How RDCs Changed Public Policy Research in Canada (and will again in the future)

Michael Baker
University of Toronto

Empirical public policy research in Canada used to be very different.

How different?

It was theoretical.

There were no data.

My solution (circa 1990):

Pool resources with Dwayne Benjamin, another new empirical public policy economist at the University of Toronto.

Purchase the microdata files of the 1986 and 1991 *Survey of Consumer Finances*.

Price: $10,000 ($15,682.00 in 2013 dollars).

Write four papers based on these data.

Observation:

Four papers may have been one too many

The paper in question is:

Baker, Michael and Dwyane Benjamin, "The Receipt of Transfer Payments by Immigrants to Canada", *Journal of Human Resources*, 30, Fall 1995, 650-676.

The paper that questions it is:

Crossley, Thomas F., J. Ted MacDonald and Christopher Worswick, "Immigrant Benefit Receipt: Sensitivity to Choice of Survey Years and Model Specification," *Journal of Human Resources*, 36, Spring 2001 379-397.

The finding in the original paper is:

> **...We also find that "assimilation" leads to greater participation in both these programs. There is a correlation of entry participation and vintage which indicates more recent immigrant cohorts have higher recipiency rates than their predecessors.**

The finding in the Crossley et al. paper is

> **Baker and Benjamin (1995) analyse the receipt of unemployment insurance by immigrant men using two years of the Canadian Survey of Consumer Finances. This study replicates their research on 13 of the annual surveys.**
>
> **Estimates are found to be sensitive to the choice of survey years. Furthermore, the standard fixed effects model of assimilation is rejected...**
>
> **Estimates from the more general model do not indicate higher incidence of benefit receipt, ceteris paribus, among more recent cohorts or that immigrants assimilate toward greater receipt of benefits.**

The implication is that by using only two years of data, Baker and Benjamin had been unable to test the validity of their empirical specification, and using the 1986 and 1991 data had happened upon conclusions that did not hold for other year/pairs in the 1980s.

This example brings up a number of important points:

1. In 1990, the “market clearing price” of a single year of data had implications for both the amount of empirical public policy research being conducted, and how it was being conducted.

   Good practices such as verifying results for additional years or in other data sets were simply too expensive.

2. Researchers in other countries were far ahead of us. Some context for this situation was US research that used 18 years of the Annual Demographic Files of the *Current Population Survey* spanning the period 1968-1984

   Murphy, Kevin M. and Robert H. Topel, "The Evolution of Unemployment in the United States: 1968-1985" *NBER Macroeconomics Annual*, Vol. 2 (**1987**), 11-58.

3. Tom Crossley, Ted MacDonald and Chris Worswick were trying to ruin my career,

Kidding.

How were Crossley et al. able to run regressions that Baker and Benjamin were not? – **The Data Liberation Initiative**

As Crossley et al. note:

> **It has subsequently become feasible to merge almost annual SCF surveys covering a 14-year time period.[3]**
>
> **[3] Statistics Canada's recent "Data Liberation Initiative" has substantially reduced the cost of microdata.**

The DLI significantly changed the landscape of Canadian empirical public policy research.

However, to a certain extent it only brought us to where researchers in other countries had been more than a decade earlier.

In 2001, Crossley et al. were able to do what Topel and Murphy had done in 1987.

In addition the world was moving on…

By the late 1980s researchers in other counties were making enormous strides using longitudinal data.

Panel surveys such as the *Panel Survey of Income Dynamics* and the *National Longitudinal Surveys* had been in operation since the 1960s.

Using these long records of people's social and economics activities there had been discoveries on the persistence of poverty and unemployment spells, the intergenerational correlation of income and economic status, and the trajectory of human capital development.

These discoveries had a transformational impact on how we think about public policies in these areas.

These panel data also facilitated the use of statistical methods such as fixed effects, which provided powerful ways to control for unobserved heterogeneity across people, and hazard models that provided new insights to the spell data.

Simply put, researchers had little to no means to make contributions to these quickly emerging literatures using Canadian data.

Existing Canadian panels, such as the Labour Market Activity Survey were simply too short (2 and 3 years) to be useful in this context.

However,

Statistics Canada initiated some new longitudinal surveys in the early 1990s, which due to privacy concerns could only be released publicly as cross sections.

The surveys included the NLSCY, SLID, NPHS and a little later the YITS, WES and LSIC.

So the following situation existed:

Statistics Canada and its partners had new, expensive panel surveys that no one outside government could access easily. *They needed researchers*.

Researchers were increasingly unable to make contributions at the highest level using Canadian data. *They needed data*.

While it does not necessarily follow that gains from trade will be realized when they are so obvious, fortunately in this case they were.

