The perspective of “limited malleability” in educational effectiveness: treatment effects in schooling

Jaap Scheerens

To cite this article: Jaap Scheerens (2018): The perspective of “limited malleability” in educational effectiveness: treatment effects in schooling, Educational Research and Evaluation, DOI: 10.1080/13803611.2017.1455286

To link to this article: https://doi.org/10.1080/13803611.2017.1455286

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 02 Apr 2018.

Article views: 59
The perspective of “limited malleability” in educational effectiveness: treatment effects in schooling

Jaap Scheerens

University of Twente, Enschede, The Netherlands

ABSTRACT
In this article, several ways to adjust gross school effects are discussed to set the stage for estimating treatment effects in schooling. Although it is quite hazardous to hypothesize realistic benchmarks for results from meta-analyses, because of the dependency of effect sizes on subject matter area, grade level, and study characteristics, a treatment effect of \( d = .20 \) can be considered as educationally meaningful. Commonly addressed effectiveness-enhancing school factors show much inconsistency across meta-analyses and frequently small effects (in the order of \( d = .10 \)). An analysis of the conceptual and methodological foundations of defining and measuring treatment effects in schooling indicates that flaws in the measurement of treatments could be an additional cause of small effect sizes in educational effectiveness research. Results of this review are analysed from the perspective of “limited malleability” in educational effectiveness.

KEYWORDS
Educational effectiveness; effect sizes; absolute and relative effects of schooling

Introduction
The effects of students’ country, educational jurisdiction, school, teacher, and educational programmes can be considered treatment effects. But since the influence of being assigned to a certain unit would be global and unspecified, it is not normally called a treatment. Treatment effects are specific aspects of schooling and teaching, or educational programmes that may be associated with, or intended to improve, student outcomes. Treatment effects include such diverse variables as student–staff ratios, school resources, administrative arrangements, teacher qualifications, and experience in addition to specific educational interventions or programmes. Educational research often separates hypothetical causes of performance differences into “given”, “contextual”, “endogenous”, or simply “prior” conditions, on the one hand, and malleable factors, or treatments, on the other hand. This distinction corresponds to the difference between “gross” and “net” or “value-added” school effects. In this article, the first two interpretations (gross and net school effects) are referred to as “unit effects”, and the term “treatment” will be reserved for the influence of malleable factors. Treatments defined in this way include both intervention programmes and policies and practices hypothesized (or believed) to enhance...
educational performance. In the most commonly used models to estimate “net” or “value-added” school effects or treatments, school effects or treatments are residual effects after adjusting for prior conditions such as students’ socioeconomic status (SES), intelligence, prior achievement, aptitude, other student characteristics, and sometimes school factors.

Generally, adjustment factors have stronger impacts on student performance than malleable factors (e.g., Organisation for Economic Co-operation and Development [OECD], 2005; Timmermans, 2012). Several contributions in this special issue provide confirming evidence at both the school level and the level of national education systems. At the school level, the most frequently applied value-added models which show strong effects for prior achievement and general intelligence leave limited scope for treatments to improve student performance. The effect size for schools, teachers, and programmes will necessarily be small given that most of the variation is accounted for by the variables indexing initial conditions. This perspective is challenged by value-added models based on progress in educational achievement, which indicate large and very large school effects (Raudenbush, 1989; Timmermans and Van der Werf, this special issue).

Given the importance of how “net” school effects are measured, Part 1 of this paper presents an overview of the most often-used value-added models, two of which are based on adjusted student performance status and two depending on progress. In the second part, various facets of the “state of the art” of establishing treatment effects in educational effectiveness research are discussed, which have implications for interpreting substantive results from meta-analyses. In the concluding part, implications for educational policy and practice and further research are discussed, and a resume is given of the way the four research papers in this special issue speak to the theme of limited malleability.

**Part 1: gross, net, relative, and absolute school effects**

In this part, distinctions are made between gross and net, and relative and absolute school effects. Relative gross school effects are the share of total variation in student achievement attributed to schools without adjustments. “Net” school effects represent the share of variation attributable to schools adjusting for prior achievement and other student background characteristics. In approaches that try to separate the effect of a certain period of going to school, in comparison to not going to school, it is attempted to estimate effects of *schooling* that do not depend on between-school variance. This is why the effects are considered as absolute rather than relative. Examples are the assessment of the effect of summer learning and comparing the achievement of same-age children that attend a higher or a lower grade.

In Part 2 of the paper, school treatment effects are discussed. School treatment effects are the effects of modelled “treatments”, as programmes or a combination of effectiveness-enhancing factors, after adjusting for prior differences between students.

**Empirical estimates of gross school effects by means of variance decomposition**

The most common way to express the contribution of schools is to express the between-school variation as a percentage of the total variation in student achievement. This corresponds to the intra-class correlation ($\rho$ or rho), which is a proportion rather than the percentage. As well as the proportion of the variation in student performance attributable to
schools, the intra-class correlation has the useful interpretation of the expected correlation in performance of two students randomly selected from the same school (Snijders & Bosker, 2012, p. 16). According to Scheerens and Bosker’s (1997, p. 79) meta-analysis, without considering other factors, schools, on average, account for 9% of the variation in student achievement; an intra-class correlation of 0.09. This percentage varies by grade level and subject matter. Between-school differences tend to be larger for mathematics and science than for reading literacy. It should be noted that between-school differences expressed as percentages of variance or the corresponding intra-class correlations are relative measures of educational output quality.

