Policy evidence by design

How International Large-Scale Assessments Influence Repetition Rates

Manuel Enrique Cardoso

Submitted in partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Policy evidence by design

How International Large-Scale Assessments Influence Repetition Rates

Manuel Enrique Cardoso

Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition[1]

Links between international large-scale assessment (ILSA) methodologies, international organization (IO) ideologies, and education policies are not well understood. Framed by statistical constructivism, this article describes two interrelated phenomena. First, OECD/ PISA and UNESCO/TERCE documents show how IOs' doctrines about the value of education, based on either Human Capital Theory or Human Rights, shape the design of the ILSAs they support. Second, quantitative analyses for four Latin American countries show that differently designed ILSAs disagree on the effectiveness of a specific policy, namely, grade retention: PISA's achievement gap between repeaters and nonrepeaters doubles TERCE's. This matters and warrants further research: divergent empirical results could potentially incentivize different education policies, reinforce IOs' initial policy biases, and provide perverse incentives for countries to modulate retention rates or join an ILSA on spurious motivations. In summary, ILSA designs, shaped by IOs' educational doctrines, yield different data, potentially inspiring divergent global policy directives and national decisions.

[1] Published by *Comparative Education Review* in November 2020. Winner of the Maureen Hallinan Graduate Student Paper Award, from AERA's Sociology of Education division. Winner of the Early Career Award from CIES's Large-Scale Cross-National Studies in Education (LCSE) Special Interest Group (SIG). Honorable Mention for the George F. Bereday Award, from *Comparative Education Review*.

When ILSAs met policy: Evolving discourses on grade repetition

This study explores phenomena of ordinalization and scientization of policy discourse, focusing on the case of grade retention in publications by OECD's PISA and UNESCO's ERCE (2007-2017), from a sociology of quantification perspective. While prior research shows these ILSAs yield divergent data regarding grade retention's effectiveness, this study shows similarities in their critical discourse on grade repetition's effectiveness. Genre analysis finds similarities in how both ILSAs structure their discourse on grade repetition and use references solely to critique it, presenting a partial view of the scholarly landscape. However, horizontal comparisons also find differences across ILSAs in the use of ordinalization (e.g., rankings) in charts, as well as differences in the extent to which their policy discourse embraces scientization. The ILSAs converge in singling out grade repetition as the policy most strongly associated with low performance; this should be interpreted in the context of one key similarity in their design.

Policymaking to the test? How ILSAs influence repetition rates[2]

Do international large-scale assessments influence education policy? How? Through scripts, lessons, or incentives? For some, they all produce similar outcomes. For others, different assessment data, shaped by different designs, and mediated by international organizations' policy directives, prompt different policy decisions. For some, participation in these assessments may be linked to lower repetition rates, as per the policy scripts hypothesis inspired by world society theory (WST). For others, assessments' comparison strategies (age vs. grade) influence repetition in participating countries, according to policy lessons or incentives hypotheses, respectively inspired by educational effectiveness research (EER) and the sociology of quantification, and particularly the notion of retroaction. Fixed-effects panel regression models of eighteen Latin American countries (1992-2017) show that participation in assessments is associated with changing repetition rates in primary and secondary, while controlling for other factors. The findings show statistically significant differences between some assessment types. The conclusions spur new questions, delineating a future agenda.

[2] A slightly different version of this article was accepted for publication by *International Studies in Sociology of Education*.

Acronyms

BRR	balanced repeated replication
CCS	comparative case study
CONFEMEN	Conference of the Ministers of Education of French speaking countries (Conférence des ministres de l'Éducation des États et gouvernements de la Francophonie)
EER	Educational Effectiveness Research
EI	Education International
ERCE	Regional Comparative and Explanatory Study (Estudio Regional Comparativo y Explicativo)
GDP	Gross Domestic Product
HCT	Human Capital Theory
HR	Human Rights
IALS	International Adult Literacy Survey
IEA	International Association for the Evaluation of Educational Achievement
IGO	International Governmental Organization
ILSA	International Large-Scale Assessment
IO	International Organization
IRR	inter-rater reliability
ISCED	International Standardized Classification of Education
LAMP	Literacy Assessment and Monitoring Programme
LLECE	Latin American Laboratory for Assessment of the Quality of Education (Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación)
NALS	National Adult Literacy Study
NER	Net Enrolment Rate
OECD	Organisation for Economic Co-operation and Development
OEEC	Organization for European Economic Cooperation
OLS	ordinary least squares
PASEC	CONFEMEN Programme for the Analysis of Education Systems (Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN)

PERCE	First Regional Comparative and Explanatory Study (Primer Estudio Regional Comparativo y Explicativo)
PIAAC	Program for the International Assessment of Adult Competencies
PILNA	Pacific Island Literacy and Numeracy Assessment
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPP	Purchasing Power Parity
RCT	randomized controlled trial
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SEA-PLM	Southeast Asia Primary Learning Metrics
SERCE	Second Regional Comparative and Explanatory Study (Segundo Estudio Regional Comparativo y Explicativo)
SES	socioeconomic status
SGER	Secondary Gross Enrolment Ratio
TALIS	Teaching and Learning International Survey
TERCE	Third Regional Comparative and Explanatory Study (Tercer Estudio Regional Comparativo y Explicativo)
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations Child Fund
WDI	World Development Indicators
WST	World Society Theory

Table of Contents

List of Tables and Figures

Tables

Figures

Acknowledgements

I would like to thank, first and foremost, my advisor Oren Pizmony-Levy, for letting me find my own answers, while always pushing me with questions. But I must confess that I'm glad that B.B., his little one, is probably making him realize what it feels like to be on the receiving end of such a persistent "Why?".

My everlasting gratitude to all my professors at Teachers College, Columbia University, and NYU, but special thanks go out to Gabriel Abend, Carol Benson, Gil Eyal, Daniel Friedrich, Aaron Pallas, and Garnett Russell.

I would be remiss if I did not single out Gita Steiner-Khamsi, for many reasons, but mostly for her warm guidance above and beyond the call of duty.

And my gratitude to the dissertation committee, composed by the abovementioned Profs. Steiner-Khamsi, Pallas and Eyal, as well as Prof. Bryan Maddox, for their invaluable guidance and input.

I would like to thank my family and friends, including former classmates, coworkers, professors, and students who listened to my dissertation-related ramblings (and even sat through presentations) for years. In Uruguay, these include Cecilia Blezio, Carmen González, Luis Eduardo González, Mario González, Carmen Haretche, Graciela Loureiro, Carlos Luján, Ester Mancebo, Felipe Monestier, Beatriz Picaroni, Pedro Ravela, and Denise Vaillant. Around Latin America: Luis Crouch, César Guadalupe, Juan Cruz Perusia, Atilio Pizarro, Liora Schwartz, and Adriana Viteri. And all over the world: Camilla Addey, Amy Jo Dowd, Bryony Hoskins, and Maya Prince.

Special thanks to Luis Benveniste, for persuading me, many years ago, to apply to a master's degree program in the US. To Fernando Reimers, for being my advisor during that master's, and ever since, and to all my other professors at the Harvard Graduate School of Education and the Kennedy School of Government. And to Sylvia Liñán-Thompson for encouraging me to apply to a doctorate. You all set me on this path, for which I am grateful.

Thanks to all my current or recent co-workers for your patience and your willingness to listen. Special thanks to Amalia Verónica Anderson, Peter de Vries, Juan Pablo Giraldo Ospino, Divya Lata, Mary Nicolas, Vina Ocampo-Borje, Saadhna Panday, Atif Rafique, Ingrid Sánchez Tapia, Mark Waltham, and Euphrates Efosi Wose. Finally, Changu Mannathoko and Josephine Bourne, then my supervisors at my full-time job, were incredibly supportive of my decision to start a doctorate; for that, please receive my gratitude.

My heartfelt thanks to my cohort: Kayum Ahmed, Lucía Caumont Stipancic (my fellow Uruguayan!), Phoebe Linh Doan, and Ji Liu. I will never forget our shared food and discussion in that odd little classroom, which in a different lighting or at a different time or day (or night) would have looked like something out of a horror movie.

Over the years, the blurry boundaries of that cohort expanded to include Anna Azaryeva, Chanwoong Baek (with whom I shared more classes than with anyone else!), Jihae Cha, Kevin Henderson, and Grace Zi Hu, among others. To all of them, my gratitude.

Finally, I would not have reached this point without the remote doctoral writing group: Abbas Abassov, Sumita Ambasta, Andrea Kennedy, Erika Kessler, and Carine Verschueren.

Dedication

A mi familia:
mis padres, Manolo y Nora;
mis herman@s, Ismael, Facundo y Sara; sus parejas, Vanina, María Noel y Nicolás;
y mis sobrinas, Victoria y Margarita

A Michael, mi pareja

A Mati

Chapter 1: Introduction

Claims of ILSAs' impact on education policy come from ILSA proponents (Schleicher 1999; Wagemaker 2013; Hanushek and Woessmann 2014) and opponents (Meyer and Benavot 2013; Smith 2016) alike. Even relative agnostics, while not necessarily focusing on policy, emphasize how ILSAs loom large for countries: "While ILSAs are 'low stakes' for test respondents, they are often 'high-stakes' for educational systems" (Maddox 2018, Kindle Edition). However, little research shows statistical links between a country's ILSA participation and the adoption of specific education policies, with a few exceptions (see Ramírez et al. 2018). Some scholars also argue that even if those associations were to be found, it would be difficult to prove causality, and particularly to demonstrate an independent effect of ILSAs on policy (Steiner-Khamsi 2019).

The overall conceptual framework for this study is inspired by the sociology of quantification (see Desrosières 2002; 2015)[1], which is linked to the sociologies of knowledge (Espeland and Stevens 2008) and expertise (Eyal 2019), and particularly by its notion that methods are socially configured and socially consequential (Williamson 2018). The methods of interest here are design features of International Large-Scale Assessments (ILSAs), socially configured by International Organizations (IOs) with different education *doctrines*, thus giving rise to different methodological approaches to the *design* of ILSAs. The study then analyzes the social consequences of these designs, starting with the *data* that they produce. Then it moves on to the policy *discourse* that accompanies the dissemination of these data. Finally, it analyzes the impact on policy *decisions*, or more precisely, their

[1] An introduction to statistical constructivism and the sociology of quantification is provided in the Theoretical Framework to the first empirical study, found in Chapter 2.

outcomes. This study explores quantitative evidence of effects of ILSA participation on education policies. But it does so by focusing on a specific policy case, namely grade retention.

As far as education policies go, few stir as much worldwide debate as grade retention. Highly controversial since at least the 1970s (Jackson 1975), repetition has been accused of lacking effectiveness (Jimerson 1999), efficiency (Eide and Goldhaber 2005), equity (Willms and Somer 2001), empathy (Byrnes and Yamamoto 1985) and evaluability (Allen et al. 2009); while fostering expulsion (Andrew 2014). As will be seen, and crucially for our approach, these various critiques align with education doctrines such as human rights and human capital theory. Critiques aside, grade repetition is still pervasive around the world, with two caveats. First, while grade repetition is nearly universal, repetition rates differ substantially across regions, being particularly high in Latin America and sub-Saharan Africa (UIS 2012). Second, repetition rates have been decreasing worldwide for some time now (UIS 2012). Therefore, while repetition is universal, it also varies substantially across geographies and cultures, as well as over time. But why do repetition rates vary across countries and, particularly, within one country, over time? Could countries' ILSA participation help explain these differences?

Geographic scope

This study focuses on Latin America for two sets of reasons. The first is linked to the availability of ILSA data on repetition; while the second one is related to the importance of repetition in Latin America, in historical and statistical terms.

This study is based on a comparison between different types of assessment that collect and disseminate data on repetition. Not all ILSAs collect data on respondents' repetition history; in fact, this criterion already rules out the two largest IEA studies, namely PIRLS and TIMSS, which do not include questions about this in their international questionnaires. This leaves PISA and the three regional assessments. In addition, the study aims to compare ILSAs while keeping countries constant to the extent possible; therefore, I have sought to maximize the geographical overlap between studied ILSAs. PASEC and SACMEQ, the two regional assessments in sub-Saharan Africa, have minimal overlap with PISA. Therefore, this study compares ERCE and PISA in Latin America, where they overlap in nine countries. As will be seen, these two ILSAs are also backed by IOs with clearly dissimilar education doctrines.

This region is also particularly interesting because of its relationship with repetition. First, Latin America has historically had high repetition rates; in fact, in 2000 Latin America had the highest primary repetition rates of any region in the world, while in 2010 it was second only to sub-Saharan Africa (UIS 2012). Additionally, at the same time, the issue of repetition has been undergoing considerable change: Latin America "has made considerable progress in reducing the number of pupils repeating a [primary] grade, which has fallen by 36% from 8.4 to 5.4 million between 2000 and 2010" (UIS 2012, 20), the sharpest reduction of any region in this period. Finally, there is a rich tradition of debate around the issue of grade repetition in Latin America that dates back nearly half a century, from early attempts to quantify it precisely (Schiefelbein 1975) to explorations of the role it plays in educational inequities in the region (Reimers 2000).

Purpose

The purpose of this dissertation is to understand the influence of ILSAs on education policy, through the lens of the sociology of quantification. The debate around the impact of ILSAs on education policy has been complicated by several factors. ILSAs' proponents argue that these have the power to steer education policy (Schleicher 1999; Wagemaker 2013;

Hanushek and Woessmann 2014), and opponents, paradoxically but perhaps unsurprisingly, tend to agree (Smith 2016; Meyer and Benavot 2013). Consequently, these two camps' debates have often shifted towards the positive or negative value of that impact, while assuming it exists, despite a dearth of empirical evidence to support the original claim.

To overcome this impasse, this study will use three strategies. First, instead of asking the general question about ILSAs' impact on education policy writ large, it will focus on a single, specific, relatively well-defined and globally quantified policy, namely, grade retention. This study aims to be only a first step towards a string of research projects, focused on different policies, that will eventually, collectively, answer the general question. Second, this study will not search for impact, but rather, more humbly, for influence: the former is "a difference in kind" while the latter "is defined as a difference in degree" (Rutkowski et al. 2020, 388). Therefore, I will ultimately be contented to find an association between participation in ILSAs and the seemingly random fluctuations of repetition rates, rather than seek the spectacular and rare arrival of an entirely novel policy in a strange land.

Finally, it will not attempt to demonstrate an *independent* effect of ILSAs on policy. The debate about the impact of ILSAs on education policy has often intersected with broader academic disputes about the role of national actors in the global educational landscape, for instance between proponents of world society theory, which emphasizes the hegemony of global scripts (Ramírez et al. 2018); and theorists of policy borrowing and lending, who underline the importance of agency of national actors (Steiner-Khamsi 2019). While this study intends to elucidate whether ILSAs have an impact on policy or not, it will not attempt to disentangle the much more complex issue of whether there is an *independent,* top-down effect of ILSAs on national policy, or whether, instead, national actors use ILSA participation to advance their own policy or reform agenda.

This study examines how a country's participation in ILSAs influences its repetition rates. However, because the proposed theoretical framework relies heavily on statistical constructivism (Porter 1996) and the sociology of quantification (Desrosières 2015), it does not treat different ILSAs as if they were all one and the same. For the sociology of quantification, the Devil is in the details, in fact, in the *methodological* details. However, it does not consider these details from a methodological or otherwise normative viewpoint, but from a sociological one, in which methods are at the same time socially configured and socially consequential (Williamson 2018).

Research questions

This study then begins by examining how different social configurations, in this case, International Organizations (IOs) with different education *doctrines*, give rise to different methodological approaches to the *design* of ILSAs. There, it focuses on a specific characteristic, the comparison strategy, which distinguishes between grade-based and age-based studies (Allen et al. 2009). It then analyzes the social consequences of these designs, starting with the *data* that they produce. Then it moves on to the policy *discourse* that accompany these data. Finally, it analyzes the observed variations in the phenomenon of interest, namely repetition, itself quantified in the form of repetition rates, which express the outcomes of the *decisions* of policymakers at different levels, from the highest spheres of government to each classroom.

In summary, the four research questions are:

(1) Are IOs' different educational doctrines, i.e., human rights and human capital, associated with different ILSA designs, namely grade- and age-based, respectively?

(2) In Latin America, does the achievement gap between repeaters and non-repeaters reported by each ILSA vary according to its comparison strategy?

(3) How does ERCE's policy discourse about grade repetition differ from PISA's?

(4) In Latin America, is the evolution of repetition rates in each country associated with its prior participation in different ILSAs?

Overview of the dissertation

This dissertation comprises three empirical research studies that address the four questions above, with the first study addressing questions 1 and 2. The studies could be published independently as book chapters or journal articles, since each contains an introduction; background; theoretical framework(s); data and methods; results or findings; discussion; and conclusions. In addition, this first chapter includes the overview, a statement of the problem, and a review of the literature on the debate around ILSAs' influence on policy. The conclusions summarize the three studies; draw links between them; and underline the implications of the dissertation, focusing on the potential to generalize its lessons so that they can be applied to similar problems.

The first article, *Policy evidence by design: International Large-Scale Assessments and Grade Repetition* (Cardoso 2020), connects the initial components: IOs' doctrines, ILSAs' designs, and data. Framed by the sociology of quantification (Desrosières 2015), it reviews the literature on human capital theory and human rights, doctrines that shape the actions of International Organizations (IOs) working in the education field, as exemplified by OECD and UNESCO, respectively. Then, it links these IOs' doctrines to their ILSAs' designs, arguing those doctrines explain each ILSA's "comparison strategy" (Allen et al. 2009), or in other words, why OECD's PISA is age-based while UNESCO's ERCE is grade-based. Finally, drawing on the literature on the impact evaluation of grade retention, and particularly on that same "comparison strategy", it shows how dissimilar designs shape data differently, leading to diverging conclusions on grade retention's effectiveness. Crucially, these conclusions confirm each IO's initial bias about the policy. Although both ERCE and PISA show statistically significant negative associations between repetition and achievement (without controlling for pre-retention achievement measures, which are not available), the PISA coefficient doubles ERCE's, a statistically significant difference.

