Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

Gándara, Fernanda; Randall, Jennifer

Investigating the Relationship between School Level Accountability Practices and Science Achievement

Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 23, 2015, pp. 1-22
Arizona State University
Arizona, Estados Unidos

education policy analysis archives
A peer-reviewed, independent, open access, multilingual journal

Arizona State University

Volume 23 Number 112 November 16th, 2015 ISSN 1068-2341

# Investigating the Relationship between School Level Accountability Practices and Science Achievement

*Fernanda Gándara*
*&*
*Jennifer Randall*
University of Massachusetts Amherst
United States

**Citation**: Gándara, F., & Randall, J. (2015). Investigating the relationship between school level accountability practices and science achievement. *Education Policy Analysis Archives, 23*(112). http://dx.doi.org/10.14507/epaa.v23.2013.

**Abstract:** This study investigates the relationship between school-level accountability practices and science achievement of 15-year-olds, across four counties: Australia, Korea, Portugal, and the United States. We used PISA 2006 data, since 2006 is the only administration that has focused on science. School-level accountability practices are here defined as activities that: (a) provide school achievement data to external stakeholders, or (b) establish consequences according to the achievement results. Using linear regression analysis, we found that school-level accountability practices varied across these four countries, albeit not all pairs of countries were significantly different from each other in this regard. Using hierarchical linear modelling, we found that school-level accountability practices had a small effect on science achievement. Importantly, this effect was not independent of schools' and students' socio-economic status.
**Keywords**: accountability; science achievement; international testing; PISA

**Investigando la relación entre medidas de rendición de cuentas por parte de las escuelas y el desempeño en ciencias**

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 16/2/2015
Revisions received: 25/6/2015
Accepted: 7/17/2015

**Resumen:** Este estudio busca comprender la relación entre prácticas de rendición de cuentas por parte de las escuelas, y el desempeño en ciencias de estudiantes de 15 años, en cuatro países: Australia, Corea del Sur, Portugal, y Estados Unidos. Este estudio usa los datos de PISA 2006 por ser la última administración enfocada en ciencias. Las prácticas de rendición de cuentas de las escuelas se refieren aquí a aquellas prácticas que: (a) proveen datos de desempeño académico a terceros, o (b) establecen consecuencias en función de dichos resultados. Usando regresión lineal, encontramos diferencias significativas en las prácticas de rendición de cuentas adoptadas por las escuelas de distintos países. Sin embargo, no todos los pares de países son diferentes entre sí bajo este criterio. Usando análisis multinivel encontramos que el efecto de estas prácticas en el desempeño en ciencias es pequeño. En particular, este efecto no es independiente del nivel socio-económico de las escuelas y alumnos.
**Palabras-clave:** rendición de cuentas; desempeño en ciencias; pruebas internacionales

**Estudo da relação entre prestação de contas das escolas e desempenho em ciência**
**Resumo:** Este estudo investiga a relação entre práticas de prestação de contas no nível escolar e o desempenho em ciências entre jovens de 15 anos de quarto países: Austrália, Coréia, Portugal, e Estados Unidos. Nós usamos dados do PISA de 2006 já que este é o único ano em que a administração teve foco em ciência. Práticas de prestação de contas são aqui definidas como atividades que: (a) fornecem dados de desempenho escolar para as partes interessadas, ou (b) estabelecem consequências relacionadas aos dados de desempenho. Usando análise de regressão linear, descobrimos que práticas de prestação de contas no nível escolar variaram entre os quatro países, mas nem todos os pares de países foram significativamente diferentes neste aspecto. Usando modelagen linear hierárquica, encontramos que estas práticas tiveram um efeito pequeno no desempenho em ciência. Importantemente, este efeito não foi independente do nível sócio-econômico das escolas e dos alunos.
**Palavras-chave**: prestação de contas; desempenho em ciencias; testes internacionais

## Investigating the Relationship between School Level Accountability Practices and Science Achievement

Educational policies across several nations have increased their focus on the learning outcomes achieved by students, to ensure that students are not only attending schools but acquiring new and relevant knowledge (Centre of Study for Policies and Practices in Education (CEPPE), 2013; Hanushek & Raymond, 2006; Women Thrive Worldwide, 2015). This current focus on students' outcomes requires greater scrutiny on the development of reliable measures and valid uses and interpretations of scores. The change in educational policies has, therefore, been accompanied by an expansion and improvement of test practices against a common set of expectations (Hanushek & Raymond, 2006). In order to assure that students meet such expectations, policy-makers have designed systems of rewards and penalties for the agents responsible for these outcomes. These practices – known as *accountability mechanisms* – have been well established in the economic field (principal-agent theory; see Laffont & Martimort, 2002 or Mas-Collel, Whinston, & Green, 1995) and have been increasingly used by educational policy-makers.

### Accountability Systems

*Accountability systems* are the sets of mechanisms and instruments that attach consequences to the accomplishment of well-established objectives, to ensure that several *agents* meet their obligations

(Hatch, 2013; Wobmann, Luderman, Schutz, & West, 2007). Theoretically, accountability in education is mainly supported by the *principal-agent theory*, which states that establishing proper incentives to agents – who perform service – will align their interests with those of the *principals* – who commission the service – and will ensure an efficient delivery (Polikoff, McEachin, & Wrabel, 2014; Wobmann et al., 2007). Under this schema, schools (agents) are conceived as providers of education (service) to students, on behalf of the state, or country, or parents, etc. (principals). Theoretically, the relationship between principals and agents is mostly jeopardized by divergent interests and by decentralized information (Wobmann et al., 2007). If the agent has control over the information about his/her performance and/or if the agent has other interests than those of the principal, there is a risk that the agent pursues his own interests at the expense of the quality of the service provided. Therefore, accountability systems rely on two mechanisms: (a) establishing rewards or penalties to differential attainment of goals, and (b) providing transparent information about the agents' performance. Clearly, accountability in education depends largely on the type and quality of the information provided to educators and "consumers" (Polikoff et al., 2014).

Accountability systems in education typically consist of three components: (a) achievement standards, (b) measures of student performance, and (c) a system of consequences attached to the latter (Wobmann et al., 2007). On one hand, achievement standards set the expectations to which agents will be held accountable. Standards can be thought of as definitions of what someone should know and be able to do to be considered competent (CEPPE, 2013). While there are different types of standards – content, assessment, and performance standards – their purpose is to make clear and explicit the learning expectations for pupils in schools (CEPPE, 2013). On the other hand, the measures of student performance provide information about the extent to which students are achieving their learning expectations. There are multiple ways to measure student achievement. Countries typically use student assessments as the measure of performance used for accountability purposes (Rosenkvist, 2010). Accountability systems that are mainly based on the results of large-scale testing programs are known as *test-based accountability systems*. Last, the system of consequences attached to the measures of student performance, refer to the rewards and penalties assigned to the agents in relation to their accomplishments. Regarding these rewards and penalties, an accountability system may be *high-stakes* – that is, when significant advantages and/or disadvantages are coupled with the test results –- or *low stakes* – when no such couplings exist (Rosenkvist, 2010).

