervention follows a pattern that, between 2^{nd} and 4^{th} grade, is practically identical to the predicted fade-out. Between 4^{th} and 6^{th} grade, the actual degree of fade-out is higher than the predicted fade-out, and therefore the overall fade-out is also somewhat higher than the predicted level. This is consistent with the predictions representing an upper bound on the extent to which the effect is expected to persist.

6 Fade-Out and Multidimensionality

So far, we have restricted our focus to a uni-dimensional measure of reading skills. This restriction may be innocuous as long as the intervention we consider did in fact only affect a single skill dimension. On the other hand, if several interrelated skills were affected, the predictions from our uni-dimensional model may not accurately correspond to the skill profile we are estimating.

As described earlier, the Danish national tests are multidimensional, with three distinct profile areas for each subject. The profile areas in all the reading tests are: language comprehension (P1), decoding (P2) and text comprehension (P3). Instead of combining them into a total reading score, we now consider these profile areas as separate dimensions of reading

Figure 5: Predicted and observed fade-out of READ



Notes: This figure compares the treatment effect for the READ intervention over time after correcting for the statistical artefact to how the initial effect is predicted to develop using the skill formation model.

Table 5: Estimated reliability ratios by profile area

Grade:	2^{nd}	4^{th}	6^{th}
Reliability ratio (P1)	0.6235	0.6633	0.6131
Reliability ratio (P2)	0.8123	0.8209	0.7915
Reliability ratio (P3)	0.7704	0.7583	0.7274
N	3886	2708	3188

Notes: This table reports the reliability ratios (the test-retest correlations) separately for each profile area.

skills. We thus re-do the equating and re-scaling separately for each profile area. As shown in Table 5, there is a lot of variation in how precisely the skills are measured across areas, making the correction even more important in the multidimensional setting.

Next, we again attempt to predict the expected fade-out by estimating the self-productivity parameters. This becomes more complicated when allowing for the fact that skills are multidimensional and may affect each other. We now have to estimate both an own- and a cross-dimensional δ-parameter for each possible combination of skill dimensions.

Hence, we specify a generalization of the model in (13), in which we allow each profile area

to affect all three profile areas in each subsequent period:

$$\theta_{p,t+1} = c_{p,t} + \delta_{1,p,t}\theta_{1,t} + \delta_{2,p,t}\theta_{2,t} + \delta_{3,p,t}\theta_{3,t} + \pi'_{p,t}X_t + \eta_{p,t}, \tag{15}$$

where $\theta_{p,t}$ denotes the level of the p^{th} reading skill component in time period t, and where $\delta_{1,p,t}$ to $\delta_{3,p,t}$ thus capture the effect of reading skills on subsequent reading skills for all pairwise combinations of profile areas. Thus, for associations within the same profile area, e.g. $\delta_{1,p,t}$, this is equivalent to the self-productivity for this skill. Additionally, the remaining δ-parameters capture the *cross-productivity*, i.e., that skills on one dimension may affect skills on the other dimensions.[11] We estimate the three equations in (15) simultaneously using a structural equation modeling approach, in which we use the equated test scores $\bar{Z}_{p,t} \equiv (Z_{p,t} - \mu_{p,t})/\lambda_{p,t}$ as proxies for the latent reading skills, $\theta_{p,t}$. Motivated by theory, we impose the estimation constraint that all self- and cross-productivity parameters must be non-negative, i.e., $\delta_{i,p,t} \geq 0$ for all $i, p \in \{1, 2, 3\}$ and $t \geq 1$.

The naive estimates of the δ-parameters in which we do not take measurement errors and covariates into account are displayed in Table 6, columns 1 and 2. We see that the persistence parameters with respect to the skill itself becomes smaller, only reaching 0.5 for decoding. At the same time, there appears to be significant spillover for all profile areas. However, such a pattern may also appear as a consequence of measurement error. We therefore again need to adjust for this bias using the reliability ratios from Table 5. Note that profile area 1 (language comprehension) has the lowest reliability. This may explain why this dimension appears to have such weak persistence.

The adjusted associations are displayed in columns 3 and 4 of Table 6. As expected, correcting for the measurement error causes all the persistence parameters to increase in magnitude. Conversely, most of the cross-productivity parameters decrease. This is expected because the different skill dimensions are positively correlated with each other, and because they are all measured with error, the skill level on one dimension will be informative about skills on the other dimensions, even though knowledge does not actually spill over. However,

[11] Potentially, skills that are not directly related to reading could also affect reading skills. However, the READ intervention had no significant effects on neither conscientiousness nor math skills. This is consistent with Aucejo and James (2021) who find that such cross-effects are not present in the production of reading skills.