The second article, *When ILSAs met policy: Evolving discourses on grade retention*, examines ERCE and PISA's discourse about that policy. The documentary analysis will focus on recurring critiques of grade retention around notions of effectiveness, efficiency, equity, empathy, evaluability, and expulsion. I explore how ILSAs' design, their data, as well as the doctrines of the IOs' behind them affect the evolution of their discourse. Regarding design, is the shared endogeneity issue linked to a convergence in the ILSAs' discourses on repetition? Or are different comparison strategies, i.e., age-based v. grade-based, linked to further discursive divergence? The study uses genre (move-step) analysis (Swales 1990) in combination with the comparative case study (CCS) approach (Bartlett and Vavrus) to identify, respectively, the similarities and differences in these ILSAs discourse on grade retention, with a particular focus on effectiveness. It also applies notions of ordinalization (Fourcade 2016) and scientization to the ILSAs' use of charts and references.

The third article, *"Policymaking to the test"? How ILSAs influence repetition rates*, focuses on the relationship between the last three components, namely ILSA data and policy discourse, and countries' policy decisions. The article ponders whether these links are best framed by world society theory (WST), educational effectiveness research (EER), or the sociology of quantification. Is participation in ILSAs always linked to lower repetition rates, as WST might predict? Or do ILSAs' comparison strategies shape the policy lessons they yield, thus affecting the behavior of national policymakers, as EER may expect? Or do ILSAs' comparison strategies incentivize participating countries to tweak their repetition rates to improve their performance in ILSAs, as the sociology of quantification, and particularly the notion of retroaction, would predict? The article uses a fixed-effects panel regression model of nineteen Latin American countries to examine the changes in repetition rates as a function of participation in PISA and grade-based ILSAs as the independent variables, while controlling for other factors like Gross Enrollment Ratio in secondary education, and GDP per capita.

Significance of the study

This study humbly aspires to make empirical, theoretical, and methodological contributions to the fields of comparative education and sociology of education. Empirically, it aims to better understand variations of repetition rates across Latin America and, within countries, over time. This is humble in two ways: it only covers one region; and the proportion of remaining unexplained variance is high. However, showing that participation in different *types* of ILSAs affects repetition rates would be a novel, significant finding. In addition, regional ILSAs like ERCE (one half of this study's focus), PASEC and SACMEQ are also understudied, as compared to cross-regional ILSAs like PISA (the other half), PIRLS and TIMSS.

The modest theoretical contribution is the proposed conceptual framework for the study of ILSAs based on the sociology of quantification (Desrosières 2015), still underused in comparative education and sociology of education. Although it cannot claim to add anything to the theory per se, it may facilitate its application to the field, by identifying aspects of the institutions (e.g., IOs, ILSAs themselves, etc.), methods (e.g., ILSA design features, analytical tools such as charts, etc.), and consequences (e.g., variations in repetition rates) in order to flesh out the notion that methods are socially configured (by institutions, but in what is actually a mutually constitutive process), and socially consequential (Williamson 2018).

Finally, the closely linked and equally humble methodological contribution relates to the specific procedures used in the different articles, which I would argue lend themselves particularly well to the theoretical framework at hand. For instance, using z statistics to compare regression coefficients is hardly revolutionary, but using them to compare the link between achievement and repetition across ILSAs with different designs is fully aligned with the central tenets of the sociology of qualification (Desrosières 2015) and the notion of statistical constructivism (Porter 1996). Similarly, using genre (move-step) analysis, a typically descriptive/prescriptive tool in the field of applied English studies (Swales 1990), to lay the ground for a comparative case study in comparative education (Bartlett and Vavrus 2017) is a relatively novel approach.

The debate on ILSAs and policy

I will devote the rest of this introduction to an update of Wagemaker's (2013) review of ILSA impacts. Next to their growth in number of studies and participating countries, Wagemaker highlights impact on these outcomes: discourse, curriculum, teaching, technical capacity, research, and policy, at the country level; as well as global and donor responses. However, for conciseness, I will focus here only on the outcomes that are most relevant to this dissertation proposal: namely, discourse, research, and policy[2]. Some outcomes are more responsive; while there is substantial evidence of ILSAs' effect on discourse, the same cannot be said of actual policy change. Therefore, nuance is needed; for instance, a useful distinction defines "impact as a difference in kind while influence is defined as a difference in degree.

[2] This section borrows heavily from the corresponding section of my specialization exam. The full set of outcomes can be found there.

(…) Policy influence can be viewed as the use of ILSA results to buttress or tweak policy settings that already exist" (Rutkowski et al. 2020, 388). This will prove a useful concept for our policy case, namely grade retention, given that this policy already exists in most countries, with room for "tweaking".

This review of scholarship about ILSA's influence on outcomes identified by Wagemaker (2013), uses three additional categorizations: political, theoretical, and methodological. The political one distinguishes scholarship that advocates for ILSAs' influence on education systems (proponents), against it (opponents), or that claims disinterest (agnostics). Meanwhile, the theoretical distinction refers to approaches in the area of comparative education or related disciplines, e.g., world culture theory or actor network theory. Finally, the field is methodologically diverse: secondary analysis of quantitative indicators and their correlation with ILSA participation (Kijima 2010); analysis of news media coverage (see Pizmony-Levy 2018) or public opinion surveys (Pizmony-Levy and Bjorklund 2018); in-depth interviews or observations with actors involved in ILSAs as practitioners, potential users (Kijima and Lipscy 2020) or respondents (Maddox, Keslair and Javrh 2018). There are occasional links between methods and theories: world culture theory scholars may rely on secondary analysis of global quantitative datasets in large-n studies (Ramírez, et al. 2018); while actor-network theory (ANT) researchers prefer interviews and ethnographies, focusing on one or a few countries at a time (Addey 2020).

Discourse

Scholars have started conducting empirical studies on the impact of ILSAs on discourse, a broad topic that demands reviewing diverse practices by a wide array of stakeholders. These include: ILSAs' own communication strategies for dissemination of their results, with a special focus on rankings (Espeland et al. 2016; Sorensen and Robertson

2020); the role of the news media, mainly in divulging results (Pizmony-Levy 2018); the discourse of national elites that either implement ILSAs or use results as inputs when designing education policies and reforms (Kijima and Lipscy 2020); and public opinion processes in education matters, with inputs from ILSAs, the media, politicians, technocrats, and other stakeholders (Pizmony-Levy and Bjorklund 2018).

The sociology of quantification is keenly interested in the devices ILSAs use to convey results to the media and the general public. Among them, one stands out: “Rankings are a compelling example of accountability measures (…) because their precise comparisons generate intense competition among those being evaluated, a competition that makes the rankings’ effects easier to see” (Espeland et al. 2016, 4). However, not all ILSAs issue them; for instance, UNESCO’s ERCE refrains from using rankings (Pizarro et al. 2016). How do rankings align with IOs’ education doctrines? They suggest “a promise of movement and social mobility, revolving around a human capital view of the world” (Sorensen and Robertson 2020, 28). Could this explain UNESCO’s reluctance? Finally, do rankings increase media coverage and impact on public opinion? “The use of ILSA data as evidence for policy is often linked to the performance of reference societies, the relative positions of countries in the rankings and the improving or declining ranking or performance of a given country” (Addey et al. 2017, 438).

News media is a key channel through which ILSAs impact public discourse: “The legitimization and institutionalization of ILSAs does not occur upon their arrival to the local domain, rather, it occurs through the publication of supportive news stories and public discourse” (Pizmony-Levy 2018, 591). This process is achieved by “indigenizing” ILSAs by referencing distinctly local ways of discussing education issues; by making them “reaction-worthy” through scandalization; and by marginalizing critiques of ILSAs (Pizmony-Levy

2018, 591). Some ILSAs, e.g., PISA, because of their recurrence, regularity, and reliance on rankings, are ready-made media stories: every third year, on "PISA day", the first Monday of December, results are likely to end up on the first page of the newspapers in participating countries, whether the news are good, bad, or no news at all.

For policy change to happen, elites are also important, particularly those involved in policymaking decisions, or in diagnosing the state of the system for those who make the decisions. A survey of these elites shows that ILSAs "generate detailed information about the shortcomings of a country's education system, which can inform reforms" (Kijima and Lipscy 2020, 176). Dissemination of ILSA results does not necessarily create the desire for reform, particularly among elites; however, it might increase receptiveness to previously existing reform initiatives (Kijima and Lipscy 2020), a pattern that has been observed elsewhere: "ILSA data are (…) often used to legitimise reform or inaction that is more closely tied to internal political agendas and participation in international communities" (Addey et al. 2017, 438).

The role of public opinion in education issues is also crucial, particularly in democratic societies (Kijima and Lipscy 2020). Media themselves matter mostly because they can help shape public opinion on the state of an education system, which, when it sours, can open the door to policy change, even sweeping reform. Crucially, PISA does not directly shape public opinion, which "is indirectly influenced by PISA performance as the scores and rankings create the opportunity to shape the portrayal of the education system which in turn could affect public opinion" (Pizmony-Levy and Bjorklund 2018, 244). But ultimately, "PISA results have a significant relationship to public trust in education" (Pizmony-Levy and Bjorklund 2018, 254), based on a large-n study that combines PISA results with trust in the education system, from nationally representative surveys.

More research is needed on how an erosion of the general public's trust in education can trigger specific policy changes or sweeping education reforms. From the perspective of the sociology of quantification, some questions emerge: Do the specifics of results and reforms matter, for instance, regarding grade repetition? Or can eroding trust in education lead the public to consider any policy change or reform? In summary, ILSAs contribute, through their data and discursive action, to shape the education discourse of the media, policymaking elites, and the general public alike, in ways that may pave the way for policy change and education reform.

Research

In terms of research, one must distinguish between research with ILSAs and on ILSAs. Research with ILSAs uses ILSA data for policy, curriculum, instruction, etc. It treats ILSAs, or one ILSA, as an objective data source on the state of school systems and the effectiveness and equity of policies. It is currently booming in developed (Agasisti and Cordero 2017; Ikeda and Garcia 2014; Willms and Somer 2001) and developing countries (Gomez Vera 2013; Rivas 2015) alike. However, its influence may be different in developing countries, especially when these share cultural traits and ownership of the assessment: "Regional comparisons are especially useful for developing countries interested in critically examining what seems to be working among neighbouring countries that may be facing similar educational issues" (Kamens and Benavot 2011, 298).

Research on ILSAs is emerging. It treats ILSAs not as a source, but as the object under study. It may: not assume they are objective; compare different ILSAs to each other; and question them from methodological (Rutkowski and Rutkowski 2016), epistemological (Gorur 2016), or pedagogical (Smith 2016) perspectives. Currently, this research is more widespread in developed countries; but developing countries may have substantial contributions to make. A review of economic research using four estimation strategies employed on data from ILSAs to evaluate the effectiveness of different interventions through causal inference techniques, the authors "longitudinal datasets at student or school level. (…) [are] required to evaluate the effectiveness of particular interventions in the long run" (Cordero, Cristóbal and Santín 2018, 900). Others claim that PISA research is "constructed and managed by a clique of researchers in dominant positions with regard to the volume and impact of their publications and the opportunities to guide the narrative surrounding the scientific debate about PISA" (Molstad and Pettersson 2017, 880).

Finally, a line of research on ILSAs that will be particularly useful to this study involves different types of responses by governments to the dissemination of ILSA results. Although in theory this could be included in the policy section, governments do not always respond with full-blown policies. Some responses can be oriented to the short- or medium-term, in an instrumental way, as in Pizmony-Levy' ILSA-oriented changes, including "short-term preparation of students in cohorts that are scheduled to be assessed by ILSAs, and increasing students' motivation to invest energy and time when participating in ILSAs" as opposed to what he calls system-oriented changes (Pizmony-Levy 2018, 588). The notion of ILSA-oriented change is clearly linked with the perverse incentives that often operate as a result of the adoption of official statistics (Desrosières 2015).

Policy

The issue of links with policy is much discussed by ILSA proponents and opponents, but the debate is often ideological or methodological, rather than rooted in empirical evidence, which is still scarce. A notable exception is a world culture theory study that finds some associations but simultaneously acknowledges an important limitation: "while international and national assessment tests are correlated with salutary outcomes in terms of educational participation and curricular breadth, we cannot infer a causal relationship" (Ramírez et al. 2018, 358; my emphasis). This quote touches upon three important questions. First, is there a correlation between ILSA participation and the adoption of education policies? If so, are those policies "salutary"? Finally, is the relationship causal? The next section will show how ILSA proponents, opponents, and agnostics, as well as the relevant subcategories under each of these, answer those questions.

Proponents

Proponents are either producers or users. Producers develop or implement ILSAs; these include, for instance, Schleicher (1999) and Wagemaker (2013). Users, without directly producing ILSA data, analyze their results, often with economic data such as GDP growth (Hanushek and Woessmann 2014). Mostly economists, users would cast themselves as the demand for ILSA data, with producers as the supply. Producers and users often agree with each other but differ in their institutional affiliations and professional cultures. Producers, in addition to being practitioners, have considerable impact on the discourse of the scholarly community. The most visible producer is Andreas Schleicher (1999), who is transparent about OECD's agenda to shape education policy through ILSAs, and through PISA in particular: while "many individuals in participating countries, including professionals and lay-persons, will use the survey results for a variety of purposes, the primary reason for developing and conducting [PISA] is to provide empirically grounded information which will inform policy decisions" (Schleicher 1999, 7).

But producers are far from monolithic in this unabashed support for the use of ILSA results for policy purposes. By comparison to OECD's PISA, "the declared explicit use of ILSAs for policy modifications are less clear for [TIMSS and PIRLS], perhaps because of the IEA's history as a research rather than policy focused organization" (Rutkowski et al.

2020, 389). IEA seems reluctant but also threatened by PISA's embrace of this new role for ILSAs: "In general, "shocks" attributed to ILSAs tend to be focused on PISA results rather than other studies" (Rutkowski et al. 2020, 391), but IEA's Wagemaker argues that the phenomenon itself preceded PISA: "Lehmann (2011) describes the "TIMSS shock" that occurred after release of the TIMSS results in Germany in the winter of 1996/1997" (Wagemaker 2013, 20).

Producers like Schleicher and Wagemaker, as well as users like Strietholt agree that ILSA regularity begets policy relevance. For PISA, the link between regularity and policy relevance seems self-evident: "OECD/PISA will produce policy-oriented and internationally comparable indicators of student achievement on a regular and timely basis." (Schleicher 1999, 7) The predictability of the PISA machinery makes it media-friendly, amplifying its visibility: "The results of the OECD assessments, to be published every three years along with other indicators of education systems, will allow national policy makers to compare the performance of their education systems with those of other countries" (Schleicher 1999, 7).

For a user like Strietholt, this represents a clear shift from the IEA studies from the 1960s to the 1980s, which "were not particularly successful at answering questions regarding the determinants of educational achievement, and the causal mechanisms involved. The primary reason for this was that the studies conducted were cross-sectional surveys, and such designs do not easily support causal inference" (Strietholt et al 2014, 12). Clearly, for this new stage in ILSAs' development, policy-relevance is not just about methodological rigor, but also about visibility in the media and among political elites; for this, predictable regularity is an asset.

As for proponents/users, the paradigmatic case is the team composed by Eric Hanushek and Ludger Woessmann. They followed in the steps of Barro and Lee (1994), who from a human capital theory perspective tried to predict GDP growth across countries, based on education data. Crucially, Barro and Lee argued that it was necessary to shift from data on education quantity, i.e., educational attainment measured in years of schooling, to education quality, i.e., educational achievement measured by assessments of learning outcomes which, for this purpose, had to be internationally comparable. As a result, "the idea of assessing what students know as opposed to how long they have been in school has diffused around the world, in part at the instigation of international development and aid agencies" (Hanushek and Woessmann 2014, 145).

Perhaps deflecting accusations of ILSAs promoting a neoliberal agenda (Smith 2016), they emphasize that ILSA data, "developed and refined over the past half century, were not collected to support any specific economic research agenda", but "have made it possible for economists to address a range of fundamental questions that previously resisted satisfactory analysis" (Hanushek and Woessmann 2014, 147). However, their work often highlights, for instance, the benefits of privatization. In response, a particularly direct critique, the article A new global policy regime founded on invalid statistics? Hanushek, Woessmann, PISA, and economic growth (Komatsu and Rappleye 2017) using Hanushek and Woessmann's own dataset, argues that their claims that ILSA scores are predictive of future GDP growth are not borne out by the evidence. Komatsu and Rappleye (2017) do not question the use of ILSAs to study education policy issues; however, perhaps even more unsettlingly for Hanushek and Woessmann, their findings undermine the view that learning outcomes as measured by ILSAs predict GDP growth, the fundamental notion that underpins the application of human capital theory to ILSAs, in the work of proponents/users.

From a methodological perspective, ILSAs' credibility as policy-relevant faces two important threats, namely their observational and cross-sectional designs. The observational nature of ILSAs represents a substantial obstacle to causal inference. However, according to some users, ILSAs can benefit from "analytical approaches that allow valid causal inferences based on observational data" (Strietholt et al 2014, 13), including propensity score matching, instrumental variables, and country-level panel regression, e.g., to ameliorate effects of endogeneity or omitted variables. Moreover, while producers praise recurring ILSAs' predictability, some users go further: cyclical ILSAs "have a longitudinal component at the system level. Unlike previous cross-sectional comparisons, such longitudinal designs allow researchers to estimate causal effects of changes in educational policies and other factors at the system level" (Strietholt et al 2014, 11). Opponents disagree: "inferences can be made but only provide a cross-sectional perspective" (Rutkowski et al. 2020, 401).

Opponents

Proponents were distinguished by their role in the ILSA market, as either producers (supply), or users (demand). Meanwhile, opponents are subdivided according to the rationale for their opposition. Ontological opponents argue that ILSAs do not influence education policy, or at least, that no evidence warrants the claims of independent effects of ILSAs on education policies (Steiner-Khamsi 2019). Ethical opponents argue that, although ILSAs do influence policy, they should not. Some ethical opponents object either to the testing practices themselves or to the specific policies that ILSAs identify as "best practices" (Smith 2016; Meyer and Benavot 2013) these are ethical/ideological opponents. Meanwhile, other ethical opponents object to allegedly unwarranted inferences about education policy that are based on ILSA data stemming from weak designs (Rutkowski and Rutkowski 2016); these are ethical/methodological opponents.