The answer was (and is) the **Canadian Research Data Centre Network**.

So what?

It was and remains possible to have a very successful career as a empirical researcher in Canada exclusively using data from other countries.

In fact in economics there is evidence that this is what an increasing number of Canadian based economists do – Simpson, W. and J.C. H Emery, "Canada Economics in Decline: Implications for Canada's Economics Journals", *Canadian Public Policy*, 38(4) 2012, 446-470.

So the private benefit of the CRDCN to researchers is not entirely straightforward.

The private benefit of the CRDCN to Statistics Canada and its partners is pretty clear, but perhaps they simply made a bad call on the investment.

Because the CRDCN came to be with government intervention in the form of SSHRC, CHIR and CFI funding, we are left to search for some sort of externalities from this investment.

While I think that there are numerous externalities from this investment, I would like to focus on two:

1) RDC data provides an opportunity to include Canada in cross country analyses
2) RDC data enables research tailored to public policy priorities in Canada.

Research from other countries is clearly not a substitute in (1), and is not always a perfect substitute for research conducted using Canadian data given Canadian policy priorities.

I would like to explore these points to try to demonstrate the importance of Canadian based research.

Based on something I know a little about – early childhood human capital accumulation.

Based on something I know even more about – my own research.

1) Inclusion of Canadian evidence in cross country studies.

Other than databases maintained by international organizations (e.g., OECD) containing quite aggregate data, the Canadian evidence is often missing from multi national studies.

Because RDC data remains difficult for foreign based researchers to access this will probably continue to be true to a certain extent.

However, the types of data available in the RDC complement those from other countries, and so Canadian based researchers have opportunities to draw cross country comparisons.

This can help us calibrate Canadian economic and social institutions to their foreign counterparts, which can be helpful when it is necessary to rely on foreign evidence.

As an example, in recent research with Kevin Milligan we have matched information in the NLSCY with information in the U.S. *Early Childhood Longitudinal Survey-Birth Cohort* and the UK *Millennium Cohort Study.*

Our focus is the time investments parents make in very young children, and how they differ between girls and boys.

One finding of interest is that in all three countries parents spend more time in teaching activities with girls than with boys.

Also the size of the boy-girl gap varies by mother's relationship status, and in all three countries in larger when a mother cohabitates.

Parental Time Spent Reading —First Born

	Canada	U.K	U.S.
Age 0-23 months	-0.114***		-0.083**
	(0.037)		(0.041)
Age 2-3 years		-0.090***	-0.118***
		(0.027)	(0.034)
Age 3-5 years	-0.162***	-0.046*	-0.119***
	(0.043)	(0.026)	0.036)

Parental Time Spent Reading by Partnership Status—First Born

	Canada	U.K	U.S.
Married	-0.081*	-0.046	-0.065
	(0.050)	(0.031)	(0.040)
Cohabitation	-0.285**	-0.191***	-0.265***
	(0.112)	(0.076)	(0.098)
Single	-0.031	-0.066	-0.160**
	(0.137)	(0.076)	(0.076)

Here I think a value of the cross country validation is that the findings are not a function of some particular characteristic of a single country (e.g., the much higher rates of out of wedlock births in the UK and especially the U.S).

2) RDC data enables research tailored to public policy priorities in Canada.

Here the U.S. plays a significant role in the story.

The U.S. has a very large, fairly well funded and productive research community.

One often marvels at the depth of research on minutiae of U.S. policy such as clauses of Medicaid legislation.

One often sees the U.S. evidence base cited for Canadian policy initiatives.

However, Canadian policymakers often can't find the evidence base for their initiatives in U.S. based research, and sometimes the evidence that they find is not appropriate.

a) In relative terms the U.S. simply does not have that much policy so the relevant research may not exist

- In the U.S. there is so much research and so little policy
- In Canada there is so much policy and ...

An example here is pay equity policies, and, increasingly, union policies.

b) In the area of social policy, targeting is the dominant system for delivery in the U.S. while in Canada it is far more likely that the delivery is universal.

The example I will explore here is early childhood education policies.

For those not in this field there is quite a large body of research in this area, which is also growing quite rapidly.

Much of this activity is in the spirit of the Barker fetal origins hypothesis that a child's environmental conditions (malnutrition) in utero, or more generally in early childhood, can have a significant impact on adult outcomes.

This research has precipitated a number of ECE policies around the world ranging from cradle to school early child education programs to less formal family support initiatives and preschool programs for older children.

Research from the U.S. reflects that fact that most of the ECE programs in that country are targeted.

Some of the best evidence is for extremely targeted programs such as the Perry Preschool Program.

Other research is for more moderately targeted programs such as Head Start.