Table 6: Estimates of the self- and cross-productivities of three dimensions of reading skills

Grade ($t=$):	(1) 4^{th}	(2) 6^{th}	(3) 4^{th}	(4) 6^{th}	(5) 4^{th}	(6) 6^{th}
	P1 (language comprehension)					
$\delta_{1,1,t-1}$	0.165	0.267	0.366	0.684	0.311	0.654
	(0.001)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
$\delta_{2,1,t-1}$	0.181	0.126	0.233	0	0.234	0
	(0.001)	(0.001)	(0.003)		(0.002)	
$\delta_{3,1,t-1}$	0.158	0.161	0	0	0	0
	(0.002)	(0.001)				
	P2 (decoding)					
$\delta_{1,2,t-1}$	0.013	0.100	0	0	0	0
	(0.001)	(0.001)				
$\delta_{2,2,t-1}$	0.468	0.512	0.680	0.807	0.658	0.808
	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)
$\delta_{3,2,t-1}$	0.129	0.149	0	0.024	0	0
	(0.001)	(0.001)		(0.003)		
	P3 (text comprehension)					
$\delta_{1,3,t-1}$	0.102	0.157	0.063	0	0.003	0
	(0.001)	(0.001)	(0.004)		(0.004)	
$\delta_{2,3,t-1}$	0.200	0.194	0	0	0	0
	(0.001)	(0.001)				
$\delta_{3,3,t-1}$	0.312	0.424	0.614	0.831	0.607	0.794
	(0.001)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
ME correction	( )	( )	(X)	(X)	(X)	(X)
All controls	( )	( )	( )	( )	(X)	(X)
N	817,241	806,933	817,241	806,933	817,241	806,933

Notes: This table reports parameter estimates from regressions of test scores for each profile area in reading in 4^{th} and 6^{th} grade to test scores on all three profile areas two years prior. In columns 3–4, the estimate is adjusted for measurement error. In columns 5–6, we additionally add the full set of controls. These are, for each parent, their educational attainment, their income and wealth ranks, their employment status, and their age at the child's birth, as well as the child's immigrant status, birth order, birth weight, gestational age, Apgar score, and whether the mother smoked during the pregnancy. Non-negativity of the persistence parameters are imposed in all estimations.

we see that a positive cross-effect remains for language comprehension (P1), which is affected by earlier decoding (P2) ability, though only in the first period.

As in the single-dimensional case, we also proceed to add the family control variables in order to limit the omitted variable bias. The estimates are displayed in columns 5 and 6 in Table 6. Again, adding the controls causes most of the point estimates to decrease, albeit by a relatively small magnitude. Although the estimates again represent upper bounds on the

persistence parameters, it is plausible that adding further controls would only slightly weaken the relationships, even more so now that we also account for the multidimensionality.

6.1 Observed and Predicted Fade-Out of READ by Subdimensions

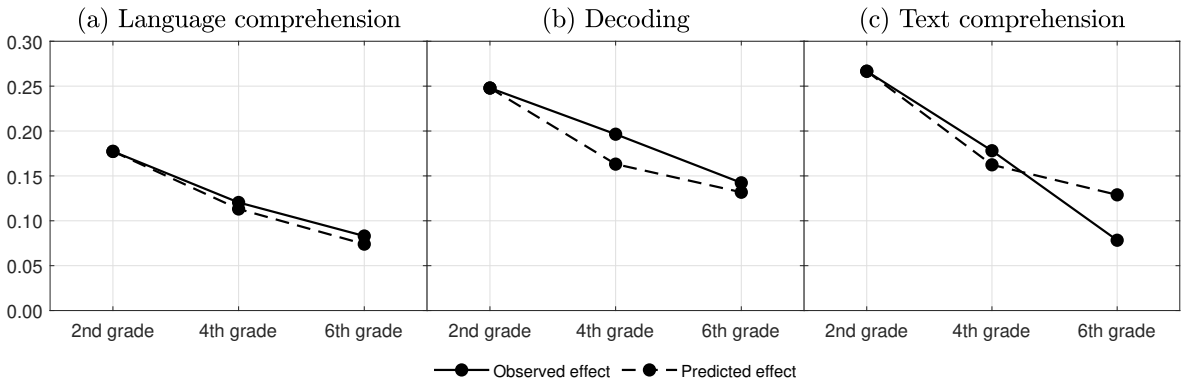
Using the associations between the different profile areas and grades, after controlling for measurement error and background characteristics, we can now predict how we should expect the READ treatment effects to develop for each reading skill dimension. This is depicted in Figure 6. It shows that the fade-out we observe for the intervention is largely consistent with the predicted fade-out for each skill dimension. For language comprehension (P1), there is very good agreement between observed and predicted fade-out throughout the period observed. The prediction also ends up close to the actual fade-out for decoding (P2). The agreement for text comprehension (P3) is good between 2^{nd} and 4^{th} grade, although the large degree of fade-out observed for this profile area after 4^{th} grade is not predicted.

While results for the three skill dimensions are largely consistent with the overall pattern seen before, two additional findings appear from this sensitivity analysis. First, to the extent that our predictions deviated from the observed fade-out for overall reading skills, this appears to be mostly a result of a deviation for a specific profile, namely text comprehension. Secondly, although language comprehension has the lowest self-productivity, predicted fade-out is not much higher for this profile area because the large initial treatment effect on decoding is predicted to have a cross-effect on language comprehension. Indeed, we see that the observed pattern of fade-out for language comprehension closely follows this prediction.

