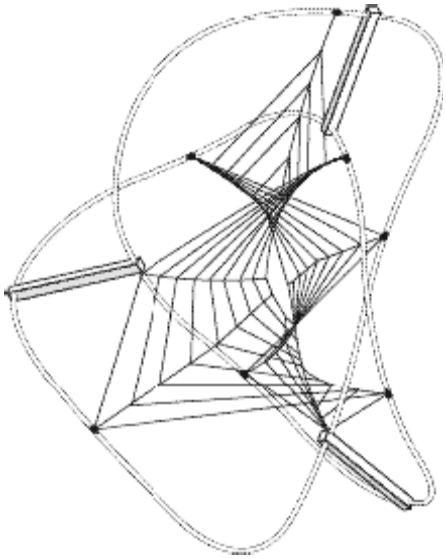
LSE CPNSS

Centre for the Philosophy of Natural and Social Science
Contingency and Dissent in Science
Technical Report 03/09

# *How to tell when efficacy will NOT translate into effectiveness*

Chris Thompson



Series Editor: Damien Fennell

The support of The Arts and Humanities Research Council (AHRC) is gratefully acknowledged. The work was part of the programme of the AHRC Contingency and Dissent in Science.

Published by the Contingency And Dissent in Science Project
Centre for Philosophy of Natural and Social Science
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE



ISSN 1750-7952 (Print)
ISSN 1750-7960 (Online)

# How to tell when efficacy will NOT translate into effectiveness[1]

Chris Thompson

**Editor's Note**

Thomspon's paper contributes to the contingency side of the Contingency and Dissent in Science project. Specifically, it draws methodological lessons for evidence based policy from the failure of the California Class Size reduction programme. He introduces an important distinction between failure of external validity due to external confounders and those that are introduced by intervention to the system. Thompson argues that the latter should be easier to predict and he describes how the California Programme is a case of this. He also develops a useful analysis of causal chains as a way of identifying potential confounders.

## Abstract

I aim to show that the failure of the California Class Size Reduction initiative highlights an important class of situations in the application of evidence to policy. There are some circumstances in which the implementation of a policy will be self-defeating. The introduction of the factor assumed to have causal efficacy into the target population can lead to changes in the conditions of the target population that amount to interfering factors. Because these interfering factors are a direct result of the policy implementation they should be relatively easy to predict, and so part of the tricky issue of judging where evidence is relevant should in some circumstances be relatively straightforward.

The failure of the California Class Size Reduction initiative also shows how important it is to identify the correct causal factor. The more accurate the attribution of causality, the less susceptible it will be to interfering factors and breaks in the causal chain.

## 1. Class Size Reduction – A case study

It seems intuitively plausible that reducing class sizes will improve the attainment of pupils. And there is a wealth of evidence that addresses this issue.[2] However the evaluation of the Californian Class Size Reduction (CSR) programme found that there was inconclusive evidence of a link between the reduced

[1] Thanks are due to Nancy Cartwright for her extensive comments on this paper, and on earlier drafts.

[2] See Blatchford, P. & Mortimore, P (1994).

class sizes in Californian schools and student attainment (Bohrnstedt, G.W. & Stecher, B.M (eds.) (2002)). It is important to account for this policy failure.

In the mid 1990's there was concern about the standard of primary school education in California. In 1996 the state legislature passed a bill that provided funding to districts for each student in a class smaller than 20 students. The cost of the programme was approximately US$1.6 billion per year. There was evidence available to the legislators[3] which, when carefully considered, lends support to the view that this programme should have been expected to be successful. This evidence included correlation studies (which seemed to show an association between *large* class sizes and improved attainment[4]); meta-analyses (of varying quality); and evidence from a randomised controlled trial (RCT) (Blatchford, P. & Mortimore, P (1994)). This last piece of evidence, the Tennessee STAR programme, was particularly significant and is frequently cited.

The Tennessee STAR programme was a four-year longitudinal study looking at the extent to which smaller class sizes lead to improved academic attainment in children. The study involved 79 schools and around 7000 children. Children and teachers were randomly assigned to small classes of 13-17 children, normal sized classes of 22-25 children and normal sized classes with an extra teacher aide. The children were of kindergarten to third grade age.