Intra-class correlations can also be expressed as the most commonly used effect size, Cohen’s $d$ (Tymms, 2004). Cohen’s $d$ effect size is about twice the correlation coefficient. Often, it is not clear which measure is being used in educational effectiveness studies.

Relative measures of school effects can also be used to assess the relative contribution of schools versus other aggregated units, such as countries, jurisdictions, classes, and teachers. School differences tend to be weaker than differences between classrooms and teachers. Kyriakides and Luyten (2009) reported percentages of variance in student achievement in Cyprus attributed to students at 70%, 18% for classrooms, and 13% for schools. Opdenakker and Van Damme (2000) found that 13% of the variance was between schools, 18% between teachers, 15% between classes, and 54% between students. They point to the fact that omitting intermediary levels (classes, teachers) in multilevel analyses tends to overestimate the effect of the next above level, in this case schools.

Hattie (2009, p. 18) provided the following comparative effect sizes ($d$ coefficients) for students, teachers, and schools: student 0.44, teacher 0.49, and school 0.23. He concluded that teachers/classrooms matter more than schools. This pattern is usually found in other studies. International studies include a variance component (or an effect size) for the country level. The OECD (2005) presented a decomposition of total between-student variation, based on Programme for International Student Assessment (PISA) 2000 data, as shown in Table 1.

The total number of OECD countries in PISA 2000 was 27. The overall pattern Table 1 shows is that a sizeable amount of variation lies between countries, although less than that between schools.

In a study comparing 11 Australian and British educational systems in math achievement progress, Tymms, Merrell, and Wildy (2005) found more variance associated with the systems (13%) than with the classroom (7%). However, the classroom and the year group, taken together, were associated with more variance (20%) than the system. On the basis of data from the Trends in International Mathematics and Science Study (TIMSS), Kyriakides (2006) found that 20% of the variance was associated with countries.

Table 1. Percentage of variance in student performance in reading and mathematical and scientific literacy in OECD countries. Results from PISA 2000, cited from OECD, 2005, p 116.

<table>
<thead>
<tr>
<th></th>
<th>Percentage at country level</th>
<th>Percentage at school level</th>
<th>Percentage at student level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>8</td>
<td>15</td>
<td>57</td>
</tr>
<tr>
<td>Mathematics</td>
<td>16</td>
<td>31</td>
<td>54</td>
</tr>
<tr>
<td>Science</td>
<td>10</td>
<td>32</td>
<td>59</td>
</tr>
</tbody>
</table>
Relative “net” school effects in terms of adjusted student performance status

As noted in the above, between-school differences and intra-class correlations are about relative differences between schools. These summary statistics do not measure the effect of schooling but rather the relative variation between schools (Scheerens, 2007, Chapter 5).

The variance decomposition methods to estimate relative school effects are more realistic if adjustments are made by including a baseline that accounts for initial differences between students. Preferably, initial differences between students are accounted for by adjusting for same-domain prior achievement, but many other variables can be, and are, used to adjust for initial differences between students, such as socioeconomic background, aptitude, and general intelligence. When adjusting for prior achievement, the percentage of between-school variation of total variation and intra-correlations can be interpreted as the extent that schools matter. Scheerens and Bosker’s (1997) meta-analysis, which included studies that used varying types of adjustment, found that the average gross school effects declined from 9% to 4%. There is a variety of adjustment models:

- “Cross-sectional” adjustment for student background conditions, such as aptitude, socioeconomic status (SES), gender, ethnicity status, and other student variables.
- Models estimating school effects net of same-domain prior achievement.
- Combining for prior achievement as well as other student background characteristics.
  This approach is often denoted as “contextualized value added”, when classroom or school composition variables such as average SES are included (Thomas, Gana, & Muñoz Chereau, 2016).

Controlling for same-subject prior achievement provides the simplest value-added interpretation, the contribution of the school net of students’ prior subject knowledge and skills. More complex models control for prior achievement scores in other subject areas. The Education Value-Added Assessment System (EVAAS) model analyses students’ scores in all subjects controlling for students’ prior test scores in all subjects across all grades and years (Wright, White, Sanders, & Rivers, 2010).

In one contribution to this special issue, the sole use of socioeconomic status of students as an adjustment variable is criticized (Marks, this issue). Adjusting for prior achievement is a more justifiable measure, but failing the availability of prior achievement adjusting for students’ intelligence or general aptitude is justifiable. The stronger the correlations between the adjustment variables and student achievement, the smaller the value-added school effects (Marks, 2015).

In a review article, Detterman (2016) concludes:

Over the last 50 years in developed countries, evidence has accumulated that only about 10% of school achievement can be attributed to schools and teachers while the remaining 90% is due to characteristics associated with students. Teachers account for from 1% to 7% of total variance at every level of education. For students, intelligence accounts for much of the 90% of variance associated with learning gains. (p. 1)

Value-added school effects tend to be small, and they are also not particularly stable. Thomas, Peng, and Gray (2007) analysed school data over a period of 11 years in the English Lancashire district. They concluded that there was some stability in school
effects. Still, when schools were categorized as average or over- or underachieving, there were many switches, and over a period of 11 years 50% of the schools had changed category. Contrary to expectations, high value-added schools did not continuously add value over the medium term. Even less stability was found in a Dutch study, where it appeared that of the highest scoring secondary schools, only 15% were still in the top category 3 years afterwards (Vermeer & Van der Steeg, 2011). Marks (2015) reported cohort correlations of school effects in five domains ranging from a very low 0.10 to 0.30 for primary schools and from 0.16 to 0.50 for secondary schools. The lack of stability undermines the argument that once schools that substantially add value are identified, they can serve as a model to improve other schools.

