

Seven Key Issues for Assessing 'Value Added' in Education

J. Douglas Willms

Canadian Research Institute for Social Policy
University of New Brunswick

This paper was prepared for a workshop sponsored by the US National Research Council and the US National Academy of Education on the use of value-added methods for instructional improvement, program evaluation and accountability. The author is grateful for funding from the Canadian Social Sciences and Humanities Research Council for its support of the collaborative research program, *Raising and Levelling the Bar*, and for its support of funding through the Canada Research Chairs program. He is also appreciative of support from the Atlantic Networks for Prevention Research and from his colleagues at the Canadian Research Institute for Social Policy.

Seven Key Issues for Assessing 'Value Added' in Education

The term 'added value' is used in manufacturing to refer to the contribution of the factors of production to increase the value of a product. The factors of production include capital goods as well as labour. In education, the term has been used to refer to the difference between a school's test results compared to some standard, usually the average score for the jurisdiction, after taking account in some way of students' family background. The *added value* to the product – student learning – is presumed to be the result of the factors of production – better teaching methods, an efficient use of teaching time, a better disciplinary climate in the classroom, and so on. In practice, the assessment of added value in education is quite complex, such that there is considerable debate among researchers as to the best approach. In this paper I discuss seven key issues regarding the estimation of value added and its use by practitioners. I argue that added value assessment could be more tightly linked to school policy and classroom practice.

1. Status versus learning.

In many jurisdictions, school average test scores are reported publicly in 'league tables', without any consideration of student's initial test scores or their family background. These unadjusted comparisons of 'status' can be very misleading and do not warrant much attention in a discussion of 'added value'. However, two points are worth noting. First, despite considerable effort by the research community in producing models for estimating added value, league tables are still the most dominant form of reporting test results. Second, there are many educators who argue that unadjusted results should be reported, as they represent students' actual 'status' as they move from one stage of schooling to the next or enter the labour market. While this is true, league table comparisons should not be confused with 'added value'.

Added value is about student learning. Therefore, any discussion of added value needs to begin with some model of what learning entails, and its estimation requires an explicit model of learning or growth trajectories (Raudenbush & Chan, 1993). There are three basic approaches that are commonly used:

- a. measuring 'gains' in student achievement with a pre-post design;
- b. measuring linear growth in student achievement using data collected on at least three occasions; and
- c. measuring linear and quadratic growth using data collected on at least four occasions.

Measuring gains in achievement is considerably better than simple 'status' comparisons. However, it is not without its problems. One important problem is that measurement error at the pretest and posttest are compounded in the estimate of the pre-post gain score, such that the estimate of student growth has low reliability. This is especially problematic when there is little variation in school average rates of growth, and the analyst wishes to discern which schools have the best or worst rates of growth. In this case, school-level gain scores have very low reliability (Willett, 1988).

Estimates of 'growth' based of data from three occasions are considerably more accurate as the errors of measurement are modeled explicitly as part of each child's growth trajectory (Raudenbush, 2001). The problem here is that growth is presumed to occur in a linear fashion, which may or may not be the case, depending on the nature of the skill or subject being learned. For many tasks, there is a critical point when learning 'takes off', similar to the pattern of children's early vocabulary development (Hart & Risley, 1995). However, for the skills required for many school subjects, there appears to be a ceiling that many students reach, and at this point learning slows down. For the purposes of this paper, I will not discuss whether this is an artifact of the way students are taught versus the way they are tested. Rather, it is suffice to say that students' test scores on scales that are intended to have equal intervals tend to rise during the early elementary and middle school years and then level off during the secondary school years (Willms, 2008).

Estimates of 'growth' based of data from four or more occasions attempt to overcome this problem by explicitly modelling the functional form of the growth process for each child. With this approach one can estimate the average growth trajectory for a school and compare schools in their average rates of learning.

To illustrate this issue, Figure 1 presents a summary of children's growth trajectories in mathematics skills from age 6 to 15 for a large nationally representative sample of Canadian children examined in the National Longitudinal Survey of Children and Youth (NLSCY). The NLSCY is a nationally representative longitudinal study of Canadian children and youth that was launched by the Canadian government in 1994 with a sample of over 22,000 children in over 13,000 families. The design included surveys administered to parents, teachers, and school principals, direct assessments of the children after age 4, and a separate self-report questionnaire for youth from age 10 onwards. The mathematics scores in the NLSCY were based on a series of tests administered to students every two years. The scores were 'vertically-equated' such that the scores from each test could be mapped onto a single continuous scale. The scores were standardized such that the age 15 results had the same mean and standard deviation as the Canadian results for 15-year olds assessed in the Programme for International Student Assessment (PISA) (OECD, 2004). For further details see Willms (2008).

