NEPC

NATIONAL EDUCATION POLICY CENTER

# Review of *Cross-Country Evidence on Teacher Performance Pay*


*Reviewed By*

Matthias von Davier

ETS


March 2011

## Summary of Review


The primary claim of this Harvard Program on Education Policy and Governance report and the abridged Education Next version is that nations "that pay teachers on their performance score higher on PISA tests." After statistically controlling for several variables, the author concludes that nations with some form of merit pay system have, on average, higher reading and math scores on this international test of 15-year-old students. Although the author lists numerous caveats, his broad conclusions do not heed these cautions. The fundamental differences among countries in the types of performance pay system are not properly considered. Nations are simply lumped together as having or not having a performance pay plan. Also, the length of time the program had been in place in each country is not addressed and the unknown intensity of program implementations argue against drawing lessons from this study. The small sample size of 28 observations requires extreme caution in interpretation. For example, the inclusion or exclusion of a single country results in large shifts in the size of the reported relationships. That is, the numbers become unreliable and invalid. With any correlational study, attributing causality is problematic; the differences among nations could be due to any number of factors. Finally, the type of regression-based analyses used to support the performance pay conclusion does not properly consider that the background variables used in these analyses can vary in terms of relationships with student scores and have different definitions across the countries under study. Therefore, drawing policy conclusions about teacher performance pay on the basis of this analysis is not warranted.

**Kevin Welner**
*Editor*

**William Mathis**
*Managing Director*

**Erik Gunn**
*Managing Editor*

**National Education Policy Center**

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: 303-735-5290
Fax: 303-492-7090

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar


This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

# Review of *Cross-Country Evidence on Teacher Performance Pay*

*Matthias von Davier,[*] ETS*

## I. Introduction

The promise of large-scale international comparisons of student outcomes is to provide valuable information for policy makers. Educators and stakeholders in policy and government look toward assessments such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) as a source of information on how their country scores in comparison with other participating countries. Correlates of these achievement scores are frequently examined for policy insights.

The primary claim of the Harvard Program on Education Policy and Governance report, *Cross-Country Evidence on Teacher Performance Pay*, also published in an abridged *Education Next* version,[1] is that nations "that pay teachers on their performance score higher on PISA tests."

The report reviewed here focuses on the relationship between teacher performance pay and student achievement. It conjectures that nations employing some form of teacher performance pay will generate a long-term positive impact on the composition and motivation of the teacher pool and thereby improve student performance on standardized tests.

To reach these conclusions, a variety of multiple linear regressions of student proficiency on student background variables, school level variables, and country level information are employed. The data for the report comes from the PISA 2003 assessment cycle using data from 28 OECD countries.

## II. Findings and Conclusions of the Report

The report demonstrates that there is a positive statistical relationship between some form of a country-level teacher performance pay variable and average country achievement after controlling for a number of country-level, school-level, and student-level variables. However, without controlling for these variables (that is, looking at the simple correlation), the

* The opinions expressed in this commentary are those of the author and do not necessarily reflect the views of the Educational Testing Service. The author would like to thank Robert J. Mislev and Shelby J. Haberman for suggestions and comments on previous versions of this document.

relationship between the presence or absence of a teacher performance pay system and country-level performance is negative.

To test the robustness of the findings, other predictors (and countries) are taken out and put back into the estimated equations to see if the results still hold. The report concludes that this conditional relationship between a country-level performance pay indicator and average country performance is robust and that the threat of omitted variable bias is thereby proven to be negligible.

## III. The Report's Rationale for Its Findings and Conclusions

The report is based on a commonly used procedure—multiple linear regression—that relates student outcomes to teacher performance pay while statistically controlling for a host of other variables that may affect the achievement scores in the various countries. The logic is that teacher performance pay has a significant influence on student outcomes if it results in a weighting that is statistically significantly different from zero after the other factors have been considered. For this rationale to be viable, all the other major relevant factors must be properly accounted for across the different nations.

## IV. The Report's Use of Research Literature

The report introduces relevant literature in support of the expected association between teacher performance pay and student outcomes. The report also provides references (in part in the footnotes of the full paper) that are relevant for the conjectured long-term effects of performance pay on the composition of the teacher pool. Missing from the report are references to literature that outline the specific analytic procedures such as those developed to account for sampling variation and measurement error in the use of large-scale educational assessment data. More specifically, analytic procedures described in technical reports of PISA, TIMSS and national assessments such as NAEP, or references to publications on analytic procedures such as the one by Rukowski and colleagues,[2] are missing. Specific effects of these omissions are addressed in the following sections.

## V. Review of the Report's Methods

How student proficiency scores are determined and how the control variables are defined and measured determines the accuracy of the conclusions. Significant data comparability issues must be examined. In addition, with a small number of cases, the reliability of the findings may be compromised and selection effects may be magnified.