**School effects in terms of student progress**

In the gross and value-added school effects models, discussed above, the dependent variable is student achievement. When student achievement and prior achievement are measured on the same scale, gain scores can be calculated as well as students’ predicted achievement and can be used as the criterion variable. Gain scores can be compared to simply adjusting for prior achievement. Based on work by Raudenbush (1989), multilevel random effects models estimate individual students’ growth curves with high levels of reliability. Various studies show substantially larger effect sizes for growth than those obtained from adjusted student performance levels: Rowan, Correnti, and Miller (2002) reported Cohen’s $d$ effect sizes of .53 for reading and .51 for mathematics in growth models. Guldemond and Bosker (2009) found intra-class correlations of 0.3 to 0.5 in growth models. Dumay, Coe, and Anumendem (2014) found that, on average, across cohorts 74% of the total slope variance was accounted for by the school-level slope variance, whereas residual gain scores analysis showed a proportion attributable to schools of just 16% on average, across cohorts. Anumendem, De Fraine, Onghena, and Van Damme (2017) showed considerable discrepancies between intra-class correlations expressed as adjusted performance status and growth; according to one of their models, this discrepancy is as high as 0.18 for performance versus 0.66 for the growth model. Palardy (2008) reported more modest discrepancies in the intra-class correlations for performance status and growth (.20 versus .23 in the full sample, .13 versus .25 and .07 versus .14 among low- and high-scoring schools, respectively). School effects obtained from growth models are only weakly associated with students’ socioeconomic status (Dumay et al., 2014; Guldemond & Bosker, 2009). Moreover, the school effects obtained from growth models were less stable than school effects obtained by adjusting for initial performance (Dumay et al., 2014). Large school effects were also found in the growth models analysed by Timmermans and Van der Werf in this special issue.

**Effects of schooling versus non-schooling**

Another approach to rule out that effects of schooling reflect external influences utilizes a counterfactual that represents “non-schooling”. Because such estimates do not depend on between-school variation, they are sometimes referred to as *absolute* effects of schooling. Examples include evaluating the effects of “summer learning” and specific applications of the regression discontinuity design. The application of the regression
discontinuity design in educational effectiveness research capitalizes on the fact that, in many countries, being assigned to a school or grade level depends on a specific age level. If the cut-off date that determines what grades students should be placed in is strictly followed, the effect of one-year schooling can be assessed by adjusting the difference in achievement between two (or more) grades in a row for the effect of age. This adjusted difference reflects the gap between the oldest students in the lower grade and the youngest in the upper grade (Cahan & Davis, 1987; Luyten, 2006). In this case, the effect of schooling is zero if the difference in achievement between grades is completely accounted for by the age of the students (Scheerens, 2007, pp. 173–174). In this approach, the “net” effect of schooling is the learning gain between adjacent grade levels that is not explained by age. Applying these methods yielded effect sizes (Cohen’s $d$) for one year of schooling, in the order of .99 (reading speed), .84 (mathematics), and .33 (when the approach of summer learning was applied) (Scheerens, 2007, Chapter 5). In another application of the regression discontinuity design, Kyriakides and Luyten (2009) found effect sizes of .87 for both mathematics and reading. Luyten, Merrell, and Tymms (2017) concluded that the net effect of schooling is usually less than half of the between-grade discrepancy effect sizes and varies between $d = .20$ and $d = .50$, depending on grade level and subject-matter area. Such effect sizes are in the range of small to medium according to Cohen’s benchmarks. On the basis of these methods, it is possible to compute the cumulative effect of schooling across grade levels; Luyten et al. (2017) reported cumulative effect sizes over six grades of primary schooling of .90 (general ability), 1.35 (mental arithmetic), 1.62 (reading), and 1.92 (general math). Ideally, the school effect, established by means of this application of the regression discontinuity design, would take care of all confounding influences from external factors, by controlling for age. This assumption can be checked by including additional student-level co-variables or interaction terms in the analyses (Perry, 2017); which would generally result in lower estimates of the school effect. This is demonstrated in the study by Luyten (2006), in which gender and number of books at home were taken into account. This caused the magnitude of the school effects to decrease, although the reduction was relatively small.

**Part 2: treatment effects in schooling**

How much difference schools, classes, or educational systems make to student achievement was discussed in the previous sections. The next question is to what extent these effects of belonging to a certain unit, or of being treated or not by schooling, can be “explained” by more specific treatments, which are malleable factors, active at specific unit levels. As stated in the introduction, treatments are both intervention programmes and factors in educational policy and practice that are amenable by agents in the educational domain and hypothesized to enhance educational performance. Of these two kinds of treatments, the emphasis in this article will not be on intervention programmes but rather on effectiveness-enhancing factors in educational practice. This implies that when discussing methodological issues and study characteristics, there will be a certain emphasis on correlational studies that exploit variance in field settings.

