International Journal of Contemporary Educational Research (IJCER)

www.ijcer.net

Evidence-Based Policy Making in Education

Margaret Wu[1]

[1] Victoria University, Melbourne, Australia

To cite this article:

Wu, M. (2014). Evidence-based policy making in education. *International Journal of Contemporary Educational Research, 1*(1), 1-8.

IJCER International Journal of Contemporary Educational Research
Volume 1, Number 1, January 2014, Page 1-8 ISSN: 2148-3868

Evidence-Based Policy Making in Education

Margaret Wu[1*]
[1] Victoria University, Melbourne, Australia

Abstract

In the current climate of “accountability” and “transparency” as demanded by the public, policy makers justify their actions by drawing on research findings and data collected by various means. There appears to be a belief that quantitative data provide more credible evidence than qualitative data. Hence the use of data has become pivotal in decision-making. More recently, education policy documents drawing on international student survey results have appeared around the world. This paper evaluates some of the evidences used by policy makers and shows that there is a great deal of uncertainty surrounding the data underlying these research findings. More importantly, the paper demonstrates that statistics alone cannot provide hard evidence. In fact, we need to draw on our own experience and a great deal of sense-making in interpreting data and drawing conclusions.

Key words: Evidence-based decision making, accountability, transparency.

Introduction

In recent years the rhetoric from politicians has often included words such as accountability, transparency and evidence-based decision-making. While accountability and transparency should be central in policy-making, there are cases where such processes are based on misguided evidence, and the proposed policies can be ineffective or even detrimental to public interests.

As an example, a New Zealand Treasury briefing paper released in March 2012 (New Zealand Treasury, 2012(a)), suggested that low student performance related to ineffective teaching. Therefore if the quality of teaching can be lifted, student performance will be raised. Further, the briefing paper recommended increasing class size to free up money to fund initiatives to raise the quality of teaching. The briefing paper cites numerous research papers and reports to justify the recommendations. Similarly, in a policy paper titled “Lessons from PISA” (Paine & Schleicher, 2011), the authors believe that investing in the improvement of teacher quality is the most important lesson for the United States from PISA (Programme for International Student Assessment). The article further cites an OECD report that an increase of 25 PISA score points will lead to a gain of $41 trillion for the U.S. economy, thus linking student achievement to GDP.

This paper examines some of the evidence used by policy papers such as the examples given above. The purpose of this paper is to bring a statistical viewpoint about data, to take a closer look at how research findings can be cherry-picked for shaping education policies for whatever reason, and to raise cautions for anyone using research findings. Some published data sources are used as examples to illustrate the pitfalls in putting too much trust in quoted research findings. These examples include findings from the OECD PISA project. The key issue relates to the complexity of the modern world where contributing factors to educational achievement are multi-faceted and intertwined. Data collected often have many caveats attached. The interpretation of data requires statistical literacy well beyond the knowledge of the layperson.

Notwithstanding the critical nature of this paper on the use of evidence, it should be made clear at the outset that the benefit of education is not disputed and that improving education should stay high on policy-makers’ agenda. The aim of this paper is to highlight the complexity and potential misuse of data, and suggest that “numbers” alone do not provide hard evidence, but numbers must be interpreted with well-grounded theories. So, in fact, sense-making is essential in using evidence to shape policies. This paper concludes that while “consumers” of educational research need to be vigilant about the applicability of research findings, the ultimate

[*] Corresponding Author: *Margaret Wu, margaret.wu@vu.edu.au*

responsibility must reside with the researchers themselves to provide clear and solidly grounded results and recommendations. Thrupp's article (Thrupp, 2010) titled "The politics of being an education researcher: minimizing the harm done by research", clearly highlights the need to be aware of how research findings could be misused, and that the responsibility also falls on the researchers to prevent the misuse of their findings. This is really where accountability and transparency start. Unfortunately, many research reports still fall short of taking such cautions, as the following examples illustrate.