## The Unclear Effectiveness of Accountability on Improving Student Achievement

Countries differ in relation to how they implement accountability systems (Rosenkvist, 2010). Among other variations, accountability systems may: (a) use one or multiple measures of performance (Polikoff et al., 2014; Rosenkvist, 2010); (b) be defined by external authorities – *external* – or be left-up to the agents – *internal* (Hatch, 2013); (c) be high-stakes or low-stakes (Rosenkvist, 2010); (d) have different agents (schools, teachers, etc.) or different principals (centralized authorities, parents, etc.) (Rosenkvist, 2010); (e) fulfill different functions (Haertel, 2013; Klenowski, 2011; Rosenkvist, 2010); and (f) have differential consequences on educational systems (Rosenkvist, 2010). In particular, standards linked to high-stakes external assessments and accountability measures represent one pole, whereas using standards as suggested guidelines and assessing their achievement through low-stakes assessments represents the other pole (CEPPE, 2013).

Several researchers have attempted to understand the effects that different accountability systems have on educational outcomes. Specifically, evidence of the effect of accountability systems on student achievement emerges from various lines of research. On one hand, researchers have looked at the effect that testing per se has on student achievement. While some show that testing

across multiple conditions (e.g. Phelps, 2012) may increase achievement, others claim that there is no conclusive evidence or convincing support to the idea that testing improves student achievement (e.g. Lee, 2008; Nichols, Glass, & Berliner, 2012). On the other hand, researchers have looked at the relationship between the level of pressure of the accountability systems – that is, their stakes – and the effect on student achievement. Again, there is no consensus among researchers, with some providing evidence that high stakes tests are particularly effective in improving achievement (e.g. Phelps, 2012), and others providing evidence that accountability pressure may be correlated with student achievement but does not explain changes in achievement over time (e.g. Nichols et al., 2012). Some make the distinction that high-stakes exit examinations increase student achievement (Bishop, 1997), but that this is only true if clear standards and goals are set (Wobmann et al., 2007).

This relationship is hard to isolate because of the multiple effects, agents, and layers that are operating within accountability systems. High-stakes test-based accountability systems have been largely criticized because they have unintended negative consequences on the curriculum, instructional practices, and other educational processes. Many researchers have pointed out that using high-stakes tests has led to an overconcentration of teaching on the content areas that are tested or to a *narrowing of the curriculum* (CEPPE, 2013; Judson, 2012; Koretz, 2008; Rosenkvist, 2010). Others have stated that making tests so predominant creates strategic behaviors to increase test scores that may be detrimental for instruction – such as the *teaching to the test* (Rosenkvist, 2010). Other detrimental consequences of high-stakes test-based accountability systems include: (a) the creation of diverse forms of practices that discriminate low ability students such as section based on prior achievement (CEPPE, 2013); (b) score inflation (Koretz, 2008); (c) incentives to cheat or an increment on the levels of cheating (Rosenkvist, 2010); and (d) an excessive deployment of resources in maintaining these systems (Rosenkvist, 2010).

Researchers recognize that the major goal of federal and state high-stakes testing policies is to improve schools (Nichols et al., 2012), and teachers - as well as other stakeholders - believe that tests are important in the educational policy agenda and tend to do more good than harm (Rosenkvist, 2010). Moreover, standards based policies are common among high performing educational systems and may carry a number of positive effects in educational systems (CEPPE, 2013; Klenowski, 2011; Schleicher, 2011). Therefore, test-based accountability systems should integrate the best of both worlds, and such policies should achieve their main purpose of improving schooling and students' performance. However, the success of accountability systems depends largely on other characteristics of the system, such as capacity-building at the school level (Hatch, 2013; Rosenkvist, 2010; Schleicher, 2011), and specifically, the adequate balance between accountability measures and those aimed at building capacities in the system (CEPPE, 2013). It is not enough to add pressure to the educational system so as to mobilize it towards improvement, but the conditions to generate this improvement need to be created or supported. The effectiveness of these systems also depends on the quality of the measures used (validity, reliability), the transparency of the incentive mechanisms, the consequences on equity and fairness, and the evidence that the positive intended outcomes surpass the effects of indirect purposes or of unintended consequences (Haertel, 2013; Polikoff et al., 2014).

The effect that an accountability program has on different levels of student performance remains unanswered (Judson, 2012). Mixed and inconclusive findings arise partly due to the fact that researchers look at different types of accountability (e.g. school, student), with differential levels of stakes, with different dependent variables and time periods, among other characteristics of the analyses (Lee, 2008). For example, the effect of accountability on student achievement is different for mathematics than for science (Lee, 2008). In light of the current prominence that accountability systems have in educational systems, more evidence about the nature of the relationship between

accountability and achievement is needed. In this study we look at the relationship between school level accountability practices and science achievement of 15-year-olds, across four countries, based on the results of a low-stakes large-scale test. We examine the relationship that certain school level accountability practices may have with achievement. Drawing on the principal-agent theory, the school level accountability practices considered refer to school practices aimed at increasing transparency and/or establishing consequences in relation to achievement results. We focus on science achievement because previous studies have mostly looked at the relationship that accountability has had on mathematics and/or reading achievement and not on science achievement. Looking at the relationship between school level accountability practices and science achievement is informative because it allows us to expand the scope of what we know about the relationship between accountability and achievement. Moreover, science education is considered critical for the development of nations in the current economic context (Carnegie Corporation of New York and Institute for Advanced Study, 2009). Therefore, focusing on science becomes relevant, as nations are increasingly interested in improving scientific literacy of their citizens, particularly, a scientific literacy that enables connections between science knowledge and societal issues (Hofstein, Eilks, & Bybee, 2011).

## Purpose

The purpose of this study is to investigate the relationship between school level accountability practices and science achievement of 15-year-old students, across four countries: Australia, Korea, Portugal, and United States. Specifically, the research questions that we answer are:

1. Do school level accountability practices vary across these four countries?
2. Is there a relationship between the extent of school level accountability practices and 15-year-olds' science achievement?

## Methodology

### Data

The data for this study come from PISA 2006 database. The Programme for International Student Assessment (PISA) is an international assessment program that measures the skills and knowledge of 15-year-olds across three domains: literacy, mathematics, and science (Organization for the Economic Co-Operation and Development (OECD), 2007). Since 2000, the assessment takes place every three years and is administered across a large number of countries and economies. PISA is described as being "forward-looking" because it focuses on the ability of students to use their knowledge and skills to meet real life challenges (OECD, 2009). PISA is not measuring the curricula of the participating countries and economies, but rather focuses on skills that are not necessarily linked to them.

Complementary questionnaires are administered to students, school principals, and parents, to gather information that is relevant to explain and contextualize students' outcomes. Two of these questionnaires are mandatory: (a) the student questionnaire and (b) the school questionnaire. The first is administered to students in order to gather information about their individual characteristics with special focus on demographic and socioeconomic aspects. The second is administered to school principals in order to gather relevant information about the schools which students attend.

