Ensaio: Avaliação e Políticas Públicas em Educação
ISSN: 0104-4036
ensaio@cesgranrio.org.br
Fundação Cesgranrio
Brasil

Goldstein, Harvey
Evaluating educational changes: a statistical perspective
Ensaio: Avaliação e Políticas Públicas em Educação, vol. 21, núm. 78, enero-marzo, 2013, pp. 101-114
Fundação Cesgranrio
Rio de Janeiro, Brasil

Evaluating educational changes: a statistical perspective

Harvey Goldstein*

Abstract

The paper explores some of the issues involved in evaluating educational policy initiatives. It gives examples of how research findings can be evaluated and draws lessons for the ways in which policymakers can interact usefully with researchers. It argues that while central government's use of research evidence is often highly selective and concerned with its own perceived short term interests, a broader view of the research process is more productive and beneficial. The issues of class size, school league tables and the effects of homework are studied in detail and the often provisional nature of research evidence is emphasised as well as the uncertainty surrounding the findings of individual studies.

Keywords: Educational research. Educational policy. Class size research. League tables. Homework.

Avaliando mudanças educacionais: uma perspectiva estatística

Resumo

O artigo explora questões que aparecem ao se avaliar iniciativas de políticas educacionais. O artigo dá exemplos de como resultados de pesquisas devem ser avaliados e tira lições de como formuladores de políticas podem interagir efetivamente com pesquisadores. O artigo argumenta que enquanto o uso que o governo faz da evidência da pesquisa é frequentemente seletivo e preocupado com seus interesses de curto prazo, uma visão mais geral do processo de pesquisa é mais produtivo e benéfico. As questões do tamanho da turma, ranqueamento de escolas e os efeitos do dever de casa são estudadas em detalhe e a natureza frequentemente provisória da evidência da pesquisa é enfatizada assim como a incerteza em volta dos resultados de estudos individuais.

Palavras-chave: *Pesquisa educacional. Política educacional. Pesquisa sobre tamanho da classe. Ranqueamento das escolas. Dever de casa.*

* Professor of Social Statistics, University of Bristol, UK. *E-mail*: h.goldstein@bristol.ac.uk

Evaluando cambios educacionales: una perspectiva estadística

Resumen

El artículo analiza cuestiones que surgen al evaluar iniciativas de políticas educacionales. Presenta ejemplos de cómo deben evaluarse resultados de investigaciones y enseña cómo formuladores de políticas pueden dialogar efectivamente con los investigadores. El artículo argumenta lo siguiente: el uso que el gobierno hace de la evidencia de la investigación es, a menudo, selectivo y preocupado con sus intereses a corto plazo, sería más productivo y provechoso tener una visión más general del proceso de investigación. En el trabajo se estudiaron detalladamente el tamaño de las clases, la clasificación o ranking de las escuelas y los efectos del deber de casa. Se enfatizó la naturaleza frecuentemente provisoria de la evidencia de la investigación así como la inseguridad sobre los resultados de estudios individuales.

Palabras-clave: *Investigación educacional. Política educacional. Investigación sobre el tamaño de la clase. Clasificación o ranking de las escuelas. Deber de casa.*

Introduction

Educational systems are in constant change. Some of this arises from their own internal processes. As new knowledge accrues, technology changes, or research suggests alternative approaches to pedagogy, so educational systems adapt both formally and informally. Thus, for example, the advent of cheap electronic computation has altered radically almost all curriculum subjects internationally. Similarly, attitudes to education as a public service will also change. Such attitudes may be driven, for example by ideology, by cost or by demographic changes. Furthermore, in a global communication structure, individual systems interact, collaboratively or competitively so that when changes occur in any one system, this may influence other systems.

The present paper is an attempt to create a general framework within which the effects of changes can be evaluated. The structure of the paper is as follows:

First I will look at some of the debates that have taken place over structural issues, focussing on the effects of school class size and the debates in the UK, and elsewhere, over school accountability in the form of school rankings or 'league tables'. I will do this in terms of the way in which research evidence has related to educational policy and seek to draw some general conclusions.

Secondly, I will look at one case history, on the efficacy of 'homework' where both policymakers and researchers have made mistakes. Again I will suggest some lessons that can be learned from this.