Linking education outcomes to economic and social benefits

As an example, the New Zealand Treasury illustrated the benefits of lifting student achievement by linking educational outcomes to economic growth. According to the Treasury:

"... if overall student achievement could be lifted by 25 PISA points ... GDP would be expected to be higher than it otherwise would be by 3-15% by 2070." (p.2, New Zealand Treasury, 2012(a))

The assertion that improving education will have economic and social impact seems plausible. However, it would appear equally believable to suggest that the causal relationship is the other way round, or in fact, bi-directional. That is, improving the economy will raise education standards. OECD PISA 2009 report (OECD, 2010a) states:

*"...*GDP per capita influences educational success, but this only explains 6% of the differences in average student performance.*" (p.3, OECD, 2010a)*

First, the PISA report casts the relationship between education outcome and GDP as one where GDP *influences* educational success, not the other way round. Second, PISA results show a tenuous relationship between education performance and GDP. Figure 1 shows a plot of GDP against PISA 2009 Reading mean score for 34 countries, where each data point in the plot is one country (Source: OECD, 2010a, p35).

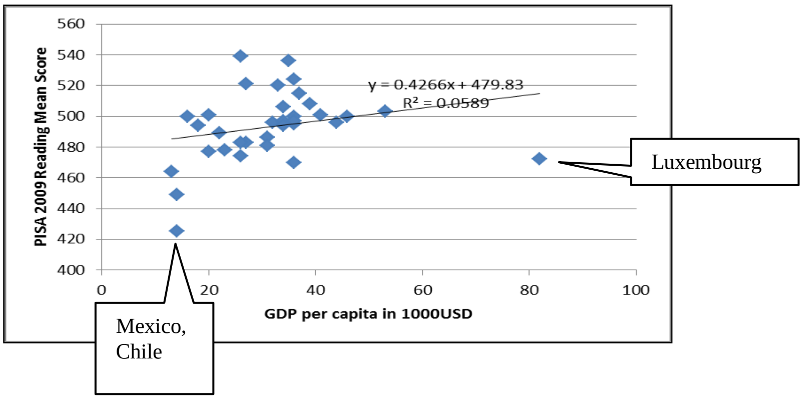

Figure 1. PISA 2009 Reading performance and GDP

Figure 1 does not show any discernible linear relationship between Reading mean score and GDP, at least not visually. For example, Mexico and Chile appear to have lower than expected Reading scores given their GDP values. Further, Luxembourg has very high GDP, but its Reading performance is not the highest in the group of countries shown. This shows that the relationship between Reading score and GDP depends very much on the set of countries being analysed, and the PISA data do not suggest a strong relationship between education outcome and GDP.

Despite the inconclusiveness of establishing a clear association between Reading performance and GDP, the PISA report (OECD, 2010a) makes the following claim on page 158:

*"...*bringing all students to Level 2 [on the PISA reading scale] could boost the combined economic output of OECD countries by around USD 200 trillion.*" (p.158, OECD, 2010a)*

But a cautionary note follows this claim:

"While such estimates will always be associated with considerable uncertainty, they suggest that the cost of educational improvement is just a fraction of the high cost of low educational performance." *(p.158, OECD, 2010a)*

So in this excerpt, the PISA report reverses the direction of the causal relationship and suggests that higher student performance could lead to much higher GDP. It is also surprising that such claims are made while stating that there is *considerable uncertainty*.

It is not surprising that policy-makers are confused by the recommendations made by OECD. Unless one pays attention to the caveats in the reports and carries out one's own analyses with the data, a person can easily begin to make big claims about how education outcomes can lead to GDP growth. In contrast, PISA data tell us that the relationship between education outcomes and GDP depends greatly on an individual country's context. Making forecasts on GDP growth based on students' test scores is extremely unreliable. Yet, policy-makers frequently made these claims, citing the OECD reports as evidence and ignoring any note of caution.