William C. Smith, an ethical/ideological opponent, sees ILSAs as part of the global testing culture, marked by pervasive international, national, and school level assessments

(Smith 2016). While acknowledging ILSAs links to human rights and equity, he emphasizes their neoliberalism, in which “Accountability measures are used to measure the efficiency of government education investments” (Smith 2016, Kindle Locations 184-185). Similarly, Heinz-Dieter Meyer sees changing relationships between education, the state, and the economy: “The orientation of public education is changing from what used to be a focus on cultural and civic socialization, to preparation for the workforce” (Meyer and Benavot 2013, Kindle Locations 136-137). Like Smith, Meyer and Benavot foresee a dystopian future, all the more terrifying in that it cloaks itself in transparency: its reliance on statistics “does not mean that it is anchored in transparent, objective, uncontestable truth. In fact, the ‘cloud of data’ produced by PISA may easily operate like a Rorschach in which anyone can find support for any preconceived idea.” (Meyer and Benavot 2013, Kindle 324-328). Paradoxically, ethical/ideological opponents agree with proponent/producers like Schleicher and Wagemaker on one important point: both camps see ILSAs’ substantial effects on education systems but disagree on whether they are positive or negative.

While ideological opponents clash with proponents, methodological opponents are more moderate. Their main concern with the policy link is that ILSAs’ designs do not warrant the conclusions inferred from their data regarding the effectiveness of education policies. An aptly titled *Call for a more measured approach to reporting and interpreting PISA results* concludes: “under the very best circumstances, PISA, by definition, is limited” (Rutkowski and Rutkowski 2016, 256). They often focus on design limitations (e.g., observational, cross-sectional) that proponents/users try to ease with assumption-laden econometric models. For methodological opponents, “the control required for making causal inferences necessitates a research question that is focused, qualified, and limited in scope” (Rutkowski 2016, 24), which is not how omnibus ILSAs are designed.

Meanwhile, ontological opponents argue that the available evidence does not warrant the conclusion that ILSAs have an independent effect on the adoption of education policies. Some align themselves with sociological systems theory, emphasizing agency and local adaptation of global scripts in opposition to world culture theory. While acknowledging the correlations between participation in ILSA and policy adoption, these scholars question whether this demonstrates an independent effect of ILSAs on policies. In doing so, they flip the core question: "Rather than analyzing what PISA does to whom, it is also possible to reverse the object of study: what do national policy actors do with PISA? How do they use the test for their own reform agendas?" (Steiner-Khamsi 2019). Kijima and Lipscy (2017) attempt to answer this question, from a perspective that I have initially labelled as agnostic.

Agnostics

In a large *n* study, Kijima and Lipscy (2017) find an association between ILSA participation and increased net secondary enrollment. Crucially for Steiner-Khamsi's point, they explore the mediating role of ongoing reforms, because "If selection bias is an important problem, we would expect to see a large difference between reform and non-reform countries" (Kijima and Lipscy 2017, 19) in this association. Their findings are nuanced. First, they find larger coefficients in reform countries, where secondary enrollment is boosted by ILSA participation and preexisting reforms (Kijima and Lipscy 2017). This aligns with Steiner-Khamsi's suspicion that national governments' agency and domestic policy agendas play an important role in this process. However, ILSA participation is linked to increased secondary enrollment also in non-reform countries (Kijima and Lipscy 2017). This could indicate an independent effect of ILSA participation on secondary enrollment, unless the intention of reform already existed in the country, without having been formalized or publicized yet.

Finally, others label themselves “agnostic as to whether there really is one best way of teaching and learning and whether standardized tests produce better learning”, but also argue that “[i]t is not clear that the global testing regime is antithetical to progressive educational expansion” (Ramírez et al. 2018, 347). They examine policy outcomes in a large *n* study with a fixed effects regression model, concluding that ILSA participation predicts some policy outcomes, but not those alleged by ILSA critics. They also find that grade retention decreases with ILSA participation, contrary to their own expectations. Crucially for the sociology of quantification’s concern with methodological detail, they recommend: “pay more attention to how different types of tests are associated with various educational outcomes” (Ramírez et al. 2018, 358).

That is exactly what this study aims to do.

Chapter 2: Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition

2.1 Introduction

International large-scale assessments (ILSAs) emerged in the 1950s in developed countries (Pizmony-Levy 2013), in the 1990s experienced a global boom (Kamens and McNeely 2010), and a shift from research to policy (Wagemaker 2013). National education debates often cite ILSAs, which have led to policy changes (Green Saraisky 2015). For instance, "because it is conducted by a transnational agency, PISA is seen as disinterested and objective, and therefore trustworthy. Countries have more faith in PISA's judgment than they do in their own" (Gorur 2016, 605). However, ILSAs are not created equal because neither are the international organizations (IOs) behind them. How do IOs' different ideologies shape ILSAs' methodologies? Do different methodologies yield diverging policy-relevant data? This article follows Porter's (1996) statistical constructivism and Desrosières's (2015) sociology of quantification, focusing on a specific policy, namely, repetition and its (in)effectiveness. From this perspective, the achievement gap between repeaters and nonrepeaters can be simultaneously real and constructed. Crucially, different ILSAs construct it differently, with potentially policy-relevant consequences.

This article contributes to the ILSA literature by showing how different assessment designs, shaped by underlying IO doctrines, yield different data. This matters for two reasons: (1) globally, ILSA results may shape IOs' policy directives, reinforcing their doctrines; and (2) for countries, these diverging results potentially create perverse incentives to join ILSAs or make policy decisions. The study examines the case of grade repetition in Latin America, where it is as widespread as in sub-Saharan Africa (UIS 2012), using data from OECD's

PISA (Programme for International Student Assessment) and UNESCO's TERCE (Third Regional Comparative and Explanatory Study) for sixth grade.[3] These assessments differ in several design characteristics, but the focus here is on grade- versus age-based. This is the research question: Does the achievement gap between repeaters and nonrepeaters vary across different types of ILSAs? Specifically, do age-based ILSAs like PISA yield larger achievement gaps than those from grade-based ILSAs like TERCE?

Structure of the Article

The *theoretical framework* section shows how key ideas by Desrosières, Porter, and others guide this study. The *background* section comprises four subsections: (1) UNESCO and OECD's different mandates; (2) key design features of ILSAs they support, including TERCE and PISA, focus of subsequent quantitative analysis; (3) ILSAs' 1990s shift from research to policy, coupled with the introduction to this study's policy case, that is, repetition; and finally, (4) a discussion of evaluations of repetition's effectiveness, and their use of two research designs (grade- and age-based). The *data and methods* section describes the PISA and TERCE data sets; selected variables; geographic scope (Latin America); and analytical methods (regression).

Subsequently, the *results* from PISA and TERCE are compared for each country. The *discussion* examines differences between the assessments' results, and potential implications for policy, based on Desrosières's retroaction (2015). The *conclusions* tie everything together: from differences in IOs' doctrines and the designs of their ILSAs; to the divergent

[3] In Spanish: "Tercer Estudio Regional Comparativo y Explicativo." PISA's target population is closer to TIMSS eighth grade, but TERCE was chosen for several reasons. Crucially, TERCE asks students about repetition, unlike TIMSS. Moreover, Latin America's high repetition makes the policy issue relevant (Schiefelbein 1975) and enables quantitative analyses (Ikeda and Garcia 2014). Also, TERCE has stronger UNESCO ties than current IEA studies (Wagemaker 2013). Finally, regional assessments PASEC or SACMEQ were not analyzed because only TERCE reaches multiple PISA countries.

data presented here regarding repetition; to potential implications for policy directives in IOs and decisions in countries.

2.2 Theoretical Framework

What is statistical constructivism? "Let us suppose for the sake of argument that scientific investigation is able to yield true knowledge about objects and processes in the world. It must nonetheless do so through social processes. There is no other way" (Porter 1996, 11). Even while believing that ILSAs can shed light on education systems, we simultaneously concede that their work, despite aspiring to objectivity, stems from a contextualized social construction. This article focuses on a crucial element in that context: the IOs that support ILSAs.

Desrosières distinguishes two classical perspectives in the history of science: internalist, which is focused on knowledge itself, tools, and results, and conducted by each discipline's own specialists; and externalist, which is focused on social conditions, including institutions and relationships with public authorities, and conducted by sociologists and historians. But in the 1970s, a new approach merges the two perspectives: "the distinction between technical and social objects, underlying the separation between internal and external histories, disappears, and sociology studies all of these objects and networks at the same time" (Desrosières 2002, 12). This article takes that approach.

Consequently, the article visits the internalists' old province of technical objects: instruments such as ILSAs and their design; their findings, specifically the achievement gap between repeaters and nonrepeaters; and the resulting knowledge regarding repetition's effectiveness. Here the article builds on impact evaluation methodology (Allen et al. 2009). But it also enters the externalists' former turf, the networks: the institutions, namely, IOs backing ILSAs (Ydesen and Grek 2019) and the participating countries (Steiner-Khamsi and

Waldow 2012). Then, since "the relation between numbers and politics is reciprocal and mutually constitutive" (Rose 1999, 198), the article emphasizes the social construction of allegedly technical objects like ILSAs and eventually hypothesizes that their results can potentially close the feed-back loop, presenting perverse incentives to countries (Desrosières 2015) and reinforcing IOs' ideologies. While most research focuses on national high-stakes assessments that create perverse incentives internally (Lingard and Sellar 2013), this study asks whether ILSAs' allegedly low-stakes can similarly impact countries.

On the institutions' side, IOs working in education fall within two camps: those with explicit education mandates (UNESCO 1990) and those whose mandate, often economic in nature, expands into education (Ydesen and Grek 2019). The former include UNESCO; the latter, the World Bank and OECD. Not coincidentally, the former adopts a human rights-based approach (Jones 1990), which does not necessitate an empirical connection between education—seen as intrinsically valuable—and other fields, while the latter adheres to human capital theory (Sellar and Lingard 2013), which sees education as instrumental to economic development. This distinction is crucial. An explicit mandate allows UNESCO to dogmatically[4] focus on education and emphasize its intrinsic value as per the declaration of human rights (HR). Meanwhile, OECD expanded its mandate toward education by emphasizing its instrumental value through establishing theoretical and empirical links between education and the field of its original mandate, namely, the economy, as postulated by human capital theory (HCT).

How do institutions (IOs) connect to instruments (ILSAs)? These IOs endorse different doctrines about the value of education, namely, HR and HCT, which shape their

[4] Understanding dogma as a "set of principles laid down by an authority," in this case, the UN, "as incontrovertibly true." https://www.lexico.com/definition/dogma?locale=en

work with ILSAs. A "dogmatic" conception of education as a human right justifies its autonomy, thus enabling a measurement approach that follows the education system's own standards, such as grades and curricula, the so-called grammar of schooling.[5] Meanwhile, a theoretical framework that values education because of its empirical ties with other fields—in this case, its role as an input into the economy—opens the door to a measurement approach that imports categories from those other fields, such as HCT's notions of stock, flows, or working age population. Because different IOs have historically supported different ILSAs, the doctrines that underlie the IOs' views on education result in ILSAs with different designs.

Another distinction mirrors the internalist/externalist divide: "For [objectivists], objects exist, and it is up to science to identify their structures. For [relativists], the objects result from the shaping carried out by scientists: other shaping would lead to other objects" (Desrosières 2002, 19). For him, this distinction is overcome by focusing on "the work of objectification," the social process that hallows knowledge as objective. This is conditioned, but not determined, by the natural world: "interested human actors make science, but they cannot make it however they choose. They are constrained, though not absolutely, by what can be seen in nature" (Porter 1996, 11). As we will see, each ILSA shapes education systems' learning outcomes in slightly different ways, for instance, by selecting a certain target population, but options are finite.

The constraints on what is observed stem partly from the objectivists' "reality" but also from the ways in which science allows itself to see the world. This relates to Porter's (1996) mechanical objectivity: some methods are deemed legitimate, others are not. The

[5] "Regular structures and rules that organize the work of instruction (...) standardized organizational practices in dividing time and space, classifying students and allocating them to classrooms, and splintering knowledge into "subjects" (Tyack and Tobin 1994, 454).

scientifically and therefore socially acceptable criteria to define the target population for ILSAs are limited to grade, age, or some combination, in either schools or households. Subsequently, however, each ILSA proclaims its approach as objective (see debate; McGaw 2008a, 2008b; Wagemaker 2008). For the relationship between repetition and achievement, these different shaping processes are particularly impactful (Allen et al. 2009; Ikeda and Garcia 2014).

If they are mutually constitutive, how do numbers link back to politics? Just like IOs shape ILSAs, do these influence IOs and national governments? The concept of retroaction, or feedback, is enlightening here. "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart 1981, cited in Desrosières 2015). This study argues that, because of the reflexive nature of human beings, a correlation observed in the wild, not subject to attempts to manage it, may not be reproducible in captivity, when it is used for purposes of social engineering. Similarly, "when a measure becomes a target, it ceases to be a good measure" (Bird 2004, cited in Desrosières 2015, 341). Porter recounts a useful anecdote: in nineteenth century France, mortality per day was proposed as a measure to monitor hospitals; a detractor argued that this method "would permit a hospital to halve its mortality rate by doubling the period of confinement" (Porter 1996, 83). In other words, *trovata la legge*, *trovato l'inganno* (found the law, found the deception).

Methodology transcends itself when its outcomes are enshrined in policy-relevant indicators. These construct social reality by determining what is worthy of becoming a "target," thus guiding actors in their purported quest for social values. Therefore, official education statistics, including ILSA results, can potentially affect the behavior of actors in the education system, and even guide education policy. Indicator design, sometimes dismissed as merely technical, can shape policy decisions and, in an unsettling twist, provoke unintended consequences and provide perverse incentives. Actors may stick to an indicator's letter while ignoring its spirit, allowing its technicalities to conceal the underlying reality it was designed to reveal.

2.3 Background

IOs Are Not Created Equal

While the *E* in UNESCO stands for "education," the *E* in OECD stands for "economic." UNESCO's mandate involves "advancing, through the educational and scientific and cultural relations of the peoples of the world, the objectives of international peace and of the common welfare of mankind" (UNESCO 1990, 5–6). Two aspects stand out: its human-rights based approach; and its relative lack of interest in education's alleged economic impact. The former is rooted in its UN-issued mandate and framed by international commitments like the Declaration of Human Rights (UN 1948; for UNESCO's human-rights focus, see Jones 1990). Crucially, if education is a human right, there is no need to prove its impact on other spheres of human life; education becomes a goal in itself. Moreover, at its inception, UNESCO did not present education as driving economic development; whenever its constitution mentions the economy, it minimizes its importance: "a peace based exclusively upon the political and economic arrangements of governments would not be a peace which could secure the unanimous, lasting and sincere support of the peoples of the world" (UNESCO 1990, 5).

As for OECD, the deputy-secretary general of its precursor, the Organization for European Economic Cooperation (OEEC), once wrote: "the aim of the organization is purely economic" (Lintott 1949, 269). However, OECD claimed a stake in education in the 1960s, partly to outlive its initial mandate, the Marshall Plan (Ydesen and Grek 2019). Since the 1990s, OECD's involvement in education increased, framed by growing interest in its effects on economic growth: Barro and Lee's influential studies (1994) used years of schooling to measure education quantity and predict GDP growth; later Hanushek and Wößmann (2007) used learning assessment results to measure education quality, for similar purposes.

Ergo, ILSAs Are Not Created Equal

This section shows how UNESCO and OECD's mandates shaped ILSAs. IOs played key roles in ILSAs' beginnings in the 1950s and in their current boom starting in the mid-1990s. Since the 1950s, UNESCO spawned several outfits that produce assessments with similar characteristics. In 1958, IEA (International Association for the Evaluation of Educational Achievement) "emerged from a working group of scholars under the auspices of the UNESCO Institute for Education in Hamburg, Germany" (Pizmony-Levy 2013, 56). After multiple one-off ILSAs, since the 1990s IEA launched TIMSS (Trends in Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study); and UNESCO cofounded regional assessments SACMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality) and LLECE (Latin American Laboratory for the Evaluation of Educational Quality).[6] All of these ILSAs are framed by the "grammar" of schooling: schools, grades, and curricula. This seems obvious until we consider alternatives based on households, age-groups, and non-curriculum frameworks.[7]

[6] UNESCO also contributed to PASEC (Analysis Programme of the CONFEMEN Education Systems) (http://www.confemen.org/479/partenaires/).

[7] One outlier in UNESCO's assessment family is LAMP (Literacy Assessment and Monitoring Programme), a household-based assessment of adults' literacy and numeracy skills.

Unsurprisingly, given its mandate, OECD's direct involvement in ILSAs came later, in 1994, with IALS (International Adult Literacy Survey),[8] which assessed adults' (16–65) skills in their households, covering three domains: prose literacy, documents literacy, and quantitative literacy. Admittedly, UNESCO also supported IALS (Bieber 2016). In 2000, OECD introduced PISA, a hybrid of two assessment traditions: IEA's school-based assessments of students such as TIMSS; and OECD's own household-based assessments of adults, for example, IALS.

Symptomatic of PISA's roots in adult skills' measurement are its domains' names: reading literacy, mathematical literacy and scientific literacy; by contrast, TIMSS assesses mathematics and science achievement. But unlike IALS, which assesses adults in households, PISA assesses students in schools, like TIMSS. However, PISA's contents are not dictated by curriculum as in TIMSS but by the skills needed in everyday life, as in IALS (Wagemaker 2013). Moreover, PISA is age-based like IALS rather than grade-based like TIMSS.