Finally, I will attempt to formulate a suitable framework within which change can take place in ways that recognise the force of both research evidence as well as ideology. I will try to identify ways in which such a framework can be encouraged as well as identifying those whose activities are likely to impede it.

Class size: does it matter?

The effects of class size on achievement have been studied since the 1920s quantitatively and qualitatively, and have certainly been debated for much longer. It remains one of the most enduring and vociferous debates in education and has been about the educational advantages of small class sizes. Opinion has been consistently polarised between those who claim that small classes lead to a better quality of teaching and learning, and those who argue that the effects are likely to be modest at best and that there are other more cost effective initiatives. There is a large number of existing studies, including observational surveys, matched designs and randomised controlled trials (RCTs). Despite the number of studies, the results are often inconclusive.

Glass and Smith (1979) first applied a meta analysis to 77 studies based on 70 years' research in more than a dozen countries. They concluded that there were positive effects for class sizes of less than 20, based on 14 of these studies which were considered to be 'well-controlled'. Slavin (1990) argued that Glass's positive finding was based on only a small number of studies and the results were largely affected by one extreme case. On reanalysis Slavin reported an effect much smaller than Glass. He also conducted an analysis of 9 randomised or matched studies.

The second view has found expression in the opinions of politicians and policy makers, worried by the enormous costs involved in hiring extra teachers. In the UK, the Government agency OFSTED (1995), on the basis of many inspectors' reports, concluded that class size made little difference and this was used by Government ministers of the day to support no change to investment in smaller classes. This sceptical view of the effect of class size has also been taken by academics like Hanushek (1999) who have argued in support of alternative uses of funding, e.g., teacher training.

In the late 1980s a major randomised controlled trial was carried out in the US state of Tennessee. The STAR research study employed a design involving random allocation of pupils entering elementary school and teachers to three classes within schools: small, 'regular, and 'regular' with teacher aide. It was found that children in small classes performed better in literacy and maths (FINN; ACHILLES, 1999; NYE; HEDGES; KONSTANTOPOULOS, 2000). These results seem to have strengthened arguments in favour of small classes and led to costly class size reduction initiatives in a number of States in the USA, notably California, as well in other countries

around the world. It is also reflected in the UK Government's commitment to a maximum of 30 in a class at reception and Key Stage 1 (KS1, 5-7 years). More recently a large, non RCT, longitudinal study in England confirmed the findings of the STAR study (BLATCHFORD et al., 2003, GOLDSTEIN et al., 2000). Broadly speaking, all the results indicate that a moderate effect size of between 0.1 and 0.2 standard deviations of achievement score is associated with a class size difference of 10 pupils, but beyond a class size somewhere between 25 and 30, there is little change. In addition while there seems to be little if any overall effect beyond the first year of schooling, there is a positive effect beyond the first year of being in a small class for the more deprived pupils and those from ethnic minorities. It would seem that this latter finding has not found its way into educational policies.

The two influential studies quoted had very different designs. The STAR study was a RCT whilst the second was an observational study with longitudinal data that allowed adjustment for cognitive development at the start of schooling. Both designs were an attempt to deal with selection effect whereby schools might have been allocating those pupils with delayed cognitive development to smaller classes so potentially masking any real effect of class size. Both studies had limitations. The STAR study randomised teachers *within* schools. Thus, it is very likely that there would have been communication across teachers, destroying one of the requirements that 'treatments' should be independent. Likewise every teacher knew which group they had been assigned to so that the study was in fact zero blind and expectations might be expected to influence results. This to some extent would also be true of the Blatchford study although the teachers would not be so aware that they were taking part in class size research. For the STAR study to avoid these problems it would have needed to adopt a cluster randomisation whereby each school was assigned to have small or large classes. Such a study would be more costly and require larger numbers of schools and classes. Furthermore, the STAR design excluded schools with only one class per year group.

The fact that both studies found similar effects suggests that, in general, evaluations for policy purpose could usefully use different designs. Where results agree then we can begin to separate the possible confounding between type of design and results. Observational studies will generally need to collect longitudinal data but will still generally be cheaper than RCTs and also likely to be more acceptable for ethical reasons. Nevertheless, the addition of even a limited RCT seems likely to add to the robustness of results. When results substantially disagree this would then suggest that careful study of why the disagreement exists can be fruitful.