The results from the programme were clear; children in the smaller classes did better in reading and maths. This was particularly true for the early grades, minority children and children living in poverty. It has been suggested that the key causal factor in improved attainment was greater pupil engagement. When classes are smaller, there is less opportunity for a child to avoid being involved. In addition, in

[3] A summary of the evidence available is provided in Blatchford, P. & Mortimore, P (1994). Illig, D.C. (1996) provides the evidence actually presented to the policy makers at the time.
[4] A variety of explanations have been put forward to explain the counterintuitive result that large classes are associated with improved attainment. For instance, it may be that more competent teachers are given larger classes, or perhaps struggling students are placed in smaller classes.

smaller classes less time is spent on class management. There was no apparent difference in the teaching practices in the smaller classes. (Finn, J.D. & Achilles, C.M. (1999)).

Given this evidence, why was the California CSR programme a failure? The answer to this question will help address the wider issue of translating evidence of efficacy into effectiveness i.e. of providing guidance for how evidence of causality seen in RCTs can be replicated in real world situations.

## 2. Validity, efficacy and effectiveness

Internal validity is an assessment of the quality of an inference of efficacy in scientific studies. An assessment of internal validity is an assessment of whether the inference that x caused y in experiment z is justified. Internal validity is an epistemological notion.

External validity is an assessment of the quality of an inference of effectiveness in a wider (universal or population based) setting. An assessment of external validity is an assessment of whether the inference that x will cause y (*simpliciter* or in a target population) is justified. Again, this is an epistemological notion.

Efficacy is the ability of a treatment to produce benefit if applied ideally. Effectiveness is the benefit that actually occurs when a treatment is used in practice. These are both metaphysical notions.

Internal validity and external validity, efficacy and effectiveness are four different but related notions.[5] Internal validity is an assessment of the warrant for claims of efficacy. External validity is an assessment for claims of effectiveness. An RCT can be used to determine the efficacy of a particular

[5] See Cartwright, N.D. (2009) p.6.

treatment. The design of the RCT can be assessed for its internal validity. The tricky thing is making use of the results of an RCT in real world settings. There is no universal algorithm for doing this.

Thus for the issue of class size and student attainment we can ask four questions:

1) Was it legitimate to infer that in the Tennessee STAR programme, small class sizes led to improved attainment in the students? (internal validity)
2) Was it legitimate to infer that the results from the Tennessee STAR programme can be replicated in California, that reducing class sizes in California will lead to improved attainment of those students? (external validity)
3) Is it the case that, given the right circumstances, small class sizes will lead to improved student attainment? (efficacy)
4) Is it the case that, given the circumstances that actually exist, small class sizes will lead to improved student attainment? (effectiveness)

Note that (2) is a different question to (4). (2) is an epistemological question about whether a particular claim was justified. (4) is a metaphysical claim, about whether causes operate in a given circumstance.[6]

The answers to the four questions above are:

1) Yes: the inferences that in the Tennessee STAR programme small class sizes led to improved attainment in the students had good foundations (Finn, J.D. & Achilles, C.M. (1999)).

2) No: the inference that the results from the Tennessee STAR programme can be replicated in California, that reducing class sizes in California will lead to improved attainment of those

[6] And because these are different questions it is possible for them to have different answers. So, for example, it is possible that a given treatment will as matter of fact be effective without us having sufficient evidence to warrant claims of external validity.

students, was shown (ex post) to be without foundation by the results of the evaluation (Bohrnstedt, G.W. & Stecher, B.M (eds.) (2002))

3) Yes: small class sizes can have causal efficacy, as shown in the results of the Tennessee STAR programme (explanations for the causal efficacy can be seen in Blatchford, P. & Mortimore, P (1994))

4) No: the real world settings of the implementation of the California CSR initiative meant that small class sizes were not effective at improving attainment.

It is this last question that I want to pursue. I will argue that there are at least two important sub-categories of effectiveness and, relatedly, that there is an important distinction that needs to be made in assessments of external validity.

**3. Efficacy and effectiveness**

If a treatment is not effective it could be due to either of two reasons, as set out below.

1) A treatment could lack effectiveness because it lacks causal efficacy i.e. even in ideal settings it has no causal impact. For example, a programme of providing antibiotics to the entire population will not reduce the incidence of polio (it is not effective) because even in a laboratory antibiotics do not destroy viruses (they lack efficacy).

2) Although the treatment has causal efficacy, it is not effective because other factors not present in the ideal experiment (or factors present in different proportions) interfere with the causal impact in the wider population. Within this category there are two broad sub categories:

2a) The treatment has causal efficacy, but there are *pre-existing* features of the treatment population which are interfering factors. For example, although experimenters can ensure that subjects take the full course of antibiotics, real patients will often forget to take their dosage, or stop part way through a course of antibiotics. So the antibiotics are efficacious but lack effectiveness because of pre-existing characteristics in people's behaviour.