At PISA's inception, OECD stated: "The focus will be on students approaching the end of secondary education" (Schleicher 1999); then why not target 18-year-olds? Because, unlike other school-based ILSAs, PISA transcends the "grammar of schooling." Instead, OECD's ILSAs, whether targeting adults in households or students in schools, form a system framed by HCT's preoccupation with stocks and flows. The working age population starts at 15 years old;[9] therefore, by targeting that age, PISA "provides a measure of human capital flows into economies" (Sellar and Lingard 2013, 193). Meanwhile, "drawing on earlier adult literacy and skills surveys [including IALS], PIAAC [Program for the International

[8] IALS, clearly inspired by the US's National Adult Literacy Study (NALS).
[9] See https://data.oecd.org/pop/working-age-population.htm .

Assessment of Adult Competencies] has been introduced to augment PISA by providing an assessment of human capital stocks among adult populations (16to 64-yearolds)" (Sellar and Lingard 2014, 925).

In summary, as per table 1, two IOs (UNESCO and OECD), with different mandates (educational vs. economic) and doctrines (HR vs. HCT) founded several ILSAs, including LLECE and PISA, respectively. These two ILSAs' designs differ regarding content (only LLECE is curriculum-based) and target population (defined by grade or age). This article focuses on how the latter difference affects ILSA results vis-a-vis the apparent effectiveness (or lack thereof) of a specific policy, namely, repetition.

TABLE 1
KEY CONCEPTS LINKING 'IOS' DOCTRINES AND ILSA DESIGNS, AS ILLUSTRATED BY THE CASES OF UNESCO'S LLECE AND OECD'S PISA

IO Cases	UNESCO	OECD
Initial mandate	Education	Economy
Education mandate	Explicit	Expansive
Value of education	Intrinsic	Instrumental
Doctrine	Human rights	Human capital theory
Design principles	"Grammar of schooling" (Tyack and Tobin 1994)	Human capital stocks and flows (Sellar and Lingard 2014)
ILSAs	SACMEQ 1995 LLECE 1997	IALS 1994 (stocks) PISA 2000 (flows) PIAAC 2012 (stocks)
Curriculum-based?	Yes	No
Comparison strategy?	Grade-based	Age-based
Specific grade/age?	Sixth grade End of primary	15 years old Start of working age population

NOTE.—IALS = International Adult Literacy Study; ILSA = international large-scale assessment; IO = international organization; LLECE = Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación; OECD = Organization for Economic Cooperation and Development; PIAAC = Programme for International Assessment of Adult Competencies; PISA = Programme for International Student Assessment; SACMEQ = Eastern and Southern African Consortium for the Measurement of Educational Quality; UNESCO = United Nations Education, Science, and Culture Organization.

From Research to Policy

This subsection discusses a crucial transformation in ILSAs' role vis-a-vis policy education debates and introduces the article's policy case. The 1990s mark a shift in ILSAs' role, from research to policy (Wagemaker 2013; Pizmony-Levy 2014), as they start having recurring cycles: the widespread expectation is that these assessments will return with updated rankings. Crucially, assessments also offer information on "what works," providing a blueprint for improvement between cycles (Gorur 2016). Thus, each ILSA turns into a device to monitor its own influence; a thermometer with the potential to become a thermostat. For instance: harshly critical of grade retention, PISA reports how this practice decreases among participating countries over successive cycles.[10] Generally, a repeater is defined as a student who "is not promoted to the next grade or does not complete an educational programme and (...) remains in the same grade the following school year" (UIS 2012, 17). This policy was controversial for decades before the current ILSA boom; according to the narrative policy framework (Schlaufer 2018), scholarly discussions often cast it as the "villain."

Historically, repetition is accused of four "crimes": inefficiency, ineffectiveness, inequity, and insensitivity. As early as 1967, OECD defined internal efficiency as "primarily concerned with (...) dropout, *repeats*, promotion from one grade to another, transition from one level to another" (emphasis added; Emmerij 1967, 19). This view of repetition as wastage soon reached Latin America (Schiefelbein 1975) and thrived at the World Bank (Psacharopoulos and Nguyen 1987), and to a lesser extent at UNESCO (Brimer and Pauli 1971). In the 1990s, repetition was deemed ineffective, as longitudinal studies failed to find lasting positive effects on achievement (Jimerson et al. 1997). By the 2000s, its inequitable nature was denounced: socially disadvantaged students were disproportionately retained, even after controlling for prior achievement (Lorence and Dworkin 2006). Recent studies

[10] See https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf.

cast repetition as insensitive, highlighting the social isolation of retained students by using network analysis (Demanet and Van Houtte 2016). This article focuses on how different ILSAs answer the effectiveness question differently. Its central tenet is that the answer will depend on each ILSA's design, specifically on whether it is grade- or age-based.

Target Population: Grade or Age?

Most school-based ILSAs—including IEA's TIMSS and PIRLS, and regional assessments LLECE, PASEC, PILNA (Pacific Island Literacy and Numeracy Assessment), SACMEQ and SEA-PLM (South-East Asia Primary Learning Metrics)—target grades. They are curriculum-based, typically sample entire classrooms, and collect information about classrooms and teachers (Wagemaker 2013). The alternative is targeting an age cohort, like 15-yearolds in PISA, which also stands out because it is not curriculum-based; here, students are not necessarily sampled with their classmates. OECD presents this population as more strictly comparable across countries, independent from national idiosyncrasies like the duration of school cycles (Pizmony-Levy 2013). The present comparison between age-based and grade-based assessments focuses on PISA and TERCE and tests this hypothesis: assessment design affects the achievement gap observed between repeaters and nonrepeaters. It shows how different designs compare different groups of repeaters and nonrepeaters. Those different comparisons, in turn, yield different estimates of the association between repetition and academic achievement. Therefore, the seemingly trivial choice of grade or age to define an assessment's target population is crucial when we examine repetition's apparent effect. Why? As seen below, the gap between repeaters and nonrepeaters stems from comparing two groups in a grade-based assessment and two different groups in an age-based assessment.

Grade-Based Assessments

In ILSAs like TERCE or TIMSS, grade is constant, but age varies. Figure 1*A* shows the age distribution of repeaters and nonrepeaters in TERCE: nonrepeaters' modal age is 12, as expected of sixth graders in most Latin American countries; meanwhile, repeaters' modal age is 13. Naturally, factors other than repetition affect the relationship between age and grade, for example, age of school entry. But typically, repeaters (treatment group) in grade-based assessments are older than nonrepeaters (control group).

Age-Based Assessments

This situation is reversed in age-based assessments like PISA. "In same-age comparisons (...) the retained students will be one grade behind their promoted classmates at the time of comparison" (Allen et al. 2009, 489). Figure 1*B* shows the grade distribution for repeaters and nonrepeaters in PISA 2015: the modal grade for nonrepeaters and for the whole population are the same. But for repeaters, the modal grade is the previous one, labeled 21. Again, other factors affect this relationship. But typically, repeaters (treatment group) in an age-based assessment are in lower grades than nonrepeaters (control group).

Should this asymmetry affect the observed achievement gap between repeaters and nonrepeaters? Once more, nonrepeaters outperform repeaters because of their different starting points. But here, repeaters lag behind literally: not only have they experienced repetition, but at the time of the assessment they have yet to reach the grade of their nonrepeating peers. This widens the gap between repeaters and nonrepeaters by compounding the effects of repetition with the difference in grade at the time of the assessment.

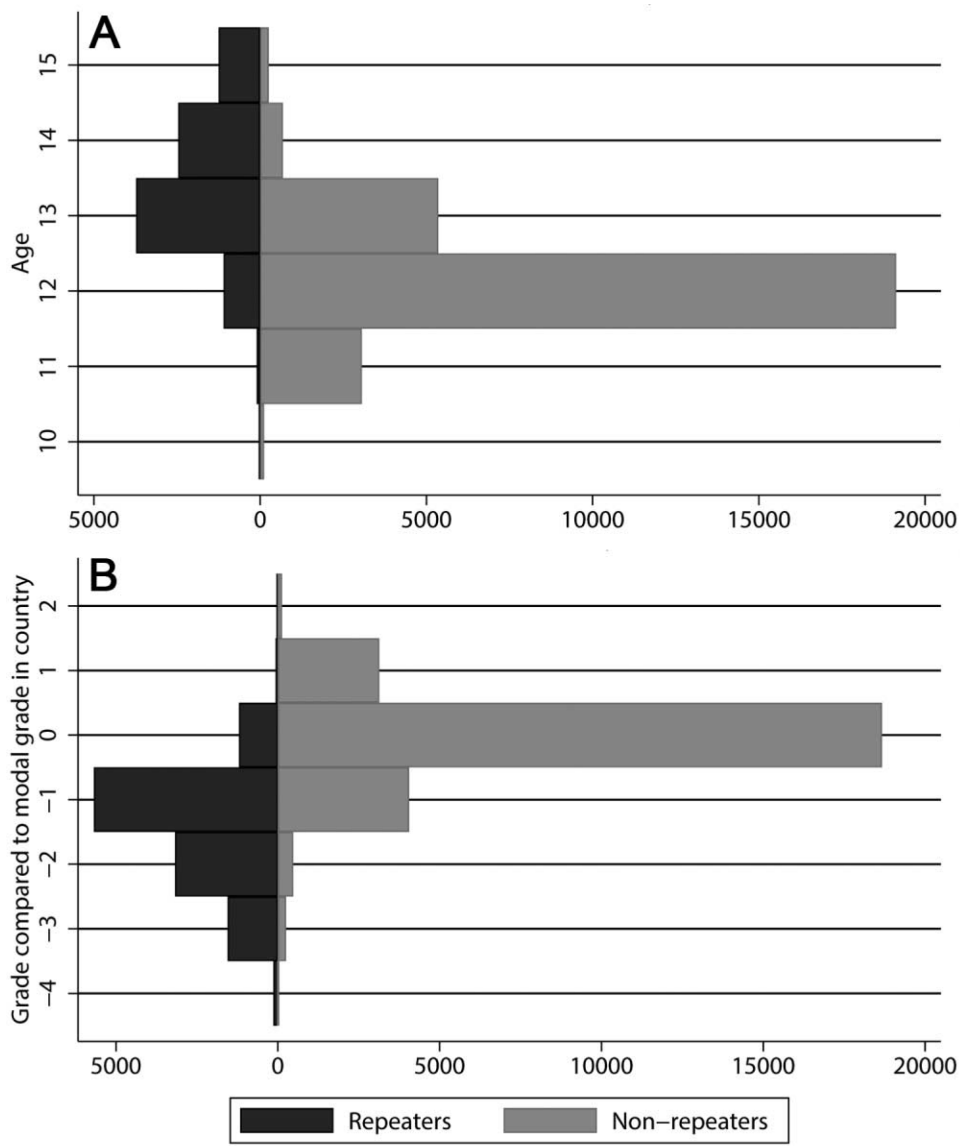


FIG. 1.—*A*, Repeaters and nonrepeaters in TERCE by age; *B*, Repeaters and nonrepeaters in PISA by grade (as compared to modal grade). SOURCES.—*A*, UNESCO; *B*, OECD.

2.4 Data and Method

This study analyzes public data from two ILSAs: PISA 2015 and TERCE. These are treated as units of analysis, alongside countries and individuals:

- PISA's 2015 round was its sixth, comprising 70 "education systems." It covers three domains: mathematics, reading, and science. PISA's sampling design involves two stages: after schools are selected, an age-based sample of 15-year-old students is drawn.
- TERCE (2013) was LLECE's third study, in 15 Latin American countries. TERCE assesses the same domains as PISA. It targets grades 3 and 6, assessing science only in grade 6. It uses a two-stage sampling design: from each selected school, TERCE samples entire classrooms. This study focuses on sixth-grade data.

TABLE 2
SAMPLE DESCRIPTIONS AND BASIC STATISTIC

	PISA (2015)				TERCE (2013)			
	Costa Rica	Dominican Republic	Mexico	Peru	Costa Rica	Dominican Republic	Mexico	Peru
N	6,866	4,740	7,568	6,971	3,520	3,669	3,622	4,801
% of girls	50.9	52.2	50.3	49.6	50.0	50.1	48.8	49.4
SES (mean)	-.79	-.88	-1.15	-1.06	.58	-.18	.29	-.35
% of repeaters	15.7	24.4	10.1	15.8	19.3	28.2	10.9	20.9
Plausible values								
Math (mean)	400	333	413	387	741	625	780	718
Math (SD)	67	69	76	81	83	58	109	109
Reading (mean)	427	363	429	398	766	637	748	699
Reading (SD)	78	83	78	88	83	78	103	107
Science (mean)	420	335	421	398	769	636	743	698
Science (SD)	69	73	71	76	91	74	97	100
	Lower Secondary				Primary			
Net enrollment rate	76.9	54.0	88.8	72.9	97.8	90.8	97.6	93.7
Completion rate	66.6	82.8	86.3	87.3	94.5	91.5	96.4	95.7

NOTE.—Results in original scales. PISA's SES scale: mean = 0; SD = 1 for OECD countries. https://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-16-Procedures-and-Construct-Validation-of-Context-Questionnaire-Data.pdf. TERCE's SES scale: mean = 0; SD = 1. https://unesdoc.unesco.org/ark:/48223/pf0000243533. PISA plausible values: mean = 500; SDs = 100 for OECD countries. https://www.oecd.org/pisa/data/PISA-2015-Technical-Report-Chapter-12-scaling.pdf. TERCE plausible values: mean = 700; SD = 100. https://unesdoc.unesco.org/ark:/48223/pf0000243532. Net enrollment rates and completion rates based on household survey data (2014). http://data.uis.unesco.org/#.

SOURCES.—For PISA, OECD; for TERCE, UNESCO.

Geographic Scope

Nine countries participated in both PISA 2015 and TERCE. This study focuses on four: Costa Rica, Dominican Republic, Mexico, and Peru. Other countries were excluded because of their sampling design.[11] Table 2 provides education indicators, sample sizes, and descriptive statistics for the "treatment" variable (repetition in primary); the outcomes: achievement in mathematics, reading and science; and key controls,[12] namely, gender and socioeconomic status (SES), ubiquitous in previous studies (Willms and Somer 2001; Gómez Vera 2013; Goos et al. 2013; Ikeda and Garcia 2014).

Analytical Approach

This study uses ordinary least squares (OLS) regression models, estimated separately for each of the three learning domains in each of the four countries and finally in each of the two ILSAs, namely, PISA 2015 and TERCE. This yields 24 models that adopt this form:

$$Achievement_{CDI} = \beta_{0CDI} + \beta_{1CDI}Non-repeater + \beta_{2CDI}SES + \beta_{3CDI}Girl + \varepsilon$$

where subindex C denotes country (one of four); D, learning domain (one of three); and I, ILSA (either PISA or TERCE). The hypothesis is that, for each country and domain, the achievement gap between repeaters and nonrepeaters (represented by b1) will be different in PISA and TERCE:

$$H_0: \beta_{1CDPisa} = \beta_{1CDTerce}$$

$$H_A: \beta_{1CDPisa} \neq \beta_{1CDTerce}$$

[11] In four countries there was at least one stratum with only one school. This peculiarity of PISA's sampling design prevented analyses using balanced repeated replication (BRR), thus excluding Brazil, Chile, Colombia, and Uruguay. Argentina's PISA data, released separately, are not analyzed here.

[12] The number of controls was deliberately kept small to reduce the impact of differences in measurement of controls across the assessments. Still, cross-assessment SES comparisons are challenging (Sandoval Hernandez et al. 2019).

That difference's statistical significance is tested by computing *z* statistics (α=.05). Since each assessment uses different metrics, the PISA socioeconomic index and plausible values[13] for achievement were converted to TERCE scales.[14] The dichotomous (or dummy) variables, namely, repetition and gender, were not transformed.

2.5 Results

This section presents each assessment's findings, deceivingly like the results of any quantitative study. However, the following section's discussion reframes these numbers as not merely the objective, inevitable result of a "technical object" but as a socially constructed product where "other shaping would lead to other objects" (Desrosières 2002). Meanwhile, table 3 shows ordinary least squares (OLS) coefficients from 24 multiple linear regression models, covering four countries (Costa Rica, Dominican Republic, Mexico, and Peru), three academic domains (mathematics, reading, and science), and two assessments (PISA 2015 and TERCE). To answer the research question, the analysis compares assessments on the relationship between repetition and achievement, while holding country and domain constant, based on this hypothesis: regression coefficients for achievement on repetition are significantly higher for PISA than for TERCE in each country/domain combination. However, one should not expect a similar pattern for the controls, namely, gender and SES.

SES predicts achievement in all domains. For instance, in Costa Rica, a single-unit increase in SES is associated with a 27-point increase in mathematics achievement in PISA

[13] The *Stata* procedure used, *mi estimate: svy: reg*, estimates regression models that (*a*) treat plausible values as multiple imputations and (*b*) account for the surveys' complex sampling design.
[14] To enable comparisons across the two assessments' regression models, each set of PISA PVs within each country and domain was linearly transformed to match its distribution parameters (mean and SD) to those of TERCE's PVs for that country/domain. Similarly, in each country, PISA's SES index was linearly transformed to match TERCE's.

and a 33-point increase in TERCE. However, the difference between coefficients is not significant, according to the z statistic (a p :05). This pattern repeats for all three assessment domains for Costa Rica and Dominican Republic. However, the differences between SES-achievement coefficients in PISA and TERCE are significant in Mexico and Peru, in all assessment domains. In summary, there is no systematic pattern of differences between assessments for the relationship between SES and achievement.

In mathematics, boys outperform girls in all countries in both assessments, except in PISA in the Dominican Republic. When comparing assessments, differences in mathematics gender gaps between PISA and TERCE are not statistically significant, again excluding Dominican Republic.

Finally, in science, gender patterns are nuanced. In Costa Rica and Peru, boys outperform girls in both PISA and TERCE. In Mexico, the boys' advantage is statistically significant in PISA but not in TERCE. Finally, in the Dominican Republic, boys outperform girls in PISA, while the opposite is true of TERCE; however, neither difference is statistically significant. When comparing assessments, the boys' advantage is generally larger in PISA than in TERCE, but the difference is only statistically significant in Costa Rica.