The history of class size research also illustrates another issue concerning the use of research evidence. For some 70 years until the meta analyses of the late 1980s and the STAR study, most studies reported little effect of class size on attainment.

Partly of course this was to do with poor design, most notably the fact that very few indeed were longitudinal. Nevertheless, both trades unions and employers remained convinced that class size did matter. Again, simply accepting the studies that exist without subjecting them to a thorough quality analysis can be misleading.

Another rather interesting example of the tentative nature of research evidence is the case of the International Adult Literacy Survey (IALS) when, in the early 1990s, France pulled out when it discovered that the results for literacy levels in France were unexpectedly low. Here, although the French unwillingness to confront unwelcome evidence may have been more influenced initially by politics than science, a perfectly reasonable *scientific* case for scepticism could have been made, and indeed was made, in order to undertake a reanalysis. In fact, the subsequent reanalysis exposed a variety of flaws in the design and execution of the study that not only supported the French stance but led to useful insights into the nature of such international comparative research (BLUM; GOLDSTEIN; GUERIN-PACE, 2001).

My next example illustrates this point with a case history from a study of the effect of homework on school achievement.

Homework and school achievement

A more detailed discussion of this case is given by Goldstein (2008), but a short summary will suffice here.

In late 2006 a well-known British educational researcher, Peter Tymms, claimed that "There was a coordinated attack and a rubbishing of my research (on homework) by the Prime Minister, by David Blunkett (then Secretary of State for Education) and by Chris Woodhead (then head of the Office for Standards in Education (OFSTED), which is the body charged by the Government to carry out inspections and publish reports on schools." The Government and its various representatives is here accused of using its power and influence to undermine the integrity of particular researchers. There are in fact two issues. The first concerns the quality of the research itself. Thus, the research claimed that increased homework was associated with poorer performance, which seems counter-intuitive and many would find surprising. In such a situation, those civil servants and others who were advising the politicians would certainly have suggested caution, and perhaps advised seeking further opinions or research. What is new, however, and in my view quite unethical, is the intervention of politicians, with no personal competence in the area, directly in a research debate.

The Tymm's report was published shortly after a previous report principally authored by Blair's then principal advisor on education, Michael Barber, which concluded that homework was associated with improved performance and lent support to the Labour Party's current policy (January 1997) in favour of mandatory periods of homework

(BARBER et al., 1977). The research had serious flaws (see www.cmm.bristol.ac.uk/team/HG_Personal/home_rep.html for a critique) but did seem to resonate with received opinion. The Tymms' report was therefore very 'high stakes'. Unfortunately, the political debate that ensued effectively prevented a proper peer-based review of the report itself as well as the currently available evidence. In fact, the Tymms' report itself was flawed since it did not take account of prior achievements in the curriculum subjects under investigation. Thus, the cross sectional sample it used did not allow a judgement about whether previous poor performance was responsible for pupils doing more homework or whether actually doing large amounts of homework depressed performance.

There is a certain irony in that, had the politicians acted responsibly, they may well have been able to substantiate their policies through critical peer review of the Tymms' research, which could have provided a more secure basis for their own homework policies.

There is an important lesson here for politicians. Taking research evidence seriously, rather than trying to 'shoot the messenger' can be, even in the short term to their advantage. Research evidence is not always good evidence. There is bad and weak research and it is only by exposing findings to critical peer review that the quality of research can be evaluated properly. More generally, the tendency to publicise research findings quickly, using the internet in particular, is not necessarily consistent with high quality. Demands from policymakers to obtain quick results or to make them 'relevant' to policy also need to be resisted if this means that the very essence of good research, namely exposure to critical review, is weakened. This is not an argument for delaying research findings, but it is an argument for ensuring that they are properly evaluated, if necessary via public debate, before being acted on.

I will now look at how we might develop a framework for educational research that avoids some of the dangers I have discussed, while enabling advances to be made in real knowledge. I shall start by discussing how, in many educational systems, schools, as well as higher education institutions, are ranked in performance or 'league' tables.