7 Horizontal Fade-out Caused by the Skill Formation Technology

In many educational systems, students need to attain a certain skill level in order to qualify for the next educational level. They may, for instance, need a certain grade point average at some qualifying exams in order to be enrolled at certain high schools or universities. If they do not reach that skill level in time, they may need to pursue other educational trajectories. Therefore, the time lead that some student may have over others in terms of their skill accumulation is

Figure 6: Predicted and observed fade-out of READ by profile areas



Notes: This figure compares the treatment effect by profile area for the READ intervention over time after correcting for the statistical artefact to how the initial effect is predicted to develop using the multidimensional skill formation model.

not just an economically meaningful measurement unit, it is also important for the individual student's educational attainment.[12]

As described in Section 2.3, when the average skill curve is concave—which we observe that it is, see Figure 3—then a reduced difference in skill levels between treatment and control group may be consistent with a maintained time lead for the treatment group over the control group. And this is close to what we find. Figure 7 compares the standard approach—measuring fade-out in terms of the (vertical) difference between treatment and control group in standard deviations (the black line)—to the horizontal perspective showing the difference in the months of time lead (the gray line). To re-express an effect size on skills ($\tau_{t+k|t}$) to a measure of time lead in months ($\phi_{t+k|t}$), we use the following crude approach to solve (11):[13]

$$\phi_{t+k|t} = \frac{\tau_{t+k|t}}{\mathrm{E}[\theta_{t+k+1}] - \mathrm{E}[\theta_{t+k}]} \times 24, \qquad k = 0, \ldots, T-t,$$

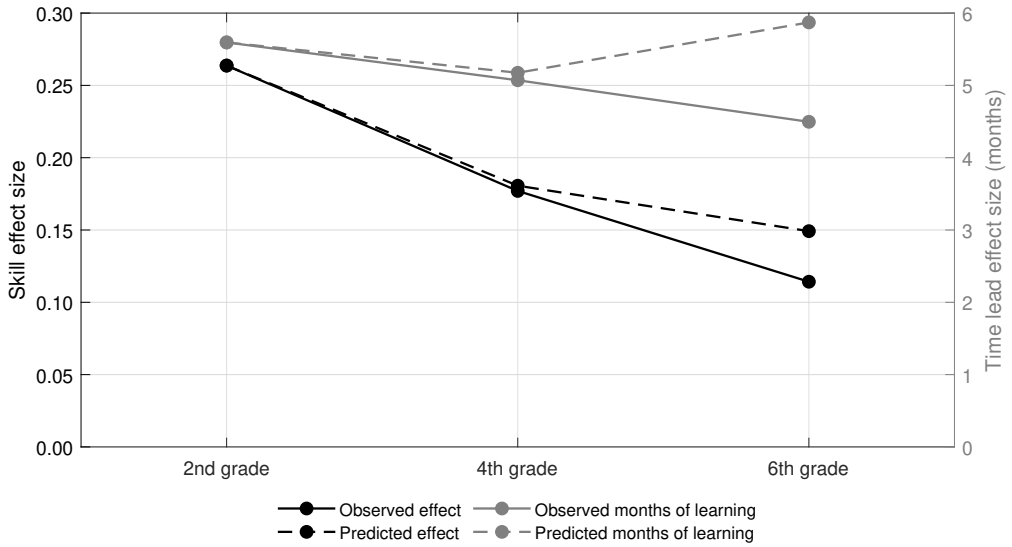
where we multiply the ratio of the effect size to the increment in expected learning by 24, because, in our application, one time period corresponds to 24 months.

The black line in Figure 7 is identical to the black line in Figure 5 and printed here for comparison. It indicates fade-out of the treatment effect, which is reduced from around 0.26

[12] Bond and Lang (2018) mitigate the arbitrariness of test score scales by tying it to an external metric: the predicted number of years of education (see also Cunha and Heckman, 2008).

[13] This approach is exact only in the case when the mapping $t \mapsto \mathrm{E}[\theta_t]$ is a piecewise linear function on all intervals $[t, t+1]$, where t is a positive integer.

Figure 7: Fade-out on different scales



Notes: The black lines, with values corresponding to the left y-axis, show the effect sizes on skills (solid line) and the predicted development of the effect size (dashed line) based on the skill formation model. The gray lines, with values corresponding to the right y-axis, show the effect sizes on skills measured in months of learning (solid line) and the predicted learning time lead (dashed line) based on the skill formation model. The intervention was implemented in 2^{nd} grade, and the effect sizes are measured at the end of 2^{nd}, 4^{th}, and 6^{th} grade.

at the end of 2^{nd} grade to about 0.12 at the end of 6^{th} grade, that is, slightly above 50% fade-out. In contrast, the gray line shows that the treatment group has a time lead at the end of 2^{nd} grade of about 5.6 months. This effect decreases much more slowly, to about 5.1 months at the end of 4^{th} grade and to 4.5 months at the end of 6^{th} grade. In other words, the treatment group maintains about 80 % of its time lead up to four years after the intervention. And because the strong concavity of the skill curve outweighs the limited self-productivity, the predicted fade-out is actually slightly negative when measured in time lead.