The figure shows the extent to which the individual growth trajectories vary around the average growth trajectory. The coloured areas capture the variation in growth trajectories: the red area shows the range from the 5th to 25th percentiles of the estimates of outcome scores for the sample at each age, the yellow from the 25th to 50th percentiles, the light green from the 50th o 75th percentiles, and the dark green from the 75th to 95th percentiles. This portrayal of the results shows that as children mature, the range of scores increases, and the distribution becomes increasingly skewed, with more children at the lower end of the score distribution. (The skewness of a distribution is an indication of whether it is symmetrical or asymmetrical; distributions that are negatively skewed have scores that extend further below the mean than high scores extend above it, while the reverse is that case for positively skewed distributions.) In this case, the distribution of scores becomes increasingly positively skewed; at age 15 the skewness is 1.03.

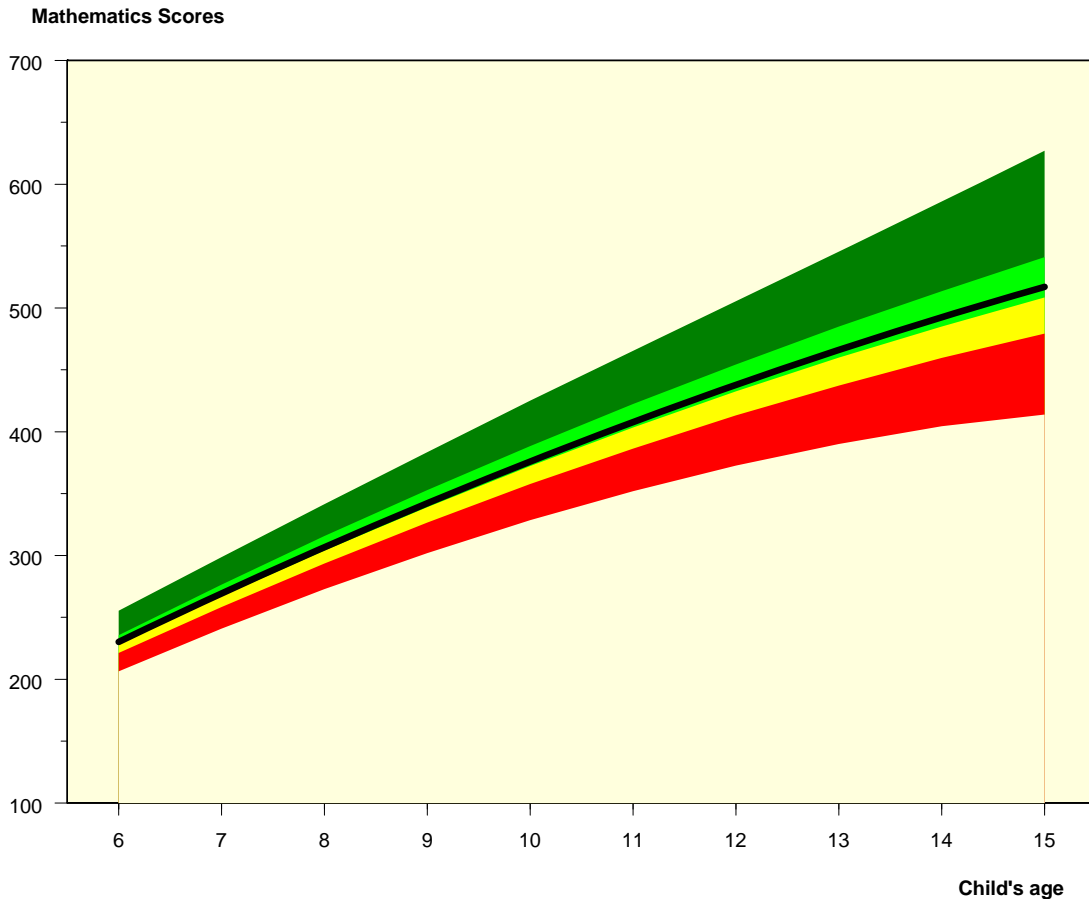


Figure 1. Variation in growth trajectories in mathematics skills among Canadian students. Source: National Longitudinal Survey of Children and Youth. Reproduced with permission from Willms (2008).

With this figure in mind, imagine a secondary school in which the majority of students arrive with mathematics skills in the bottom quartile. It is likely that this school would find it more difficult to achieve ‘gains’ in skill proficiency than one in which the majority of students arrived with skill levels in the top quartile.

To summarize, there is a clear hierarchy here with respect to how accurately ‘added value’ can be estimated. Two-time point estimation is the minimum requirement for considering ‘added value’; three-time point estimation is considerably more accurate, and four-time point estimation is better still. But there is also an hierarchy in terms of the costs associated with implementation and analysis. The ‘best’ or right approach depends on the nature of the skill or subject area being assessed and the kinds of decisions to be made from the results.