Making conjectures about the direction of causal relationships is not just confined to OECD reports. A report by Hanushek and Woessmann (2009) about education outcomes and economic growth also suggests that education achievement increases GDP. Hanushek and Woessmann attempted to establish that the relationship between international testing results and GDP was causal. They did this by controlling for possible mediating variables, and concluded that the association between test results and GDP was not likely due to other mediating variables, but there was a *real* association. Nevertheless, Hanushek and Woessmann never established the direction of the causal relationship. In fact, the test data used in the Hanushek and Woessmann report were mostly obtained in the past 20 years, while the GDP growth data were for the past 50 years. Any association established between test scores and GDP will likely suggest that higher GDP leads to higher test scores rather than the other way round. If high test scores lead to high GDP, then one would expect a time lag for the association, since today's students are tomorrow's workforce. If the students have high test scores, their impact on GDP will only be apparent after these students participate in the workforce. Notwithstanding these arguments, Hanushek and Woessmann suggest that high test scores lead to high GDP, as reflected in the title of their report.

The above example is not used to refute the claim that education level is related to economic growth. The relationship between education and economy may very well exist, but the particular set of OECD data shown in this paper does not lead to such conclusions. This could be because the analysis is too crude – perhaps other contextual factors need to be taken into account, or perhaps the sample size is too small to provide the power to establish a relationship. Nonetheless, there is an inconsistency between the data and the findings. In particular, putting a dollar figure (USD 200 trillion) based on a very tenuously established regression line is making a very far-fetched inference.

There have been many studies examining the link between education and economic growth. Bredt and Syez (2007) carried out an extensive literature review on this topic and reported that a large number of studies have found a positive association between education and economic growth, but the empirical work in these studies has not been able to establish any causal link between education and economy (p. 7). Bredt and Syez suggest a bi-directional relationship:

> *Not only will more technologically advanced economies require a higher skilled workforce, but they will also have the resources to invest in expanding their educational institutions and research sectors. (p.7, Bredt & Syez, 2007).*

The point of the above discussions is not so much about determining exactly how education will affect economy, but the examples demonstrate that, at least in some cases, researchers have not been particularly prudent in making claims of causal relationships. When an association is observed between two variables, it is tempting to draw conclusions of causal inference and make claims on the direction of the causal inference. The next section shows why statistics alone cannot be used to establish causal relationships. This is a common source of misunderstanding among people who are not familiar with statistical methods.

Establishing causal relationships

It is unfortunate that in regression analysis in statistics, the variables are termed explanatory (X) and dependent (Y) variables, in the regression equation $Y = a + bX$. Such nomenclature suggests a causal relationship, i.e., X has an impact on Y. But in fact if we reverse the equation and fit the model $X = a + bY$, we obtain exactly the same statistical significance result, as illustrated in the example below. Table 1 shows the data of GDP and Reading mean scores for 34 countries (data source: OECD (2010a), Table I.2.20, Annex B1, p219).

Table 1. PISA 2009 country Reading mean score and GDP

Country	PISA 2009 Reading	GDP (x1000USD)	Country	PISA 2009 Reading	GDP (x1000USD)
Australia	515	37	Japan	520	33
Austria	470	36	Korea	539	26
Belgium	506	34	Luxembourg	472	82
Canada	524	36	Mexico	425	14
Chile	449	14	Netherlands	508	39
Czech Republic	478	23	New Zealand	521	27
Denmark	495	36	Norway	503	53
Estonia	501	20	Poland	500	16
Finland	536	35	Portugal	489	22
France	496	32	Slovak Republic	477	20
Germany	497	34	Slovenia	483	26
Greece	483	27	Spain	481	31
Hungary	494	18	Sweden	497	36
Iceland	500	36	Switzerland	501	41
Ireland	496	44	Turkey	464	13
Israel	474	26	United Kingdom	494	34
Italy	486	31	United States	500	46

The results for a regression analysis with Reading as the dependent variable and GDP as the explanatory variable are given in Table 2

Table 2. Regression Reading = a + b GDP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	479.828	10.310		46.542	.000
	GDP	.427	.301	*.243*	*1.416*	*.167*

a. Dependent Variable: Reading

For those who are not thoroughly familiar with regression analysis, the results (highlighted in *italics* in Table 2) of the analysis show that the (standardized) coefficient, Beta, in the regression equation is 0.243 and is not statistically significantly different from zero (Sig. is 0.167, greater than 0.05 for the 95% confidence level). So OECD’s claim that GDP explains 6% of the variance of Reading mean scores should be revised to *there is no strong evidence that GDP is correlated with Reading mean scores*. Note the 6% comes from $(0.243)^2=0.06$, but 0.243 has been found to be not statistically different from zero.