8 Conclusion

It is often argued that early interventions are much more effective investments in human capital than investments in young adults. However, this view is challenged by observations that the effects of interventions often fade out within a few years, making early interventions appear futile. Therefore, understanding fade-out of interventions aimed at improving the skills of young children is crucial to questions about human capital development and education policy.

Previous research has clearly demonstrated that in the absence of common metrics for

measuring skills at different age levels, it is almost impossible to estimate fade-out validly. Depending on the assumptions used to equate across different tests at different age levels, the same data may support conclusions ranging from complete fade-out to fade-in (Wan et al., 2021).

We provide a method for equating a set of reading tests taken in different school grades such that they measure the underlying skill on the same scale. Our method accounts for the possibly different reliability of each test as well as the potentially changing distribution of underlying reading skills in the population. We find that the variance of reading skills is decreasing over time, which implies that conventional estimates of fade-out that standardize test scores to a standard deviation of 1 at each grade level will underestimate the degree of fade-out.

We also estimate a skill formation model which predicts that some fade-out follows naturally from the (lack of) persistence of reading skills over time. We find the self-productivity parameter to be less than one, which implies that an initial skill difference between otherwise similar students (such as the effect of an intervention) will tend to decrease over time. This predicted fade-out pattern is very consistent with what we observe for the READ intervention. This is not due to the specific intervention being ineffective, but rather is a natural result of future reading skills being influenced by other factors beyond the current skill level of the student. For example, an intervention may successfully raise student skill levels, but it will not raise educational attainment of the student's parents. In so far as parental education matters for future skill formation, the treated students will still be at a deficit compared to other students who arrived at a high skill level not as a result of the intervention but as a result of having highly educated parents.

Finally, our results also demonstrate that the skill growth curve is concave. The older students become, the longer it takes to increase their skills. Thus, although students in the control group tend to catch up over time in terms of the vertical skill gap, treated students are predicted to maintain their time lead four years after the intervention. In many educational systems, time is crucial, because if students do not reach the required skill level in time (as measured by their grade point average or some other test), they will not qualify for further education and may never get the chance to further improve their skills. The conversion of test scores from the arbitrary test score scale to the time scale makes the treatment effect

economically meaningful. Whenever possible, it would be useful to convert test scores to a scale of time progress rather than the often used standardized test scores–in order to make meaningful comparisons of effect sizes across different studies (cf. Kraft, 2020; Lipsey et al., 2012).

In sum, our results show that fade-out is a substantive phenomenon. In the setting we consider here, it is present to such an extent that, all else equal, an intervention in 6^{th} grade is at least as good as an intervention four years earlier even if the early intervention has twice as large an effect initially. However, all else is probably never equal for two interventions aimed at students four years apart. In particular, the returns to one year of schooling appear higher for younger students, which suggests that interventions will also tend to be more effective early on. Nevertheless, our findings caution against the conclusion that earlier is better if such conclusions are based only on effects observed shortly after an intervention.

References

Agostinelli, F. and M. Wiswall (2023). Estimating the Technology of Children's Skill Formation. *Journal of Political Economy*, Forthcoming.

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *Quarterly Journal of Economics 134*(3), 1121–1162.

Andersen, S. C. and H. S. Nielsen (2016). Reading intervention with a growth mindset approach improves children's skills. *Proceedings of the National Academy of Sciences 113*(43), 12111–12113.

Aucejo, E. and J. James (2021). The path to college education: The role of math and verbal skills. *Journal of Political Economy 129*(10), 2905–2946.

Bailey, D., G. J. Duncan, C. L. Odgers, and W. Yu (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness 10*(1), 7–39.

Bailey, D. H., G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest 21*(2), 55–97.

Bond, T. N. and K. Lang (2018). The black–white education scaled test-score gap in grades k-7. *Journal of Human Resources 53*(4), 891–917.

Burgess, S., S. Rawal, and E. 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.

Cascio, E. U. and D. O. Staiger (2012). Knowledge, tests, and fadeout in educational interventions. Technical report, National Bureau of Economic Research.

Cunha, F. and J. J. Heckman (2007). The Technology of Skill Formation. *American Economic Review 97*(2), 31–47.

Cunha, F. and J. J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Cunha, F., E. Nielsen, and B. Williams (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics 13*, 487–513.

Del Bono, E., E. Del Bono, and R. Pavan (2022). Identification of dynamic latent factor models of skill formation with translog production. *Journal of Applied Econometrics 37*, 1256–1265.

Duncan, G., A. Kalil, M. Mogstad, and M. Rege (2023). Investing in Early Childhood Development in Preschool and at Home. In *Handbook of the Economics of Education*, Volume 6. Elsevier.

Dweck, C. S. (2006). *Mindset: The New Psychology of Success* (Reprint edition ed.). Random House.