2b) Although the treatment has causal efficacy, application of the treatment to a wider population *generates* other factors not present in the ideal experiment which interfere with the causal impact in the wider population. For instance, cautious prescription of antibiotics can be efficacious in treating bacterial infections. But widespread usage of antibiotics can lead to evolved resistance in bacteria. This bacterial resistance was not a pre-existing phenomena, the introduction of the antibiotics to the community *generated* the resistance, it was a self-defeating move.[7]

Careful considerations of the methodology should be able to rule out situations such as (1). Policy makers should be attempting to establish efficacy before they assume or test whether a factor will be effective. And this is a matter of internal validity, whether the conclusion that there is efficacy is a legitimate inference to draw. Situations of type (2a), where there are pre-existing interfering factors in the wider population, may be difficult to predict. This is because we may not have a detailed understanding of the population to which we are applying the treatment, we may not understand precisely how the treatment works and we may not understand what might oppose or dilute the treatment. Even if we have a reasonable understanding of the causal mechanisms at play and the

[7] I am not insisting that the categories of (2a) and (2b) are mutually exclusive, that an interfering factor is *either* a pre-existing feature *or* it is introduced by the treatment (but not both); it may be a matter of degree. There may be some cases where the lack of effectiveness is the result of an interaction between pre-existing and introduced interfering factors. All I need for there to be a useful distinction between (2a) and (2b) are cases in which the interfering factors would not exist in the treatment population were it not for the introduction of the treatment itself.

potential interfering factors for these causal mechanisms, we may not know whether or not the interfering factors exist in the treatment population (or whether there are some factors present in different proportions in the treatment population which interfere with the treatment) . However, situations of type (2b), where interfering factors are a direct consequence of the policy itself, should be much easier to predict. This is because policy makers have some control over the introduction of the policy, and from their understanding of how the policy works in ideal circumstances they should be able to see how their policy can introduce interfering factors.[8]

## 4. Explaining the policy failure in California

The next task is to establish which of these categories the California Class Size Reduction (CSR) initiative fits into. Given that class size was *not* effective at improving attainment in California was this because (1) class size is not efficacious for improving attainment; or (2a) class size is efficacious but there were some pre-existing interfering factors in California; or (2b) class size is efficacious at improving attainment, but the introduction of the policy in California also introduced some interfering factors?

But before I address this question I want to set to one side two possible complications. It is arguable that policy makers gave insufficient consideration to the evaluation of the policy. The policy was rolled out so rapidly that there were insufficient 'control' schools (i.e. schools with large class sizes) to compare results with. Thus the evaluation was *inconclusive* and did not show that the policy was a failure or success. Nevertheless, the expectation of the policy was that, even given these complications, there should have been at least some evidence of improved attainment as a result of small class sizes.

[8] We can of course investigate the population more thoroughly before implementing a treatment to see if there are any potential pre-existing interfering factors. We can also investigate the population to see if there are factors present in different proportions to the experimental population, such that they may interfere with the effectiveness of the treatment. But we are at somewhat of an 'advantage' if potential interfering factors will be generated by the introduction of the treatment itself. Because we know that the treatment has causal efficacy we should have some idea of how the mechanism works and so we should be able to predict and counter the generated interfering factors.

The expectation was that small class sizes would lead to improved attainment, and that if there was improved attainment then it would be observable.

It could similarly be argued that small class sizes are not causally efficacious at improving attainment in students per se. Rather, small class sizes are causally efficacious at improving attainment of the early grades, minority children and children living in poverty. So the policy in California was perhaps based on the wrong kind of evidence. The evidence from the Tennessee STAR programme might be relevant if California wanted to improve attainment for the early grades, minority children and children living in poverty. But given that California apparently wanted to improve attainment across the board, or at least wanted to roll out the policy of smaller class sizes across the board, then the Tennessee STAR evidence was irrelevant. However in the implementation of small class sizes in California, children in early grades, minority children and children living in poverty were also the beneficiaries of the treatment. And their gains should have been seen in the wider results – either as an impact on the average gain in attainment or at least in the disaggregated data. So again, given that there are good grounds for efficacy, this should have translated into effectiveness.

There is at least some evidence[9] that small class sizes have causal efficacy, and that as a consequence the failure in the California CSR cannot be explained as an instances of kind (1) above.