The discrepancy between unit (system, school, classroom) effects and treatment effects is illustrated by Hanushek (2011) with regard to teacher effects:
Literally hundreds of research studies have focused on the importance of teachers for student achievement. Two key findings emerge. First, teachers are very important. No other measured aspect of schools is nearly as important in determining student achievement. Second, it has not been possible to identify any specific characteristics of teachers that are reliably related to student outcomes. (p. 467)

Similarly, Rivkin, Hanushek, and Kain (2005) concluded that “teachers have powerful effects on reading and mathematics achievement, though little of the variance in teacher quality is explained by observable variables, such as education or experience” (p. 449). Similar conclusions had been reached by Nye, Konstantopoulos, and Hedges (2004). Rockoff, Jacob, Kane, and Staiger (2011) used a broad set of teacher characteristics, which included, in addition to experience and formal qualifications, personality traits as predictors of overall teacher effects. They found limited effects for observable teacher measures. Brandsma (1993) found that only 10% of the student variation was explained by the malleable school variables included in his study on Dutch primary schools. In contrast, Baumert et al. (2010) found sizable effects for teachers’ “pedagogical content knowledge” ($d = 0.30$) on mathematics achievement net of prior knowledge of mathematics, reading literacy test scores, mental ability, and SES (which had no effect).

The general point is that, although we have ample evidence that “teachers and schools matter”, in terms of variance components, the question on how they matter is more complex. In a subsequent section, we will look at empirical evidence on the size of treatments and malleable conditions based on meta-analyses.

Assuming an average effect of a year’s schooling in the order of $d = .40$ to $d = .50$ (Hattie, 2009; Luyten et al., 2017), this suggests that individual treatment variables in schooling that show effects in the order of $d = .20$ are important, while effect sizes around $d = .10$ would be small. Cohen’s typology would denote these effects as small to very small. The What Works Clearinghouse (2008) in the US considers programmes effective if they produce effect sizes of at least .25. However, average effects may hide considerable variation, as estimates of school effects vary between subject-matter areas and grade levels. Moreover, study characteristics may also influence average effect sizes (see the next section).

Before looking further at substantive results in terms of effect sizes for effectiveness-enhancing malleable variables, some methodological issues are addressed. These methodological issues may at least partially explain inconsistencies of treatment effect estimates in meta-analyses.

**Methodological issues**

**Study characteristics and treatment effect sizes**

Cheung and Slavin (2016) showed that higher effect sizes are associated with researcher-constructed tests rather than independently designed standardized tests, small- rather than large-scale studies, quasi-experimental rather than experimental designs, elementary rather than secondary schools, and published rather than not published work (see Table 2).

Explanations for these differences are a better control of treatments in smaller scale studies, selection bias in quasi-experiments (e.g., in the sense that subjects in treatment groups may be more motivated), and researcher-made instruments more tailored for the treated group. Publication bias occurs because journals are more likely to reject
studies with small or statistically insignificant effects and authors are reluctant to write up and submit articles with small educational effects.

Educational effectiveness research frequently uses “correlational” non-experimental research designs, given the difficulties in administering true and quasi-experimental designs. Longitudinal studies can separate the effects of malleable conditions from “given” antecedents, such as prior achievement, but in cross-sectional studies, such as international assessment studies, this is not possible. The advantages of using school effectiveness models in non-experimental causal modelling is that effects of specific school effectiveness-enhancing factors can be identified. In this way, correlational studies could be seen as having potentially high practical value, as they depend on on-going educational practice. The major problem is the low internal validity, as it is extremely difficult to rule out all disturbing internal and external conditions in such studies. When considering the distinction between experimental and correlational as a study characteristic that might influence effect sizes, low internal validity might create upward as well as downward bias in effect-size estimation.

### Relatively underutilized theories and conceptual models

In the field of educational effectiveness, there is broad consensus about the strategies and malleable conditions that matter. This is illustrated in Table 3, which shows lists of effectiveness-enhancing factors from different sources. The first source are research reviews on educational effectiveness (Hopkins, Stringfield, Harris, Stoll, & Mackay, 2014; Muijs et al., 2014; Reynolds et al., 2014), the second source are components in Comprehensive School Reform (CSR) Programs in the US, and the third source behavioural categories that are part of teacher observation instruments in the Measuring Effective Teaching (MET) study (Kane, McCaffrey, Miller, & Staiger, 2013).

The way these sets of factors developed over time is an eclectic process, involving common-sense thinking about school organization and teaching, supported by accumulating empirical research evidence, but rarely driven by theory or conceptual models. The influence from more established theories can be discerned in teaching strategies originating from behaviourism or cognitive theory (constructivism), or by organization and planning models (Scheerens, 2013a, 2015). Several conceptual multilevel educational effectiveness models have been around for about three decades (Scheerens, 1992; Stringfield & Slavin, 1992) and were also updated in more recent versions (Cree- mers & Kyriakides, 2008; Scheerens, 2016). What these models show are a choice of the factors, very much along the lines of what is illustrated in Table 3, ordered across aggregation levels (system, school, classroom/teachers, students), a distinction between malleable variables and background conditions at each level, assumptions on how factors interact, and sometimes reference to more general underlying models.