Before moving on to repetition, let us summarize the differences between assessments. For the SES slopes, the differences between PISA and TERCE were statistically significant in two countries out of four. For the gender gaps: in reading, the differences between assessments were statistically significant in two countries out of four; and in mathematics and science, only one difference was significant. In summary, although analyses of SES slopes and gender gaps show occasional discrepancies between PISA and TERCE, there is no clear, systematic pattern of statistically significant differences. As we will see, the situation changes with repetition.

TABLE 3
REGRESSION COEFFICIENTS FOR OLS MODELS

	Mathematics			Reading			Science		
	PISA	TERCE	z	PISA	TERCE	z	PISA	TERCE	z
Costa Rica:									
Nonrepeater	62.17	29.41	5.53	68.42	37.05	4.74	70.33	38.85	5.08
SES	27.12	33.16	-1.75	31.99	34.49	-.72	33.00	34.08	-.28
Girl	-22.95	-15.68	-1.45	14.78	7.19	1.59	-25.54	-11.85	-2.85
Dominican Republic:									
Nonrepeater	37.29	19.63	4.43	59.77	26.52	6.03	48.98	16.83	6.63
SES	21.51	25.89	-1.30	31.03	38.86	-1.76	30.46	35.01	-1.05
Girl	.05	-10.91	3.08	22.81	11.72	2.10	-5.74	2.71	-1.61
Mexico:									
Nonrepeater	82.07	49.64	3.29	79.85	47.13	3.64	73.37	36.88	4.20
SES	24.26	40.59	-4.68	27.47	44.45	-5.13	24.31	38.38	-4.07
Girl	-12.63	-16.62	.74	17.74	7.68	2.21	-12.25	-4.28	-1.68
Peru:									
Nonrepeater	64.70	48.17	2.21	63.89	49.56	2.19	57.92	43.18	2.36
SES	37.70	45.71	-2.19	41.54	48.77	-2.36	36.15	42.36	-2.07
Girl	-12.02	-22.13	1.91	8.93	1.45	1.47	-11.89	-8.31	-.82

NOTE.—Bold font indicates statistical significance (a = :05). Intercepts not presented. SOURCES.—For PISA, OECD; for TERCE, UNESCO.

2.6 Discussion

The preceding results show that, even within the boundaries of legitimate scientific methods, the relativists' view holds: other shaping leads to other objects. Why? This section tackles that question and explains how these findings may matter for policy. In a nutshell, the divergent results across these assessments are a methodological artifact. But let us remember: "reality is constructed from artifice" (Porter 1996, 5). The discrepancy was expected given the "comparison strategy" issue, but it is more than a subtlety of research methodology. After all, US studies on repetition's impact adopt longitudinal designs and apply techniques that ameliorate the impact of endogeneity and the "comparison strategy." But when this methodological artifact encounters the oversimplifying design of ILSAs coupled with the amplifying power of their media presence and political influence, it becomes something altogether different.

ILSAs are cross-sectional and not specifically designed to evaluate the effectiveness of any policy (Rutkowski 2016), especially one as challenging as repetition. In a show of transparency, ILSAs share data sets online with researchers who may disregard how data are produced and how this affects results. ILSAs themselves lack incentives to emphasize their own limitations and recommend caution in interpreting results. They sometimes flaunt their ability to appraise the effectiveness of education policies and influence them; that is one reason why countries participate (Addey et al. 2017). This combination of factors may lead to naïve comparisons between the results of repeaters and nonrepeaters (Gomez Vera 2013), which ignore both endogeneity and comparison strategy issues. This is crucial because ILSA results often become official statistics in participating countries.

When that happens, ILSA findings acquire important policy implications. Should Latin American countries continue to implement repetition? According to these results, the answer depends on which ILSA is consulted.[15] Or which ILSA should a country join? The answer may depend on the country's stance on repetition.

Countries strive to "perform" in ILSA rankings, thus improving their standing vis-a-vis their peers. This may provide perverse incentives for countries to adopt policies to improve their educational performance as measured by the indicators prioritized by IOs. Or conversely, could this incentivize a government to join an assessment to justify policies that it was already planning to implement? For instance, a government aiming to eliminate repetition and "borrow" a different policy (Steiner-Khamsi and Waldow 2012), for example, tracking, might join PISA to have an "evidence-based" rationale to justify this policy. This is feasible because "while numbers seem indispensable to politics, they also appear to

[15] The magnitude of repeaters' achievement gap varies according to comparison strategy. But then why do both TERCE and PISA, despite different comparison strategies, yield negative slopes for achievement on repetition? Because their cross-sectional design fails to address the endogeneity of grade retention: in Latin America, on average, repeaters have lower achievement than nonrepeaters before being retained. Consequently, the answer to the "objectivist" question of repetition's effectiveness cannot come from cross- sectional studies, like current ILSAs; but from longitudinal studies that consider learners' starting points.

depoliticize whole areas of political judgment" (Rose 1999, 198); and ILSAs, as seemingly external to national political systems, may appear particularly "depoliticizing." This may amount to "policy-based evidence" (Marmot 2004), a "tendency of policy makers to formulate, announce, and implement policy, then encourage research engagement that could retrospectively support the policy initiatives" (Yore et al. 2010, 597).

What would these perverse incentives look like? This study did not analyze countries' response to the "comparison strategy" phenomenon. However, they may be incentivized in two ways: all else equal, high-repetition countries may have incentives to join grade-based assessment like TERCE and low-repetition countries to join PISA;[16] or countries that already participate in ILSAs may modify their grade retention policy to move up in the rankings. For instance, PISA countries have incentives to reduce repetition, even without improving achievement. By contrast, TERCE countries have incentives to prevent low-performing students from reaching the assessment's target grade, potentially even increasing repetition rates. Whether countries lower, maintain or increase repetition to improve ILSA standings, this is a perverse incentive. Future research can test this hypothesis: repetition rates should decline faster in PISA countries than in TERCE or TIMSS countries. Following Desrosières, when a measure (e.g., the average mathematics score in PISA or TERCE) becomes a target, it ceases to be a good measure.

Implications for IOs' doctrines should be further explored. For HCT, PISA's larger gap reinforces traditional discourses about repetition as ineffective and inefficient (Emmerij 1967). For the HR approach, TERCE's narrower gap minimizes differences between

[16] However, countries undergoing reforms may join the ILSA that rewards the type of country they aim to become, e.g., a high-repetition country aiming to lower repetition may join PISA. Moreover, for these countries, becoming a "top improver" from one cycle to the next may be a more realistic goal than becoming a "top performer."

repeaters and nonrepeaters, thus downplaying effectiveness and, particularly, efficiency concerns, dismissed by this doctrine as part of the production function approach to education. This creates instead a space for an allegedly universalistic, humanistic, and holistic conception of learners, which transcends viewing them as (human) capital. Would this approach be more receptive to criticisms focused on repetition's alleged inequity and insensitivity?

2.7 Conclusion

IOs with different mandates embrace different educational doctrines: UNESCO's explicit education mandate leads to a focus on education's intrinsic value as a human right; OECD's expansive mandate from the economy into education leads to an emphasis on education's instrumental value in producing human capital. These IOs' diverging education doctrines shape the design of the ILSAs they support: UNESCO, by supporting IEA in the 1950s and regional assessments in the 1990s, directly or indirectly shaped TIMSS, PIRLS, LLECE, and SACMEQ; consistently with UNESCO's view of education as intrinsically valuable and therefore autonomous from the economy, these ILSAs follow "the grammar of schooling": schools, grades, curricula. Meanwhile, OECD approached ILSAs through adult literacy, with IALS in the 1990s, focused on human capital stocks, administered in households and oblivious to curricula or grades; in 2000, its PISA, focused on human capital flows into the economy, is less constrained by the "grammar of schooling" than other school-based assessments: neither curriculum- nor grade-based, it is age-based instead, like IALS and later PIAAC.

Subsequently, the difference in assessment design known as the "comparison strategy" leads to differences in data: specifically, for four Latin American countries, in age-based PISA, the achievement gap between repeaters and nonrepeaters doubles that of grade-based TERCE; this difference between the assessments, unlike others, is systematic, statistically significant, and substantial. Undoubtedly, this study has limitations. Four countries and two control variables in a single-level analysis of two ILSAs is only the beginning of this line of research. Confounders like curriculum orientation and target population specifics (PISA's students are older than TERCE's) should be examined. Still, these findings enable the next stages of inquiry into this field.

Diverging data emerging from different ILSA designs can potentially shape IOs' policy directives, influence countries' policy decisions, and ultimately, coming full circle, reinforce IOs' education doctrines. As ILSA results become official statistics, countries are incentivized to follow the letter of indicators rather than their spirit, thus acquitting themselves on technicalities rather than pursuing the underlying social values. This study identifies incentives without claiming, yet, that countries engage in the expected behavior; this should be explored.

Globally, IO doctrine links to ILSA design, data, and back to doctrine. Findings from HCT-inspired ILSAs like PISA reinforce this doctrine' historically critical views on repetition; meanwhile, ILSAs informed by HR approaches, like TERCE, are more forgiving of repetition's shortcomings. This can be "neither coincidence nor conspiracy, but culture," to borrow Garfinkle's (2015, 30) words. It is too systematic to be dismissed as random. Conversely, conspiratorial interpretations underestimate how much organizational and professional cultures shape how we learn about the world.

Chapter 3: When ILSAs met policy: Evolving discourses on grade repetition

3.1 Introduction

International large-scale assessments (ILSAs) are highly controversial, but both their advocates (Schleicher 1999) and critics (Smith 2016) argue they substantially influence education policy. While evidence of this influence remains elusive, recurring ILSAs like PASEC, PISA and TIMSS have produced substantial discourse on education policies since the mid 1990s (Wagemaker 2013), while educational effectiveness research (EER) scholars saluted ILSAs as they elevated that agenda to a global level (Strietholt et al. 2014; Hanushek and Woessmann 2014). Research on discourse from and about ILSAs is gaining momentum (Pizmony-Levy 2018; Pizmony-Levy and Torney-Purta 2018; Baroutsis and Lingard 2021), but it does not yet tackle specific policies. That said, grade retention stands out, since the mid 2000s, among policies that have drawn ILSAs' attention, as will be seen.

Grade retention is virtually universal, except in a few countries including Japan and Norway; however, repetition rates have been decreasing around the world (UIS 2012). Highly controversial since at least the 1970s (Jackson 1975), it has been accused of lacking efficiency (Eide and Goldhaber 2005), effectiveness (Jimerson 1999), equity (Schielfelbein 1979), empathy (Byrnes and Yamamoto 1985) and evaluability (Allen et al. 2009), while promoting expulsion (Andrew 2014). These debates pervade not only academia but also international organizations (IOs) such as the World Bank (see Haddad 1979), OECD (see Emmerij 1967) and UNESCO (see Brophy 2006). As will be seen, some controversies stem from challenges posed by attempts to evaluate its impact. While literature on grade repetition is plentiful, including some that uses ILSA data (Willms and Somer 2001; Ikeda and Garcia 2014; Gomez Vera 2013), research on ILSAs' discourse around it is lacking.

The gap this study aims to address is the intersection of those two spaces: on the one hand, the long tradition of research on grade repetition, and discourse around it; and on the other, fast-growing research on discourse emanating from ILSAs. Framed by the sociology of quantification, the study examines the discourse of OECD's PISA and UNESCO's ERCE[17] on grade retention, as presented in their publications, specifically, their results reports. It builds on research showing that these ILSAs' designs differ due to their respective parent IOs' educational doctrines, and consequently yield divergent data regarding grade retention's effectiveness, with PISA producing more unfavorable results (Cardoso 2020).

The present study contributes to the field by exploring whether all these elements, namely IOs' education doctrines and ILSAs' design and data, jointly shape these ILSAs' policy discourse on grade retention. The research questions are: (1) What are the similarities and differences between discourses on grade repetition in OECD's PISA and UNESCO's ERCE publications from 2007 to 2017? (2) How has each ILSA's discourse changed over time? and (3) How do these similarities and differences in policy discourse relate to IOs' doctrines and ILSAs' designs and data?

Structure of the Article

The *Theoretical framework* briefly presents the sociology of quantification and introduces the notions of genre, ordinalization and scientization. The *Background* revolves around critiques of grade retention. The *Data and methods* section specifies the corpus (ERCE and PISA results reports); chronological scope (2007-2017); and analytical methods (genre analysis and comparative case studies). The *Findings* are presented in four subsections: increased prominence of grade retention; genre (move-step) analysis and

[17] Linked to UNESCO/LLECE, and known in the past, in successive cycles, as PERCE, SERCE or TERCE.

quantitative (horizontal and transversal) comparisons; analysis of references in the Justification and Critique; and analyses of charts on grade repetition's effectiveness. The *Discussion* builds on the answers to the first two questions to elucidate the third. The *Conclusions* tie it all together: IOs' doctrines with ILSAs' designs, data, and discourse.

3.2 Theoretical framework

This section presents the sociology of quantification to then focus on the notions of genre, ordinalization and scientization. The sociology of quantification is linked to the sociologies of knowledge (Espeland and Stevens 2008) and expertise (Eyal 2019). It emphasizes the production of official statistics (Desrosières 2010; 2015), thus intersecting substantially with politics and policy. (For an application to ILSAs and IOs, see Cardoso 2020). Historically, two perspectives had been running on parallel tracks: the "internalist" perspectives on so-called hard science, focused on methods and driven by the practitioners themselves; and the "externalist" perspectives, focused on institutions and driven by historians and sociologists (Desrosières 2010). In the 1970s, the field transcended that distinction, drawing the necessary connections between institutions and methods.

Simultaneously, Desrosières' acknowledges the challenge in his ambition: "it is hard to think at the same time that the measured objects exist, and that this is nothing but a convention"[18]. This theoretical and methodological framework hones in on methods as its subject. I will approach these, however, not from a methodological perspective, but instead from a sociological perspective: "methods are seen as socially configured, but also socially consequential insofar as methods 'do things' to make their objects of study visible,

[18] "il est difficile de penser en même temps que les objets mesurés existent bien et que cela n'est qu'une convention" (Desrosières 2010, 8).

intelligible, and even possible to act upon or intervene-able – that is, methods are performative" (Williamson 2018, Kindle Edition).

The methods of interest here are of two types: ILSAs design features, and particularly two of them, namely their cross-sectional nature and the comparison strategy (Allen et al. 2009); and the rhetorical devices used in ILSAs' discourse on grade retention, as seen in their results reports. ILSA design features are shaped by IOs doctrines, and shape ILSA data on grade retention (Cardoso 2020). Do ILSA design features, which shape their data on grade retention, also shape their discourse on this practice? Now, considering methods in isolation would defeat the purpose, since the sociology of quantification transcends the distinction between the traditional internalist perspective focused solely on ILSA methods and the equally traditional externalist view focused on social institutions; the two perspectives must be mutually complemented and integrated, which brings me to the next question.

How are these methods socially configured? Which social institutions and organizations shape ILSAs' designs? World society theorist Chabbott (2003) analyzes international development as a discursive, organizational, and professional field, where global scripts on education are disseminated, and multilateral and bilateral organizations contribute to the dissemination of global scripts by providing financial aid and technical assistance to countries. However, "IGOs and bilateral organizations also compete to establish preferred models within sectors that better fit their own national experience or the structure of the assistance they are able to provide." (Chabott 2003, Kindle Locations 2012-2013).

Therefore, even from a WST perspective, IOs are not a monolithic bloc operating in a fully coordinated manner, partly because they play different roles in a global division of labor marked by diverging mandates. While world society theorists "see education, simultaneously human capital and a human right, is constructed as a means to fulfill both visions of development" (Chabott 2003, Kindle Location 451); these twin views of education as human capital and human right are not equally distributed across IOs. There is nuance in how various IOs perform these scripts, e.g., in dealing with trade-offs, real or imagined, between efficiency and equity. This study explores how IO's positions on this continuum between education as human right or human capital matter for their ILSAs' discourse.

Genre as persuasive strategy

IOs and ILSAs do not just quantify education phenomena, they present them to the public in structured narratives, e.g., in publications, through which they aim to influence policy debates. Genre theory provides a framework to reveal these publications' underlying structure, which, as will be seen, is one of the sources of their power to persuade: "A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and therefore constitute the rationale for the genre" (Swales 1990, 58).

While Swales (1990) applies genre analysis to academic uses of English, others apply it to ILSAs' policy discourse: "Publications within the genre share a common communicative purpose, namely: the identification of education 'best practices' (i.e. 'what works') to support policy transfer" (Auld and Morris 2016, 203). The 1990s' shift in ILSAs' focus "from research to policy […] and educational quality and accountability" (Pizmony-Levy 2014, 321) demanded a reorientation of this particular use of genre analysis: "Given the move from research to advocacy, we turned to political science to aid interpretation of the strategies, focusing on the role of narratives in persuasion" (Auld and Morris 2016, 203).

Following the advocacy coalition framework (see Sabatier & Weible 2007), Auld and Morris (2016), distinguish between *deep core*, *policy core*, and *secondary beliefs*. These scholars focus on the *policy core*, in their case "generalisable causal relationships, located within school systems' practices and structures"; meanwhile, this study tackles *secondary beliefs*, "represented in advocates' specific policy preferences" (Auld and Morris 2016, 204), in particular, on grade repetition, as a tradition of discourse on this policy, briefly reviewed in the *Background* section, is reappropriated by ILSAs' policy-focused results reports. Auld and Morris (2016) expect secondary beliefs to vary substantially across actors.

Ordinalization as a rhetorical device

For the sociology of quantification, the specific devices through which ILSAs convey their results to the media, and through them, to the public, are crucial. Among them, one stands out: "Rankings are a compelling example of accountability measures both because they are so common in contemporary society and because their precise comparisons generate intense competition among those being evaluated, a competition that makes the rankings' effects easier to see" (Espeland et al. 2016, 4). In fact, rankings disregard within-country variation while creating artificial differences between countries. Two countries' average scores may not be statistically different from each other; but rankings sort all countries. Consequently, meaningless changes across ILSA cycles become sensational media stories about one top performer overtaking another.