Was the lack of effectiveness of small class sizes in improving attainment in California then a type (2a) or type (2b) situation i.e. if the failure of the policy was a result of interfering factors, were these factors pre-existing in California or did the CSR policy itself introduce the factors? To shed light on this we need to look at the facts of the policy implementation. The policy implementation of the CSR programme required that districts first reduce all first grade class sizes in schools, followed by all second grades and finally by either kindergarten or third grade classes. The programme was rolled out in California rapidly, with little time for schools to prepare. This meant there was a big increase in

[9] Finn, J.D. & Achilles, C.M. (1999).

demand for teachers and for classrooms, a demand that well exceeded supply. Implementation lagged in schools serving minority and low income students, in part because they lacked adequate classroom space. As a consequence most of the unqualified teachers ended up in the schools with the most disadvantaged students.

As Blatchford (2003) states: "The evidence from the California Class Size Reduction Program is that teacher quality was vital. It is now recognized that the haste in finding teachers to implement class size reductions meant that inexperienced teachers were hired and this made the initiative less effective than it should have been."[10]

Poor teacher quality could then count as an interfering factor. Small class sizes are efficacious and will be effective provided that the teachers are of adequate quality. The question then is whether poor teacher quality was a pre-existing condition in California or whether poor teacher quality was introduced along with the CSR. If poor teacher quality was already present in California then this would be a situation akin to (2a). If poor teacher quality was a factor introduced alongside the small class policy, then this is a situation akin to (2b). The analysis points to (2b) – the drop in teacher quality (the interfering factor) was a direct consequence of the decision to decrease class sizes (the treatment).[11] Note that so far this issue has been treated as a metaphysical one, we have established why small class sizes are efficacious but not effective and improving attainment. But the implementation of policy needs to address issues of validity, in particular external validity: when is it legitimate to infer that a relationship between small class size and attainment seen in one setting will also be seen in a different setting?

[10] Blatchford, P. (2003) pp.151-152

[11] Or more particularly, the drop in teacher quality was a direct consequence of the decisions around policy implementation, of the decision to decrease class sizes rapidly.

**5. External validity**

Arguably, because we have a situation in which it was the interfering factors introduced by the CSR that were the reason the smaller class sizes were not effective, this should have been predicted. Note that I am not providing a criticism of the actual CSR policy as implemented in California, nor of the actual decisions made; there may well have been those who did foresee the problems and there may well have been good reasons for the way in which the policy was actually implemented. I am instead making normative methodological points in this paper, I am suggesting how the evidence on small class sizes should have ideally been used. It should have been clear that the inference that small class sizes would improve attainment in California lacked external validity (given the particular way in which the policy was to be implemented there).

Evidence was available at the time that if the number of children in the California school system remains constant and class size decreases, then there will be a need for more teachers. If the pool of fully qualified teachers is limited (or at least cannot be expanded rapidly), then less qualified teachers will have to be hired. Therefore evidence was available at the time that the CSR programme would lead to reduced teacher quality. The only remaining issue is whether policy makers should have been aware that poor teaching quality is an interfering factor.

Blatchford, P. & Mortimore, P (1994) cite eight possible factors associated with small class sizes that might explain the link with improved attainment. These factors include the quality of teaching – small class sizes being associated with a greater variety and imagination in activities, better assessment of individual needs and more in depth teaching of content. Teachers also tended to have better control of classroom discipline. Thus the policy makers should have been aware of the importance of quality teaching and that absence of quality teaching would amount to an interfering factor.

In sum, there was evidence that the CSR policy would introduce interfering factors into the treatment population, and so would not be effective. Therefore the judgement of external validity should have been relatively straightforward.

**6. The importance of identifying the correct causal factor**

So far I have argued that if a treatment will not be effective because of interfering factors introduced into a population by the treatment itself, then this should be relatively easy to predict, and so in some circumstances assessing whether evidence is relevant should be relatively straightforward. With the example of the California class size reduction initiative the interfering factor was poor teacher quality. Had poor teacher quality been a pre-existing condition in California then policy makers cannot have been expected to be aware of it. However, poor teacher quality was introduced by the policy itself, it was a predictable consequence.

The second claim I want to advance is that it is of vital importance to 'capture' the causal relationship as closely as possible. The more accurately a causal relationship is captured (identified, then manipulated), the less likely it is to be susceptible to interfering factors.