When the dependent and explanatory variables are swapped, the results of a regression with GDP as the dependent variable and Reading as the explanatory variable are shown in Table 3.

Table 3. Regression GDP = a + b Reading

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-36.464	48.202		-.756	.455
	GDP	.138	.098	*.243*	*1.416*	*.167*

a. Dependent Variable: GDP

Note that the significance level of the regression coefficient in Table 3 is exactly the same as that in Table 2.

In fact, the computation of the correlation coefficient produces exactly the same statistical significance results, as shown in Table 4.

Table 4. Correlation between Reading and GDP

Correlations

		GDP
Reading	Pearson Correlation	*.243*
	Sig. (2-tailed)	*.167*
	Sample size	34

Table 2, Table 3 and Table 4 show that regression analysis tells us only about correlations, despite the fact that we *hypothesise* a model with a causal relationship. When a statistically significant result is obtained from a regression, all we can say is that there is an association (correlation) between two variables. The regression analysis does not provide evidence of any causal relationship. Similarly, for more sophisticated regression analyses, such as path analysis and structural equation modelling, the causal relationship model is *hypothesised*, but the results are really estimates of correlations. The direction of a causal relationship is an interpretation made by the researcher, not suggested nor verified by the statistics.

It is difficult to estimate the number of researchers who misunderstand results of regression analysis. But judging from the number of papers and reports where the authors make assertions of causal relationships based on regression analysis, it is worthwhile highlighting these statistical facts.

There are further complications in establishing causal relationships between variables. A classic (and somewhat absurd) example in statistics for demonstrating the invalidity of drawing conclusions about causal relationships is that ice cream sales have been found to be positively correlated with the crime rate. It has been found that there is more crime in summer, and, of course, there is more ice cream sold in summer. In fact there is no real association between crime rate and ice cream sales, although there is a statistical correlation. In this case, “the time of the year” is called a mediating variable in the sense that crime rate and ice cream sales are mediated by “the time of the year”.

So how does one identify causal relationships if statistics alone cannot? Let’s consider the case of cigarette smoking and lung cancer. If a statistical correlation is observed between smoking and the incidence of lung cancer, three propositions can be made. The first is that lung cancer causes smoking; the second is that smoking causes lung cancer; the third is that there is no real relationship between smoking and lung cancer as there are mediating variables at play. The first proposition is clearly illogical. The second and the third are likely. Since there are other evidences to demonstrate that smoking damages the lungs, the second proposition is entirely a reasonable one, so we conclude that (it is most likely) smoking causes lung cancer. In other words, the conclusion is not based on statistics alone. Additional evidences and a great deal of reasoning are required in making causal inferences. So, it is essential to have a theoretical model that underpins the interpretation of statistical results. Not only that the hypothesised models need to be plausible, there need to be well-grounded theories to explain data. Currently, a problem with the use of evidence is that people often believe “numbers” have more status as evidence and they draw conclusions based on statistics alone, even when the conclusions are contrary to their good judgment. The proposal to increase class size is such an example.