Gensowski, M., R. Landersø, D. Bleses, P. Dale, A. Højen, and L. Justice (2020). Public and parental investments and children's skill formation. Working Paper 155.

Houmark, M., V. Ronda, and M. Rosholm (2020). The nurture of nature and the nature of nurture: How genes and investments interact in the formation of skills. IZA Discussion Paper 13780.

Kolen, M. and R. Brennan (2014). *Test Equating, Scaling and Linking: Methods and Practices*. Springer.

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

Lipsey, M. W., K. Puzio, C. Yun, M. A. Hebert, K. Steinka-Fry, M. W. Cole, M. Roberts, K. S. Anthony, and M. D. Busick (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research*.

Lockwood, J. and D. F. McCaffrey (2020). Recommendations about estimating errors-in-variables regression in Stata. *The Stata Journal 20*(1), 116–130.

Nicoletti, C. and V. Tonei (2020). Do parental time investments react to changes in child's skills and health? *European Economic Review 127*, 103491.

Rege, M., I. Størksen, I. F. Solli, A. Kalil, M. M. McClelland, D. Ten Braak, R. Lenes, S. Lunde, S. Breive, M. Carlsen, et al. (2021). The effects of a structured curriculum on preschool effectiveness: A field experiment. *Journal of Human Resources*, 0220–10749R3.

Rossin-Slater, M. and M. Wust (2020). What is the added value of preschool for poor children? long-term and intergenerational impacts and interactions with an infant health intervention. *American Economic Journal: Applied Economics 12*(3), 255–86.

Wan, S., T. N. Bond, K. Lang, D. H. Clements, J. Sarama, and D. H. Bailey (2021). Is intervention fadeout a scaling artefact? *Economics of Education Review 82*, 102090.

APPENDIX

A Additional Theory

A.1 Relationship Between Fade-out and Self-Productivity

Given the skill production function

$$\theta_{t+1} = \gamma_{0,t} + \gamma_{1,t}\theta_t + \gamma_{2,t}I_t + \gamma_{3,t}\theta_t I_t + \pi_t' X_t + \eta_t, \tag{A.1}$$

where the production shock, η_t, is assumed to be independent of both θ_t, I_t, and X_t in any time period, the treatment effect in period $t+1$ is given by:

$$\begin{aligned}
\tau_{t+1|t} &= \mathrm{E}[\theta_{t+1} \mid D_t = 1] - \mathrm{E}[\theta_{t+1} \mid D_t = 0] \\
&= \tau_{t|t}\gamma_{1,t} + \tau_{t|t}\gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}(\mathrm{E}[\theta_t I_t \mid D_t = 1] - \mathrm{E}[\theta_t I_t \mid D_t = 0]) \\
&= \tau_{t|t}\gamma_{1,t} + \tau_{t|t}\gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}(\mathrm{E}[\theta_t \mid D_t = 1]\,\mathrm{E}[I_t \mid D_t = 1] \\
&\quad - \mathrm{E}[\theta_t \mid D_t = 0]\,\mathrm{E}[I_t \mid D_t = 0] + \mathrm{cov}(\theta_t, I_t \mid D_t = 1) - \mathrm{cov}(\theta_t, I_t \mid D_t = 0)).
\end{aligned}$$

Because treatment only affects I_t through θ_t, we have that $\mathrm{cov}(\theta_t, I_t \mid D_t = 1) = \mathrm{cov}(\theta_t, I_t \mid D_t = 0)$, and thus

$$\begin{aligned}
\tau_{t+1|t} &= \tau_{t|t}\gamma_{1,t} + \tau_{t|t}\gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}(\mathrm{E}[\theta_t \mid D_t = 1]\,\mathrm{E}[I_t \mid D_t = 1] - \mathrm{E}[\theta_t \mid D_t = 0]\,\mathrm{E}[I_t \mid D_t = 0]) \\
&= \tau_{t|t}\gamma_{1,t} + \tau_{t|t}\gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\Big((\mathrm{E}[\theta_t \mid D_t = 0] + \tau_{t|t})\,\mathrm{E}[I_t \mid D_t = 1] \\
&\quad - \mathrm{E}[\theta_t \mid D_t = 0]\Big(\mathrm{E}[I_t \mid D_t = 1] - \tau_{t|t}\frac{\partial I_t}{\partial \theta_t}\Big)\Big) \\
&= \tau_{t|t}\Big(\gamma_{1,t} + \gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\Big(\mathrm{E}[I_t \mid D_t = 1] + \frac{\partial I_t}{\partial \theta_t}\,\mathrm{E}[\theta_t \mid D_t = 0]\Big)\Big).
\end{aligned}$$

Given that treatment is randomly assigned with $P(D_t = 0) = P(D_t = 1) = \frac{1}{2}$, we obtain that

$$\tau_{t+1|t} = \tau_{t|t}\Big(\gamma_{1,t} + \gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\Big(\mathrm{E}[I_t] + \frac{\partial I_t}{\partial \theta_t}\,\mathrm{E}[\theta_t]\Big)\Big)$$