Knowledge may be created but also destroyed

The following brief account is of the history of school league tables in England, but there are parallels in other systems, especially in the United States and Australia. It will serve to introduce alternative ways of gaining understanding.

It was in 1986 that the regime of Prime Minister Margaret Thatcher, building upon work carried out by the Inner London Education Authority[1], first tentatively

[1] This was the body that organised primary and secondary education for inner London up to its disbanding by the Government in 1990 and the allocation of education management to the component local authorities in the area.

decided to publish secondary school average examination results and thus provided the means for ranking schools on what were claimed to be measures of school 'quality'. This policy was strengthened over the next few years. During this time the Government introduced 'key stage tests' at the ages of 7, 11 and 14, and by the time of the New labour Government in 1997 the 11 year old (key stage 2) test results were also being published and parents encouraged to use the rankings in their choice of schools. The Royal Statistical Society report on performance indicators (www.rss.org.uk) provides a broad review of the issues surrounding the use of league tables in a number of areas including Education, Health and Crime. A technical discussion of the statistical issues is given by Goldstein and Spiegelhalter (1996). Briefly, the main issues are as follows.

The first rankings to be published were simply percentages of pupils in each school achieving the highest grades in the GCSE and A level examinations[2], these being the certification examinations generally taken at age 16 and 18 respectively. A scoring system based upon all the examination grades was also used with the rankings based upon the average score for each school. From the outset, many in education had pointed out the shortcomings of the league table policy, citing research findings that demonstrated the need to adjust results for school intake characteristics (the value added model) and also the need to provide uncertainty (confidence) intervals[3] for the mean scores based on relatively small sample sizes. Nuttall et al., (1989) provided an early critique that demonstrates the inadequacy of these rankings by using research data where intake measures were available. They showed that after adjustment, rankings of many schools were changed, and that when confidence intervals were placed around the school estimates, most schools could not be statistically distinguished from the average. This was later reinforced by Fitz-Gibbon (1997) in an officially commissioned report.

In response the Government was able to point out in the early days that the data were not available to carry out such adjustments for the whole population , and indeed these data only really become available towards the end of the 1990s. Nevertheless, it was in 1995 that the then Conservative Government first officially committed itself to value added 'performance tables' and by 2007, thanks partly to the existence of the National Pupil Database (http://www.bris.ac.uk/Depts/CMPO/PLUG/) containing longitudinal pupil data, these have become regular publications although there remains a considerable reluctance to embrace the crucial notion of confidence intervals.

[2] The GCSE is the General Certificate of Secondary Education taken by 16 year olds at the end of compulsory schooling in year (grade) 11. The A level is the advanced level General Certificate of Education examination that is taken at the end of year 13 and principally serves as a university entrance qualifying examination.

[3] A confidence interval provides a range of values that, with a given probability – typically 0.95, is estimated to contain the true value of the school score. If such an interval includes the value zero then an equivalent statement can be made that the true value is not significantly different from zero at the 5% level.

There is a whole set of key issues about the ways in which the imposition of a 'high stakes' national testing regime affects the behaviour of schools, pupils and parents. These include incentives by all players to maximise their test scores at the expense of longer term learning goals; the stress that testing imposes upon pupils; the encouragement to 'gaming' by institutions all the way up to outright cheating. In short, the tables cease to be, if they even ever were, an 'objective' measure of school quality. Indeed it seems strange that anyone ever really believed that they could be. They are based on limited information and highly unreliable. Yet decisions about whether to 'punish' a school and the children and parents associated with it, even to close it, relies heavily on its position in the league table. The inspections of schools by OFSTED largely acts to legitimise, for example, a 'failing' school that happens to lie towards the bottom of its league table. Most importantly, however, judging a whole school in this way adds almost nothing to our knowledge of what makes a good school, or more importantly what constitutes effective education, along all the many dimensions against which it should be judged. Worse, it detracts from high quality, typically long term, research that attempts to understand what works and why. Let me say a little more about this.