However, not all ILSAs issue rankings: "in the first three decades of its existence, IEA was oriented toward education research and avoided the publications of ranking tables and "horse race" discourse more generally" (Pizmony-Levy 2013, 161). In the context of ILSAs' aforementioned shift from research to policy (Pizmony-Levy 2014; Wagemaker 2013), however, this changed, at least for some ILSAs, although not necessarily the emerging regional ones. "Since the mid-1990s, as the field transformed and reoriented itself toward educational audit and accountability, reports published by IEA and OECD (PISA) have included a variety of ranking tables and semi-report cards" (Pizmony-Levy 2013, 69).

Even PISA itself avoided rankings in 2018 for its controversial global competency domain that many OECD countries, including the US and the UK, opted out of. Why? Perhaps because "ordinalization is ideally sustained by the inclusion of the largest population possible; the more comprehensive the scale, the more significant and legitimate in terms of objectivity, rational judgment, and as a policy tool" (Sorensen and Robertson 2020, 33). Similarly, TALIS[19], linked to PISA in its 2013 cycle, also omitted rankings: "the drive toward ordinalization in TALIS has remained muted so far" (Sorensen and Robertson 2020, 23); perhaps due to the involvement of Education International, a global federation of teachers' unions.

As a method, how are rankings socially configured? Specifically, how do they align with IOs and their education doctrines? "[O]rdinalization suggests a promise of movement and social mobility, revolving around a human capital view of the world" (Sorensen and Robertson 2020, 28). Therefore, OECD embraces rankings, unlike UNESCO. And how are these methods socially consequential? Some signs point to rankings as a factor that increases media coverage:

> *With respect to content, research shows that media outlets focus their attention on ranking tables and evoke casual explanations for ranking and policy implications, even if these are not fully supported by the official reports of international assessments* (Pizmony-Levy 2013, 132).

First, rankings showcase top performers as suitable models for other countries, thus increasing the prominence of ILSA-related discourse. Second, rankings give every participating country a horse in the race; Olympic podia, best-seller lists and top-forty radio

[19] Teaching and Learning International Survey.

shows have primed the public for this (Fourcade 2016). Finally, movements up and down the ranking, while often statistically nonsignificant, give the media a story that laypeople relate to. In other words: "The use of ILSA data as evidence for policy is often linked to the performance of reference societies, the relative positions of countries in the rankings and the improving or declining ranking or performance of a given country" (Addey et al. 2017, 438).

The scientization of policy discourse

Scientization can be understood, for instance, as "an explicit emphasis on the ultimate value of the rational analysis of all physical and social phenomena (…) [e.g.] indicated by an increase in scientific publications" (Zapp 2017, 3). The scientization of politics, while not new (Steiner-Khamsi et al. 2020), has lately received considerable attention in the field of comparative education, partly because of PISA's advent. It is accompanied by the politicization of science in "a sort of recursive process of mutual support, mutual shoring up, but also mutual pollution and collateral damage, between state agencies and scientific institutions" (Eyal 2019, 97).

In these disputes, some territories are more contested than others: "The middle lane of regulatory science is thus contentious and crisis-prone precisely because it serves as the interface between scientific research, law and policy" (Eyal 2019, 8). I argue here that education policy discourse occupies a similarly disputed space, and that ILSAs, partly because, as alleged outsiders, they enjoy a special status in these debates, reinforced by a mix of institutional and discursive factors: "Policymakers and government officials often find the assumed objectivity of ILSAs useful in pushing forward reforms" (Gorur 2017, 343).

Therefore, understanding how ILSAs convey this objectivity, especially on policy matters, becomes crucial. I will focus here on scientized policy discourse, which plays a role in the balance of power among different groups involved in policy debates. Particularly, if scientized, so-called “empirical” models of education prevail, practitioners like teachers may see their autonomy curtailed and their professional status diminished. This is not lost on the likes of Education International (EI), which commissioned the aptly titled report *Alternative models for analysing and representing countries' performance in PISA* (Mortimore, 2009).

What does this scientized policy discourse look like? The importance of references in scientific discourse -or discourse that adopts its trappings to achieve legitimacy- cannot be overstated: “a reference is validation of evidence. (…) [R]eferences are used to provide authoritative status to the evidence presented in policy documents” (Steiner-Khamsi 2021, 2). Bibliometric network analyses reveal, for instance, how PISA looms large in some countries' policy documents, as citers seek to legitimate their own discourse (Baek 2020). As for PISA's own legitimacy, it stems from the IO behind it, from its own data, but also from the references it cites; the same applies to ERCE.

While references make the connection with allegedly established academic knowledge, it is equally important to examine how publications introduce new evidence. Some sociologists have noted -and decried- the scientization of discourse via certain analytical techniques and chart types: “a quasi-natural aura of scientific realism through the compulsory use of such textual formats as the abstract and the literature review (…), the regression table and the scatter plot” (Kemple and Mawani 2009, 235). Their initial use in the natural sciences (Porter 2010) lends a veneer of scientific credibility to its more recent use in sociology and comparative education.

Finally, academic research in social sciences may look different across geographic and cultural contexts. Nonquantitative articles in US and Mexican journals differ in their epistemological approaches (Abend 2006), although one wonders what the differences might have been had the comparison focused on quantitative articles instead. Nonetheless, this issue matters because differences in social science discourse across cultures may shape scientized policy discourse as well. ILSA reports may differ from each other in their epistemology and rhetoric not necessarily because of variations in the degree of their scientization, but because of cultural differences in the scientific models they intend to mimic.

3.3 Background: academic discourse on grade retention

Grade retention has been intensely debated over the years, approached differently by several IOs, and studied through different designs, both in the US and internationally, with diverging results. Partly because it is widespread in Latin America, its scholars have long played a central role in these debates, both in Spanish and English (Schiefelbein 1975; Reimers 2000). Repetition is conceptualized differently by educational doctrines like human rights and human capital theory (Cardoso 2020), and by the IOs that embrace them, e.g., UNESCO (Coleman and Jones 2004) and OECD (Ydesen 2019), respectively. Study design moderates ts estimated impact on learning, particularly due to the comparison strategy, the difference between age- and grade-based comparisons between repeaters and non-repeaters (Allen et al. 2009); this applies to US studies, but also to ILSAs, e.g., PISA and ERCE (Cardoso 2020), not designed to answer specific research questions (Rutkowski 2016).

Upon reviewing the academic English-language literature on grade retention, two camps emerge: those who attack the policy; and those who, without defending it, attack its attackers' methods, sometimes questioning whether this policy's impact can be evaluated. Since at least the 1960s, grade repetition is consistently criticized, due to its alleged inability to deliver on values expected of education systems and policies, as seen in these questions:

1. Effectiveness. Does grade repetition improve student achievement? (Shepard and Smith 1990).
2. Efficiency. "What are the costs and benefits?" (Eide and Goldhaber 2005, 195).

3. Equity. “What characteristics are associated with being retained?” (Jimerson 1999, 244).
4. Empathy. “What are the emotional effects of retention?” (Shepard and Smith 1990, 85).
5. Evaluability. Is it possible to objectively evaluate the impact of retention? (Allen et al. 2009).
6. Expulsion. “Is grade retention associated with dropping out?” (Jimerson 1999, 244).

While my review covered all questions, due to space constraints I will focus here on effectiveness and evaluability. On the *effectiveness* question, the 1970s saw “no reliable body of evidence to indicate that grade retention is more beneficial than grade promotion”, an oft-misinterpreted statement, despite the caveat that the evidence is “so poor that valid inferences cannot be drawn concerning the relative benefits of these two options” (Jackson 1975, 627). Contemporaries emphasized that statement’s first half, ignoring the second (see Haddad 1979). However, 1980s’ standards-based reform in the US yielded better empirical studies with longitudinal designs. By the 1990s, two camps emerged. Each saw itself as Jackson’s heir, equally gifted with healthy skepticism. But they were skeptical about different things: Shepard and Smith (1990) and Jimerson (1999) were skeptical of repetition’s effectiveness; meanwhile, others doubted the evidence against it, for methodological reasons.

Given its focus on methods, *evaluability*[20] appears secondary, but pioneers deemed most research “quite inadequate for making valid inferences about grade retention effects” (Jackson 1975, 614). Eventually, the question transcended the research quality issue: underlying was the notion that the policy itself poses methodological challenges (Allen et al.

[20] Admittedly, the term has a different meaning in evaluability assessment (Wholey 1987).

2009). Grade retention clashes with randomized controlled trials (RCTs), impact evaluation's gold standard, on several issues, two of which I address here: endogeneity and comparison strategies. This matters because "we are witnessing a hierarchization of information (very often with randomized controlled trials on the top and qualitative data on the bottom), rendering some types of evidence more relevant than others" (Steiner-Khamsi 2021, 4).

Unlike in RCTs, "principals and teachers are reluctant to assign treatments randomly to pupils" (Jackson 1975, 628) in repetition studies since the 1940s, thus creating endogeneity issues. Among other issues, endogeneity complicates causal inference. Innovative techniques aim to alleviate endogeneity's effects: comparisons to low-performing non-repeaters (Alexander et al 2003); propensity score methods (Andrew 2014); instrumental variables (Jacob and Lefgren 2009); etc. But most of all, researchers need pre-retention measures of academic achievements and socio-emotional variables, which means that virtually all grade retention studies, since at least the 1980s, are longitudinal.

Regarding the comparison strategy (Allen et al 2009), the typical RCT collects data pre- and post-intervention, simultaneously across treatment and control groups. But with grade retention, the post-intervention measurement raises a dilemma: When should retainees be assessed? And since repetition shifts age-for-grade, who are the controls? The retainees' new classmates, or their age-cohort peers? Practices vary: same-grade (Lorence and Dworkin 2006), same-age (Jimerson 1999), or both comparisons (Alexander et al 2003). This affects findings: "Results based on same age comparison groups yielded larger effects of retention on school achievement than results based on same-grade comparison groups" (McCoy and Reynolds 1999, 273); Alexander et al (2003), less critical of repetition, agree. In a nutshell, if a grade repetition study uses only one comparison, it is not telling the whole story.

Paradoxically, consensus unites grade retention's critics and their methods-minded challengers: conclusions on effectiveness depend more on researchers' long-held beliefs than on evidence. Retention's critics claim it persists "in the face of negative evidence because teachers and parents cannot conduct controlled experiments. Without controlled comparisons, retention looks as if it works, especially if you believe that it does" (Shepard and Smith 1990, 85). Their intellectual antagonists agree: repeating "has not helped these children (…) but can it really be said that retention is the source of their problems? Any such conclusion takes us out of the realm of evidence and into the realm of opinion, preconception, or worldview" (Alexander et al 2003, 242).

Around the 1990s, increasingly prominent equity and empathy questions joined decades-old concerns with efficiency and effectiveness. Longitudinal studies reinforced this by collecting information from students about ascriptive and socio-emotional characteristics, respectively linked to sociologists' equity concerns and psychologists' empathy interests. The 1970s question on whether repetition worked for the system shifted by the 1990s to whether it was fair and sympathetic to students: system-centered efficiency and effectiveness stand in contrast to student-centered equity and empathy, paralleling the opposition between education as human capital and human right. Finally, since the 2000s, evaluability concerns academia's (in)ability to understand repetition's effects; I deem it a science-centered concern.

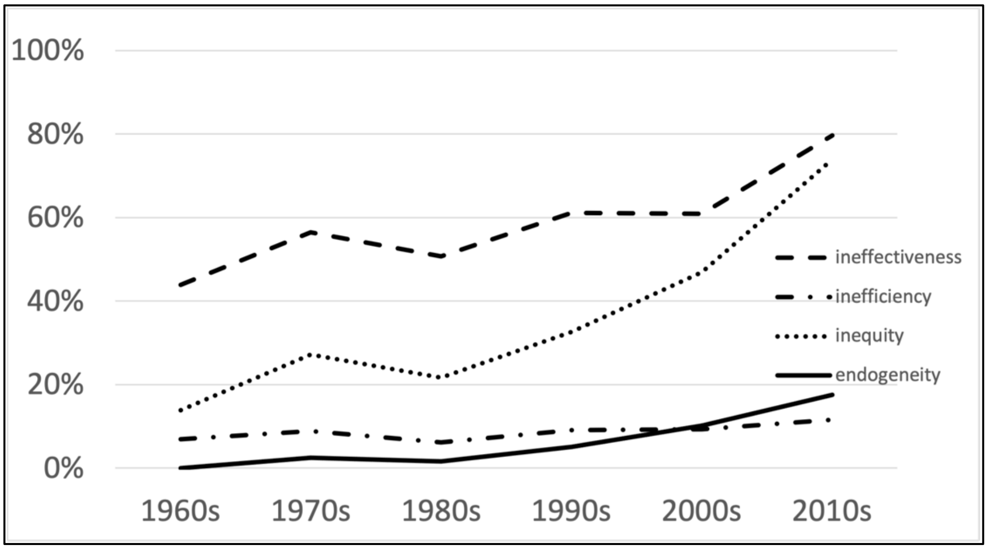


Figure 1: System-centered (inefficiency, ineffectiveness), student-centered (inequity) and science-centered (endogeneity) critiques of grade retention. Author's calculations based on Data extracted from Google Scholar on July 3rd, 2021.

Figure 1 presents sources (Google Scholar, July 3rd, 2021) that included "grade retention" or "grade repetition" by decade. I calculated percentages of sources including these additional strings: "inefficiency"; "ineffectiveness"; "inequity"; and "endogeneity", as an *evaluability* proxy. (*Empathy* is excluded due to the complex array of outcomes it represents.) A stagnant "inefficiency" critique hovers around 10% for five decades, while "ineffectiveness" grows from less than 50% to about 80%. Student-centered "Inequity" grows even faster, from 14% to 74%. Finally, science-centered "endogeneity" emerges, going from 0% to 18%.

3.4 Data and methods

This section covers the chronological scope, the criteria for inclusion or exclusion, the corpus, and the methods, including genre analysis (Swales 1990) and the comparative case study (CCS) approach (Bartlett and Vavrus 2017).

Data

Chronological scope. ERCE's first round (i.e., PERCE) preceded PISA's by three years, and is not comparable with subsequent rounds, even targeting a different grade. Moreover, qualifying documents from PERCE were not found. Given that, and the fact that SERCE coincided with PISA 2006, I set 2007 as the lower bound for the period. Since PISA has 3-year cycles, and ERCE rounds are currently separated by 6-7 years, by setting 2017 as the upper bound I was able to compare two ERCE cycles to four PISA cycles.

Inclusion criteria. Documents meeting all the following criteria were included:

1. Drawn from either https://unesdoc.unesco.org or https://www.oecd-ilibrary.org .
2. Published between 2007 and 2017.
3. Published solely[21] by the ILSA's parent IO (i.e., UNESCO or OECD).
4. With the ILSA's name (in ERCE's case, PERCE, SERCE or TERCE) in the title or cover.
5. With the words "repetition" or "retention"; or "repetición" or "repitencia".

Exclusion criteria. The following types of documents were excluded:

1. With "technical report" or "reporte/informe técnico" in the title.
2. With a country name in the title. (This exclusion was done manually.)
3. Stand-alone *Executive summaries* were manually excluded, since every *Full report* was available.

[21] OECD co-published two excluded documents with other IOs: Lockheed et al. (2015) and OECD/UIS (2003).

Corpus. This yielded 22 documents: 9 from ERCE; 13 from PISA. Two ERCE documents were then excluded for using criterion #5 strings in a different sense. A final filter preserved only documents with the strings in the title, or 10+ combined occurrences in the text. This excluded four PISA publications with 2-6 occurrences, and three ERCE publications with 1-3 occurrences. This yields a final corpus of 13 publications: 4 from ERCE; 9 from PISA. Selected publications have codes with a letter (E for ERCE, P for PISA); year of publication; and a lowercase letter when needed, e.g., E2015a, P2016d. (Corpus description in Appendix 3.A.)

The publications fall under three categories: 8 results reports (159-546 pp.); 3 monographic leaflets (4-6 pp.); and 2 monographic reports (212-224 pp.). Only leaflets were analysed in their entirety. For reports, in order to avoid isolated mentions of grade repetition, I identified excerpts of at least two-pages long with an average density of two or more occurrences per page. Each excerpt is up to six pages long; each publication contains up to three of them. (While some had self-explanatory headings like "Grade repetition", others did not.) Consequently, for reports, the number of examined pages ranged between 2 and 11.

Methods

This article uses genre (move-step) analysis (Swales 1990) and combines it with the comparative case study approach (Bartlett and Vavrus 2017). Genre analysis "represents academic research articles in terms of hierarchically organized text made up of distinct sections; each section can be subdivided into moves, and each move can be broken down into steps" (Kanoksilapatham 2004, 271). My purpose here is to uncover this subgenre's rhetorical structure, and the devices used with a persuasive purpose (Auld and Morris 2016), e.g. the use of references (Baek 2020; Steiner-Khamsi 2021). Moreover, a genre is a shared template that facilitates the ensuing comparisons across ILSAs and over time. The codebook appears in Appendix 3.B.

An inter-rater reliability analysis (IRR) was conducted on the application of the genre (move-step) analysis. Details of the result of this analysis appear in Appendix 3.C. The results from these IRR analyses led me to introduce two changes that led to a more streamlined coding rubric, and in a way, to a simplified analysis. First, it was difficult to distinguish *Recommendations* about grade repetition from *Alternatives* to grade repetition; therefore, these two steps were merged into a more general category, the move *Directives*.

Second, while it was relatively straightforward to distinguish the grade repetition themes (*Effectiveness, Efficiency, Empathy, Equity, Evaluability, and Expulsion*) from each other, many statements were ambiguous in terms of their inclusion into either the *Critique* or the *Results*. Therefore, the findings presented here (see Table 3 below) merge the *Effectiveness* step under the *Critique* move with the *Effectiveness* step under the *Critique* move; and similarly for the other themes.[22] While this limits the granularity of the analysis, this ambiguity between what constitutes a *Critique* based on the literature or a *Result* of the ILSA itself could be deemed a finding in its own right, and further explored.