### Table 2. Study characteristics affecting effect sizes, after Cheung and Slavin (2016, pp. 287–289).

<table>
<thead>
<tr>
<th>Study characteristic</th>
<th>Effect sizes bigger</th>
<th>Effect sizes smaller</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence of instruments</td>
<td>Researcher-made:.40</td>
<td>Independent:.20</td>
</tr>
<tr>
<td>Scale of the study</td>
<td>&lt; 250:.30</td>
<td>&gt;250:.16</td>
</tr>
<tr>
<td>Study design</td>
<td>Quasi-experimental:.23</td>
<td>Experimental:.16</td>
</tr>
<tr>
<td>Grade level</td>
<td>Elementary:.20</td>
<td>Secondary:17</td>
</tr>
<tr>
<td>Publication statues</td>
<td>Published:.30</td>
<td>Grey literature:.16</td>
</tr>
</tbody>
</table>
dimensions of the specific factors. A basic notion is that malleable conditions at a higher organizational level facilitate malleable factors at lower levels, and an example of a more general underlying dimension is the issue of alignment versus loose coupling of factors within and across levels. Causal modelling in which moderators and mediators are distinguished and explored by means of path analytic techniques is a beginning of empirically testing some of these model assumptions (Desimone & Hill, 2017; Scheerens, 2013b). Best practices in the development of school improvement programmes in which effectiveness-enhancing school and teaching conditions are combined are the earlier cited Comprehensive School Reform (CSR) Programs in the United States and applications of the dynamic model developed by Creemers and Kyriakides (2008). Still, consequently thinking through how combinations of factors are expected to stimulate student learning is a relatively underdeveloped area in this field. On the other hand, we see theoretically ungrounded and empirically unsupported ideas getting a great deal of publicity. In particular, there are currently high expectations on the effects of

### Table 3. Effectiveness-enhancing factors, from reviews, CSR, and observation of effective teaching.

<table>
<thead>
<tr>
<th>Research reviews, Scheerens (2014)</th>
<th>CSR (Borman, Hewes, Overman, &amp; Brown, 2003)</th>
<th>MET observation categories (Kane, Kerr, &amp; Pianta, 2014, p. 246)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective Leadership</td>
<td>Integrative school management; Shared leadership</td>
<td>The majority of students were on task throughout the class.</td>
</tr>
<tr>
<td>Academic focus</td>
<td>Measurable goals for student academic achievement</td>
<td>The teacher used formative assessment effectively to be aware of the progress of all students</td>
</tr>
<tr>
<td>A positive orderly climate</td>
<td></td>
<td>An appropriate amount of time was devoted to each part of the lesson.</td>
</tr>
<tr>
<td>High expectations</td>
<td></td>
<td>Content communicated through direct and non-direct instruction by the teacher is consistent with deep knowledge and fluency with the mathematics concepts of the lesson.</td>
</tr>
<tr>
<td>Monitoring progress</td>
<td></td>
<td>The lesson allowed students to engage with or explore important concepts in mathematics (instead of focusing on techniques that may only be useful in exams).</td>
</tr>
<tr>
<td>Parental involvement</td>
<td></td>
<td>The lesson was well organized and structured.</td>
</tr>
<tr>
<td>Staff professional development</td>
<td>Meaningsful involvement of parents and the local community</td>
<td>The lesson included an investigative or problem-based approach to important concepts in mathematics</td>
</tr>
<tr>
<td>Effective teaching (time)</td>
<td>High-quality and continuous teacher and staff professional development</td>
<td>The classroom environment encouraged students to generate ideas, questions, conjectures, and/or propositions that reflected engagement or exploration with important mathematics concepts.</td>
</tr>
<tr>
<td>Opportunity to learn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structuring and scaffolding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching learning strategies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pupil involvement/active teaching</td>
<td>Employed proven methods for student learning, teaching, and school management that are founded on scientifically based research and effective practices and have been replicated successfully in schools</td>
<td></td>
</tr>
</tbody>
</table>
school organization and management on student performance due to several reports by McKinsey (e.g., Moursheed, Chijioke, & Barber, 2010).

**Few standardized instruments to measure effectiveness-enhancing conditions**

Despite the fact that the list of input and process factors considered to enhance student achievement has been remarkably consistent for about four decades, there are few standardized measures. This lack of standardized instruments is a weakness (e.g., Muijs & Brookman, 2016; Scheerens, 1992). The dominant practice is the development of new instruments for each new study. Two examples of standardized instruments are: the Hallinger scale, which measures instructional leadership and has been applied in many contexts and studied in about a hundred research studies (Hallinger & Wang, 2015), and the International Comparative Analysis of Learning and Teaching (ICALT) instrument, which measures teaching (Van de Grift, Chung, Maulana, Lee, & Helms-Lorenz, 2017).

**Possibly reactive and confounded measurement of factors**

In contrast to the attention given to the outcome measures, the quality of the instruments that measure treatment variables is rather neglected (lack of conceptual rationales, few standardized instruments, and each study reinventing the wheel).

Self-reports from teachers are vulnerable to social desirability and confounding of process and outcome measures. This is illustrated by the variable “high expectations of student performance”. This variable can be understood as a positive and optimistic attitude that stimulates student engagement and performance. On the other hand, it may simply reflect teachers’ knowledge about actual student performance (Brophy, 1983). Furthermore, little attention is given to the interdependence of factors, when several are included in one study. Because of multi-collinearity, the estimate of the effect sizes is dependent on the set of other factors included in the model (Scheerens, 2014).

**Results from meta-analyses**

Empirical results on the effects of treatments in schooling comprise a research literature denoted as school or educational effectiveness research. As already indicated above, reviews of educational effectiveness are generally consistent in listing which treatments and effectiveness-enhancing factors are effective. These include instructional leadership and cooperation between teachers, while a range of other variables are more closely associated with teaching and learning: opportunity to learn, time investment, achievement orientation, parental involvement, structured teaching, school and classroom climate, and frequent monitoring and evaluation.