Gaining useful knowledge

I will first look at alternative ways of understanding school and teacher performance since there is a legitimate general demand for accountability from institutions that are publicly important, including those that directly receive public funds such as schools, hospitals, police forces, transport systems and also those whose activities affect the welfare of citizens and public life more generally, such as financial institutions, transport systems and so forth.

Broadly speaking we can consider two types of data gathering to inform us about institutional performance. The first is that based around administrative systems that routinely collect information about, for example, examination results, school composition or student characteristics. Such statistics form the basis of most of the accountability systems we currently see. The second consists of specific studies that collect information related to just those aspects of interest. These would include the international comparative studies of performance such as IEA (SCHMIDT; MCKNIGHT, 1995) and PISA (GOLDSTEIN, 2004) and studies that evaluate a single educational programme.

Using routinely collected data for gaining understandings has strengths and weaknesses. An obvious strength is that the data are already collected and often part of a long term policy so that continuity is reasonably well assured. It does have problems, however. One is that, because of its generality and design for particular administrative purposes, the data may not be the most appropriate. They may not be collected at the level of detail required to answer specific questions and the range of measurements will tend to be narrow. On the other hand, some routinely

collected data, often that collected by national statistical offices, is in the form of surveys that researchers can contribute to and these can also allow flexibility to adapt to policy events. Where administrative data are used for accountability further problems arise. Thus, for example, league tables of schools or teachers based upon examination results or routine test scores encourage optimisation behaviour or 'gaming' that results in the results reflecting aspects of institutional behaviour that will contaminate the very things intended to be measured. Such contamination might be small but the problem is that the extent of the contamination is generally unknown, although in some cases independent studies have demonstrated large effects that cannot be ignored. Thus Haney (2000) discusses the so called Texas miracle where a 'high stakes' test whose results were used to rate schools in ways that affected their income and could even lead to closure. He shows how the resulting test scores were distorted by the requirements for accountability and became too unreliable for use in any evaluation – an example of how a requirement for accountability succeeded in destroying rather than creating knowledge.

The international comparative studies mentioned above are an example of non-administrative studies that collect data intended for research purposes but with a very general remit rather than being tied closely to any specific educational programme or initiative. Their strength is that they are typically large scale, across educational systems and repeated at regular intervals. Their weakness is that they are typically compromises between the need to collect system specific information and data that is comparable across systems with difficult issues such as translations of test items. Such studies, since they are carried out on a sampling basis also do not provide information about individual institutions.

The conclusion I draw from this is that it is important to separate the issue of institutional accountability from research into the effectiveness of learning programmes or other innovations. While administrative data can be used to assist institutional accountability, this is only really sensible within a low stakes system where public rankings of schools are not available to be used for making judgements, either by policymakers or, for example, parents of children. Yang et al. (1999) discuss how such a system can operate. Large scale repeated surveys, especially where they are longitudinal, may be useful for generating interesting hypotheses or studying general trends over time, but are not particularly appropriate for evaluating any given initiative or programme. I now turn to ways of designing studies that can be used directly to do this.

Studies for evaluation

One of the best known educational evaluations is that of the 'Head Start' programme that started in the 1960s in the USA and still continues. Its aim is to provide resources to promote optimal development of children up to five years of age from low income families and has federal support and is locally administered.

A series of evaluations of this have been carried out, the most recent extensive one by a US Government agency (UNITED STATES, 2005). The study measured Head Start's effectiveness as compared to a variety of other forms of community support and educational intervention. Study participants were assigned to either Head Start or other *parent-selected* community resources for a year. The first report showed consistent small to moderate advantages in head start for 3 year old children including pre-reading, pre-vocabulary, and parent reports of children's literacy skills. Fewer positive benefits were found for 4 year olds. The benefits also improved with early participation and varied among racial and ethnic groups. Other evaluations similarly found mixed effectiveness and any effects that were detected tended to diminish after the children left the programme. The report was carefully designed and subject to a sophisticated analysis using multilevel modelling.

As in the class size research, a major finding is that early interventions tend to have little if any long lasting effects once the intervention programme has ended. If this is a very general finding then the policy implications are important and run somewhat counter to received wisdom, which is that early intervention can have long lasting effects and that the most productive use of limited resources lies in early intervention. While other research often appears to support the role of early life circumstances for later development, this is largely based upon observational studies where 'causation' is difficult to determine. This suggests that, while observational studies are important for suggesting causal links, it is only with well-planned evaluations that we can begin to base our knowledge upon secure foundations.