Every PISA administration focuses on a single domain. This focus has two implications. On one hand, more achievement data are collected for that subject. On the other hand, the

complementary questionnaires collect information around variables that are exclusively related to the given subject. For that reason, this study uses 2006 data, as PISA 2006 is the most recent administration to focus on science. Key variables such as the interest in science, are not available in more recent datasets (2009 & 2012).

**Case Selection**

PISA 2006 was administered to 57 countries and economies. For parsimony and interpretability, we selected four countries based on their differences on the following criteria: (a) gross domestic product per capita (GDP per capita), (b) the percentage of GDP annually spent in research and development (GERD), (c) geography (continent), and (d) language. These variables were selected in order to represent different socio-economic realities (GDP per capita, geography), cultural realities (language, geography), and the importance that science has in each (GERD). While we acknowledge that GERD relates to all resources put into research and development regardless of their subject area (i.e.: the research could be in humanities or in physics), we also know that natural sciences and engineering are among the most expensive disciplines and typically take the larger share of such expenditures (OECD, 2014).

The database used for this selection was built by one of the authors using the OECD and World Bank databases. Values for 2006 were used for the GDP per capita and GERD values. Among those countries and economies for which we had full information, we selected the four countries which differed the most among these variables. The GDP per capita and GERD variables were dichotomized (low, high) and the geography and language variables were nominal. If two countries differed on one variable, their distance increased by one.

The selected countries were Australia, Korea, Portugal, and USA. These countries have similarities and differences in terms of these variables. Australia was among the countries with high GDP per capita and high GERD; Korea was among the countries with low GDP per capita and high GERD; Portugal was among the counties with low GDP per capita and low GERD; and USA was among the countries with high GDP per capita and high GERD. Except for Australia and USA, the official languages are different for these countries. Last, all of them are located in different continents.

These countries also present similarities and differences in terms of national level testing practices. By 2006, Australia did not have a national testing program or national standards and only used locally developed tests for evaluation purposes (Klenowski, 2011). However, in 2009 Australia established a National Assessment Program (NAP) which is an annual assessment for students in grades 3, 5, 7 and 9 (Rosenkvist, 2010). It measures four domains: reading, writing, language conventions, and numeracy (Australian Curriculum, Assessment and Reporting Authority (ACARA), 2014b). In addition, from 2003-on it administers sample assessments in three subjects for students in grades 6 and 10, including science literacy (ACARA, 2014b). This assessment is only administered to grade level 6. The results are not used to punish schools or teachers but are publicly available.

Korea has a national level system of diagnostic assessments for primary and secondary education, namely, the National Assessment on Educational Achievement (NAEA). The program was launched more than 60 years ago, but has been fully implemented for only 16 years. In particular, before 2008, the NAEA was administered to a representative sample of students from grade levels 6, 9 and 10[1]. However, after that year, Korea began administering the NAEA to all students enrolled in primary and secondary education. From that year on, assessment results became

[1] Middle school 3rd graders and high school 1st graders.

publicly available and began being used for certain accountability purposes (Korean Educational Development Institute, 2010). Notwithstanding, at the time when students took PISA these policies were not yet in place. The subject matters assessed include Korean, mathematics, English (some grade levels), social sciences, and science.

On the other hand, Portugal has a national standardized test (summative) administered to students in Grade 4 in two curricular areas: Portuguese and mathematics (Santiago, Donaldson, Looney, & Nusche, 2012). The objective of this low-stakes assessment is to inform stakeholders about students' performance. Other national examinations are used at the end of Grades 6, 9 and 11 or 12. These are not dominant for the final mark and also measure achievement in Portuguese and mathematics. Overall, in Portugal there is an emphasis on internal formative assessments led by teachers. In fact, external assessments were implemented partly to support internal based evaluation systems. Also, teachers and schools have considerable autonomy to define the assessment criteria (Santiago et al., 2012). These tests were in place before 2006.

The United States has one national assessment, the National Assessment of Educational Progress (NAEP). For more than 40 years, NAEP collects and reports information about academic achievement on nationally representative samples. The results are widely reported and inform government evaluation of the condition and progress of education (Rosenkvist, 2010). Science is one of the many subjects assessed. However, science is not assessed every year. Between 2006 and 2014 there were only two years in which science was one of the subjects nationally assessed (National Center for Education Statistics, 2015). Despite having one single and low-stakes national level assessment, the United States currently operates a mandatory nation-wide state accountability test-based system. Under the No Child Left Behind (NCLB) Act of 2001, schools and districts are held accountable for student achievement. States must measure student progress in reading, mathematics, and science achievement, to all students. In the cases of mathematics and reading, assessments are administered for grades 3 through 8. In the case of science, assessments have to be administered at least once during grades 3-5, grades 6-9, and grades 10-12. For accountability purposes, several outcomes and performance indicators associated with these tests have to be reported and met. Consequences are attached to the results, making these tests high stakes to schools and districts. States may choose to include science achievement for this accountability function or not. By 2012, 11 states have chosen to use science achievement in the accountability calculations (Judson, 2012).

The states' testing programs and accountability systems are embedded in a political context that permits some flexibility and therefore, variation across them. From 2011, states were allowed to apply for waivers of the key requirements of the NCLB (CEP, 2012). To date, most states have approved or extended waivers in place, and these waivers have brought several changes to state-level accountability systems. For example, states with NCLB waivers have modified some of their goals, such as the 100% proficiency in reading and mathematics for all students. Some have also expanded their range of annual measurable objectives (AMOs) or targets of performance which schools have to meet in order to make adequately yearly progress. Some states have established different AMOs for different students' sub-groups. Or for instance, some states have implemented new and complex performance indexes to evaluate the yearly progress of schools (CEP, 2012). In particular, one of the important changes that these "waiver states" could face is related to the achievement standards to be measured by the large-scale tests used in the accountability process. The current policy is such that these states had/have to adopt "college and career-ready standards" and use assessments aligned to these standards (CEP, 2012). To be college- and career-ready means that "a high school graduate has the knowledge and skills in English and mathematics necessary to qualify for and succeed in entry-level, credit-bearing postsecondary coursework without the need for remediation" (Achieve, 2014).

Some of the waiver states adopted the Common Core State Standards (Council of Chief State School Officers (CCSSO) & the National Governors Association Center for Best Practices (NGA Center, 2014) which are meant to provide college and career readiness high-quality standards in mathematics and English language arts/literacy. It is expected that these states adopt the assessment systems developed by either Smarter Balanced (Smarter Balanced, 2014) or Partnership for Assessment of Readiness for College and Careers (PARCC) (PARCC, 2014), two state consortia that are developing high quality assessment systems aligned to the common core state standards. Beyond the differences across the states in terms of waivers, the adoption of common assessment systems could create "multi-state" accountability system. However, some legislators are currently pushing for a reauthorization of the NCLB act which in case of approval, might change the accountability scenario importantly. In particular, the current draft of the NCLB reauthorization – called the Every Child Achieves Act of 2015 – advocates for limiting the faculties of the Secretary of Education in states' decisions on how to assess and what to assess (Ravitch, 2015; Schneider, 2015). Notably, one of the prohibitions to the Secretary is that of requiring a State to enter into a voluntary partnership as a condition of approval of a waiver, among other conditions (Schneider, 2015). Therefore, it is possible that some waiver states decide to drop out from assessment consortia if the new Act is approved and its essential dispositions are kept. In sum, the accountability and testing practices will continue to undergo important changes, depending on the modifications of the educational policies that define them. To date, the United States educational system has one national testing system and overall, is and remains a test-intensive country.