2. Control for SES

The assumption is often made that if we are assessing ‘learning’ rather than ‘status’, then controlling for students socioeconomic status (SES) is less important. It is certainly the case that measures of SES by themselves are inadequate for estimating ‘added

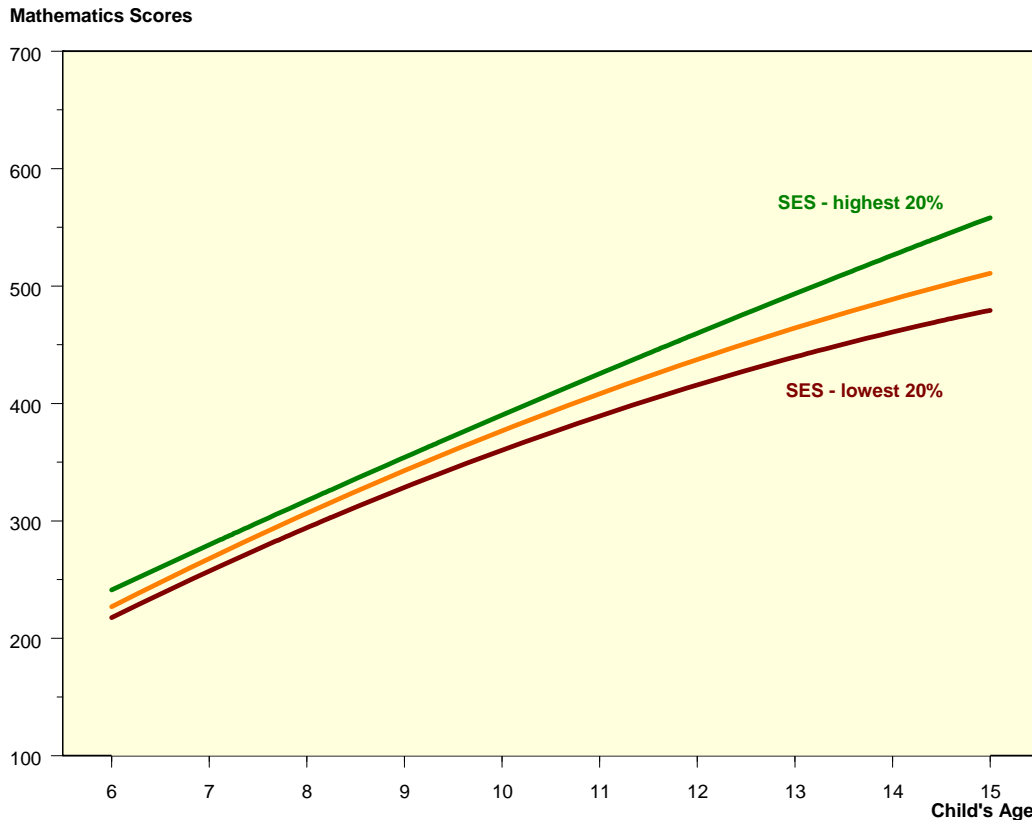


Figure 2. Average growth trajectories in mathematics skills for Canadian students in the bottom, middle and top SES quintiles. Source: National Longitudinal Survey of Children and Youth. Reproduced with permission from Willms (2008).

value' in the assessment of students' academic achievement (Willms & Kerckhoff, 1995). However, there is also compelling evidence that prior test scores by themselves are insufficient. Extending the example shown above, Figure 2 shows the average growth trajectories for students in the bottom, middle, and top SES quintiles. Although the trajectories are close to being parallel up to age 10, thereafter the gaps between the groups widen, such that by age 15 there is a considerable achievement gap associated with SES.

The estimation of school's added value is further complicated by the fact that children from low SES backgrounds tend to lose skills over the summer months while those from high SES backgrounds maintain or even increase their academic skills (Alexander, Entwisle, & Olson, 2001; 2007). However, the evidence also indicates that high SES children are learn at a faster pace in school than low SES children, and therefore SES needs to be taken into account in assessments of added value.

3. Controlling for student composition and contextual effects

I use the term ‘composition effect’ to refer to the effects on student learning associated with the schools’ demographic composition, such as the mean SES of the school, or the distribution of student ability upon intake. In contrast, ‘contextual effect’ refers to “the environment in which teaching and learning takes place. It includes, among other factors, school and classroom resources, interactions among peers, the relationships between teachers and students, the disciplinary climate of the classroom, and the norms for academic success. Thus, it comprises factors that characterize or *describe* the learning environment – its physical features and its culture.” (Willms, in press).

The issue with respect to assessing ‘added value’ is that composition and context are correlated; for example, schools that enrol higher SES students tend to have a more favourable learning context – more resources, fewer discipline problems, and more positive student-teacher and peer interaction for example. In the assessment of added value we would like to separate the effects of composition from those associated with those aspects of context that are reasonably under the control of teachers. The problem, though, is that many contextual factors, such as maintaining a positive disciplinary climate, are correlated with composition; for example, it is easier to maintain a positive disciplinary climate in schools that have fewer students from disadvantaged backgrounds.

Raudenbush and I noted this issue and advocated estimating two measures of school effects – ‘Type A effects’ that included control for prior ability and SES, but not school composition, and ‘Type B effects’ that included control for prior ability, SES, and the mean SES of the school (Willms & Raudenbush, 1989; Raudenbush & Willms, 1995). We argued that Type B effects provide a fairer measure of the added value associated with school policy and classroom practice. However, the model for estimating Type B effects may be over-specified, such that some of the effects of good educational practice that one wants to be included in an assessment of added value is removed. For example, if more talented and dedicated teachers tend to be found more often in high SES schools, then the Type B estimate would not fully account for teachers’ skills and effort.