While the Head Start, and similar studies, did not carry out a full scale randomisation with control groups, unlike the STAR study, it was specifically aimed at evaluating the Head Start programme itself. In fact, the randomised control trial (RCT) is relatively rare. It tends to be expensive and needs to be designed at the beginning of any study so that its findings are robust. As we saw for the STAR study, the level at which randomisation occurs is important to avoid contamination effects. There may also be problems with 'contextual effects. Thus, if an effect only occurs when a minimum proportion of students in a classroom have a particular characteristic, such as belonging to an ethnic minority, then considerable care needs to be taken with randomisation to ensure that such classrooms are selected or created by the randomisation process. An observational study, on the other hand, will tend to include such classrooms if they exist and this strengthens the importance of carrying out observational studies as a part of the research process. Goldstein (2002) discusses this issue in more detail. The principle of designing the evaluation of any new programme or policy at the outset is an important one and can be done in different ways.

In designing an evaluation study we need to distinguish between a study that is designed to evaluate the *implementation* of an intervention[4] and one designed to understand whether the outcomes match the ones intended, for example to enhance learning. Studies designed to evaluate the implementation of a new programme are very important. They are needed to establish that the programme is implemented according to plan and can provide insights into overall feasibility, difficulties etc. They are not, however, a substitute for an outcome evaluation study that seeks to decide how successful an intervention may have been. Such an outcome study can have many forms and I now look at some of the necessary criteria for success.

The first consideration is that an evaluation study should be independent of those carrying out the programme. This ideally should extend to the funding source. In practice, of course, an evaluation will often be funded by the institution responsible for the programme, but where this occurs there needs to be a clear understanding that the evaluation is independently contracted and its reports are not subject to any kind of censorship by the programme funding institution. In the case of Government, which is typically the funder of interventions, funds for evaluation should be channelled through a research council or similar 'arms-length' body that can help to ensure independence. It is important that an evaluation study is openly published and also that it is subject to peer review before publication.

The second consideration is that a potential evaluation study should be envisaged at the outset and a programme advised so that an evaluation study is feasible. Thus, for example, many evaluations will be strengthened by taking measurements or carrying out surveys prior to the implementation of a programme, in which case the evaluation may have to start before the programme itself. Similarly, a programme design should incorporate suitable comparison groups for an evaluation to compare and ensure that sample sizes are adequate to detect likely effects. A further consideration is that the evaluation should be part of the overall process of deciding whether to extend a programme, for example by incorporating it in a national curriculum. This may require a period of waiting for an evaluation to be completed, and in some cases this may take some time, especially if any findings are subject to varying interpretations. It is very important to institute peer review but this will take time. This may be difficult for some policymakers to accept, but to ignore it may well make the evaluation somewhat pointless.

Thirdly, the design of a quantitative evaluation is difficult for a number of reasons. As I have discussed in the case of class size, attempts to set up control groups within schools have problems associated with lack of independence.

[4] We use the term 'intervention' to include any planned introduction of a new programme with a clear protocol and set of aims designed to alter behaviour, learning etc.

Allocation at the institution level is generally a better strategy but will typically require larger sample sizes. In practice, allocation at institutional level is often carried out in a non-random fashion and this creates particular problems. Thus, for example, the introduction of the 'literacy hour' in the late 1990s in English primary schools used an 'opportunity sample' of local authorities and an evaluation of this (MACHIN; MCNALLY, 2004) had to attempt to obtain a suitable matching comparison group. The comparison was in terms of changes over time (pre and post introduction of the programme) for the programme and 'control' schools. The problem is that there is a limited range of variables that can be used to 'match' local authorities and schools and to make adjustments for any pre-existing differences. Furthermore, as would be the case with a randomised evaluation, the institutions involved may change in ways that are difficult to measure, as a result of demographic changes.