By differentiating (A.1) with respect to θ_t and taking the expectation, we can equivalently express this relation as

$$\tau_{t+1|t} = \tau_{t|t} \, \mathrm{E}\left[\frac{\partial \theta_{t+1}}{\partial \theta_t}\right],$$

i.e., as the treatment effect that remains one period later is completely determined by how the initial effect carries over to the next period. Thus, we can also express the degree of fade-out as:

$$1 - \frac{\tau_{t+1|t}}{\tau_{t|t}} = 1 - \mathrm{E}\left[\frac{\partial \theta_{t+1}}{\partial \theta_t}\right] \equiv 1 - \delta_t,$$

where we use δ_t to denote the *self-productivity* of skills in period t, which in this model is given by

$$\delta_t = \gamma_{1,t} + \gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\Big(\mathrm{E}[I_t] + \frac{\partial I_t}{\partial \theta_t}\,\mathrm{E}[\theta_t]\Big).$$

A.2 Estimating Self-Productivity without Investments

Because investments are themselves a function of skills, we can estimate the self-productivity of skills without accounting for investments explicitly. Measuring investments are only required in so far as we want to break down the total self-productivity into its three components (direct skill persistence, compensating/reinforcing investments, dynamic complementarity/substitutability). To see this, note that the self-productivity parameter estimated by the regression

$$\theta_{t+1} = c_t + \delta_t \theta_t + \pi_t' X_t + \eta_t,$$

is given by

$$\delta_t = \frac{\mathrm{cov}(\theta_{t+1}, \theta_t)}{\mathrm{var}(\theta_t)},$$

which following the skill formation model in (10) is equivalent to

$$\delta_t = \frac{\mathrm{cov}(\gamma_{0,t} + \gamma_{1,t}\theta_t + \gamma_{2,t} I_t + \gamma_{3,t}\theta_t I_t + \pi_t' X_t + \eta_t, \theta_t)}{\mathrm{var}(\theta_t)}.$$

This can be rewritten as

$$\delta_t = \gamma_{1,t} + \gamma_{2,t}\frac{\partial I_t}{\partial \theta_t} + \gamma_{3,t}\big(\,\mathrm{E}[I_t] + \mathrm{E}[\theta_t]\frac{\partial I_t}{\partial \theta_t}\big),$$

where the last equality follows under the assumption that investments depend linearly on the skill level, such that $\frac{\mathrm{cov}(I_t,\theta_t)}{\mathrm{var}(\theta_t)} = \frac{\partial I_t}{\partial \theta_t}$. This assumption is also adopted by Cunha et al. (2010) and Agostinelli and Wiswall (2023). Hence, we have that

$$\delta_t = \mathrm{E}\left[\frac{\partial \theta_{t+1}}{\partial \theta_t}\right],$$

i.e., under correct specification the δ_t parameter from the regression without investments is equivalent to the causal effect of skills on next period skills, that is, the self-productivity. Of course, with an unbiased measure of investments, one can additionally estimate the relative contribution of $\gamma_{1,t}$, $\gamma_{2,t}$ and $\gamma_{3,t}$ to δ_t .

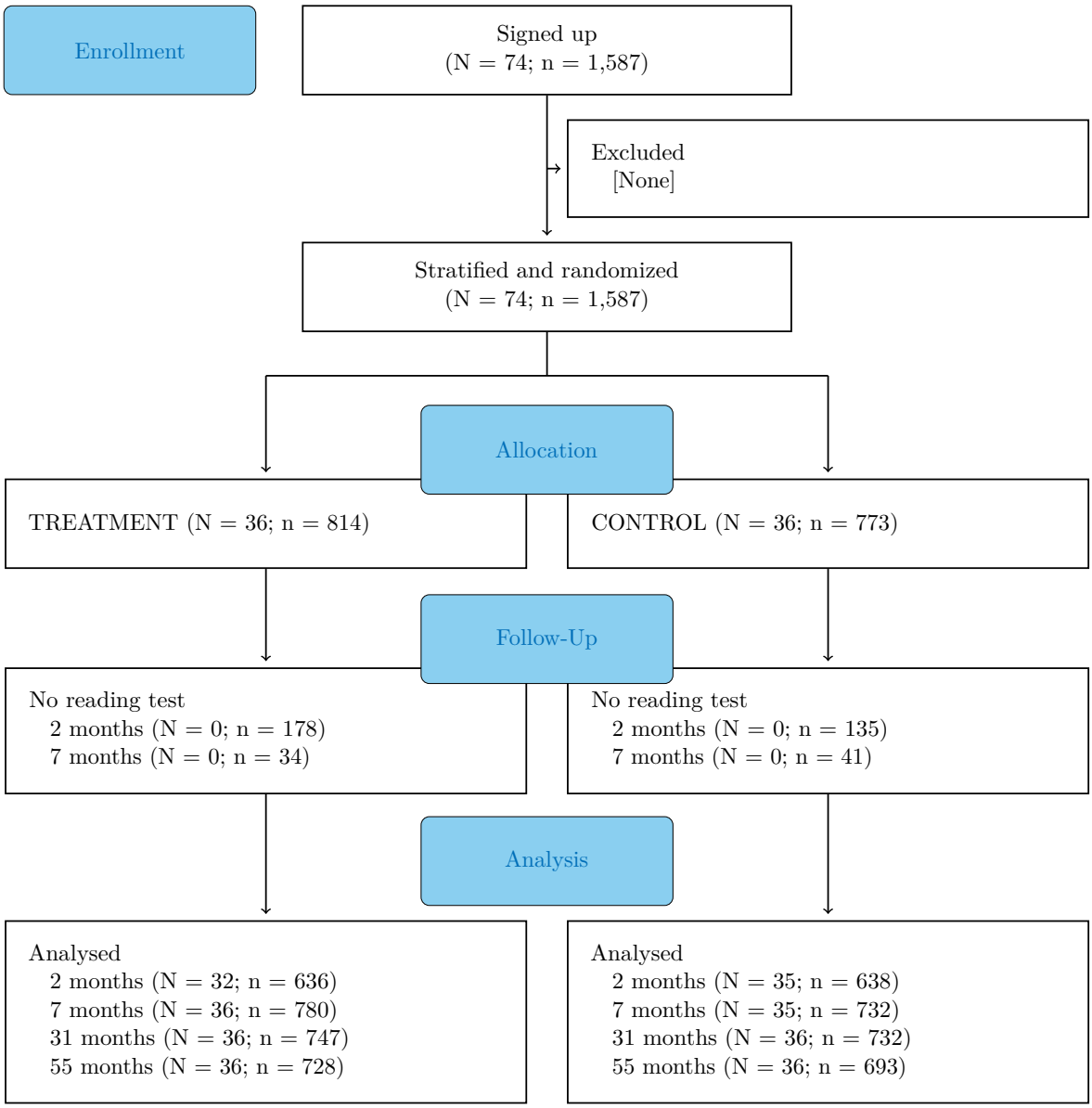
B Supplementary Information on the READ trial