If you unpack the research on seemingly more universal programs such as preschool it is often the case that the initiatives are in school boards that serve disadvantaged populations.

In Canada, most ECE initiatives are universal – e.g., the Quebec Family Plan, Ontario's full day kindergarten program.

There are many rationales for ECE programs (work/life balance, gender equity) but from a child development point of view universal delivery can be a bit tricky.

If more advantaged children derive a greater benefit from the program than less advantaged children then the program becomes an instrument that promotes social and economic inequality.

If advantaged children do not benefit from the program at all then the program might promote social equity but is really an instrument of, we hope, benign childcare for these children.

So, some would argue, that we need advantaged children to gain some significant benefit from the program but not as much as disadvantaged children.

Clearly theory isn't going to tell us whether this is true.

In Canada we are fortunate that more advantaged children outnumber less advantaged children.

But this means that more advantaged children are the majority of children treated by universal ECE programs.

Therefore what we need to know is what is it like to grow up advantaged in Canada and how ECE programs affect more advantaged children.

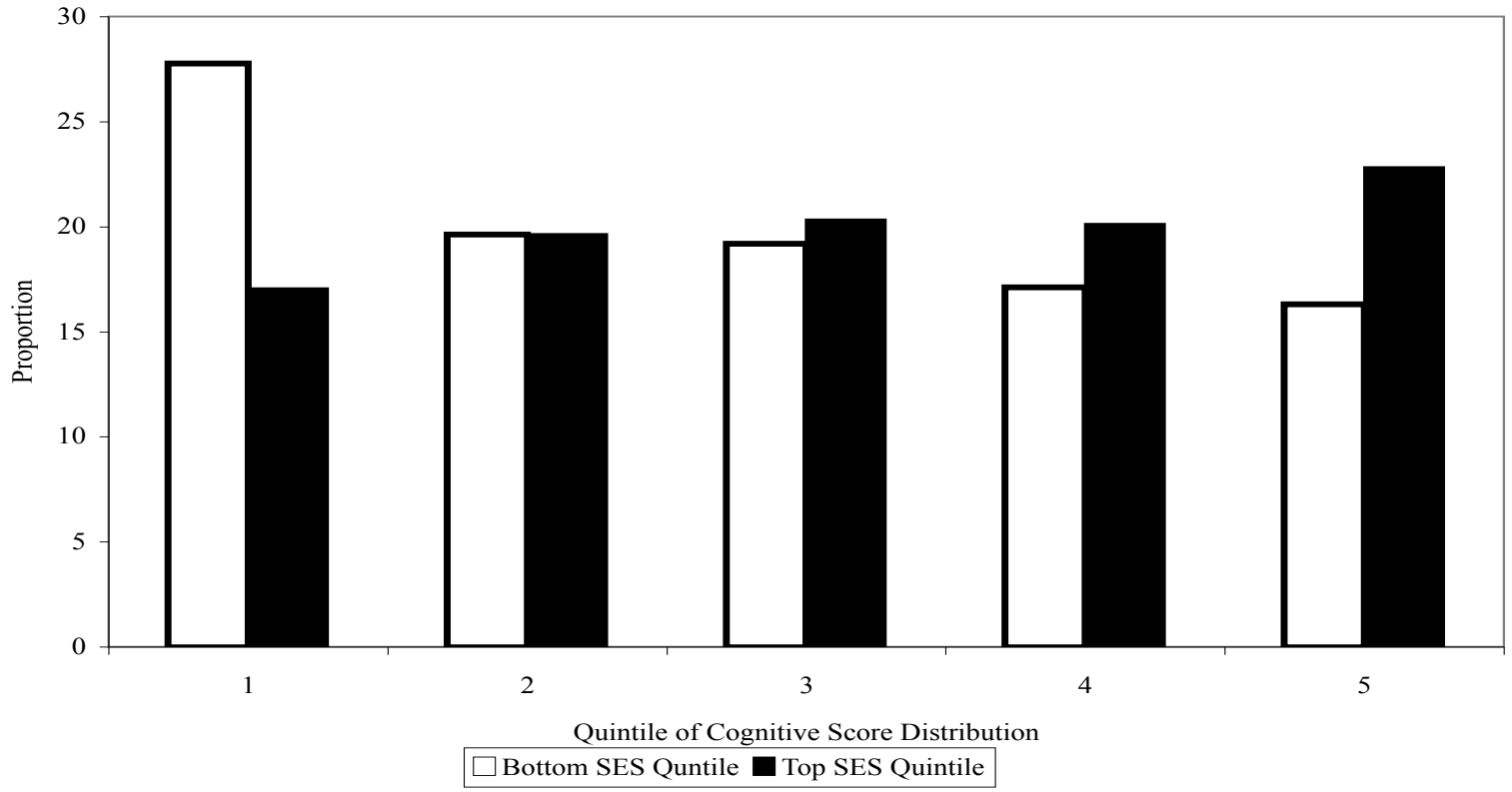
U.S. research doesn't really answer these questions.