### The Accurate Use of Large Scale Proficiency Data

Beginning with the test scores themselves, the databases for the major international assessments (PISA, TIMSS, and PIRLS) do not contain simple test scores of students, but are instead imputation-based[3] quantities referred to as *plausible values*.[4] These plausible values are

calculated based on a measurement model utilizing item response theory (IRT)[5] and a multiple regression model[6] using individual student and parent information as well as other school- and system-level background data collected in questionnaires.

This procedure ensures that statistical analyses of relationships between background data and student proficiency data yield unbiased results.[7] If plausible values were not used, but rather just one out of the set of values was used, or if the mean of plausible values was used instead of the proper procedures,[8] then the analysis would no longer lead to accurate results. Unfortunately, the report fails to provide clarity on this essential issue.

## Imprecise and Viscous Measurement of "Teacher Performance Pay" and Control Variables

While some background variables are well defined in the sense that they mean exactly the same in each country, other variables are subject to substantial local varieties of definition. Variables such as gender are well-defined, but even variables such as the grade level are not unambiguously defined. For example, some countries have almost universal preschool or kindergarten attendance, with at least some academic content, while other countries may have much more variability in attendance and in the content of preschool programs.

There are, however, other cases where the same umbrella definition covers a whole range of

*Causal interpretations from this analysis should be avoided.*

meanings that, on the surface, appear the same but on closer inspection are substantially different. The certification of teacher qualification, for example, takes quite different forms across countries. The lack of teachers during some periods of time in past decades has led some countries to adapt a policy that allows applicants from relevant academic fields to enter the teaching profession. Some countries consider a teacher certified based on passing a written or oral exam. Other countries may require some level of residency or teaching apprenticeship before a teacher gets *fully certified*. Yet other countries may consider someone a certified teacher only if he or she graduated from a pedagogical, as well as a content-oriented, institution. Thus, a person with a master's degree or even a Ph.D. in economics who was hired by a vocational school to teach may not be considered a fully certified teacher in some countries, since he or she had no degree from a teachers college but rather a graduate degree in economics. Is this person more or less qualified than a person in another country who used to be a substitute teacher without a college degree, but was hired due to a lack of certified teachers and now teaches classes independently?

The same issue is indeed discussed in the report. The report's Table 1 shows three different ways performance pay is determined, either by the school principal or by a local or a national authority. This is, however, only one part of the story. There are many ways in which these three granting parties could determine performance pay: Is it based on the last academic year only, or on multiple years? Is it based on observational protocols in one or multiple classrooms, is it

based on student test scores at the end of the year, or on pre-/post-test differences, and so on. The report addresses these complexities by subsuming them under measurement error. Measurement error describes the level of uncertainty or imprecision with which a variable is measured. This approach is problematic because these differences in how performance pay is determined should be considered as different methods of assessing teacher quality. If there are single-year and multiple-year criteria, and if there are status scores versus gain scores used in

*It is not surprising that the inclusion or deletion of a single case has dramatic effects on the findings.*

different countries, the performance pay determination is different both in methods and in criteria applied. These differences lead to differences in reliability and validity among the various approaches to performance pay. These differences exist in addition to differences in measurement error. Different reliabilities and validities in the assessment of teacher quality are critical, since every educational system applying teacher performance pay must do so with the highest possible degree of accuracy.

It is important to note that there will most likely be variability across countries in how performance pay is determined, and that this variability in definitions makes the interpretation of the results (and attempts to use them) somewhat problematic. If a country considers adopting performance pay, what type of criteria for performance pay would be the most advantageous for the given system? And more importantly, would performance pay have the intended effect in the educational system under consideration, given the certainty that any given country-specific system differs from the systems in other countries on a multitude of factors?

## Interpreting Country-Level Correlations

The main finding presented in the report is based on two country-level variables with a sample size of only 28 countries, and the main result is based on the statistical relationship of a binary (performance pay – yes/no) with an average student performance variable (PISA 2003 country mean). This is problematic on several fronts.

### *Alternate Causal Interpretations*

For example, some may argue that the situation is essentially one of spatial correlation and that this would lead to serious concerns about any attempts to even discuss causal relationships: The geographical distance between countries, or clusters of countries, may already indicate that countries closer to each other are more similar on a number of variables. Countries that cluster geographically may be more similar not only in terms of performance pay and level of student performance, but also with respect to a host of other variables not available in the data, so that the conjectured effect of performance pay in terms of long-term teacher pool sorting and motivation is just one out of a multitude of other possible explanations based on other

commonalities between countries. The use or the absence of teacher performance pay in certain groups of countries may be a result of other shared features of the educational systems involved.