Table 1.
*Characteristics of the selected countries*

| | | | | | Assessments | |
|---|---|---|---|---|---|---|
| **Country** | **GERD** | **GDP per Capita[a]** | **Continent** | **Language** | **National** | **Science** |
| Australia | High (1.97%) | High (36 M) | Oceania | English | Publicly available without direct punishment | Yes (sample) |
| Korea | High (3.01%) | Low (21 M) | Asia | Korean | Publicly available with increasing uses | Yes |
| Portugal | Low (1.02%) | Low (20 M) | Europe | Portuguese | Low-stakes | No |
| USA | High (2.61%) | High (46 M) | America | English | High-stakes | Yes, but not on regular basis (representative sample) |

Note. Current information. GERD corresponds to the Gross Expenditure in Research and Development. GDP per capita corresponds to the Gross Domestic Product divided by the total population.
[a]Current USD. These values were updated using the current World Bank database and are very close to those used to select the countries.

## Variables

To answer the first research question, we needed a variable that could provide varying levels of accountability practices at schools. Because we were interested in accountability practices

generally, and did not care about specific accountability practices, we built an overall accountability index from questions 15 and 17 of the school questionnaire administered by PISA (see Appendix A). Question 15 explored aspects of accountability to parents, and was composed of three yes/no items. Question 17 explored the ways in which schools used achievement data, and was composed of five yes/no items. Under the assumption that all items contributed equally to the overall accountability level of a school, we added the eight yes/no items and created our index. The coding of questions 15 and 17 was such that for every "yes" answer, a "1" was assigned to the data, and for every "no" answer, a "2" was assigned. Therefore, we reversed the index and created the variable *inverse overall accountability (ACC)*. The inverse overall accountability is an ordinal variable, ranging from eight to sixteen.

In addition, we created three dummy variables for the different countries. The dummy variables were Australia (AUS), Korea (KOR), and Portugal (POR). Other variables that were of interest in the context of the first question, were those variables that could importantly influence the overall accountability level of schools. We only created variables for which we had information for the four countries. Therefore, we could not use the type of funding of school (private vs. public) as a variable in our analysis, because that information was not available for Australia. However, we were able to build variables that referred to the levels of influence and pressure that different stakeholders exerted on different school decisions. In particular, using data from questions 12 and 16 of the school questionnaire (see Appendix B), we built five variables: (a) internal influence on budgeting and staffing (IBS), (b) internal influence on instructional content and assessment practices (IIA), (c) external influence on budgeting and staffing (EBS), (d) external influence on instructional content and assessment practices (EIA), and (e) parental pressure (PP).

In relation to our second research question, we needed a measure for science achievement. We defined science achievement as the average of the five plausible values that PISA reports for each student. PISA uses a sampling method in which students are not exposed to all science questions. Plausible values provide an estimate of student achievement as that student would have answered all of the questions.

Other variables that were relevant to our second research question were those student level variables that have been consistently found significant in previous studies on science achievement, and/or that have been used in similar studies: (a) gender (Gilleece, Cosgrove, & Sofroniou, 2010; OECD, 2007; Tomul & Sebile Savasci, 2012), (b) interest in science (Krapp & Prenzel, 2011), (c) SES (Gilleece et al.,2010; Nichols et al., 2012; OECD, 2007; Tomul & Sebile Savasci, 2012), and (d) immigration status (OECD, 2007). PISA provided information about gender and immigration status. Gender was re-coded as a dummy variable where 0 corresponded to males and 1 to females (therefore, this was a dummy code for females). Immigration status was recoded into a dummy variable where 0 corresponded to native students and 1 corresponded to either first or second generation of students (i.e. "non-native"). Moreover, PISA also provided a single continuous measure for both interest in science and SES, which were left unchanged.

In addition, we were interested in those school-level variables that have been found significant on these set of student level variables and/or that are have been consistently used in similar studies: (a) gender composition of schools (e.g. Sullivan, Joshi, & Leonard, 2010), and (b) school SES (Gilleece et al., 2010). These variables were built from the survey and student data, and were included in our analysis.

**Procedures**

To answer our first research question, we used linear regression. As factors and covariates we included: (a) the three country dummy variables – which is our independent variable (IV) of interest, (b) internal influence on budgeting and staffing (IBS), (c) internal influence on instructional content and assessment practices (IIA), (d) external influence on budgeting and staffing (EBS), (e) external influence on instructional content and assessment practices (EIA), and (f) inverse parental pressure (PP). The final model is shown in Fig. 1.

Where:
= coefficient for Australia
= coefficient for Korea
= coefficient for Portugal
= coefficients for Internal Influence on Budgeting and Staffing
= coefficients for Internal Influence on Instructional Content and Assessment Practices
= coefficients for External Influence on Budgeting and Staffing
= coefficients for External Influence on Instructional Content and Assessment Practices
= coefficients for Parental Pressure

Fig. 1. Regression Analysis

To answer our second question, we used hierarchical linear modelling. This technique has been used in similar analyses (e.g. Gilleece et al., 2010; Koretz, McCaffrey, & Sullivan, 2001; OECD, 2007) because data are nested and applying general linear models might inflate type I error rates by underestimating the standard error of estimates. In our analysis, the dependent variable was science achievement.

PISA provides data that are nested in three levels: students nested in schools, and schools nested in countries. To determine if a two- or three-level model was most appropriate, we conducted an unconditional three-level model to understand the structure of the explained variance. We decided to use a two-level model because it is easier to interpret and it was reasonable given that the variance explained by the country level was only 5%. A 5% of variance explained is low and is almost half of what was found for all the countries according to OECD (2007).

The hierarchical linear model was built using a multistep approach. This approach allows us to consider all the relevant variables while keeping the model at its simplest. First, we fit a fully unconditional two-level model to estimate the variance components at each level. Second, we added the four level-1 predictors previously selected: gender, SES, interest in science, and immigration status. The two dummy variables (gender and immigration status) were un-centered, and the two continuous variables (SES and interest) were group centered. We only retained those predictors that contributed significantly to the explanation of the level-1 variance. Third, we added the accountability variable as a level-2 predictor for the intercept and slopes. We retained the predictor in the intercept and in the slopes where it contributed significantly to the explanation of the school variance. Fourth, we added the remaining predictors for the corresponding intercept and slopes. Not all level-2 predictors were included in all the slopes' equations. The level-2 predictors were selected

based on the conceptual relation existing between the variables. For the intercept we added all level-2 variables. For the gender slope, we only added the percentage of females in the corresponding school. For the immigration status, we only added the country dummy variables. For the SES level, we only added the mean SES of the school attended. For the interest in science we only added the percentage of females in the corresponding schools. The final model only retained those predictors that were relevant in explaining the variance across levels, after adding these predictors. A summary of all the continuous variables considered in this model is presented in Table 2.