Finally, it is may well be the case that a programme has an observable effect in terms of its stated objectives, but there may be associated negative effects in other areas. Thus, in the evaluation of the literacy hour it appeared that a positive effect occurred for both literacy and also mathematics, but other aspects of the school curriculum were not studied. These include other subjects as well as things such as attendance and behaviour, and these were not envisaged in the design of the evaluation. It is not enough to consider any one intervention in isolation and designs of evaluation studies should routinely seek to study as wide a range of outcomes as possible.

Conclusions

I have attempted to describe the basic conditions needed for carrying out successful evaluations of educational programmes. It is clear from experience that designing a good evaluation is no easy task. It often requires extensive resources, foresight in planning and careful analysis of data and attention to any unplanned consequences. All too often programmes are introduced on the basis of limited, or even non-existent, evidence about their efficacy and arguably therefore involve a waste of resources and public money. Sound evaluation is the only sure way of establishing what works and should be seen as an integral component of educational initiatives.

References

BARBER, M. et al. *School performance and extra-curricular provision.* London: Department of Education and Employment, 1977.

BLATCHFORD, P. et al. Are class size differences related to pupils' educational progress? Findings from the Institute of education class size study of children aged 5-7 years. *British Educational Research Journal*, London, v. 29, p. 709-730, 2003.

BLUM, A.; GOLDSTEIN ,H.; GUERIN-PACE, F. International adult literacy survey (IALS): an analysis of international comparisons of adult literacy. *Assessment in Education*, Oxon, v. 8, p. 225-246, 2001.

FINN, J. D.; ACHILLES, C. M. Tennessee's class size study: findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, Washington, v. 21, p. 97–109, 1999.

FITZ-GIBBON, C. T. *The Value Added National Project Final Report:* feasibility studies for a national system of Value Added indicators. London: SCAA Publications, 1997.

GLASS, G. V.; SMITH, M. L. Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy analysis*, Washington, v. 1, n. 1, p. 2-16, 1979.

GOLDSTEIN, H. *Designing social research for the 21st century*. Bristol: University of Bristol, 2002.

_____. International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education*, Oxon, v. 11, p. 319-330, 2004.

_____. Evidence and education policy - some reflections and allegations. *Cambridge Journal of Education*, Cambridge, v. 38, n. 3, p. 393-400, 2008.

GOLDSTEIN, H.; SPIEGELHALTER, D. J. League Tables and Their Limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*: Series A (Statistics in Society), London, v. 159, n. 3, p. 385-443, 1996.

HANEY, W. The Myth of the Texas Miracle in Education. *Education Policy Analysis Archives*, [S.l.], v. 8, n. 41, 2000. Available from: <http://epaa.asu.edu/ojs/article/view/432>.

HANUSHEK, Eric A. Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis,* Los Angeles, v. 21, n. 2, p. 143-163, Summer, 1999.

MACHIN, S.; MCNALLY, S. *The Literacy Hour*. London : Centre for the Economics of Education: London School of Economics, 2004. Available from: <http://cee.lse.ac.uk/ceedps/ceedp43.pdf >.

NYE, B.; HEDGES, L. V.; KONSTANTOPOULOS, S. The effects of small classes on academic achievement: the results of the Tennessee class size experiment. *American Educational Research Journal*, Thousand Oaks, CA, v. 37, p. 123-151, 2000.

OFSTED. *Class Size and the Quality of Education.* London: HMSO, 1995.

SCHMIDT, W. H.; MCKNIGHT, C. C. Surveying educational opportunity in mathematics and science: an international perspective. *Educational evaluation and policy analysis*, Los Angeles, v. 17, p. 337-349, 1995.

SLAVIN, R. Class size and student achievement: is smaller better? *Contemporary Education*, Terre Haute, IN, v. 62, n. 1, p. 6-12, 1990.

UNITED STATES. Department of Health and Human Services. *Head Start Impact Study and Follow-up, 2000–2012*. Washington: [s.n.], 2005. Available from: <http://www.acf.hhs.gov/programs/opre/hs/impact_study/>.

YANG, M. et al. The use of assessment data for school improvement purposes. *Oxford Review of Education*, Oxon, v. 25, p. 469-483, 1999.

Recebido em: 19/12/2012
Aceito para publicação em: 18/04/2013