Research from Europe is potentially more helpful but different mitigating circumstances are also at play.

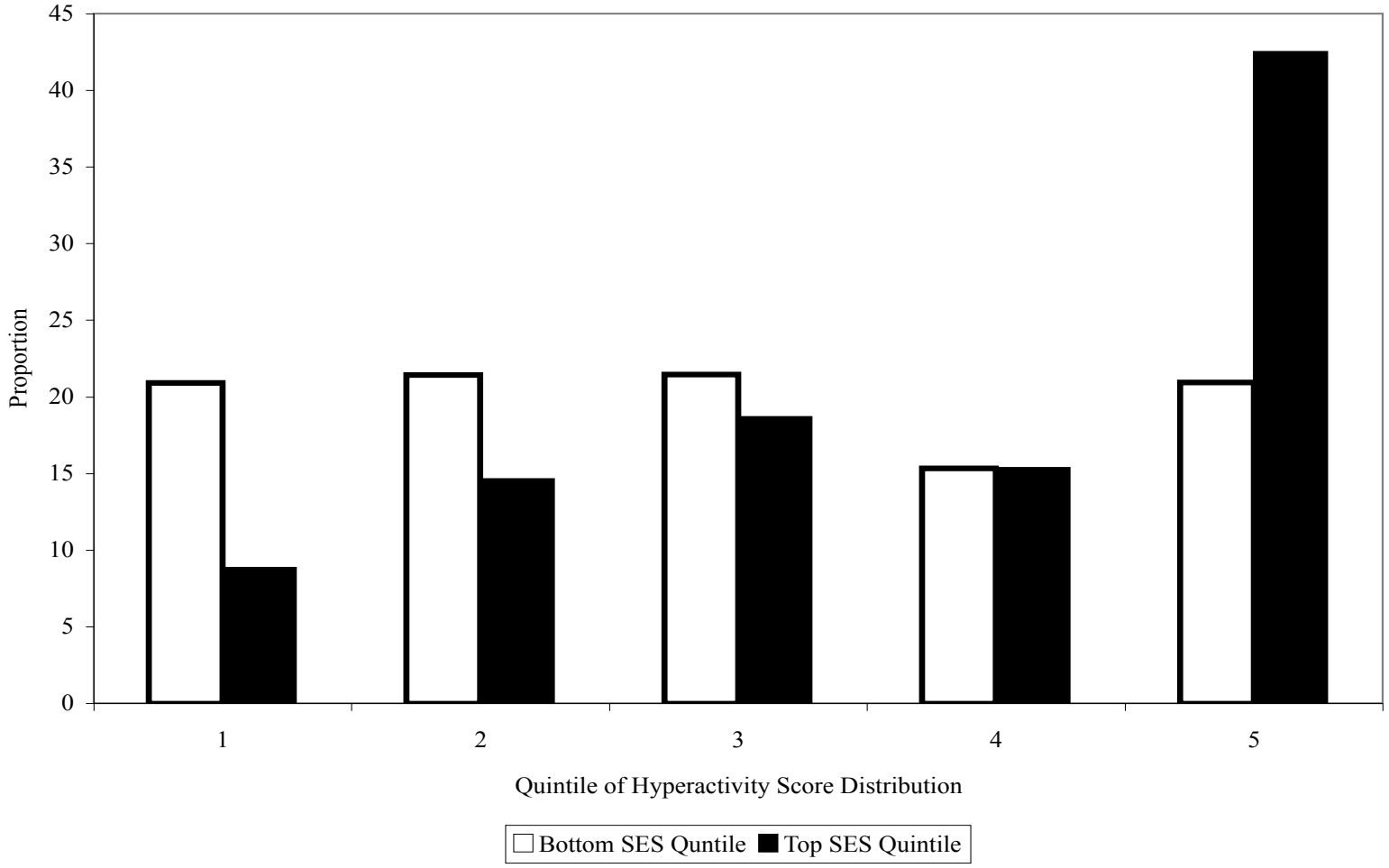
Fortunately longitudinal data available in the RDCs can help us answer these questions.

Using NLSCY data I have performed some analysis that starts to address these questions.