Table 2
*Descriptive Statistics for Continuous Variables*

| Measure | No. | Min. | Max. | M | S.D. | Skewness |
|---|---|---|---|---|---|---|
| | | Student-level | | | | |
| Science Achievement | 30,066 | 130.30 | 841.04 | 508.92 | 97.48 | -.141 |
| Interest | 30,066 | 50.35 | 904.82 | 488.95 | 98.76 | -.269 |
| SES | 29,822 | -3.90 | 3.35 | .01 | .96 | -.344 |
| | | School-level | | | | |
| Accountability | 794 | 8.00 | 16.00 | 12.10 | 2.00 | .053 |
| School SES | 849 | -2.06 | 1.59 | -.03 | .59 | -.520 |
| Percentage of Girls | 831 | 0.00 | 1.00 | .49 | .22 | .120 |

We used full maximum likelihood procedures to estimate the parameters. All level-2 predictors, except for the countries' dummy variables, were grand-mean centered. Missing data were eliminated from the analyses. The software used to conduct the analyses was HLM.

## Results

In relation to our first research question, the estimates for the regression analysis are displayed in Table 3. The r-squared for the model was 0.24. We observe that only one of the five covariates had a significant coefficient: external influence on instructional content and assessment (EIA). This result could suggest that accountability practices are externally driven. All three coefficients for the dummy variables corresponding to different countries were significant and negative. USA was the reference category, so this result means that USA has significantly higher levels of school accountability, in relation to the other selected countries. To determine if accountability rates differed across the countries, we examined the 95% confidence intervals for their coefficients, which are displayed in Table 4.

As seen in Table 4, the coefficient for Korea is not significantly different from that of Australia or that of Portugal; the confidence intervals for their estimates overlap. However, the coefficients from Australia and Portugal are significantly different. Therefore, there is a difference in the levels of school accountability practices for most pair of countries. However, there are no significant differences between Korea and Portugal, and Korea and Australia. A summary of these results are presented in Table 5. Significant differences are indicated by a "yes" in the pairwise table.

Table 3
*Predictors and Hypothesis Tests*

| Parameter | β | S.E | *t* | Sig. |
|---|---|---|---|---|
| Intercept | 15.155 | 0.681 | 22.252 | <0.001 |
| | | | | |
| Australia | -1.790 | 0.188 | -9.567 | <0.001 |
| Korea | -2.030 | 0.215 | -9.424 | <0.001 |
| Portugal | -2.849 | 0.226 | -12.589 | <0.001 |
| EBS | -0.073 | 0.075 | -0.967 | 0.334 |
| EIA | -0.208 | 0.071 | -2.916 | 0.004 |
| IBS | -0.059 | 0.068 | -0.870 | 0.385 |
| IIA | 0.113 | 0.067 | 1.693 | 0.091 |
| PP[a] | 0.051 | 0.100 | 0.511 | 0.610 |

Note. EBS=External Influence on Budgeting and Staffing. EIA=External Influence on Instructional Content and Assessment Practices. IBS= Internal Influence on Budgeting and Staffing. EIS= External Influence on Instructional Content and Assessment Practices.
[a]Inversed

In relation to our second research question, the results of the hierarchical linear regression are presented in Tables 6(fixed parameters) and 7 (random parameters). There are five models presented in this table: Model 1 corresponds to the fully unconditional two-level model; Model 2 includes the level-1 predictors; in Model 3 we added the accountability predictor (level-2), after retaining the significant level-1 predictors; Model 4 includes all the level-2 predictors; Model 5 corresponds to the final model, after removing the non-significant level-2 predictors.

Table 4
*Confidence Intervals for Countries' Parameters*

| Parameters | 95% C.I. | |
|---|---|---|
| | Lower Bound | Upper Bound |
| Australia | -1.978 | -1.602 |
| Korea | -2.245 | -1.815 |
| Portugal | -3.075 | -2.623 |

Table 5
*Pairwise Comparison of Countries' Coefficients*

| | USA | Australia | Korea | Portugal |
|---|---|---|---|---|
| 1. USA | - | Yes | Yes | Yes |
| 2. Australia | | - | No | Yes |
| 3. Korea | | | - | No |
| 4. Portugal | | | | - |

As observed in Table 6 under Model 1, the predicted mean science achievement is 508, which is higher than the mean for all students who took the assessment across all countries. The average science plausible value for the complete PISA database varies between 19 and 913 points, with a mean of 476 points. Also, the intra-class correlation coefficient (ICC) for Model 1 is 28%; this means that 28% of the variance at the student level is explained by school level factors. The gender, SES, interest in science, and immigration status predictors were added in Model 2.

Table 6
*Model Comparison: Fixed Effects*

| Parameter | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Intercept | 507.6 **(1.9) | 523.3** (3.0) | 523.3** (3.1) | 505.2** (6.0) | 528.4* (5.1) |
| *Level-1* | | | | | |
| Gender | | -2.4* (1.2) | -2.4* (1.2) | -1.9 (1.1) | -2.13 (1.2) |
| SES | | 23.3** (0.8) | 23.4** (0.8) | 23.6** (0.8) | 23.53** (0.8) |
| Interest | | 0.1** (0.0) | 0.1** (0.0) | 0.1** (0.0) | 0.10** (0.0) |
| Immigration | | -13.0** (2.0) | -13.0** (2.0) | -18.0** (4.3) | -23.6** (4.2) |
| *Level-2* | | | | | |
| Intercept | | | | | |
| Accountability | | | 0.2 (1.5) | 0.2 (0.7) | -1.4* (0.7) |
| Mean SES | | | | 67.6** (2.4) | 64.8** (2.5) |
| Girls (%) | | | | -11.7 (6.5) | -- |
| Australia | | | | 14.4* (6.8) | -7.9 (6.1) |
| Korea | | | | -54.9 (80.4) | -- |
| Portugal | | | | 53.1** (9.6) | 26.2** 9.0) |
| Gender | | | | | |
| Accountability | | | 0.3 (0.6) | -- | -- |
| Girls (%) | | | | 0.7 (5.9) | -- |
| SES | | | | | |
| Accountability | | | 1.3** (0.4) | 0.9* (0.4) | 0.9* (0.4) |
| Mean SES | | | | 7.1** (1.4) | 7.0** (1.4) |
| Interest | | | | | |
| Accountability | | | 0.0 (0.0) | -- | -- |
| Girls (%) | | | | 0.0 (0.0) | -- |
| Immigration | | | | | |
| Accountability | | | 0.6 (0.9) | -- | -- |
| Australia | | | | 9.7* (4.8) | 15.4** (4.8) |
| Korea | | | | 87.7 (80.2) | -- |
| Portugal | | | | -20.6** (7.4) | -14.8* (7.3) |

Note. "- -" denote the parameters that were not estimated for the corresponding model.
*p<.05, **p<.01

From Table 6 we note that the coefficients corresponding to these variables are all significantly different from zero. We observe that for a given school, the predicted science achievement of a native male with average SES and interest is 523 points. This average decreases to 508 for a first or second generation female. And these values increase in 23 points for each unit increase in the difference between SES and the mean SES for the corresponding school. The effect of the interest in science, albeit significant, is limited. Last, the percent of level-1 variance explained by these predictors is 10%. Table 7 shows that including the level-1 predictors results in a 10% reduction in the unexplained variance (from 6,689 to 6,021).