4. The critical transition

Perhaps the most important control measure regarding school composition and added value is the proportion of students that have successfully made the critical transition from ‘learning-to-read’ to ‘reading-to-learn’. For most students this occurs at about age 8, typically by the end of the third grade. If children are not able to read with ease and understand what they have read when they enter fourth grade, they are less able to take advantage of the learning opportunities that lie ahead.

Figure 3 shows the socioeconomic gradient for the reading skills of 15-year olds in the US, based on the 2003 PISA. The small blue dots are students’ scores on the PISA reading test plotted against their family’s SES, for a representative sample of 5,000 US students. The results are shown on the OECD scale for Reading proficiency, which was scaled for PISA 2000 to have a mean of 500 and a standard deviation of 100 for all OECD countries. The coloured bands indicate the six proficiency levels used in PISA.

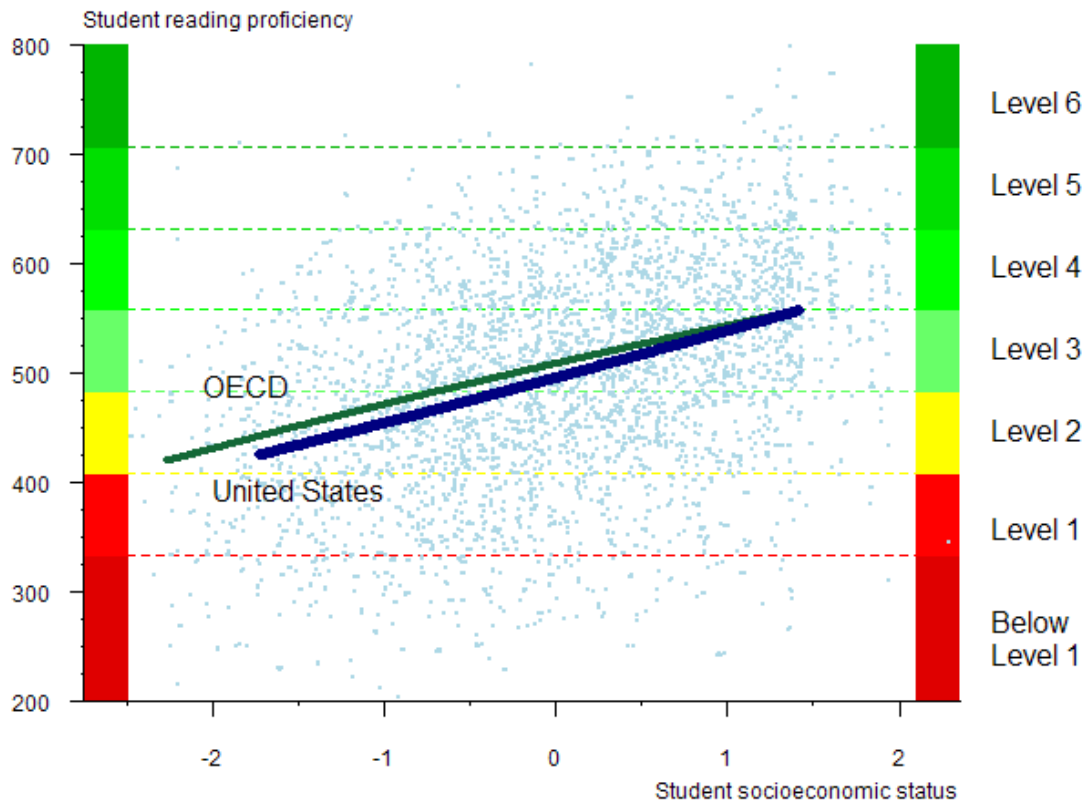


Figure 3. Socioeconomic gradient in reading proficiency for the US. Source: 2003 Programme for International Student Assessment.

Students that score at Level 1 or lower have very minimal reading skills; we could say that these children are ‘struggling readers’ who have failed to make the transition from ‘learning-to-read’ to ‘reading-to-learn’. Generally these students lack the basic skills necessary for pursuing post-secondary education or attaining jobs in the knowledge economy. PISA uses a measure of SES derived from students’ reports of their parents’ education and occupation, and the material and cultural possessions in the home. The measure was scaled in this analysis to have an average of zero and a standard deviation of one for all US students.

There are three important results evident from the socioeconomic gradient for the US. First, its gradient is slightly below that of the OECD, and is slightly steeper, with a slope of 42.2 versus 36.0 for the OECD. Therefore, the gap in performance is greater for low SES students with OECD norms than there is for high SES students. Second, there is a high proportion of students scoring at Level 1 or ‘below Level 1’. 19.4% of US students scored at this low level in reading, and a further 22.7% scored at Level 2. These results are consistent with Shaywitz’s (2003) conclusion that that about 20% of children have reading disabilities, and with estimates based on the National Assessment of Reading Proficiency that indicate about 40% do not achieve basic levels of proficiency (Fletcher & Lyon, 1998). Third, as one might expect, the figure shows clearly that there are disproportionately more students from low SES backgrounds.