Some reported causal relationships in Education

The OECD PISA study produces a set of student outcome measures in Reading, Mathematics and Science. In addition, PISA provides measures of student and school characteristics in each country. Linking the two sets of measures is not an easy task and it is always a conjecture to draw conclusions about which educational characteristics influence student education outcomes. In fact, OECD states the following:

> "While PISA cannot identify cause-and-effect relationships between inputs, processes and educational outcomes, it can highlight key features in which education systems are similar and different, sharing those findings with educators, policy makers and the general public." *(Vol. V, p18, OECD, 2010c)*

If PISA data are to be used for shaping policy, there need to be a great deal of reasoning and other evidences to form propositions for good policies. The problem is that PISA data also have limitations. In particular, not all salient factors to the success of education outcomes are included. For example, the prevalence of coaching schools and parental pressures in the top performing Asian countries has not been captured as part of the education environment. Conclusions are drawn based on what are provided by the PISA data set. It is quite possible that the key reasons for success in achieving, say, high mathematics test scores, have a great deal more to do with the sheer number of study hours outside schools, than with teaching methods and school management. Students' lives outside the school are not typically captured in many educational surveys.

Bearing in mind the above cautions about establishing causal relationships, we turn our attention to the suggested impact of "prioritising teachers' salaries over smaller class sizes" by OECD:

> "…many successful school systems share some common features: … spending in education that prioritises teachers' salaries over smaller classes." *(Vol. IV, p29, OECD, 2010b)*

There is actually no statistical evidence of the above statement from the OECD PISA data. Table 5 shows a summary of class size and teacher salary data from PISA (Vol. IV, p85, OECD, 2010b, Figure IV.3.7).

Table 5. Number of Countries by Class Size and Teacher Salary

	Small class size and/or *low* teachers' salaries	*Large* class size and *high* teachers' salaries	Number of countries performed higher than OECD average in reading / Total number of countries
Low cumulative expenditure on education	3 out of 31 countries performed higher than OECD average in reading.	3 out of 12 countries performed higher than OECD average in reading	6/43
High cumulative expenditure on education	8 out of 20 countries performed higher than OECD average in reading	2 out of 2 countries performed higher than OECD average in reading	10/22
Number of countries performed higher than OECD average in reading / Total number of countries	11/51	5/14	16/65

To test whether countries with large class size and high teachers' salaries tend to perform better in reading, a binomial test can be carried out. That is, if, overall, there are 16/65 (about 1 in 4) countries with higher than OECD average performance, is it unusual to have 5/14 (about 1 in 3) countries with higher than OECD average performance in the "large class/high salary" category? The significance level from this test has a *p* value of 0.11, which is not statistically significant at the 95% confidence level.

(As an aside, in contrast, a statistical test for whether countries with high cumulative expenditure in education perform better in reading has a *p* value of 0.01, which is highly statistically significant. That is, about half the countries with high expenditure in education performed above the OECD average, while only a quarter performed higher than the OECD average for all countries in the OECD study.)

In short, based on this data set, there is no strong statistical evidence that "prioritising teachers' salaries over smaller class sizes" leads to a better educational outcome. Further, it is unclear why the OECD report directly links teacher salaries and class size and regards this as a trade-off choice made consciously by the countries. The countries with large class size and high teachers' salaries are Brazil, Chile, Colombia, *Hong Kong*, Indonesia, *Japan*, Jordan, *Korea*, *Macao*, Mexico, *Shanghai*, *Singapore*, *Taipei* and Thailand. The high performing east Asian countries are all in this group (italicised country and city names). These countries generally have high population density and share similar cultural backgrounds. The teaching styles in these countries are in general compatible with large class sizes. Further, the demands on resources (such as land, buildings and infrastructure) of highly populated countries may necessitate large schools and large classes, and it is not a matter of choice to have large class size for the single purpose of increasing teachers' salaries.

Thus we need to question the claim of the impact of "prioritising teachers' salaries over smaller class sizes", and further question whether such a policy is compatible with the valued pedagogy of Western countries in particular, noting that no Western country has yet "chosen" to take on this priority.