### *The Effect of Small Sample Size*

To understand the large impact small differences in the sample data of 28 observations have at this level of aggregation, a few correlations are given here based on the data available in Table 1 in the report:

1. The correlation between teacher salary and average student performance per country is 0.43. By removing a single data-point (Luxembourg—the country with highest teacher salary among the 28, but only average performance), this correlation increases to 0.54. The same is true for country GDP per capita.

2. The correlation between average student performance by country and "any teacher performance pay" is negative, -0.27. By changing a single data point (Mexico has performance pay—let us assume for a moment it does not have it), the correlation changes to being close to zero, -0.02.

3. In contrast, the correlation of one of the specific teacher pay variables, "performance pay decided by school principal," and average student performance per country is positive: 0.26. This is, in absolute terms, almost as high as the negative correlation reported above. If we assume that one country (Turkey) had "school principal decides performance pay" instead of "national authority decides about performance pay," the correlation would drop to almost zero: 0.08.

### *Instability Due to Selection Effects*

The unconditional relationship of variables such as teacher salary, GDP per capita (both correlate 0.43 with average country performance), and performance pay most likely depends on the particular selection of countries we examine. Such correlations change dramatically if one or two countries are taken out or are replaced. PISA not only involves the 28 OECD countries used in the report, but was also administered in 11 PISA-plus countries in 2003 and 27 PISA-plus countries in 2006, which are not OECD members. Even though it is understood that some of the country-level data might be harder to come by, it would be helpful to involve these countries as well.

### *Inferring a Long Term effect From a Single Data Point*

A system could decide within a short window of time to adopt or remove teacher performance pay. Systems could reform how performance pay is determined within short periods of time. The variable "any performance pay" as well as the more specific ones depend on what is a snapshot in time: Does that particular country have performance pay at that point (2003) in time when the data were collected? In contrast, the suggested interpretation is long term; the report argues that performance pay may sort the pool of teachers and may have long-term motivation effects on current teachers. This indicates that other variables, such as the length of time a country has teacher performance pay in place, would have been a more valid way to test these conjectured long-term effects on the teacher pool.

If performance pay has long-term effects, the type of performance pay as well as the performance pay indicator should be tracked over time, along with the proportion of the total salary affected and other relevant system-level variables affecting teacher sorting and motivation. This shortcoming in the available data and the continual concern of omitted variables in these types of analysis prevents conclusive findings.[9] As a consequence, causal interpretations from this analysis should be avoided.

## Are the Variables and Their Relationships Consistent Across Countries?

An example of why we need to carefully consider *what* we measure and *why* is connected to the attempts of PISA and other assessments to collect data on what one may call a proxy of socioeconomic status. For obvious reasons, students cannot be asked to provide information about their parents' income: Either they do not know it, or it is illegal to ask, or they may (substantially) over- or underestimate what their parents earn. Therefore, proxies or what one may call *indicators* of wealth or social status are collected instead of direct measures of parents' earnings. One of these measures, and one that is pretty indicative in the United States, is the number of full bathrooms in the parents' home(s). Note however, that this is a variable that may say little to nothing about parents' socioeconomic status in other societies. Even within a country, this variable may likely not be distributed in the same way in suburban, rural, and urban areas.

The number of books in a student's home is a similar case, and is one that is actually included in the regression presented in the report. Obviously, in some countries, for example, those with excellent public library systems, the number of books at home may be less predictive than in

*Drawing conclusions about performance pay from this analysis cannot be validly sustained.*

other countries. Most certainly, there is no reason to believe that the average number of books should be the same across countries. For example: Are graphic novels in book form commonly found in large numbers in some countries equivalent to the same number of nonfiction or scientific books in other societies? Are books relatively cheap in some countries, while relatively expensive in other countries? Will the advent of e-book readers completely change this measure in some countries but not in others?

To control for these types of differences, statistical methods that address the issue of hierarchically organized systems (in this case, students sampled within schools, and schools [potentially within states or regions] sampled from within countries) are available. Commonly referred to as multilevel[10] or hierarchical (linear) models,[11] these approaches allow the relaxation of the criterion that the same measures of student proficiency on background data hold in all countries. Using such a hierarchical approach would show whether proxies—such as the number of books at home—for the student background variables we want to get at are equally good predictors of student proficiency across participating countries. If by using these models, we find differences in the variable weights from one country to another, this could reflect not only

differences in the definition or meanings of predictors from one country to the next, but would imply different relationships among these variables and student performance. If such methods had been employed and if differences were found, this would cast serious doubt on the interpretation of the analyses employed in the paper.