In relation to Model 3, two things are noted. First, accountability was only significant in explaining the variance of the slopes for the SES variable. The interpretation is that for each unit increase in the accountability level, in relation to the grand-mean, the effect of SES on science achievement increases on average, 1.3 points. Second, the inclusion of the accountability predictor

accounted for a small portion of the between-school variance components. Again, as shown in Table 7, all the variance components for the random effects were significantly different than zero.

On the other hand, Model 4 shows that, in terms of predictors, neither the dummy variable for Korea nor the percentage of females in the classrooms were significant in their corresponding equations. School level accountability practices remained irrelevant to explain the variance of the level-1 intercept. However, the rest of the coefficients (Portugal, Australia, and Mean SES) were significant in this regard. Indeed, as shown in Table 7, the percent of between-school variance explained by this model is considerably better than the previous one. Again, all variance components for the random effects were significantly different from zero. The percentage of variance explained for the rest of the level-1 predictors was not very high, yet some improvements were observed in relation to the gender, SES, and immigration predictors.

Table 7
*Model Comparison: Random Effects*

| Parameter[a] | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Level -1 | | | | | |
| Intercept | 6,688.84 | 6,021.74 | 6,021.94 | 6,027.89 | 6,026.55 |
| Level -2 | | | | | |
| Intercept | 2,658.47** | 3,542.63** | 3,554.97** | 2,047.22** | 2,168.50** |
| Gender | | 133.21** | 130.05** | 126.05** | 127.54** |
| SES | | 170.32** | 163.46** | 153.16** | 152.68** |
| Interest | | 0.01** | 0.01** | 0.01** | 0.01** |
| Immigration | | 361.49** | 363.29** | 328.48** | 324.08** |

Note. "- -" denote those parameters that were not estimated for the corresponding model.
[a]Variance components. Covariance estimates are not included in this table.
*p<.05, **p<.01

Model 5 corresponds to the final model. The predictor for accountability was retained on the basis that removing the other non-significant predictors, made it significantly different to zero in the equation for the level-1 intercept. On the other hand, we note that the percent of level-2 variance explained by this model is similar but slightly higher than that of the previous model with all level-2 predictors (model 4). Again, all the variance components for the random coefficients were significantly different than zero.

The data from this last model suggests several points, although they are not straightforward given the number of predictors at each level. For an average student attending an average school (as defined by the selected variables) from the U.S. or Korea, the mean predicted science achievement is 528 points. This predicted mean reduces to 526 for a similar female student, and to 503 if she is either a first or second generation in the corresponding country. Students who have a one-unit higher SES level than that their peers, obtain on average scores that are higher by 23 points. This is a relevant result given that the standard deviation for the individual SES variable in this sample was approximately 0.96 points. Moreover, this effect is higher for schools with higher mean SES. On the other hand, despite that the coefficient for the interest in science is small, the scale of this variable is similar to the scale of the science achievement and so one standard deviation increment in the interest in science, produces nine points of difference in science achievement.

In relation to the results pertaining to accountability, one unit increase in the school level accountability practices, in relation to the grand mean for this sample, decreases the expected mean science achievement by 1.44 points. However, there is a positive interaction between school level accountability and individual SES. Students with higher SES than their peers would get additional

0.88 points for each unit increase in their individual SES in relation to the school mean (beyond the 23 points increase they get for such difference). Therefore, these results indicate that: (a) the effect that school accountability practices have on science achievement, is small and negative; (b) the effect of accountability is not independent of the SES level of the student.

## Discussion

The purpose of this study was to gain insight on the relationship between accountability and science achievement. Science achievement is often considered critical to nations, yet it is rarely the focus of studies that look at the relationship between accountability and students' outcomes. In particular, we focused on the relationship that a number of school level accountability practices had with science achievement, across four countries: Australia, South Korea, Portugal, and the United States. To that end, we first sought to determine if there were significant differences between the extents of school level accountability practices across these countries at the time when data were collected. Using linear regression analysis, we concluded that there were differences in this regard between pairs of countries, but not all pairs of countries were significantly different from each other. On average, the United States had higher levels of school accountability practices than the rest of the countries. Australia had higher levels of school accountability practices than Portugal, but not from Korea. In fact, Korea was not significantly different from either Australia or Portugal under this metric.

These school accountability practices did not necessarily emerge from accountability systems in place. However, these differential levels matched the expectations entailed by such systems for each country at that time. By 2006, the United States had already in place a system of high-stakes state accountability and it is therefore expected that schools endorsed accountability practices such as the ones that we considered in this analysis. Despite not having a national accountability system by that time, Australia had several systems of school level accountability in place. These systems varied by territory/states and typically included school monitoring practices carried internally and externally (Gurr, 2007). On the other hand, Korea had a national testing system in place but its reach or stakes were not transformed into a census high-stakes accountability system until 2008. And Portugal had a low-stakes national testing program operating at the time.

Beyond the differences in school level accountability practices, we looked at the relationship that these practices had with science achievement, using multilevel analysis. We did not hypothesize any result because the relationship between accountability and achievement is "varied, limited, and relatively inconclusive" (Nichols et al. 2012, p. 26). Moreover, potentially stable relationships identified in the literature do not refer to science achievement. Our results indicate that the relationship that school level accountability practices have with science achievement is small, and mostly, negative. In particular, school level accountability practices were relevant in explaining mean science achievement[2] and the effect that individual SES had on this predicted mean. The results from this study suggest that the effect that school level accountability practices have on science achievement, as measured by PISA data, is not independent from the SES of the students. In particular, these accountability practices may be more beneficial for students with high SES, when compared to their school peers.

Other outcomes are worth noting. The model confirms that the effect that individual SES has on science achievement is huge, as indicated by innumerable studies that have looked at this relationship across different subjects. A result that was not in line with our expectations was that the

[2] For non-immigrant men with SES and interest levels as the mean from their school peers.

percentage of females in a classroom did not affect the mean science achievement. This does not support studies that have found that the sex-composition of schools is related to differential achievement outcomes for girls (e.g. Gándara & Silva, 2015). A possible explanation is that the effect that sex-composition has on science achievement varies across countries. Last, being an immigrant was on average, associated with lower science achievement. This finding was true for all countries, but it was considerably better for immigrants in Australia, a result that may deserve further exploration.