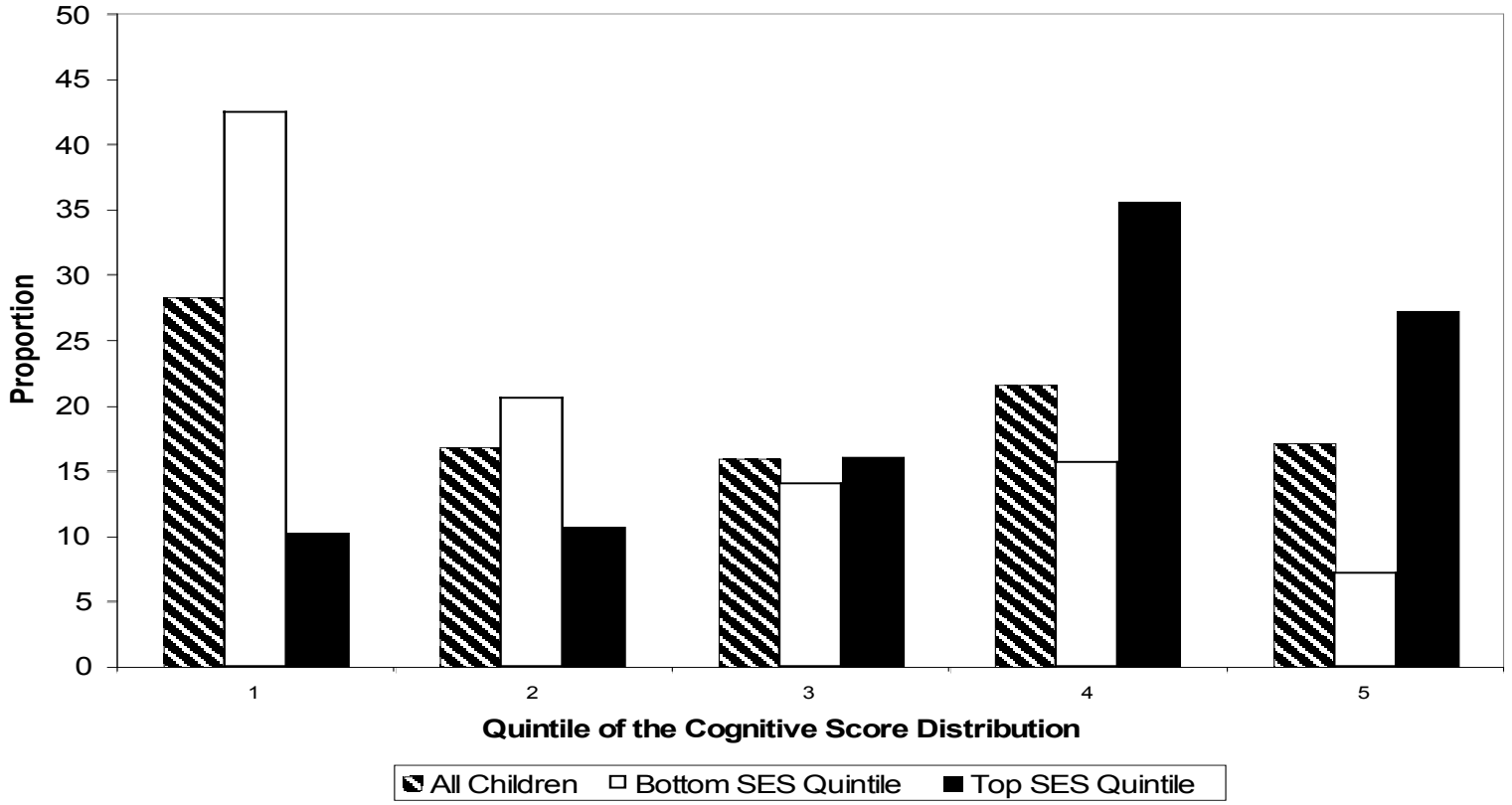
The distribution of cognitive scores for children aged 0-5 from the top and bottom quintiles of family SES