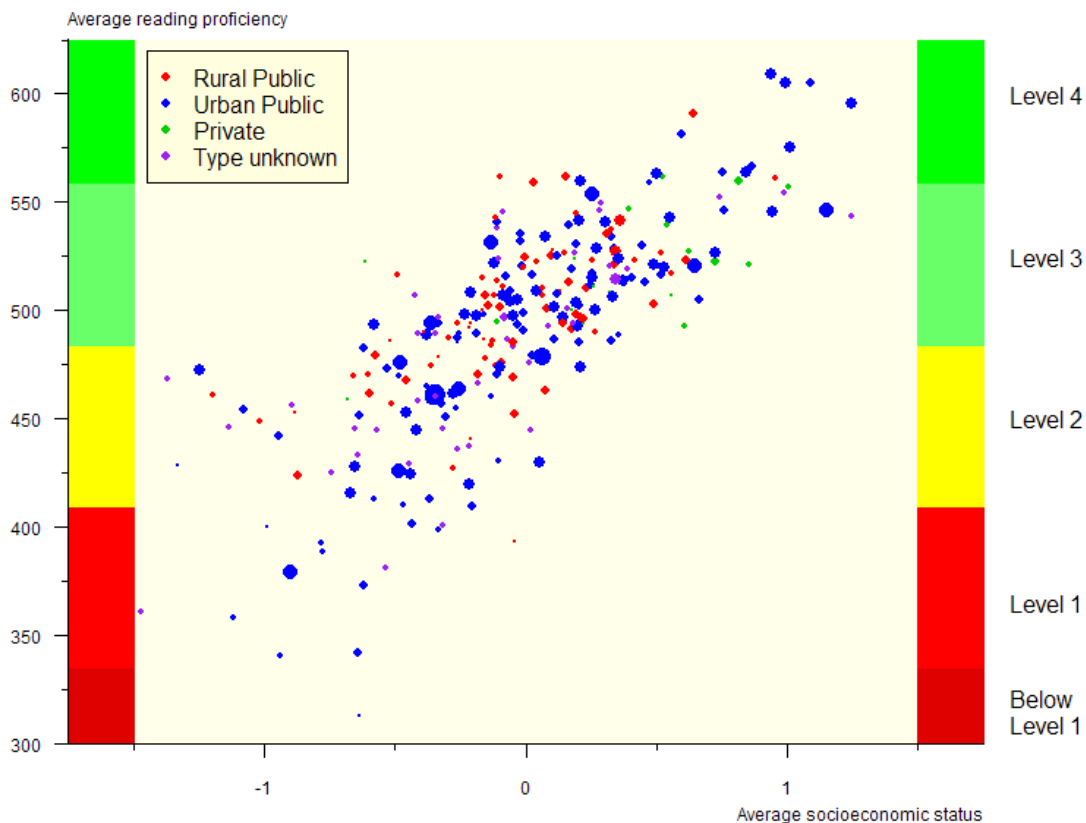


Figure 4. School profile in reading proficiency for the US. Source: 2003

Programme for International Student Assessment.

Figure 4 displays the ‘school profile’ in reading proficiency for US schools serving age 15 students, which like Figure 3 is based on data from PISA 2003. In this case, mean reading scores are plotted against mean SES with each dot representing one of the participating schools. Rural public, urban public, and private (rural and urban combined) are displayed with red, blue, and green dots respectively; the relative size of the dots corresponds to total school enrollment. The most prominent finding evident from the school profile is that there are several US schools where the *average* level of school performance is at or below Level 2. Most of these schools serve students from low socioeconomic backgrounds. In earlier work, based on PISA 2000, Willms (2004) found that the majority of these schools are large inner-city schools.

These results have profound implications for the assessment of added value in US schools. First, about 40% of youth at age 15 have poor reading skills. We can presume from other research that the majority of these students fell off track at about age 8, and thereafter their progress in all other subject areas was severely impeded by their poor reading skills. Second, a large proportion of these students are concentrated in a small number of the nation’s schools. Taking these two facts together, the question facing many schools is whether their efforts to ‘add value’ should be *singularly* directed at improving basic literacy skills, or directed at trying to improve scores as much as possible in the subject curricula for each grade. These findings also show that the concept of ‘added value’ has a fundamentally different meaning for educators in low SES urban public schools than it does for those in high SES public and private schools.

5. The right outcomes at the right level.

Multilevel studies that have partitioned the variance in student outcomes into student, classroom, and school components indicate that there is greater variation among classes within schools than among them. For example, the study of primary schools in Inner London Educational Authority by Mortimore, Sammons, Stoll, Lewis, and Ecob (1988) found that the variation in mathematics performance among classrooms ranged from 2 to 16%, while the variation among schools was only 3 to 8%. Hill and Rowe (1996) found that for a sample of grade 3 and 5 Australian students about 46% of the variation was among classes and 7% was among schools. Willms (2000) found that for a sample of grade 6 students in New Brunswick, Canada, about 7% of the variation in mathematics performance was among classes within schools, while only 5% was among schools. It is more difficult to estimate classroom components of variance at the secondary level, but it is most likely the case that the variation among classrooms is also greater than the variation among schools. Scheerens, Vermeulen, and Pelgrum's (1989) study of student performance at the end of the second year of secondary school found that in the United States 46% of the variation was among classes within schools, while only 10% was among schools.