Small class sizes do not in and of themselves lead to improved attainment, rather as we have seen, small class sizes may lead to improved class discipline, which (for the sake of argument) may lead to more time spent on learning, which in turn may (through other causes) lead to improved attainment. We can talk about an intervention occurring closer to the effect (improved class discipline intervenes *closer* to the effect of improved attainment than an intervention of small class sizes does); and we can talk about an intervention occurring further from the effect (so intervention of small class sizes operates *further* from the effect of improved attainment than the intervention of improved class discipline does).

There are a number of reasons why policy makers may choose to intervene at a position in the causal chain that is some distance from the desired effect that is further up the causal chain. It may simply be too difficult to intervene at a closer distance to the effect, or it may be relatively easy to intervene at a further distance. We may have a detailed understanding of how causes operate at a distance, but be less sure of the causal forces at a local level. We may know *that* a distal intervention is effective, but not know *how* it is effective; if we were to intervene at a lower level we run the risk of not capturing the causal path at all. And finally there may be a number of more contingent reasons for the choice of a particular upstream policy intervention, for example, that the policy is politically popular or that it fits well with other policy initiatives.

However, here I want to argue for why, ceteris paribus, we should try to intervene as close to the effect as possible.

The objective of the CSR initiative was to improve the reading and maths skills of students, as measured by some standard test. We cannot intervene directly, in students, to improve their reading and maths ability. We have to improve their reading and maths ability via some distal lever. For the sake of argument, there may be evidence that learning targeted to the specific needs of particular students is effective in improving their attainment. There may also be evidence that if teachers have more time to give to individual students then they can tailor their teaching to the needs of the individuals. But how do we ensure that teachers have sufficient time to provide individual attention? One way is to cut down on the administrative work required of them, another way is to reduce the amount of time they have to spend on class management and student discipline. Small class sizes are then one way of reducing the amount of time teachers spend on class management, with the ultimate

aim that they will be able to target specific teaching to individuals, thereby improving the reading and maths skills of those students.[12] This can all be presented in Figure 1 below:[13]

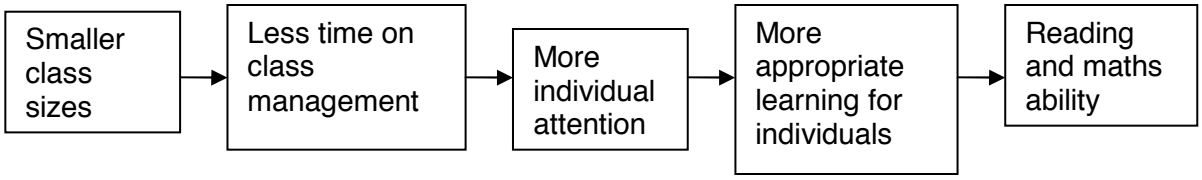



**Figure 1** – A possible causal chain.

But smaller class sizes are not the only way of encouraging more appropriate learning for individuals. For instance we could have a policy of improving teacher quality so that although the amount of time spent on individual students is the same, the quality of that interaction is higher. Or we could have a policy of encouraging parents to be more involved in the education of their children, for instance by providing them with guidance on how to help with maths homework. Arguably, a good rule of thumb in choosing a policy solution is that (ceretis paribus[14]) we should choose the policy that is in the shortest causal chain from the actual cause so that we minimise the chance of breaking the causal chain. Any step in the causal chain is susceptible to interfering factors. A higher step in the causal chain will be at risk of all the interfering factors of a lower step, along with its own set of interfering factors. Therefore the lower the step chosen for policy intervention, the lower the risk of interfering factors. This can be illustrated in the figure below:

[12] Individualised teaching, individual attention and less time on classroom management are all put forward in Blatchford, P. & Mortimore, P (1994) as processes that might explain the link between class size and educational outcomes.

[13] This picture involves obvious simplifications. Complications would arise if the causality were not linear. For example, it could be that by intervening at the macro level we are influencing a number of causal factors that we are not aware of. If we instead choose to intervene at a more micro level we risk ignoring these necessary causal factors, and so overall our intervention will lack efficacy.

[14] Other factors, such as cost, feasibility, political acceptability or side effects could in some situations be of far more significance in choosing a policy intervention than the factor of the shortness of the causal chain.