Figure B.1 shows the flow of participants in the READ trial.

C Robustness and Sensitivity Analysis

In this section, we conduct a series of robustness and sensitivity analyses to further corroborate the findings in the main text. First, in Section C.1, we argue for the choice we have made regarding the maximum number of days between the two voluntary tests in the Equating- and Reliability Sample. Second, in Section C.2 we provide a test to justify that the condition of Assumption 3 are met in our empirical application. Third, we analyse how sensitive our main results are to the use of the reliability ratios calculated by Rasch algorithm instead of our own estimates in Section C.3. Finally, in Section C.4, we re-estimate the equating parameters using weighted samples to account for the potential differences between the population and the selected subsamples.

Figure B.1: Participant flow diagram



N: Number of classrooms. n: Number of students

C.1 Sensitivity Analysis of Duration Between Reading Tests

As described in Section 4, the underlying skill trajectory can be identified if each pair of tests are taken simultaneously by the same individual. Naturally, any empirical application will have to approximate this. In our main analysis, the Equating Sample includes children who complete two different tests within a maximum of 30 days, and 5 days for the Reliability Sample. The results are not sensitive to these exact thresholds, although using tests taken many months apart does change the results, as one would expect.

There are two reasons why we use different thresholds for the two samples. First, for any given threshold, the Reliability Sample is substantially larger than the Equating Sample. Since we face a bias/variance-tradeoff, this difference implies that we should choose a lower threshold for the Reliability Sample. Second, by deviating from the theoretical simultaneity, we introduce an order in which the two tests are taken, and this may be consequential in the case of the Reliability Sample.[14] Although the tests themselves are identical, we may worry that the second test measures something differently as a result of being taken later.

Indeed, we do observe that individuals score slightly higher on the second test. There are two potential explanations for this: (i) either the student has become better at reading during the time elapsed between the two tests (an increase in θ_t), or (ii) the student now better "knows" how to take the test (captured by an increase in μ_t). Importantly, only the former explanation is a violation of the identifying assumptions (because μ_t does not enter into the covariances between the test scores). To reduce the concern that θ_t is changing, we therefore choose the threshold of only five days, making it unlikely that any actual learning happening between the two tests will be substantial. Furthermore, the next section shows a test result that is consistent with θ_t being constant in this interval.

C.2 Validating Assumptions

In this section, we test for violations of the identifying Assumption 3. Our identification strategy depends on the assumptions that $\lambda_t^{(1)} = \lambda_t^{(2)}$ and $\text{var}(\epsilon_t^{(1)}) = \text{var}(\epsilon_t^{(2)})$. An implication of this assumptions is that $\text{cov}(Z_t^{(1)}, Z_{t+1}) = \text{cov}(Z_t^{(2)}, Z_{t+1})$, where $Z_t^{(1)}$ and $Z_t^{(2)}$ denotes the

[14] The order may also matter for the Equating Sample, but here we can simply make sure that the sample is balanced in the sense that the easier of the two tests comes last as often as it comes first. Our results are robust to weighting the observations so that this balance is achieved.