Within that move-step analysis, I also analysed the presented charts, employing the notion of ordinalization (Fourcade 2016; Sorensen and Robertson 2020). A critical analysis of PISA 2006 reports identifies four data presentation formats: “alphabetic lists”; “ranked lists”; “scattergrams”; and “multilevel charts” -tables with outputs from multilevel regression models (Mortimore 2009, 10). I found four similar types of charts on grade repetition in this study’s corpus: (1) alphabetic bar charts/tables; (2) ranked bar charts; (3) scatterplots; and

[22] *Expulsion*, in theory, should only appear in the *Critique*, because these ILSAs are school-based and therefore cannot draw inferences about links between repetition and dropout.

(4) charts without country-level information. The first three represent, respectively, the nominal, ordinal, and interval scales of measurement (Stevens 1946).

Finally, the comparative case study (CCS) methodology identifies three comparison axes (Bartlett and Vavrus 2017). Here, the crucial horizontal axis compares PISA and ERCE as though they were sites. The transversal axis compares an ILSA's publications over time, especially those in series that present similar data each cycle, exceptionally useful to monitor evolving discursive trends. The vertical axis lies beyond the scope of this article; however, future research could focus on relationships between each ILSA and its parent IO, as well as on links to national discourses.

3.5 Findings

This section comprises four subsections. First, I situate grade repetition's increasing prominence in ERCE and PISA's publications within their increased overall attention to education policies. Then I present the two-level (move and step) rhetorical structure of ERCE and PISA's treatment of grade retention in their publications, alongside brief quantitative comparisons across ILSAs (i.e., horizontal) and over time (i.e., transversal). The third subsection focuses on the transition between two elements of that structure, namely *Justification* and *Critique*, highlighting the role of academic references. Finally, I explore the different uses of charts, across ILSAs and over time, to illustrate the issue of effectiveness.

Increasing prominence of grade repetition

Policies are increasingly visible in ERCE and PISA publications during this period. In PISA 2006, no report is devoted specifically to policy issues, which are included in the "Analysis" volume (P2007). But already in PISA 2009 (P2010) and 2012 (P2013), Vol. IV of the Results focuses on "What makes a school successful? Resources, policies and practices". PISA 2015 upgrades it to Vol. II (P2016d), with a more concise, focused title, and without question marks, perhaps to convey certainty: "Policies and Practices for Successful Schools". Similarly, ERCE introduces "Associated Factors" in 2010 (E2010), after SERCE; and "Policy Recommendations" in 2016 (E2016), after TERCE, alongside updated "Associated Factors" (E2015a). Repetition gains prominence during this period, according to several metrics (see Table 1).

TABLE 1
"REPETITION" OCCURRENCES IN ILSA PUBLICATIONS BY CYCLE

ERCE cycle	Qualifying publications	Combined occurrences	Average occurrences per publication
SERCE (2006)	1	24	24.0
TERCE (2013)	3	86	28.7
PISA cycle	Qualifying publications	Combined occurrences	Average occurrences per publication
2006	1	9	9.0
2009	2	46	23.0
2012	2	105	52.5
2015	4	210	52.5

Qualifying publications = included in the corpus.
Combined occurrences = occurrences of "repetition" across qualifying publications.

I also separately examined the most policy-focused publication in each cycle [23]: occurrences grow over tenfold for PISA (Table 2). To account for increasing length, I calculated number of occurrences per page; this also grows over tenfold. For ERCE, it increases slightly, from 0.15 in SERCE's 2010 report to 0.26 in the most comparable TERCE reports, published in 2016. The number of academic references related to grade repetition in policy-oriented PISA

[23] "Analysis" (P2007); "What makes a school successful? Resources, policies and practices" (P2010); "What makes schools successful? Resources, policies and practices" (P2013); "Policies and Practices for Successful Schools" (P2016d).

reports grew from one (1) in the 2006 cycle to 22 in the 2015 cycle; an "explosive growth of references" like the one observed overall in Norway's policy documents (Steiner-Khamsi et al. 2020, 20). Multiplying reference tokens outpace "repetition" occurrences.

TABLE 2
"REPETITION" OCCURRENCES AND USE OF ACADEMIC REFERENCES
IN EACH ILSA CYCLE'S POLICY-ORIENTED REPORT

	Policy-oriented report				
ERCE cycle	Occurrences	Occurrences/ page	Reference tokens	Reference tokens/ Occurrence	Unique references
SERCE (2006)	24	0.15	0	0.00	0
TERCE (2013)	44	0.26	4	0.09	3
PISA cycle	Occurrences	Occurrences/ page	Reference tokens	Reference tokens/ Occurrence	Unique references
2006	9	0.02	1	0.11	1
2009	32	0.10	4	0.13	2
2012	89	0.16	15	0.17	13
2015	108	0.23	22	0.20	18

Policy-oriented report = the report in each cycle that is most concerned with policy issues. (For details, see Appendix A.)
Occurrences = occurrences of "repetition" within the policy-oriented report.
Reference tokens = appearances of references, including repeated appearances of a unique reference.
Unique reference = number of unique references within each cycle's policy-oriented report.

Genre analysis and quantitative comparisons

Table 3 shows the two-level (move-and-step) rhetorical structure of publications in the corpus.[24] The first level involves five moves: *Opening, Critique, Quantification, Results,* and *Directives*: the first four are "conventional", appearing in 60% or more publications; falling below that threshold, *Directives* (presenting recommendations about, and alternatives to, grade repetition) is "optional" (Kanoksilapatham 2004). The fact that the *Directives* moves is the only optional one is particularly meaningful if one takes into account Auld and

[24] Appendix 3.D tabulates the presence of moves and steps by publication.

Morris's (2016) view, according to which these ILSA publications focus on persuasion. That said, this persuasive aim does not necessarily entail the presence of explicit policy directives. In fact, genre analysis enables the identification of persuasive strategies embedded in the rhetorical structure, as in the example presented in the section *From justification to critique*.

Each conventional move is further split into 3-6 steps. The *Opening* and *Quantification* moves have three steps each that appear in a conventional order, hence their numbering. The *Critique* and *Results* moves are inspired by the grade repetition themes presented in the *Background*, which in this table are merged to account for the aforementioned difficulty to distinguish, occasionally, between review and results. This phenomenon deserves further study. This occasional ambiguity in identifying whether a given statement refers to the literature or presents new evidence derived from the current cycle of the ILSA in question may play a role in the persuasion strategy, by blurring the boundaries between "received" knowledge and new evidence that allegedly confirms it.

The Opening move includes a conventional step, *Justification* (presenting arguments for grade retention), and two optional ones: *Introduction* (situating repetition within the bigger picture) and *Definition* (providing a theoretical definition); when present, they precede the *Justification*. Similarly, *Quantification*'s core is the conventional step *Level* (quantifying the extent of repetition's presence), sometimes preceded by *Collection* (describing the ILSA's collection of grade repetition data), or followed by *Evolution*, (monitoring changes in repetition levels), both optional. Regarding the *Critique*/*Results* moves, *Effectiveness, Efficiency and Equity* step are conventional; all other steps are optional.

While the rhetorical structure is remarkably stable, horizontal comparisons show systematic variation across ILSAs. Table 3's last two columns show the number of publications from each ILSA that feature each move and step; those with sizable quantitative differences between ILSAS are highlighted in **bold**. First, 3 out of 9 PISA publications define repetition, or describe the data *Collection* procedure; ERCE publications do neither. Second, and similarly, *Evaluability* is absent from ERCE, while sometimes present in PISA's *Critique* or *Results*. Third, *Efficiency* is the norm for PISA, but the exception for ERCE. Finally, all ERCE publications feature *Directives*, while only 3 out of 9 PISA publications do.

TABLE 3
GENRE ANALYSIS: TWO-LEVEL (MOVE-STEP) RHETORICAL STRUCTURE, AND QUANTITATIVE HORIZONTAL COMPARISON FOR SELECTED MOVES AND STEPS

Nodes	Publications coded /13	Conventional move/step	ERCE /4	PISA /9
1 Opening	11	C	3	8
1.1 Introduction	7		2	5
1.2 Definition	3		**0**	**3**
1.3 Justification	9	C	2	7
2 Critique (see Results for steps)	12	C	4	8
3 Quantification	11	C	3	8
3.1 Collection	3		**0**	**3**
3.2 Level	11	C	3	8
3.3 Evolution	5		1	4
4 Results	13	C	4	9
Effectiveness (C, R)	13	C	4	9
Efficiency (C, R)	9	C	**1**	**8**
Empathy (C, R)	7		2	5
Equity (C, R)	10	C	3	7
Evaluability (C, R)	4		**0**	**4**
Expulsion (C)	7		2	5
5 Directives	7		**4**	**3**

Digits in each column indicate the number of publications in which a move or step appears, out of the total indicated on the top of each column.
Conventional move/step = Appears in 60% or more of coded publications.
All other moves or steps are either optional, or not applicable.

Regarding transversal comparisons, only two ERCE cycles fall within the period, and the upcoming qualitative analyses show little change in ERCE's discourse on repetition. Conversely, perhaps due to "a lot of internal pressure to try something new" (OECD officer cited in Zapp 2020, 10), PISA's four examined cycles show meaningful changes (see

Appendix 3.D). First, *Evaluability* initially appears in 2013 and becomes recurring. Second, while the *Directives* move is pervasive in ERCE, it first appears in PISA publications in 2014, becoming recurring after that. The upcoming qualitative sections deepen the genre analysis and comparisons, focusing on two resources that legitimize evidence: references and charts.

From justification to critique: Folk wisdom v. science

This section examines the transition from the *Justification* step, which typically closes the *Opening* move, to the *Critique* move, focusing particularly on the use of references or lack thereof. Most PISA publications (7 out of 9) and half of ERCE's include a *Justification* of repetition that typically precedes the *Critique*. Horizontal comparisons reveal stunning qualitative similarities across PISA and ERCE; and the transversal comparison shows similar resources across assessment cycles. Both ILSAs use one set of rhetorical devices for the *Justification* and another for the Critique:

> *Grade repetition is a policy through which school systems try to meet students' educational needs. By repeating a grade, slower students are given a second chance to master their coursework. Grade repetition also serves a motivational purpose because it is sometimes also used as a way to penalise students who do not perform well or do not put forth the necessary effort in school. With the prospect of repeating a grade – and thus not moving forward with their peers – students at risk may decide to put more effort into their studies to avoid retention. In practice, however, grade repetition has not been shown to benefit student learning (Allen et al., 2010; Alexander et al., 2003). Moreover, grade repetition may have adverse system-level effects as retained students are more likely to drop out, stay longer in the school system, or spend less time in the labour force (Rumberger, 2011; OECD, 2011b)* (P2013, 88)

The *Justification*, in the first half, does not include any references, while the *Critique* abounds in them: "Arguments can take a dialogical form, i.e. the form of two or more people arguing. But it is also useful to analyse 'monological' arguments such as this one in a dialogical way" (Fairclough 2003, 82). The transition between the two imaginary debaters is aided here by the contrastive connector "however", preceded by "In practice" to stress that statements in the *Critique* are supported by real-world evidence, while the references in it show the backing of the academic community; the *Justification* has neither. While ERCE uses fewer references, it uses them in similar ways, always in the *Critique*, never in the *Justification*:

> *Grade repetition is employed as a remedial measure that is supposed to improve students' performance. However, this study's results show that it does not fulfill that purpose, and the regional evidence about this is robust (Treviño et al., 2010)* (E2015a, 55)

In addition to the lack of references, the *Justification* employs two resources that dilute the identity of the stakeholders that support and implement grade retention: prosopopoeia and passive voice. First, the policy is anthropomorphized: "Grade repetition (…) **seeks** to (…)" (P2010, 62; variant in P2013, 73); "repeating a grade **gives** students…" (P2016d, 160-1). Elsewhere, the passive voice evokes vanishing educational actors: "Sometimes the prospect of grade repetition, itself, **is seen** as a source of motivation" (P2013, 34); "grade repetition is still commonly used" (P2010, 62); "**is employed** as a remedial measure" (E2015a, 55); "**used** to group students" (P2011). By whom? Research on teachers' beliefs about repetition (e.g., Byrnes & Yamamoto 1986; Tomchin & Impara 1992; Range et al. 2012) is not cited.

In a legal analogy, the justification is repetition's defense, while the critical literature review is the prosecution. The evidence each side presents is fundamentally different. The defense's evidence would be considered hearsay in a court of law: it is never clear who is doing the talking, or whether they have knowledge of the matter at hand. Conversely, the prosecution presents expert testimony when it cites scholars' writings published in peer-reviewed journals. Grade repetition's defense uses the language of folk wisdom, rumor, and similarly antiquated, delegitimized forms of knowledge, while the prosecution's case flaunts the trappings of science. Ultimately, for both ILSAs, the case for repetition is nothing but a straw man, a house built only to be demolished.

Whereas the similarities in the use of references -or lack thereof in the *Justification*- are remarkable, a horizontal comparison across the two ILSAs yields several differences. First, there is no overlap between their references, since PISA cites only English-language sources, and ERCE, Spanish-language ones[25]. Second, PISA publications include more reference tokens on grade repetition (a mean of 8.11 per publication) than ERCE publications do (mean=2.25). Third, PISA sometimes attributes a given finding to multiple sources, e.g., "students from socio-economically disadvantaged backgrounds appear to be hurt most by grade repetition (Hauser, 2004; Alexander, Entwisle and Dauber, 2003)" (P2010, 35); ERCE never does this. Finally, PISA sometimes appears to attribute several findings to several sources; this ambiguity could be rhetorically effective.

This example is unambiguous about which sources back each statement: "Students who have repeated a grade often also show more negative attitudes and behaviours towards school (Finn, 1989; Gottfredson, 1994; Ikeda and García, 2014) and are more likely to drop out of school (Jacob and Lefgren, 2004; Manacorda, 2012)" (P2016d, 161). Some sources back the *empathy* statement, some the *expulsion* statement. But backing a finding with multiple sources still creates the impression –and some, like Alexander et al. (2003) and Allen et al. (2009), would say, the illusion– of scholarly consensus. Moreover, that same report later lumps together statements and sources, resulting in ambiguity in the attribution of findings to sources:

[25] Curiously enough, the published English translations of ERCE results reports drop the Spanish-language references included in the original Spanish-language publications.

> *Students who have repeated a grade tend to display more negative behaviours and attitudes towards school, are more likely to drop out and may be stigmatised among their classmates (Ikeda and García, 2014; Rumberger and Lim, 2008; Thompson and Cunnigham, 2000; West, 2012)* (P2016d, 232)

Five out of the nine examined PISA publications feature uses of ambiguous attribution that resemble this one.[26] In this example, all sources appear to back all statements about empathy and expulsion; this intertwining of themes could also be read as implying a causal link between the stigmatization, the development of negative attitudes and the dropout.

Worst practices: visualizing repetition's (in)**Effectiveness**

Effectiveness is the only step that appears in every publication, from either ILSA, throughout the period, as well as in 13 out of the 29 examined graphs. Despite ERCE and PISA's shared concern about grade retention's ineffectiveness, a deeper horizontal comparison reveals two important qualitative differences. First, rankings (Espeland et al. 2016) or "ordinalization" (Fourcade 2016) are central to PISA's *Effectiveness* discourse but absent from ERCE's. Second, ERCE analyzes differences between groups *within* countries, while PISA increasingly focuses on differences *between* them. Their use of charts (e.g., bar charts, scatterplots, etc.) illustrates both differences.

Ordinalization divides these two ILSAs: while ERCE evades it, PISA embraces it. ERCE publications feature four *Effectiveness* graphs, all alphabetic bar charts (type 1, see Figure 2). They showcase "Differences in performance among students who have repeated a grade and who have not repeated" (E2015, 3) within each country. Arguably, alphabetic

[26] For additional examples of both ambiguous attribution of findings to sources, see Appendix 3.E.

charts reflect nominal judgments, "oriented to essence" (Fourcade 2016), resisting ordinalization: while countries can be compared[27], they should not, or cannot, be ranked. Why not?

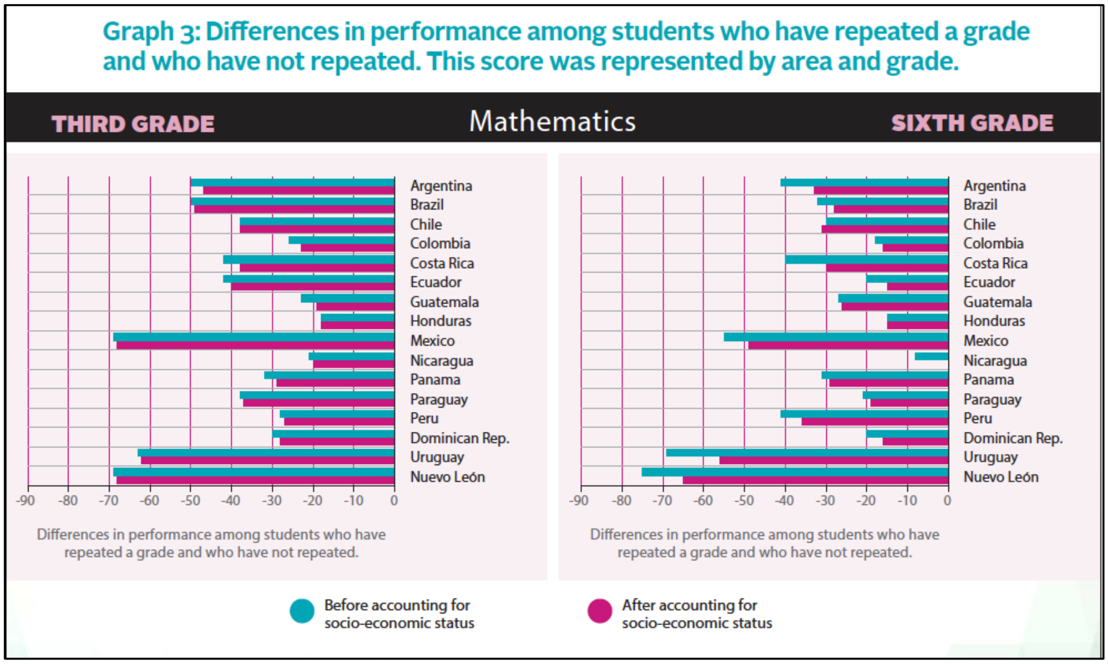


Figure 2: Example of alphabetic bar chart on grade repetition *Effectiveness* (E2015B).