Some comments about trading priorities

The difficulty about prioritising policy measures is that the trade-offs are not quite as easily evaluated. For example, the effect of having larger classes is not only a reduction in the number of teachers, but there will also be an increased workload on current teachers. Teachers will need to spend more time marking assignments, for example. If individualised learning is implemented, there must be more work for the teacher if there are more students in a class. Even if there is funding (saved from hiring fewer teachers) for teacher professional development to raise teaching quality, teachers will have less time to participate in such professional development owing to the added workload. Further, many quality teaching practices are not compatible with large class size. Thus the idea that priorities can be easily traded is flawed, since many policy measures are interrelated. One danger of the misuse of statistical information is the belief that net effect size of factors contributing to educational achievement can be computed using simple arithmetic. For example, if class size has a smaller effect size than teaching quality, then it is assumed that we can prioritise teaching quality over class size, and have a net positive effect. This is not to mention that effect sizes of various factors depend greatly on the contexts in which they are measured, and there are wide variations from different studies, so that in fact effect sizes are not reliably measured in the first place.

Conclusion

This paper has demonstrated that data often do not have the reliability and validity the public generally perceive, particularly when inferences are made from the data about causal relationships. In fact, statistics alone should not be used to draw conclusions about causal relationships. We need to bring a great deal of well-grounded theories to bear in using data. Simplistic interpretations of data can only lead to ineffective or even damaging policy measures.

Politicians, like the public, are end-users of research findings. One can hardly put blame on their misconstrued interpretations of data if the authors of research reports are careless in drawing conclusions. While there may be cautionary words about the reliability and validity of data, sometimes authors are not always consistent with heeding their own cautions in making claims of research findings. Consumers of research findings are not all familiar with statistical procedures. They will have little chance of unravelling the complexities of the data so they can be easily misled by such research reports. Like "Chinese whispers", when a major report has been used and quoted by several researchers in succession, the cautions and caveats stated in the original report are often lost, and the claims of research findings become unequivocal.

Consequently, the ultimate responsibilities must reside with the researchers to ensure that only valid conclusions are drawn. Policy-makers and the public need to understand that not all data present evidence.

References

Bredt, J., & Syez, C. (2007). *Education and Economic Growth: A Literature Review.* Labour Market Research Unit, Department of Education, Training and the Arts, Queensland Government. Working Paper No. 50. Retrieved September 29, 2012, from http://training.qld.gov.au/resources/employers/pdf/wp50-education-economic-growth.pdf

Hanushek, E., & Woessmann, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes and causation.* NBER Working Paper No. 14633. Retrieved May 15, 2012, from http://www.nber.org/papers/w14633

New Zealand Treasury (2012(a)). *Treasury's Advice on Lifting Student Achievement in New Zealand: Evidence Brief.* Retrieved May 15, 2012, from http://www.treasury.govt.nz/publications/media-speeches/speeches/economicleadership/sanz-evidence-mar12.pdf.

New Zealand Treasury (2012(b)). *Economic Leadership: When Business Isn't Usual. Speech delivered by Gabriel Makhlouf, Secretary to the Treasury.* Retrieved May 15, 2012, from http://www.treasury.govt.nz/publications/media-speeches/speeches/economicleadership/sp-econlead-20mar12.pdf

OECD (2010a). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)* Retrieved May 15, 2012, from http://dx.doi.org/10.1787/9789264091450-en

OECD (2010b). *PISA 2009 Results: What Makes a School Successful? – Resources, Policies and Practices (Volume IV).* Retrieved May 15, 2012, from http://dx.doi.org/10.1787/9789264091559-en

OECD (2010c). PISA 2009 Results: Learning Trends: Changes in Student Performance Since 2000 (Volume V). Retrieved May 15, 2012, from http://dx.doi.org/10.1787/9789264091580-en

Paine, S.L., & Schleicher, A. (2011). *What the U.S. can learn from the world's most successful education reform efforts.* Policy Paper. McGraw-Hill Research Foundation. Retrieved November, 20, 2013, from http://www.mcgraw-hillresearchfoundation.org/wp-content/uploads/pisa-intl-competitiveness.pdf

Thrupp, M. (2010). The politics of being an educational researcher: minimising the harm done by research. *Waikato Journal of Education 15(2)*, 119-133.