## VI. Review of the Validity of the Findings and Conclusions

*Cross-Country Evidence on Teacher Performance Pay* addresses a significant issue in contemporary policy debates: how performance-based pay of teachers may be related to student outcomes. It is important to keep in mind that the finding is based on a sample of only 28 countries, and that countries that apply performance pay may be different in a variety of other ways from countries that do not apply performance-based pay differentiation. Thus, it is not surprising that the inclusion or deletion of a single case has dramatic effects on the findings. The various definitions and applications of performance pay vary widely from country to country Thus, although the same term (performance pay) may be applied to these diverse definitions and practices, it is unlikely that the same thing is being measured across countries. Further, the differences in the definitions of the control variables across the countries is unresolved. Thus, drawing conclusions about performance pay from this analysis cannot be validly sustained.

## VII. Usefulness of the Report for Guidance of Policy and Practice

It is important to note that, besides performance pay, there are a number of other relevant variables with a significant relationship to the average performance of a country that are not considered in this analysis. It may be, for example, that high-scoring countries experiment with performance pay rather than high scores being caused by merit pay. The definitions and the range of performance pay as well as the process of how an increase or decrease in pay is determined will most certainly vary by country. In addition, if other forms of payment adjustment exist alongside performance pay, this will moderate the overall effect of performance related pay increases on teacher salary. Such factors argue against generalizing from this report.

More analyses and more in-depth studies of differences among educational systems are needed before a decision to adopt performance pay should even be discussed. As stated above, both the size of the regression weights and the error terms can be distorted when hierarchal structures are not taken into account. Unless the key results of this study are borne out in such models and replicated using data from other assessments as well as in relationships of performance pay and achievement data at the country and school level, they will not constitute an adequate basis for policy recommendations.

It would be helpful if large-scale assessment databases were designed so that cross-country variables in the background data at the student, parent, school and systems level considered comparable were highlighted. Even in the case where background variables may be considered comparable across countries, however, the differences among groups defined by this variable may be larger in one country and smaller in another country. Therefore, analyses of cross-country data should take the possibility of variations in relationships of background variables and student outcomes into account.

The study presented in the Harvard Program on Education and Governance report, and abridged in *Education Next*, is a step in the right direction. However, its findings are based on a series of regression analyses using cross-sectional data. As a consequence, the study does not yield results powerful enough to support the adoption of a policy of performance pay. Its use of a flat rather than hierarchal model of analysis leaves too many questions unanswered for it to be used as the basis for policymaking.

## Notes and References


1 Woessmann, L. (2010). *Cross-Country Evidence on Teacher Performance Pay*. Cambridge, MA: Harvard Program on Education Policy and Governance. Retrieved March 29, 2011, from http://www.hks.harvard.edu/pepg/MeritPayPapers/Woessmann_10-11.pdf

An abridged version can be found in the journal *Education Next*:
Woessmann, L. (2011, Spring). Merit pay international. *Education Next 11* (2). Retrieved March 29, 2011, from http://educationnext.org/merit-pay-international/

2 Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010, March). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39* (2), 142-151.

3 Little, R. J. A., & Rubin, D. B. (1987*). Statistical analysis with missing data*. New York, NY: J. Wiley & Sons.

4 von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 2,* 9-36.

5 Lord, F. M., & Novick, M. R. (1968*). Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

6 Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

von Davier, M. Sinharay, S., Oranje, A., & Beaton, A. (2006) Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26. Psychometrics.* Amsterdam, the Netherlands: Elsevier.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57–76). New York, NY: Springer Verlag.

7 Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 2,* 9-36.

8 Little, R. J. A., & Rubin, D. B. (1987*). Statistical Analysis with Missing Data*. New York, NY: J. Wiley & Sons.

9 Braun, H., Jenkins, F., Grigg, W., & Tirre, W. (2006). *A Closer Look at Charter Schools Using Hierarchical Linear Modeling*. Washington, DC: U.S. Department of Education.

10 Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed.). London, England: Oxford University Press.

11 Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

| | |
|---|---|
| DOCUMENT REVIEWED: | **Cross-Country Evidence on Teacher Performance Pay** (Also published in an abridged form titled **Merit Pay International**) |
| AUTHOR: | Ludger Woessmann |
| PUBLISHER/THINK TANK: | Harvard Program on Education Policy and Governance; *Education Next* |
| DOCUMENT RELEASE DATES: | PEPG: June 3-4, 2010;<br>*Education Next:* Spring 2011 |
| REVIEW DATE: | March 31, 2011 |
| REVIEWER: | Matthias von Davier, ETS |
| E-MAIL ADDRESS: | mvondavier@ets.org |
| PHONE NUMBER: | (609) 734-1717 |

SUGGESTED CITATION:

von Davier, M. (2011). *Review of "Cross-Country Evidence on Teacher Performance Pay."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-pisa-performance-pay.