With regards to the limitations, it is important to remember that this is a correlational analysis and we are not attributing a causal effect from accountability on science achievement. Any cross-cultural interpretation of students' performance on international assessment needs to take into consideration the educational realities of the countries (Kim, Lavonen, & Ogawa, 2009). Because of this, we included several covariates identified as relevant by previous research. However, we could not include all the variables we intended, because there was incomplete information. For example, it is said that Korean success in scientific literacy is partly due to the large investments in private education, sector in which parents play a fundamental role (Kim et al., 2009), but we could not include the type of school funding (private vs. public) because the data were not available for one of the countries. Another limitation is that science achievement measured by PISA is not linked to the curriculum and therefore, results may be different using other outcome, specially, one with high-stakes attached. Moreover, we used data for 15-year-olds only, and it is possible that the stability of the results change across grade levels and ages. Research shows that the correlation between measures related to accountability and achievement change not only by subject but also by grade (Nichols et al., 2012). Also, there are many sources of incomparability related to international assessments (Ercickan, Roth, & Asil, 2015), and we did not validate the comparability of the measures but assumed them as comparable. Last, the school level accountability practices herein considered are limited to what was asked in questions 15 and 17 from the PISA school questionnaire. We recognize that these questions do not encompass all possible school level accountability practices, and we cannot state that they are representative of all school level accountability practices. For all of the above reasons, our results must be interpreted with caution. Notwithstanding, the skills measured by PISA are critical to any national curriculum, and we did control an important amount of variance with appropriate covariates, so our results should not be discarded on the basis of its limitations.

Since 2006, PISA has not focused on science achievement. The next battery that will focus on science will be administered in 2015. Further research should follow up this study using the 2015 data. In that context, the new question that arises is: given that Australia and Korea have increased their national accountability practices since 2008, how has this impacted the average school-level accountability practices reported in PISA? It is likely that many countries modified their national evaluation systems after 2006, not only Australia and Korea. Further research may look at the relationship between the changes in the amount of school-level accountability practices and the gains in science achievement. PISA's school questionnaire has increased the number and detail of questions related to accountability (e.g. OECD, 2011) and we expect that the 2015 administration will provide rich data in this regard. Also, further research should pay closer attention to other school-level characteristics that could explain achievement differences across countries. For example, curricular differences across countries and/or the different assessment practices at a school level. Digging deeper into these matters may improve our understanding around the relationship of school level accountability practices and science achievement.

## References


Achieve. (2014, December). *College and career readiness*. Retrieved from http://www.achieve.org/college-and-career-readiness

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2014, March 01). *NAP Sample Assessments*. Retrieved from National Asssessment Program: http://www.nap.edu.au/nap-sample-assessments/nap-sample-assessments.html

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2014, March 01). *NAPLAN*. Retrieved from National Assessment Program: http://www.nap.edu.au/naplan/naplan.html

Carnegie Corporation of New York and Institute for Advanced Study. (2009). *The opportunity equation: Transforming mathematics and science education for citizenship and the global.* New York, NY: Carnegie Corporation of New York and Institute for Advanced Study.

Centre of Education Policy (CEP). (2012). Accountability issues to watch under NCLB waivers. Retrieved from http://www.cep-dc.org/displayDocument.cfm?DocumentID=411

Centre of Study for Policies and Practices in Education (CEPPE). (2013). Learning standards, teaching standards and standards for school principals. *OECD Education Working Papers, No. 99*. http://dx.doi.org/10.1787/5k3tsjqtp90v-en

Council of Chief State School Officers (CCSSO) & the National Governors Association Center for Best Practices (NGA Center). (2014). *About the standards*. Retrieved September 2014, from Common Core State Standards Initiative: http://www.corestandards.org/about-the-standards/

Ercikan, K., Roth, W., & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teachers College Record, 117*, 1-28.

Gándara, F., & Silva, M. (2015). Understanding the gender gap in science and engineering: Evidence from the Chilean college admissions test. *International Journal of Science and Mathematics Educaiton (Online)*. doi:10.1007/s10763-015-9637-2

Gilleece, L., Cosgrove, J., & Sofroniou, N. (2010). Equity in mathematics and science outcomes: Characteristics associated with high and low achievement on PISA 2006 in Ireland. *International Journal of Science and Mathematics Education, 8*, 475-496. http://dx.doi.org/10.1007/s10763-010-9199-2

Gurr, D. (2007). Diversity and progress in school accountability systems in Australia. *Educational Research for Policy and Practice, 6*, 165-186. http://dx.doi.org/10.1007/s10671-007-9021-2

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives, 11*, 1-18. http://dx.doi.org/10.1080/15366367.2013.783752

Hanushek, E. A., & Raymond, M. E. (2006). School accountability and student performance. *Regional Economic Development, 2*(1), 51-61.

Hatch, T. (2013). Beneath the surface of accountability: Answerability, responsibility and capacity-building in recent education reforms in Norway. *Journal of Educational Change, 14*, 113-138. http://dx.doi.org/10.1007/s10833-012-9206-1

Hofstein, A., Eilks, I., & Bybee, R. (2011). Societal issues and their importance for contemporary science education - A pedagogical justification and the state-of-the-art in Israel, Germany, and the USD. International Journal of Science and Mathematics Education, 9, 1459-1483. http://dx.doi.org/10.1007/s10763-010-9273-9

Judson, E. (2012). When science counts as much as reading and mathematics: An examination of different state accountability policies. *Education Policy Analysis Archives, 20*(26), 1-21. http://dx.doi.org/10.14507/epaa.v20n26.2012

Kim, M., Lavonen, J., & Ogawa, M. (2009). Experts' opinions on the high achievement of scientific literacy in PISA 2003: A comparative study in Finland and Korea. *Eurasia Journal of Mathematics, Science, & Technology Education, 5*(4), 379-393.

Klenowski, V. (2011). Assessment for learning in the accountability era: Queensland, Australia. *Studies in Educational Evaluation*, 78-83. http://dx.doi.org/10.1016/j.stueduc.2011.03.003

Korean Educational Development Institute. (2010). *Country background report for Korea.* OECD Publishing.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Koretz, D., McCaffrey, D., & Sullivan, T. (2001). Predicting variation in mathematics performance in four countries using TIMSS. *Education Polciy Analysis Archives, 9*(34), 190-217.

Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods, and findings. *International Journal of Science Education, 33*(1), 27-50. http://dx.doi.org/10.1080/09500693.2010.518645

Laffont, J. J., & Martimort, D. (2002). *The theory of incentives: The principal-agent model.* Princeton University Press: Princeton, NJ.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research, 78*(3), 608-644. http://dx.doi.org/10.3102/0034654308324427

Mas-Callel, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory.* Oxford University Press: NY.

National Center for Education Statistics (February, 2015). Timeline for National Assessment of Educational Progress (NAEP) Assessments from 1969 to 2017. Retrieved from: http://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives, 20*(20), 1-35. http://dx.doi.org/10.14507/epaa.v20n20.2012

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report.* OECD Publishing.

Organisation for the Economic Co-Operation and Development (OECD). (2007). *PISA 2006: Science competencies for tomorrow's world.* OECD Publishing.

Organisation for the Economic Co-Operation and Development (OECD). (2011). *School Questionnaire for PISA 2012.* OECD Publishing.