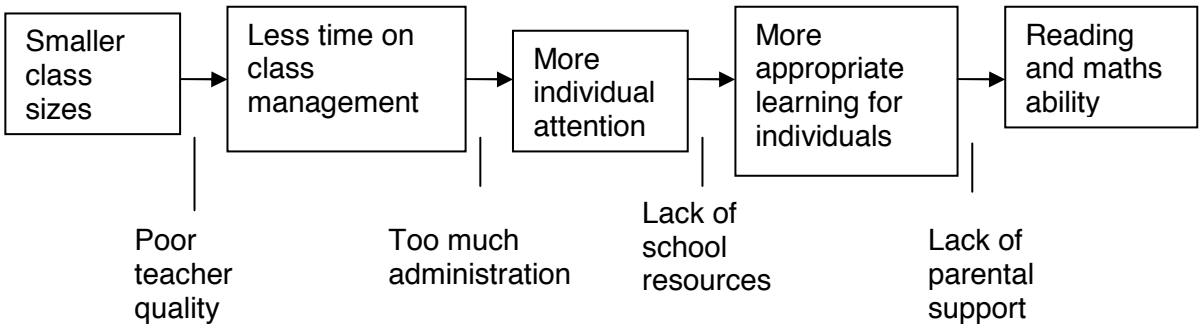



**Figure 2** – A possible causal chain, with interfering factors.

Even if we were able to ensure that students received more individual learning, a lack of encouragement from parents could put them at a disadvantage. If we were to intervene at a level above and ensure that students received more individual attention from teachers, then this may not translate into more individual learning if, say, the school did not have a variety of books to cater to the different reading abilities of students. We could try to ensure that teachers spend less time on class management so that they can provide more individual attention to students. But if the administrative tasks such as testing and marking papers are too great then any extra time could be consumed by these tasks rather than by being spent on actual teaching. Finally, we could reduce class sizes in the hope that teachers will need to spend less time on class management, with a view to ultimately improving attainment. But teachers with less experience and fewer qualifications may be less able to control discipline even in small classes.

If we choose to intervene by encouraging more appropriate learning for individuals,[15] then we risk an interfering factor of a lack of parental support. If we choose to intervene by reducing class sizes, we risk interfering factors of a lack of parental support, a lack of school resources, too much administration and poor teacher quality. In sum, interventions higher up in the causal chain risk all the interfering

[15] Perhaps by streaming classes or by providing teachers with a variety of teaching strategies as part of their own training.

factors further down the causal chain, along with additional interfering factors all their own. The more closely the causal relationship is captured, the less likely it is to be susceptible to interfering factors.

The risk with using policy levers far back in the causal chain is that the relationship between the policy lever and the actual causal mechanism can break down more readily. For instance, in choosing small class sizes as a policy lever for improving attainment there was always a risk that the relationship seen between small class sizes and attainment seen in circumstances such as the Tennessee STAR programme would break down elsewhere.

Whereas in the Tennessee STAR programme there is a strong relationship between small class size, pupil attention and attainment, this relationship could have broken down in California. It could be that, perhaps because of the interfering factor of poor teacher quality, there is no longer a strong relationship between class size and improved class discipline. Had policy makers been able to capture a more immediate causal mechanism leading to improved attainment then they would have reduced the risk of the causal chain breaking down.

**Conclusion**

The failure of the Californian CSR initiative highlights important lessons in the application of evidence to policy. The consensus view seems to be that poor teacher quality was an interfering factor. But what is particularly interesting is that this interfering factor was not a pre-existing feature in the target population but one introduced by the policy itself. As such it should have been relatively easier to predict. Secondly, the Californian CSR shows how important it is to identify the causal relationships as closely as possible. The tighter the causal relationship, the less susceptible it will be to interfering factors.

**References**

Blatchford, P. (2003) "The Class Size Debate: is smaller better?", Open University Press

Blatchford, P. & Mortimore, P (1994) "The issue of class size reduction for young children in schools: what can we learn from research?" *Oxford Review of Education*, Vol.20, No.4, pp.411-428

Bohrnstedt, G.W. & Stecher, B.M (eds.) (2002) "What Have We Learned About Class Size Reduction in California" (CSR Research Consortium capstone report)

Cartwright, N.D. (2009) 'What is This Thing Called Efficacy?', to appear in *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*, , C. Mantzavinos (ed), Cambridge: Cambridge University Press

Finn, J.D. & Achilles, C.M. (1999), "Tennessee's Class Size Study: Findings, Implications, Misconceptions", *Educational Evaluation and Policy Analysis,* Vol.21, No.2

Illig, D.C. (1996), "Reducing Class Size: A Review of the Literature and Options for Consideration" (a report prepared for the California State legislature)

Kim, J.S. 'The relative influence of research on class-size policy' *Brookings papers on Education Policy: 2006/2007*