Despite consensus on which factors matter, meta-analyses show considerable divergence in mean effect sizes for each of these factors. For example, at the school level, Marzano (2003) reports an average effect size for “opportunity to learn” of 0.88, while Hattie (2009) reports an effect size of 0.39 for a related variable “enrichment programmes for gifted children”. For “monitoring”, Hattie reports an average effect size of .64, while Marzano reports 0.30. “School leadership” has a small effect size of around 0.10, according to Marzano (2003) and Scheerens (2007), whereas Hattie’s (2009) estimate is 0.34. The pattern across these three meta-analyses is similar for “cooperation” but with much smaller coefficients (.06, .04, and .18). At the classroom level, Seidel and Shavelson (2007) reported an effect size as low as .08 for a combination
of learning time and opportunity to learn, while Hattie’s estimate was 0.36. A recent set of related meta-analyses found small effect sizes of 0.10, 0.10, 0.14, 0.10, and .12 for “learning time”, “homework”, “monitoring”, “assessment”, and “school leadership”, respectively (Hendriks, 2014; Scheerens, 2012, 2014). The small effect size for leadership stands in sharp contrast to the optimistic accounts in the school improvement literature (e.g., Day et al., 2009).

Other evidence of treatment effects
The Measuring Teacher Effectiveness study (MET), funded by the Bill & Melinda Gates foundation, investigated different methods to assess teacher effectiveness (Kane et al., 2014). Three types of measures were investigated and compared:

1. past performance of teachers based on value-added achievement (VAM) of students taught in previous school years;
2. student ratings of teacher quality;
3. observation of teachers’ teaching.

The three methods to assess teaching effectiveness are fundamentally different. VAM studies exhibit much greater predictive validity than the other two approaches. The VAM method is the most statistically defensible approach but does not reveal anything about which aspects of teaching are more effective. With respect to teaching effectiveness, this measure says about as much or as little as the studies that, by means of variance decomposition, established a relevant teacher effect, in the sense of the difference it makes to be taught by one teacher or the next. The student rating measures are subjective ratings of teachers aggregated for each teacher. The observation method is based on observers categorizing behaviours that are considered as stimulating achievement, and is therefore covering specific effectiveness-enhancing teaching factors. Results on the predictive validity of each of these three methods are relevant to our discussion on the size of effective treatments in schooling. Results indicate that the past performance measure based on earlier VAM is by far the best predictor, with standardized regression coefficients in the order of .40, consistent over subject-matter areas and school level (elementary and secondary). Coefficients for student ratings and classroom observation are considerably lower, in the order of .15 and .10, respectively. Moreover, the rating and observational coefficients are much less consistent across subject-matter areas and grade levels (Kane et al., 2014). The practical implication is that observed teacher behaviour is considered to have a predictive validity that is too low to support application of this method as the sole procedure to assess teacher effectiveness. In the context of the present analysis, these findings, particularly the relatively low predictive validity of behavioural observation of effectiveness-enhancing teaching conditions of .10, corroborate previously discussed results all pointing to small treatment effects.

A second interesting case are the results of meta-analyses of Comprehensive School Reform (CSR) Programs (Borman et al., 2003). CSR programmes are based on best-evidence syntheses of educational effectiveness research on specific effectiveness-enhancing conditions. As such, evaluations and meta-evaluations of the results of these programmes are crucial tests of the available knowledge base. According to Borman (2009):
Our various analyses suggest that… CSR schools can be expected to score between nearly one-tenth and one-seventh of a standard deviation… higher than control schools on achievement tests. The low-end estimate represents the overall CSR effect size of $d = .09$ for third-party studies using comparison groups, and the high-end estimate represents the effect size of $d = .15$ for all evaluations of the achievement effects of CSR. (p. 55)

He concludes that large-scale reform is capable of widespread, but only modest, improvements in student achievement.

**Discussion**

*Relatively small effect sizes of treatments, long lead time of educational reforms to take effect, and equifinality of treatments representing classroom teaching*

Overall effects of schooling, expressed as the cumulative effect of attending 6 years of primary school, are by no means small. Moreover, multiplier effects of school-level conditions over large numbers of students should be considered as well. But the effects of malleable aspects of schooling, or treatments, are small when compared to the influence of stable student characteristics. This is underlined in recent studies where adjustments are made for intelligence, aptitude, and/or previous achievement, as demonstrated in other contributions to this special issue. Treatment effects vary across subject-matter area and grade level, so that benchmarks for interpreting effect size should ideally be specific for subject-matter, grade-level combinations. Effects tend to be larger for subject matter that is more exclusively offered at school (e.g., mathematics), as compared to subjects that are supported by “the informal curriculum” of home and other societal contexts (e.g., reading). Results of meta-analyses on the effectiveness of malleable school variables and treatments vary not only because of subject and grade differences but also because of variation in study characteristics, such as externality of achievement measures, scale of the study, and study design.

Despite broad consensus on the kind of factors that “work” in schooling, explicit rationales and models, let alone theory-based conjectures, are often lacking in educational effectiveness research. Besides, there are few standardized instruments to measure the effectiveness of school programmes and interventions, and the wide-spread use of self-reports in those studies is likely to result in respondents providing socially desirable answers.