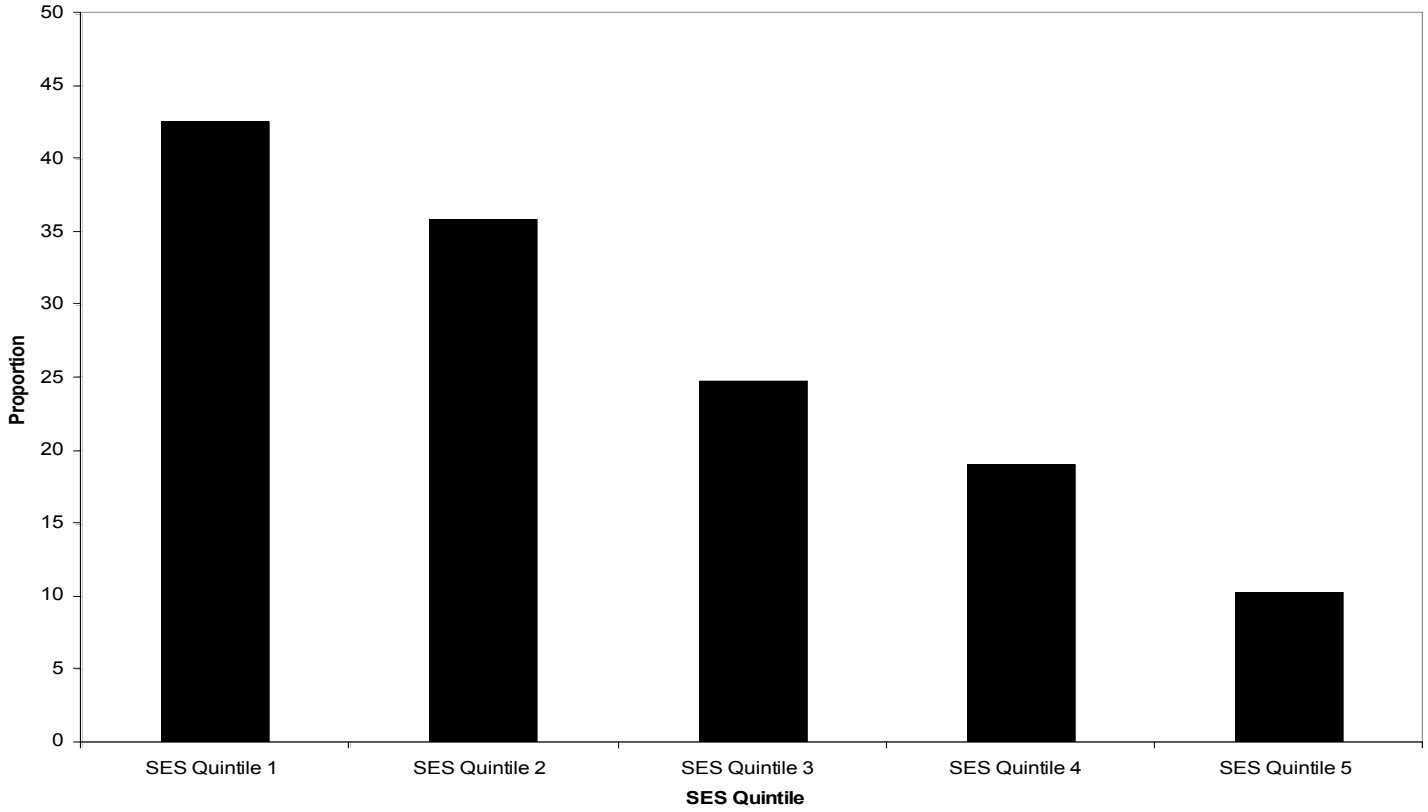
The distribution of the hyperactivity scale for children aged 2-5 from the top and bottom quintiles of family SES