These studies suggest that it is the work of the classroom teacher that matters more than the overall learning climate of the school. Given that in most middle and secondary schools students have different teachers for certain school subjects, the findings suggest that schools can have a different 'added value' depending on the outcome being considered. However, analyses of large-scale data bases using multilevel, multivariate techniques indicate that there is a high correlation among outcomes at the school level, even when student-level SES is taken into account. This is even the case when one considers the correlations among affective outcomes such as self esteem and academic achievement (Willms, 2006). These paradoxical results suggest that to tease out the differential effects of classroom instruction on different outcomes requires longitudinal designs with repeated measures over the course of a school year.

6. Attending to process

Measures of 'added value' attempt to assess the effect on students' learning associated with attendance at a particular school, net of the effects associated with student family background, and wider social and economic factors that lie outside the control of teachers or school administrators (Raudenbush & Willms, 1995; Willms & Raudenbush, 1989). In simple terms we could say that learning is a function of tractable factors (e.g., teaching quality, school and classroom learning climate) and intractable factors (e.g., students' ability, family SES, community support).

Learning = Function (tractable school factors + intractable family factors)

What we are really interested in, the 'added value', are the *tractable influences on student learning*. Value added models typically approach this indirectly. The equation is reversed:

Added Value = influence of tractable school factors = Learning | intractable family factors

There are three problems here. First, as discussed above, it is difficult to specify all of the intractable family influences, at both the individual and school levels. Second, not all

influences on student learning are clearly tractable or intractable – there are ‘shades of tractability’. Consider parent involvement in student learning for example; one could consider this an intractable factor, yet there are some schools that make a concerted effort to involve parents in their child’s schooling (Epstein, 1987). Third, schools want and need information on the tractable factors that affect student learning in their school. Without this, the value added exercise simply tells schools that they are doing well or not-so-well, without offering any direction on how to improve.

7. Links to intervention

A related point is that there is a very weak link between value-added assessment and the efforts by educators to reform their schools. The proponents of value added assessment could argue that effective schools and school systems generally have strong systems for monitoring performance (Bishop, 1997, 1999; Lezotte, 1991; Scheerens, 1992). Some economists maintain that large-scale testing improves the quality of teaching because teachers are “motivated to perform well in order to gain non-monetary rewards like reputation or acceptance among colleagues, parents, and students” (Juerges, Richter, & Schneider, 2004, p. 1). However, educational systems are often described as “loosely-coupled” (March & Olsen, 1976; Weick, 1976): the lines of authority and levels of the system are responsive to each other, but each element strives to preserve its own identity and autonomy. I believe that most teachers would maintain that their primary rewards come through witnessing student success in learning and emotional development. Thus, it is not surprising that many teachers are quick to dismiss value-added assessment as a top-down accountability tactic, and direct their efforts to preserving their autonomy.

Response from Educators

The drive for ‘standards-based reforms’ and ‘value-added’ models has been upon us for at least a decade, yet test scores in the US have declined over this period. In PISA, for example, the US scores in mathematics declined from 493 in 2000, to 483 in 2003, and 474 in 2006, while science scores fell from 499 in 2000, to 491 in 2003, and 489 in 2006. Reading scores fell from 504 in 2000 to 495 in 2003. (Data on reading scores for 2006 were not released due to an error in the way the test booklets were printed.) At the macro level, there is no strong evidence that top-down efforts based on test scores and ‘added value’ drives student performance.

The acid test of whether ‘value-added’ approaches work is whether teachers change their day-to-day classroom practice. Firestone, Mayrowetz, and Fairman (1988) found that reforms in Maryland and Maine aimed at improving test scores through standards-based reforms did not cause teachers to significantly alter their teaching strategies. Rather, they persisted in teaching approaches that emphasized individual practice by students on many small problems, rather than the exploration of new topics which require testing hypotheses, logical reasoning, and solving larger problems. Teachers have also expressed concern that estimates of value-added are unstable from year-to-year, especially for smaller schools (Jones & Whitford, 1997). Some researchers argue that ‘high stakes’ testing places students under undue stress (Smith, 1991).