Organisation for the Economic Co-Operation and Development (OECD). (2014). Dépense intérieure brute de R-D par secteur d'exécution et par domaine scientifique. Retrieved December 20, 2014. Retrieved from http://stats.oecd.org/Index.aspx?DataSetCode=ONRD_COST

Partnership for Assessment of Readiness for College and Careers. (2014). *About PARCC.* Retrieved from http://www.parcconline.org/about-parcc

Phelps, P. R. (2012). The effect of testing on student achievement, 1910-2010. *International Journal of Testing, 12*(1), 21-43. http://dx.doi.org/10.1080/15305058.2011.602920

Polikoff, M. S., McEachin, A. J., & Wrabel, S. L. (2014). The waive of the future? School accountability in the waiver era. *Educational Researchers,43*(1), 45-54. http://dx.doi.org/10.3102/0013189X13517137

Ravitch, D. (2015, April). Senate committee reaches agreement on new ESEA. Diane Ravitch's blog. Retrieved from http://dianeravitch.net/2015/04/07/senate-committee-reaches-agreement-on-new-esea/

Rosenkvist, M. A. (2010). Using student test results for accountability and improvement. *OECD Education Working Papers, No. 54*, 1-50.

Santiago, P., Donaldson, G., Looney, A., & Nusche, D. (2012). Student assessment. In P. Santiago, G. Donaldson, A. Looney, & D. Nusche, *OECD reviews of evaluation and assessment: Portugal* (pp. 49-66). OECD Publishing. http://dx.doi.org/10.1787/9789264117020-en

Schleicher, A. (2011, October). Is the sky the limit to education improvement? *Phi Delta Kappan, 93*(2), pp. 58-63. http://dx.doi.org/10.1177/003172171109300213

Schneider, M. (2015, April). Some details on the senate-proposed ESEA reauthorization. Mercedes Schneider's EduBlog. Retrieved from https://deutsch29.wordpress.com/2015/04/07/some-details-on-the-senate-proposed-esea-reauthorization/

Smarter Balanced Assessment Consortium (Smarter Balanced). (2014). *About.* Retrieved from http://www.smarterbalanced.org/about/

Sullivan, A., Joshi, H., & Leonard, D. (2010). Single-sex schooling and academic attainment at school and through the lifecourse. *American Educational Research Journal, 47*(6), 6-36. http://dx.doi.org/10.3102/0002831209350106

Tomul, E., & Sebile Savasci, H. (2012). Socioeconomic determinants of academic achievement. *Educational Assessment, Evaluation, and Accountability, 24*, 175-187. http://dx.doi.org/10.1007/s11092-012-9149-3

Wobmann, L., Luderman, E., Schutz, G., & West, M. R. (2007). School accountability, autonomy, choice, and the level of student achievement: International evidence from PISA 2003. *OECD Education Working Papers, No.13*, 1-85.

Women Thrive Worldwide (June, 2015). Open work group brief: Sustainable development goals on equitable learning. Retrieved from: http://womenthrive.org/sites/default/files/docs/resources/request_for_sign-on_-_owg_call_to_action_for_equitable_learning_updated_3.28.14.pdf

**Appendix A**
**School Accountability Index**
**School Questionnaire**

**Q15 This set of questions explores aspects of the school's <accountability> to parents.**

*(Please tick one box in each row)*

| | *Yes* | *No* |
|---|---|---|
| a) Does your school provide information to parents of students in <national modal grade for 15-year-olds> on their child's academic performance relative to other students in <national modal grade for 15-year-olds> in your school? | ☐ 1 | ☐ 2 |
| b) Does your school provide information to parents of students in <national modal grade for 15-year-olds> on their child's academic performance relative to national or regional <benchmarks>? | ☐ 1 | ☐ 2 |
| c) Does your school provide information to parents on the academic performance of students in <national modal grade for 15-year-olds> as a group relative to students in the same grade in other schools? | ☐ 1 | ☐ 2 |

Question 15 from PISA's School Questionnaire

**Q17 In your school, are achievement data used in any of the following <accountability procedures>?**

*Achievement data include **aggregated** school or grade-level test scores or grades, or graduation rates.*

*(Please tick one box in each row)*

| | *Yes* | *No* |
|---|---|---|
| a) Achievement data are posted publicly (e.g. in the media) | ☐ 1 | ☐ 2 |
| b) Achievement data are used in evaluation of the principal's performance | ☐ 1 | ☐ 2 |
| c) Achievement data are used in evaluation of teachers' performance | ☐ 1 | ☐ 2 |
| d) Achievement data are used in decisions about instructional resource allocation to the school | ☐ 1 | ☐ 2 |
| e) Achievement data are tracked over time by an administrative authority | ☐ 1 | ☐ 2 |

Question 17 from PISA's School Questionnaire

**Appendix B**
**Influences**
**School Questionnaire**

**Q12** **Regarding your school, which of the following bodies exert a direct influence on decision making about staffing, budgeting, instructional content and assessment practices?**

*(Please tick as many boxes as apply)*

| | *Area of influence* | | | |
|---|---|---|---|---|
| | *Staffing* | *Budgeting* | *Instructional content* | *Assessment practices* |
| a) Regional or national education authorities (e.g. inspectorates) | ☐ 1 | ☐ 1 | ☐ 1 | ☐ 1 |
| b) The school's <governing board> | ☐ 1 | ☐ 1 | ☐ 1 | ☐ 1 |
| c) Parent groups | ☐ 1 | ☐ 1 | ☐ 1 | ☐ 1 |
| d) Teacher groups (e.g. Staff Association, curriculum committees, trade union) | ☐ 1 | ☐ 1 | ☐ 1 | ☐ 1 |
| e) Student groups (e.g. Student Association, youth organisation) | ☐ 1 | ☐ 1 | ☐ 1 | ☐ 1 |
| f) External examination boards | ☐ 1 | ☐ 1 | ☐ 1 | ☐ 1 |

Question 12 from PISA's School Questionnaire

**Q16** **Which statement below best characterises parental expectations towards your school?**

*(Please tick only one box)*

| | |
|---|---|
| There is *constant pressure* from many parents, who expect our school to set very high academic standards and to have our students achieve them | ☐ 1 |
| Pressure on the school to achieve higher academic standards among students comes from a *minority of parents* | ☐ 2 |
| Pressure from parents on the school to achieve higher academic standards among students is *largely absent* | ☐ 3 |

Question 16 from PISA's School Questionnaire

## About the Authors

**Fernanda Gándara**
University of Massachusetts Amherst
mgandara@educ.umass.edu
Fernanda Gándara is a doctoral candidate in psychometrics at the University of Massachusetts Amherst. Her research is oriented towards the improvement of assessment practices across educational contexts. She is mostly interested in the intersection between assessment and policy, as well as the assessment practices that affect linguistic minorities and students in need. Fernanda has worked in Chile and the U.S., and has been involved in international education projects. Until recently, she conducted research for the Center of Educational Assessment at the University of Massachusetts Amherst. She is currently a Research Associate at School-to-School International.

**Jennifer Randall**
University of Massachusetts Amherst
jrandall@educ.umass.edu
Jennifer Randall is an associate professor of education at the University of Massachusetts Amherst. Her research interests include the assessment and grading practices/philosophies of classroom teachers, the utility and appropriateness of test accommodations for special populations, as well as scale development for difficult to measure constructs.

education policy analysis archives
Volume 23 Number 112 November 16th, 2015 ISSN 1068-2341