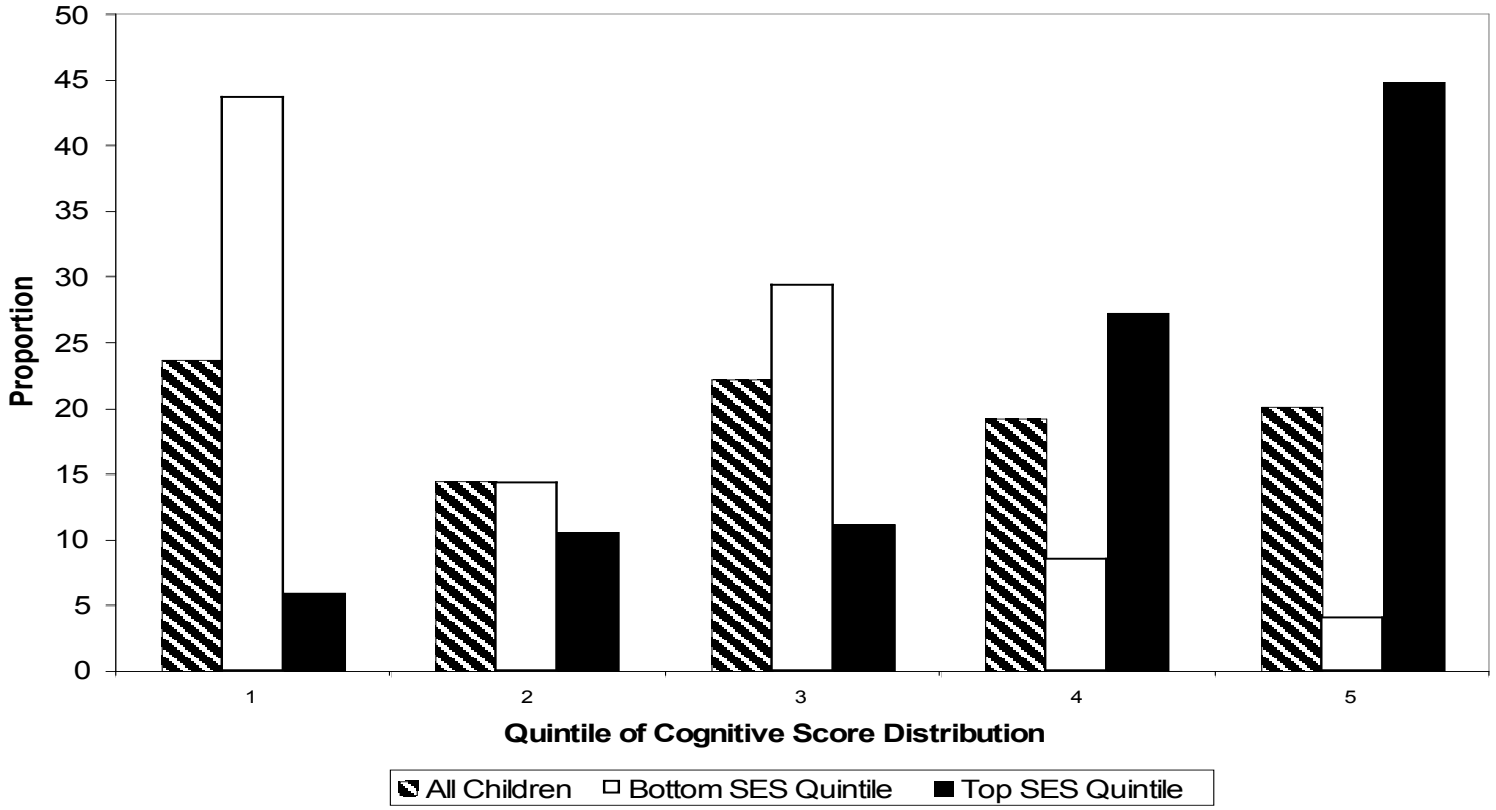
Transition probabilities from the bottom quintile of the cognitive score distribution at ages 0-3 to the quintiles of the cognitive score distribution at ages 12-15, by quintile of family SES at ages 0-3



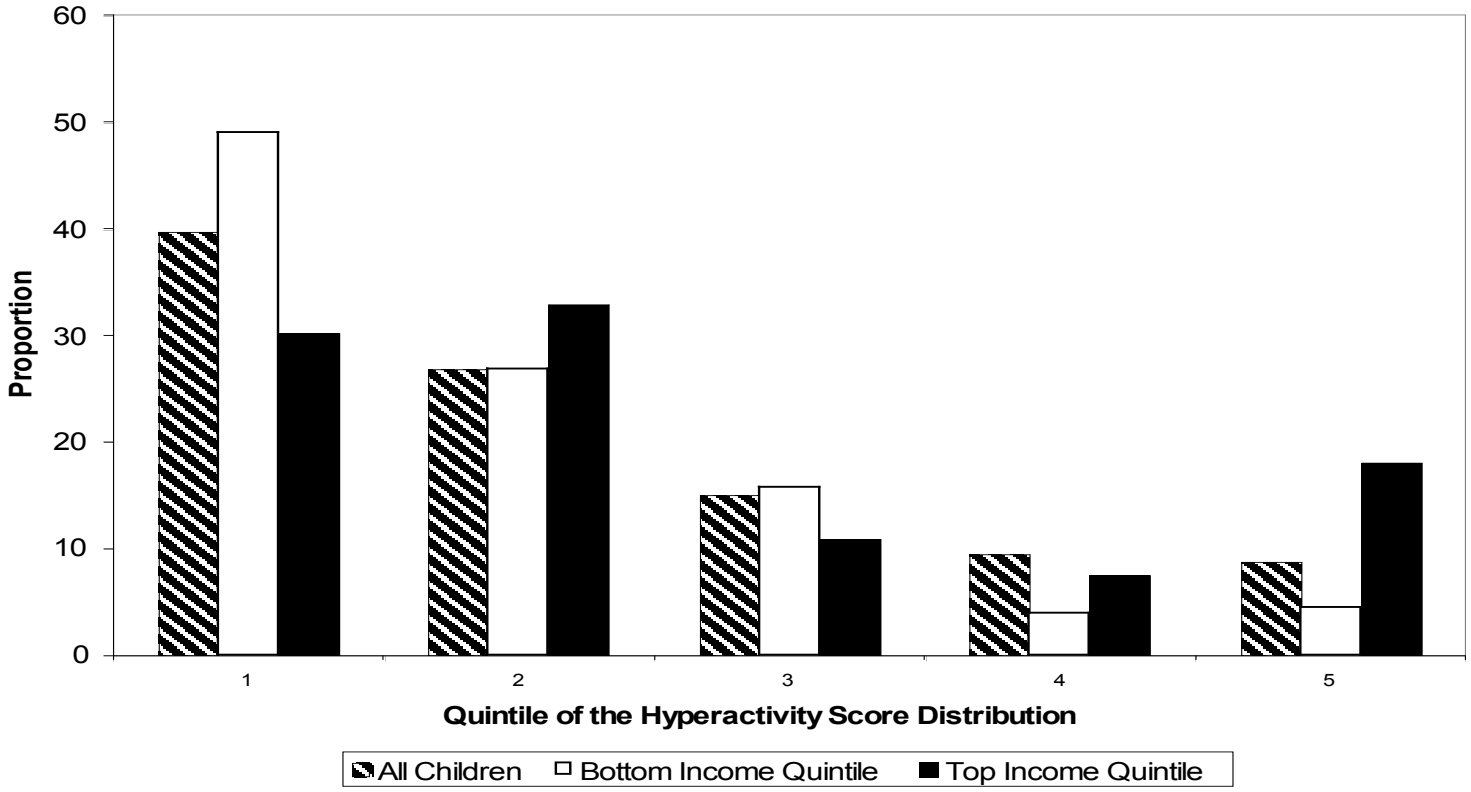
Transition probabilities from the bottom quintile of the cognitive score distribution at ages 0-3 to the bottom quintile of the cognitive score distribution at ages 12-15, by quintile of family SES at ages 0-3