Some Alternatives

National, state, and district monitoring systems that assess added value are usually based on *trailing indicators* of student performance measured after a fixed period of schooling. These systems are important as they also allow school administrators to understand how school outcomes are distributed within and between schools, and whether there are inequalities in educational outcomes among ethnic and social class groups and between the sexes. However, many systems do not include the collection of data describing classroom and school processes, and when they do the turnaround time from data collection and the reporting of results is too long – typically four to six months – making it difficult to link assessment results to school improvement efforts. District administrators, principals and teachers need *leading indicators* that provide a framework for intervention, can be used to guide school policy and practice, and can help staff identify issues relevant to particular students or groups of students. Recently, as part of the call for schools to establish ‘professional learning communities’ (Dufour, 2004; Newman *et al.*, 1996; Hord, 1997), principals and teachers have been pressed to use data for evidence-based decision-making. However, many of them lack the expertise to correctly interpret from assessment results, and need a process for using data to inform instruction (Murnane, Sharkey, & Boudett, 2005).

In Canada, over 300 schools are using an assessment system called *Tell Them From Me* (www.thelearningbar.com) which provides leading indicators of student engagement and wellness, and classroom and school climate. The aim is to meet the needs of school and district staff for timely, reliable, transparent results with a system that is cost-effective, user-friendly, extensible and affordable. *Tell Them From Me* includes a dynamic web-based student survey and optional teacher and parent surveys, which together assess 16 student outcomes pertaining to student engagement and wellness and 15 aspects of classroom and school learning climate that are known to affect learning outcomes. The climate measures are consistent with Lezotte’s (1991) correlates of school effectiveness, and with the features of a preventive, whole-school approach to supporting positive student behaviours as advocated by Sugai and Horner (2002).

A unique feature of *Tell Them From Me* is that student demographic data is entered at the beginning of the school year, and students are then randomly assigned to a week of the school year. Each week a different group of students completes the survey. The results are updated daily, creating a system for continuous feedback that enables school and district staff to respond immediately to specific concerns, and assess whether school reforms and interventions are having their intended effect. A form of ‘added value’ is included in the system: as each student completes the survey they are matched to students in the previous year’s database that have similar demographic characteristics using a procedure similar to propensity matching. As results are unfolding throughout the school year the school staff can compare their results to its ‘replica school’ comprised of similar students, as well as to school’s results for their school from the previous year and to the national average (Willms & Flanagan, 2007).

Another alternative to ‘added value’ approaches which entails leading indicators is the ‘response to intervention’ (RTI) model. RTI requires the continuous assessment of student progress in the regular classroom setting, accompanied with a tiered approach

to intervention for students with learning difficulties or behavioural challenges. The first tier involves universal strategies for improving student learning with extra support for those experiencing difficulties. The second tier involves targeted interventions for those who do not respond to tier 1 strategies. The interventions would typically include further, more detailed assessment, a different strategy for delivering instruction, and possibly a modified curriculum. The third tier aims to address the needs of students that do not respond to tier 2 interventions. At this level, students would normally receive a more intensive individualized program (Canter, Klotz, & Cowan, 2008).

This approach holds promise for ‘raising and levelling the learning bar’ as it shifts the focus away from comparing schools to improving the literacy skills of struggling readers and other vulnerable students. Large-scale implementation of RTI in the US would require a dramatic change in the allocation of resources among schools. For example, if we consider students with Level 2 PISA scores as requiring tier 2 RTI strategies, and those with Level 1 scores as requiring tier 3 RTI strategies, then the findings presented in Figure 4 suggest that the *primary* mission of about one-quarter to one-third of US schools would be on improving the literacy scores of their students using tier 2 and 3 strategies.

In New Brunswick, Canada, children are assessed in their early literacy skills at age 4 with a direct assessment called the Early Years Evaluation (EYE) (KSI Research International Inc., 2008). The evaluation identifies children who may require tier 2 and tier 3 interventions when they enter kindergarten the following year. In each school district, a ‘transition-to-kindergarten’ coordinator meets with the parents of these children to discuss the child’s learning needs and plan activities that will better prepare them for school. Some schools follow-up this evaluation with a teacher-based version of the EYE as well as further testing using the Dynamic Indicators of Basics Early Literacy Skills (DIBELS) (Kaminski & Good, 1996). This provides schools with the tools necessary for implementing a RTI program (Sloat, Beswick & Willms, 2007).

The use of *leading indicators* for identifying children for outcomes-based interventions and for guiding school policy and practice is not incompatible with value-added models based on annual testing. The critical elements in the examples above are that outcomes and processes are measured continuously with short assessments, results are immediate and transparent, and there is an explicit link to teacher practice. If value-added assessments are to be successful, they need to be based on comparisons to fixed standards rather than on comparisons with other schools, they need to be accompanied with leading indicators of policy and practice, and they need to place greater emphasis on the ‘value-added’ that comes through bringing students from learning-to-read to reading-to-learn.