When taking results from meta-analyses as a basis for assessing the state of the art of treatment effects in schooling, the considerable variation in average effect sizes is not surprising given the various sources of this variation, namely, different model specifications, study characteristics, and no standardization of treatment measures. The diversity in results of meta-analyses makes it difficult to draw overall conclusions, but the norm in school effectiveness studies is small effect sizes, below the benchmark of $d = .20$, which is seen as indicative of “educational significance”. Meta-analyses on effective teaching strategies tend to show higher effect sizes than school effectiveness studies (e.g., De Boer, Donker, & Van der Werf, 2014; Donker, De Boer, Kostons, Dignath-Van Ewijk, & Van der Werf, 2014; Hattie, Biggs, & Purdie, 1996), in the order of $d = .40$ to .80. At the same time, there are also meta-analyses on teaching factors that show very small effect sizes (Creemers & Kyriakides, 2008; Scheerens, 2007; Seidel & Shavelson, 2007). A hypothetical explanation for the difference in effect sizes between school and teaching effectiveness
studies could be the larger scale of field studies that take schools and school-level strategies as a focus in comparison to smaller scale experimental teaching effectiveness studies.

In teaching effectiveness studies, quite different approaches may show similarly sized positive effects (Hickendorff et al., 2017), sometimes interpreted as “everything works” (Hattie, 2009, p 15). A striking example are studies that find no significant differences between structured teaching strategies and constructivist strategies (Kirschner, Sweller, & Clark, 2006; Louis, Dretzke, & Wahlstrom, 2010). These kinds of results can be subsumed under the concept or equifinality, as defined in general systems theory (von Bertalanffy, 1968, p. 40). Equifinality is the principle that in open systems a given end state can be reached by many potential means. Equifinality adds to an overall impression of relative “indifference” and “bluntness” in the way schools operate in producing learning progress. If “many roads lead to Rome”, a general sense of direction might be sufficient to reach the destination, albeit slowly. If basic conditions of good schooling are in order (knowledgeable teachers, an orderly environment, general sense of direction), it may not make much difference which specific methods are applied. This conclusion could be taken as urging the search for more basic underlying conditions of good schooling (Scheerens, 2016). Relatively small treatment effects in large-scale field studies and equifinality in teaching effectiveness should temper the high expectations associated with reform agendas expressed in some of the literature on systemic reform and school improvement (Day et al., 2009; Mourshed et al., 2010)

**The contribution of the four empirical studies reported in this special issue**

The review presented in this paper tends to conclude that malleability in educational effectiveness is limited. The conclusions on “limited malleability” are partly confirmed and partly challenged in the four empirical studies that make up this special issue.

The first paper in this special issue, by Marks, demonstrates small “net” school effects, when adjustments are made for subject-specific and across-subject prior achievement. The outcomes of this study were placed in the perspective of research that examined the impact of heredity and intelligence on educational achievement.

In the second paper, by He, Van de Vijver, and Kulikova, a broad range of “non-educational” contextual influences was correlated with student achievement. It appeared that not all plausible educational policy levers that were considered had the expected positive correlation. This paper can be seen as innovative for studying system-level educational effectiveness, as it lays the foundation for a fuller treatment of cultural conditions that affect educational performance.

In the third paper, by Aloisi and Tymms, average performance of countries shows much stability according to six waves (2000–2015) of OECD’s PISA study. In addition, the paper illustrates that in jurisdictions where relatively large changes occurred, these changes were to a large extent due to contextual and not to educationally malleable factors. Finally, the one educationally malleable variable that was studied in depth, namely, curriculum innovation, showed hardly any effect. The methodology that was used in this study has several interesting features, such as the use of growth curve analyses of country-level educational performance and the way “curriculum change” was defined and longitudinally measured in the study. Using country-level longitudinal data to measure progress in educational
performance makes it possible to define and empirically address the temporal aspect of treatment measurements, in the sense of defining the initiation and the effective duration of a treatment.

In the fourth paper, by Timmermans and Van der Werf, growth curve analysis was applied to a longitudinal data set based on the monitoring of student progress in Dutch primary schools. The results appear to challenge the hypothesis of “limited malleability”, as relatively large “value-added” effect sizes were established using growth models. Similar large school effects have been found in other school effectiveness studies that had used growth curves analyses. Dumay et al. (2014, p. 77) argue that the internal validity of the growth model is high, because the growth estimates are more efficient to isolate the part of the between-school variance in performance that is less associated with exogenous factors such as the school’s social composition. At the same time, they prove to be less stable, and therefore less reliable to predict future success. Given the relatively limited set of studies on progress using growth curve modelling, there are still open questions about the impact of adding additional control variables. Moreover, high “net” school effects are “unit effects” and not treatment effects as defined in this article. High “net” unit effect sizes do not guarantee high treatment effect sizes, which is illustrated in the study by Rowan et al. (2002). These authors found net school effects in the order of $d = .50$, and a treatment effect for “opportunity to learn” of .10. Another facet of the specific behaviour of growth curves estimates is the way they should be interpreted. The more traditional variance composition models facilitate a relatively simple additive interpretation of the impact of contextual variables and treatment effects. Such interpretations could be seen as useful in practical applications such as school self-evaluations and external school evaluation. The relatively low stability of the growth curves coefficients over time, as compared to estimates based on adjusted performance status (Dumay et al., 2014; Timmermans & Van der Werf, this issue), would make this method more problematic for such practical applications. But if growth curves analysis would indeed prove the most internally valid way of assessing net effects of schooling, it might become the preferred method in educational effectiveness research.