Transition probabilities from the top quintile of the cognitive score distribution at ages 0-3 to the quintiles of the cognitive score distribution at ages 12-15, by quintile of family SES at ages 0-3



Transition probabilities from the bottom quintile of the hyperactivity scale distribution at ages 2-5 to the quintiles of the hyperactivity scale distribution at ages 7-11 by quintile of family SES at ages 2-5



This evidence suggests that any cognitive and behavioural delay of more advantaged children at young ages is either transitory, or remediated by their parents' investment.

On the other hand, early life cognitive and behavioural delay of less advantaged children is more persistent and the early life achievement is more perilous.

If true these findings pose challenges for the universal delivery of ECE in a world of finite budgets.

In particular, if in the attempt to serve all children the quality and type of ECE delivered to less advantaged children is compromised, then these programs will likely not have the intended effects on the distribution of social and economic outcomes.

The Future

While the CRDCN has facilitated significant strides in policy research using Canadian data, researchers in other countries are moving on.

“Big Data” has become rallying cry for a coalition of policy makers, politicians, advocates and researchers.

While there are numerous workshops and conferences exploring the implications of large administrative databases for policy making, researchers in other countries are already putting these ideas into practice.

The Data Effect: Our 21st Century Opportunity

Session 1: Curating the Data: Why Good Data is as important as Big Data

Session 2: More with Less: Data in the Public Sector

Session 3: The New Partnerships: How shared research is re-shaping innovation in health & governance

Session 4: Health Care in the Data Age

Leaders in the analysis of administrative data are European researchers who have been using immense population registries for research for at least a decade.

However, administrative data is also becoming a norm for research in the U.S.

Kevin Milligan has noted that at a conference of leading labour economists held over the summer, 19 out of the 24 papers presented used administrative data, while only 2 relied on survey data as the primary information source.

What's the big deal about big data?

1. Very large sample sizes allow the use of certain, very popular, statistical techniques that facilitate causal inference in a non-experimental context (e.g., regression discontinuity).

   As has been the case in the past (e.g., Heckman correction, fixed effects, etc.) these techniques become the "gold standard" for analysis, and therefore a "barrier to entry" to the top journals for those that do not use them.

2. Very large sample sizes allow the analysis of very detailed spatial and demographic heterogeneity.

   This is particularly important in a country in which provinces have responsibilities in many social and economic areas.

Therefore, as in 1990s we are at a juncture where there must be a transformation of the data landscape for researchers using Canadian data to full participate in their respective intellectual communities.

So the following situation exists:

Government and their agencies have vast and rich administrative data sets that could be exploited to support their jurisdictional responsibilities and that no one outside government can access easily. *They need researchers*.

Canadian researchers are increasingly unable to make contributions at the highest level using Canadian data. *They need data*.

While it does not necessarily follow that the gains from trade will be realized when they are so obvious...

If the RDCs are to continue to act as an agent for the promotion of research useful to Canadian public policy, getting administrative data into the hands of researchers must be a priority.

Big data is the organizing principle of the RDC Network's planning for the next granting cycle.