References

- Alexander, K. L., Entwisle, D.R., and Olson, L. S. (2001). Schools, achievement and inequality: A Seasonal perspective." *Educational Evaluation and Policy Analysis* (23), 171–91.
- Alexander, K. L., Entwisle, D. R. & Olson, L. S. (2007). Lasting Consequences of the Summer Learning Gap. *American Sociological Review*. 72(2), 167-180.
- Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *American Economic Review, Papers and Proceedings*, 87, 260-264.
- Bishop, J. H. (1999) Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review*, 6, pp. 349-398.
- Cantor, A., Klotz, M. B., & Cowan, K. (2008). Response to intervention: The future of secondary schools. *Principal Leadership, February*, 12-15.
- Dufour, R. (2004). What is a "Professional Learning Community"? *Educational Leadership, May*, 1-6.
- Epstein, J. L. (1987). Parent involvement: What research says to administrators. *Education and Urban Society*, 19(2), 119-36.
- Firestone, W. A., Mayrowetz, D. & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, 20(2), 95-113.
- Fletcher, J. M., & Lyon, G. R. (1998). Reading: A research-based approach. In W. M. Evers (Ed.), *What's gone wrong with America's classrooms?* (pp. 49-90). Stanford, CA: Hoover Institution.
- Hart, T., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.
- Hill, P., & Rowe, K. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, 7(1), 1-34.
- Hord, S. M. (1997). *Professional learning communities: Communities of continuous inquiry and improvement*. Austin, TX: Southwest Educational Developmental Laboratory.
- Jones, K. and Whitford, B. L. (1997). Kentucky's conflicting reform principles. *Phi Delta Kappan, December*, 276-281.
- Juerges, H., Richter, W. F., & Schneider, K. (2004). Teacher quality and incentives: theoretical and empirical effects of standards on teacher quality. CESIFO Working Paper No. 1296. Available on-line at: www.CESifo.de.
- Kaminski, R. A., & Good, R. H., III. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25(2), 215-227.
- KSI Research International Inc. (2008). *Early Years Evaluation – Direct Assessment* (Revised June 2008). Fredericton, NB: KSI.

- Lezotte, L. W. (1991). *Correlates of Effective Schools: The First and Second Generation*. Okemos, MI: Effective Schools Products Ltd.
- March, J. G. & Olsen, J. P. (1976). *Ambiguity and choice in organizations*. Bergen, Norway: Universitetsforlaget.
- Mortimore, P., Sammons, P. Stoll, L., Lewis, D. & Ecob, R. (1988). *School Matters*. Wells: Open Books.
- Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2005). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for students placed at risk*, 10(3), 269-280.
- Newmann, F. M., et al. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco: Jossey-Bass.
- Organisation for Economic Cooperation and Development (OECD) (2004). *Learning for Tomorrow's World: First results from the PISA 2003*. Paris: OECD.
- Raudenbush, S. W. & Chan, W. (1993). Application of hierarchical liner models to the study of of adolescent deviance in overlapping cohort designs. *Journal of Clinical and Consulting Psychology*, 61(6), 941-951.
- Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. Collins & A. Sayer (Eds.), *New Methods for the Analysis of Change*. Washington, DC: American Psychological Corporation.
- Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioural Statistics*, 20(4), 307-335.
- Scheerens, J. (1992). *Effective schooling: Research, theory, and practice*. London: Cassell.
- Scheerens, J., Vermeulen, C., & Pelgrum, W. J. (1989). Generalizability of instructional and school effectiveness indicators across nations. *International Journal of Educational Research*, 13(7), 789-799.
- Shaywitz, S. (2003). *Overcoming dyslexia*. New York: Alfred A. Knopf.
- Sloat, E.A., Beswick, J.F., & Willms, J.D. (2007). Using early literacy monitoring to prevent reading failure. *Phi Delta Kappan*, 88(7), 523-529.
- Smith, M. L. (1991). Put to the test: The effects of external testing in students. *Educational Researcher*, 20(5), 8-12.
- Statistics Canada. (2005). *National Longitudinal Survey of Children and Youth Microdata User Guide - Cycle 5*. Ottawa.
- Sugai, G. & Horner, R. (2002). The evolution of Disciplinary Practices: School-wide positive behavior supports. In James K. Luiselli & Charles Diament (Eds.), *Behavior psychology in the schools: Innovations in evaluation, support, and consultation*. New York: Hawwoth Press.

- Weick, K. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21, 1-9.
- Willett, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education, Volume 15* (pp. 345-422). Washington, DC: American Educational Research Association.
- Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26(3), 209-232.
- Willms, J.D. & Kerckhoff, A.C. (1995). The challenge of developing new social indicators. *Educational Evaluation and Policy Analysis*, 17(1), 113-131.
- Willms, J. D. & Flanagan, P. (2007). Canadian students "Tell Them From Me". *Education Canada*, (47)3, 46-50. Also see (www.thelearningbar.com).
- Willms, J.D. (2000). Monitoring school performance for "standards-based reform". *Evaluation and Research in Education*, 14(3&4), 237-253.
- Willms, J.D. (2004). *Reading Achievement in Canada and the United States: Findings from the OECD Programme for International Student Assessment*. Ottawa, ON: Human Resources and Skills Development Canada.
- Willms, J. D. (2006). *Learning divides: Ten policy questions about the performance and equity of schools and schooling systems*. Montreal, QC: UNESCO Institute for Statistics.
- Willms, J. D. (2008). *Developmental trajectories of Canadian children and youth*. Ottawa, ON: Human Resources and Skills Development Canada.
- Willms, J. D. (in press). Contextual effects on student outcomes. *Teachers College Record*.