**Conclusion: making up the balance for the “limited malleability” thesis**

At the end of the line, it has to be acknowledged that, in this paper, the hypothesis of “limited malleability” in educational effectiveness cannot be confirmed or rejected. The exploration that was carried out could not meet the broad range of phenomena and the sheer magnitude of relevant research that are subsumed under the thesis. As compared to a fully representative review, this paper’s status is more like a position paper and a possible upbeat for future studies that focus on facets of the overall thesis. The research that was reported leads to some results that support the thesis and other results that challenge it. Notwithstanding the nuances in the summary of main results presented below, there is sufficient ground to keep the hypothesis of “limited malleability” in educational effectiveness on the table for further research and analyses, hopefully also inspiring a more realistic and efficient attitude to system-level educational reform and school improvement.

Value-added school effects tend to be small when co-variance adjustments are made for prior achievement. The argument for the persuasiveness of this method is
corroborated by studies that found high correlations between student background variables, such as intelligence, and educational achievement. Another supportive indicator for this conclusion about small school effects is high stability of student achievement across time.

According to the logic of value-added analyses, the size of treatment effects depends on the share of total student-level variation that remains unexplained by previous achievement and other student background variables. This sets an upper limit for the size of specific effectiveness-enhancing (treatment) variables and is therefore a straightforward explanation for modest treatment effects, assuming strong influence of the adjustment variables.

Estimates of school effects expressed as student achievement growth over more than two points in time by means of growth curves analyses are astonishingly high. Conceptually, the coefficients yielded by this method are interesting as they can be thought of as efficiently isolating the part of the between-school variables from exogenous variables (Dumay et al., 2014). Yet, the slope ratios on which the school effects are based may be lacking a sound statistical rationale. The evidence on how the growth effects “behave”, in terms of stability and sensitivity to student background variables, is sparse and, to the degree that it exists, not quite convincing.

Effects that are computed by comparing the learning progress of same-age students in different grades are sometimes indicated as “absolute” school effects, because they do not depend on between-school variance (and are therefore described as effects of schooling). Although they are not strictly to be seen as “value added”, one could say that the comparison is controlled for age. Effect sizes for a year of schooling are about $d = .50$, on average. The feasibility of this method of estimating effects of schooling for assessing treatment effects of malleable school factors depends on there being sufficient eligible students per school, which might prove to be difficult in practice.

The field of educational effectiveness research manifests a relatively strong consensus on which factors matter. Meta-analyses show strongly diverging effect sizes per factor. Effect sizes for treatment variables differ for subject matter area, grade level, study characteristic, and for school organizational as compared to teaching factors. This diversity makes it extremely difficult to draw average conclusions about effect sizes being high or low, although the logic of “narrow margins” resulting from properly conducted value-added analyses makes modest effects more plausible.

Effect sizes of school-level malleable conditions tend to be lower than factors that measure teaching characteristics. The fact that school effectiveness studies are more often large-scale field studies with non-experimental designs and teaching effectiveness studies are frequently small-scale experimental or quasi-experimental studies might partially explain this.

The impression that, in teaching effectiveness studies, approaches based on strongly different schools of thought (structured teaching vs. constructivism) show quite similar effect sizes can be seen as proof of “equifinality”, different methods equally effective in attaining a fixed goal. Equifinality could be seen as indicative of schooling being relatively robust with respect to all kinds of methodological variations, conditional on the presence of basic conditions of well-functioning. Equifinality is another way of saying that educational systems are difficult to change in ways that matter, and as such just another manifestation of limited malleability.
Conservation and stability is also what is demonstrated in studies that compared growth or decline of country-level achievement in international data sets. The study by Aloisi and Tymms (this issue) clearly demonstrates this, as have other analyses. Empirical studies that have analysed the effects of system-level levers of educational performance represent a relatively young field, so far not showing convincing effects, as effect sizes tend to be very small and inconsistent between studies. The two studies in this special issue that carried out system-level analyses (Aloisi & Tymms, this issue; He et al., this issue) underline the importance of conditions and influences outside the educational province (such as cultural factors) that may complicate assessing the impact of educational policy measures.

Relatively small treatment effects in large-scale field studies, high stability in the performance of educational systems over time, and equifinality in teaching effectiveness call for a more conservative attitude toward educational reform and school improvement. Little is to be expected of governments turning out a continuous stream of short-lived policy initiatives and “innovations”. A similar kind of overproduction of change in organizational conditions at the school level under the heading of “continuous improvement” is likely to have disappointing results as well. Maintaining and incrementally improving educational performance is best served by consistent support of basic conditions of good schooling close to the primary processes of teaching and learning, in the realm of teacher training, curriculum alignment, performance monitoring, and formative assessment.

Notes
1. The effect sizes in the section “Results from meta-analyses” are $d$ coefficients.
2. Comprehensive school reforms are based upon scientifically based research and effective practices that include an emphasis on basic academics and parental involvement so that all children can meet challenging academic content and academic achievement standards.

Acknowledgement
The author is indebted to Gary N. Marks for input and comments on previous versions of this article.

Disclosure statement
No potential conflict of interest was reported by the authors.

References
Borman, G. D. (2009, March). National efforts to bring reform to scale in America’s high-poverty elementary and secondary schools: Outcomes and implications. Paper commissioned by the Center on


Detterman, D. K. (2016). Education and intelligence: Pity the poor teacher because student characteristics are more significant than teachers or schools. The Spanish Journal of Psychology, 19(E93), 1–11. doi:10.1017/sjp.2016.88


Hendriks, M. A. (2014). The influence of school size, leadership, evaluation, and time on student outcomes: Four reviews and meta-analyses (Doctoral thesis). Enschede: University of Twente.


Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of elementary schools. *Teacher College Record, 104*(8), 1525–1567. doi:10.1111/1467-9620.00212