Table C.1: Comparison of covariances

Grade:	2^{nd}	4^{th}	6^{th}
$\text{cov}(Z_t^{(1)}, Z_{t+1})$	0.661	0.746	0.675
$\text{cov}(Z_t^{(2)}, Z_{t+1})$	0.685	0.716	0.710
Difference (p-value)	*0.492*	*0.391*	*0.273*
N	1522	1085	1126

Notes: This table reports the empirical covariance between voluntary and subsequent mandatory reading tests. The p-values are from a series of tests testing for equality between $\text{cov}(Z_t^{(1)}, Z_{t+1})$ and $\text{cov}(Z_t^{(2)}, Z_{t+1})$.

two tests of difficulty level t, and Z_{t+1} is another test designed for the next difficulty level. The restriction is testable, and the failure to reject this hypotheses increases the likelihood that this identifying assumption is satisfied.

To perform this test, we use the succeeding mandatory test as the benchmark. For example, if $Z_t^{(1)}$ and $Z_t^{(2)}$ are voluntary 2nd grade tests, Z_{t+1} is the mandatory 4th grade test. The empirical covariances are displayed in Table C.1 together with p-values from a test testing the null hypotheses of equality of covariances. For all pairs of covariances, we do not find any statistically significant differences at any conventional significance level.

C.3 Using SEM from the Rasch Algorithm

In this section, we illustrate the sensitivity of the results where we correct for differences in measurement error using the theoretical standard errors of measurement calculated by the Rasch algorithm during the individual reading tests instead of our own estimates of the reliability ratios. These theoretical reliability ratios are displayed in Table C.2. Below the ratios, we compare the fade-out correction that we make in the paper to the same correction but based on the theoretical ratios. We see that either method arrives at practically the same degree of fade-out. The same is true for the fade-out that we predict if we run the regressions in Table 4, column 5, but with the measurement error correction based on the theoretical SEM.

C.4 Weighting the Subsamples

A potential concern is that the subsamples used to identify the location and scale parameters and the reliability ratios are selected subsamples of the full population with potentially different

Table C.2: Reliability ratios (theoretical SEM)

Grade:	2^{nd}	4^{th}	6^{th}
Reliability ratio	0.8225	0.8158	0.8385
Observed fade-out (empirical)	0.2637	0.1770	0.1142
Observed fade-out (theoretical)	0.2637	0.1835	0.1141
Predicted fade-out (empirical)	0.2637	0.1806	0.1492
Predicted fade-out (theoretical)	0.2637	0.1719	0.1472
N	311,763	305,118	283,568

Notes: This table reports the reliability ratios and observed and predicted fade-out when using the standard error of measurement provided by the Rasch algorithm underlying the Danish national reading tests.

skill profiles. To accommodate this concern, we reproduce all the parameters where we weight each observation based on the estimated probability that the individual is part of the subsample in question. To calculate these probabilities, we run a probit model of a dummy for being in the subsample on the full set of covariates that we use as controls in, e.g., Table 4.

The weighted parameters are displayed in Table C.3. As evident, none of the means and standard deviations of the equating subsample, nor the reliability ratios of the reliability subsample, are significantly different from their unweighted counterparts. For the equating subsample, this comparison even overstates the difference in the implied parameters because, to the extent that the means and standard deviations do change, they tend to do so in the same direction for each pair of tests. Thus, the weighted and unweighted parameters are practically identical. The weighted μ_t's imply that the average reading skill level increases by 3.369 SD between grade 2 and 8, as opposed to 3.326 based on the unweighted parameters. Similarly, the λ_t's imply that a standard deviation difference in grade 8 only corresponds to 0.783 standard deviations in grade 2, as opposed to 0.792 standard deviations based on the unweighted parameters. Hence, all of the conclusions drawn in the paper would clearly be identical if we used the weighted versions of the parameters instead.

Table C.3: Weighted parameters

	Equating					
	2^{nd} grade	4^{th} grade	4^{th} grade	6^{th} grade	6^{th} grade	8^{th} grade
Mean	0.329	-1.000	0.322	-0.712	0.321	-0.471
p-value	*0.691*	*0.939*	*0.602*	*0.567*	*0.409*	*0.573*
SD	0.966	1.159	0.914	0.939	0.990	1.026
p-value	*0.657*	*0.894*	*0.359*	*0.260*	*0.323*	*0.538*
Obs.	1518		2140		1800	

	Parameters			
	2^{nd} grade	4^{th} grade	6^{th} grade	8^{th} grade
μ_t	0	-1.407	-2.442	-3.423
λ_t	1	1.237	1.238	1.323
r_t	0.786	0.836	0.793	0.843
p-value	*0.931*	*0.994*	*0.690*	*0.805*
Obs.	3886	2708	3188	2360

Notes: This table reports means and standard deviations for the optional test scores by the equating subsample, and the reliability ratios for the reliability subsample. In both cases, the samples are re-weighted by the estimated probability that an individual is part of the subsamples. These probabilities are estimated using a probit model using all the covariates that we control for in Table 4, column 5. The p-values are from a test of equality between the weighted parameters and their unweighted counterparts.