There are at least two possible interpretations. Ostensibly, ERCE does not rank countries on outcomes or even effectiveness of their policies because they have different contexts and starting points. ERCE's extensive presentation of background information on the countries' economic, human, and social development backs this interpretation (E2015a, 24-37). Instead, a detractor might argue that countries cannot be ranked because "Ordinalization techniques differentiate people and things in a world that presumes fundamental, underlying equivalence" (Fourcade 2016, 188). Naturally, this equivalence is not inherent to countries; comparability is made by using metrics (Cardoso & Steiner-Khamsi

[27] ERCE, after all, stands for "Estudio Regional *Comparativo* en Educación".

2017). If ERCE's achievement metric was not comparable across countries, then comparing countries would be dubious, but ranking them would be simply wrong.

Meanwhile, PISA presents nine *Effectiveness* charts, none of them alphabetical: three ranked bar charts (type 2); one scatterplot (type 3); and five without country-level information (type 4). Arguably, ranked bar charts (type 2), such as the one shown in Figure 3, reflect ordinal judgments "oriented to relative positions" (Fourcade 2016). Why should it matter that PISA ranks countries on outcomes or policies?[28]

> *Rankings express the drama of social life in an ideologically egalitarian society, playing and replaying the competition between individuals and the contest between organizations on a seemingly leveled playing field (Espeland and Sauder 2007). Rankings have enough stability over time that they do indeed order the world. But they contain enough movement that they can preserve the perception of mobility, the possibility of change* (Fourcade 2016, 184).

In other words, rankings paint a picture of egalitarianism with (fair) competition, and of mobility within stability. Simultaneously, rankings have a surprising ability to legitimize themselves: if ranking countries requires a comparable metric, the acceptance of rankings by countries must mean that the underlying metric is, indeed, comparable. By ordering countries according to its own criteria, and having countries buy into those rankings, an ILSA can reinforce the notion that the data it produces are truly on an equivalent scale. This real-world influence is potentially more powerful than any jargon-laden technical argument: The proof is in the pudding!

[28] ERCE presents a ranked bar chart regarding grade repetition, but not on effectiveness. It is not ERCE data, but UIS data on repetition levels for ERCE countries and Caribbean countries that do not participate in ERCE (E2015b).

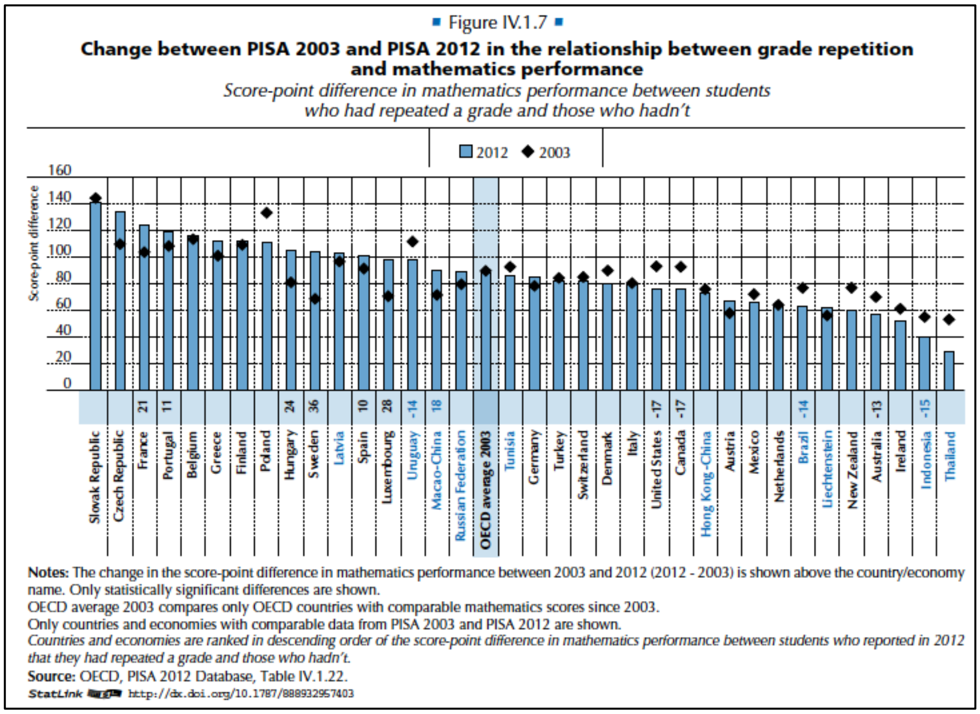


Figure 3: Example of ranked bar chart on grade repetition *Effectiveness* (P2013, p. 38). OECD (2013), *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices (Volume IV)*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201156-en

These charts have, however, some rhetorical limitations. Figure 2 focuses on how, within each country, non-retainees outperform retainees, a point that may appear obvious. Meanwhile, it does not inform readers about how prevalent grade retention is in each country. Even more importantly, it does not remind the reader how Slovak Republic (far left) or Thailand (far right) do in the achievement ranking. Finally, it does not give the reader any clues about the links between grade repetition and achievement at the system level. But could all this be done with just one image?

The second difference between the two ILSAs is the focus on within-country v. between-country effectiveness analyses. Despite their differences, Figure 2 (ERCE's alphabetic bar chart) and 3 (PISA's ranked bar chart) both focus on within-country differences between retainees and non-retainees, although arguably, by ranking countries, PISA starts to shift the focus towards differences between countries. But then it shifts this focus even further: PISA 2009 states that "school systems with high rates of grade repetition are also school systems that show lower student performance, whatever the causal nature of this relationship" (P2010, 35).

Figure 4, a scatterplot, and its accompanying analysis, both illustrate this pivot towards systems: "Grade repetition is more prevalent in school systems where students score lower in the PISA science assessment" (P2016d, 162). PISA embraces this approach, partly, because it can: if including large numbers of countries legitimizes rankings (Sorensen and Robertson 2020), this is even truer of this type of analysis. With over 70 participating school systems, PISA can get away with scatterplots, regression lines, and determination coefficients; ERCE must be more cautious[29]. This begs the question: "with enough data, do numbers speak for themselves?" (Gorur 2019, 53). Surely not, but with enough data, however this is defined, PISA can claim to examine the issue "not at the individual student level but at the system level in order to avoid selection bias (Heckman and Li, 2003)" (P2013, 34).[30]

[29] ERCE uses scatterplots with countries as units of analysis, but never on grade repetition, and typically not on policy. It uses them to stress the links between achievement and development indicators beyond education policymakers' control: GDP per capita; Gini; HDI, etc. The closest to a policy variable is education expenditure (E2015a).

[30] An endnote presents a definition: "Selection bias in this case refers to how to separate the effect of grade repetition from differences in achievement due to the selection of students who must repeat grades." (P2013, 66).

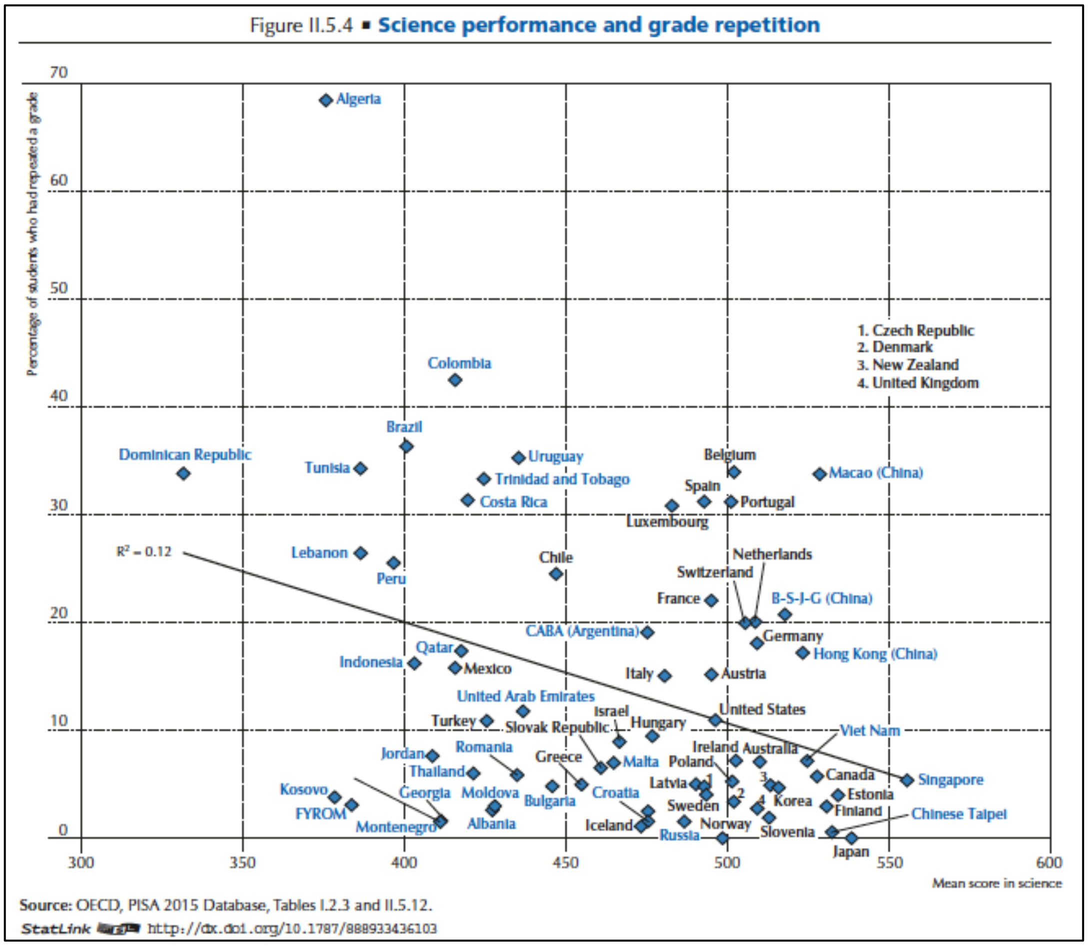


Figure 4: Example of scatterplot on grade repetition *Effectiveness* (P2016d, p. 162). OECD (2016), *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264267510-en

There is precedent for trying to solve selection bias issues at the individual level by aggregating them to the country level, particularly among human capital theory scholars using ILSAs for educational effectiveness research (EER): "In cross-country analyses, one can aggregate the institutional variable of interest up to the country level, thereby circumventing the selection problem" (Hanushek and Woessmann 2011, 112). However, these scholars recognize that this is not fail/safe: "The main cost to this—apart from the limited degrees of freedom at the country level—is that unobserved heterogeneity at the country level may introduce new forms of omitted variable bias" (Hanushek and Woessmann

2011, 112). These omitted variables could include cultural factors, which have been shown to influence repetition levels (Agasisti and Cordero 2017; Wong 2013).

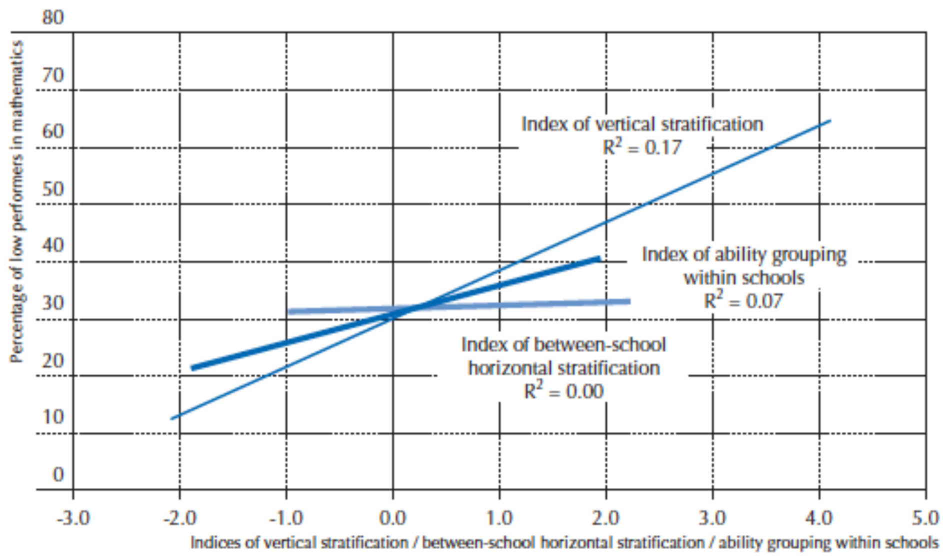
Crucially, in addition to allegedly solving the selection bias issue, this pivot to the system level also somehow turns the studies, from cross-sectional, into longitudinal: "Recent assessments such as PISA and TIMSS are repeated every few years and thus have a longitudinal component at the system level" which would "allow researchers to estimate causal effects of changes in educational policies and other factors at the system level" (Strietholt et al 2014, 11). This seems particularly fitting for an ILSA with a short, 3-year cycle, like PISA.

All this offers PISA three strategic advantages. First, PISA can produce methodologically correct analyses of the link between grade repetition and achievement that are allegedly unaffected by selection bias. Other ILSAs, e.g., ERCE, cannot do this with few participating countries, becoming then delegitimized to address grade repetition's effectiveness, thus giving PISA the monopoly. Second, PISA can estimate regression models and show scatterplots that evoke analyses in the natural sciences (Porter 2010) or economics, adding scientific credibility. And PISA can argue that, at the system level it monopolizes at least for this policy, its analyses are no longer cross-sectional, but longitudinal instead.

Finally, PISA goes from cross-national differences to erasing countries altogether in a scatterplot without a scatter: once its messy uncertainty is concealed, the certainty of each straight line is revealed (Figure 5). This lets PISA compare repetition's association with achievement to other "selecting and sorting students" approaches: "System-level analysis shows that more vertical stratification, including grade repetition, is related to a greater incidence of low performance in mathematics" (P2016a, 184) as compared to within- and between-school horizontal stratification.

■ Figure 5.11 ■

Sorting/selecting students and percentage of low performers in mathematics



Notes: A significant relationship (p < 0.10) is shown by a darker line.
The *index of vertical stratification* is based on the degree of variation in the grade levels in which 15-year-old students are enrolled.
The *index of ability grouping within schools* is based on the prevalence of ability grouping within schools across the school system.
The *index of between-school horizontal stratification* is based on five indicators: the number of education tracks, the age at which students are selected into those tracks, the prevalence of vocational programmes, the academic selectivity of the school, and school transfer rates.
Only countries and economies with available data are included.
Source: OECD, PISA 2012 Database, Table 5.4.
StatLink http://dx.doi.org/10.1787/888933315908

Figure 5: Example of "scatter-less" scatterplot on grade repetition *Effectiveness* (P2016a, p. 183). OECD (2016), *Low-Performing Students: Why They Fall Behind and How to Help Them Succeed*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264250246-en.

Consequently, PISA also ranks policies, starting with grade repetition: "Of all variables analysed in this report, repeating a grade has the second strongest association with low performance, after socio-economic disadvantage" (P2016a, 184). However, while ERCE refuses to rank countries and does not share PISA's ordinalization/scientization journey, it lands in a shockingly similar spot: "After considering socioeconomic status, repeating a grade is the second variable that has the strongest relationship with learning achievements."

(E2015, 2).[31] This, however, aligns with Latin America's scholarly tradition about -and for the most part, against- grade repetition (Schiefelbein 1975; Reimers 2000).

Based on prior research, two aspects of this convergence stand out. The first is deemed typical of PISA discourse: "contextual elements such as family background are marginalised by distinguishing them from features that are amenable to control, which are described as 'operational elements' (OECD 2012, 23)" (Auld and Morris 2016, 216). The second one, however, is quite unexpected: while secondary beliefs, understood as advocates' policy preferences, "hold the greatest potential for variation within the community" (Auld and Morris 2016, 204), this particular belief about grade repetition seems to cut across these two ILSAs.

While the ideology of the IO behind PISA, and grade retention's history in ERCE's region surely played a part in shaping this convergence, an additional aspect should be considered. There are differences in these ILSAs designs that produce nuances in their data that turned out to be statistically significant (Cardoso 2020); however, there is also one crucial design similarity that may elucidate the similarities in the data and discourses:

> *The magnitude of repeaters' achievement gap varies according to comparison strategy. But then why do both TERCE and PISA, despite different comparison strategies, yield negative slopes for achievement on repetition? Because their cross-sectional design fails to address the endogeneity of grade retention: in Latin America, on average, repeaters have lower achievement than nonrepeaters before being retained. Consequently, the answer to the "objectivist" question of repetition's effectiveness cannot come from cross- sectional studies, like current ILSAs; but from longitudinal studies that consider learners' starting points.* (Cardoso 2020, 613).

While both -and in fact, all- ILSAs share the same design feature, only one of these two addresses the issue in their results reports. PISA offers unexpected nuance, with a caveat that

[31] Interestingly, the ERCE publication, from November 2015, briefly precedes the PISA report, from February 2016.

dances around this very issue, without using the word cross-sectional (as opposed to longitudinal), but speaking instead about a related concept, namely "correlational" (as opposed to experimental or quasi-experimental):

> *Caution is advised when interpreting the link between grade repetition and low performance, however, because determining the direction of the association is particularly difficult. On the one hand, grade repetition in the earlier grades makes a student more likely to perform poorly in a later grade, because teachers have lower expectations for these students, because these students might have greater difficulties in integrating themselves into peer and school cultures, or for other reasons (Kaplan, Peck and Kaplan, 1997; Roderick, 1994). But the association might run in the opposite direction if students repeat a grade simply because they are chronic low performers. PISA data provides only correlational evidence, so no causal inferences should be drawn from this analysis. (*P2016a, 88)

A genre perspective can also help explain this display of methodological caution: "Genre knowledge is further demonstrated by highlighting limitations and including caveats relating to culture, context, the complexity of social interactions, the problem of causality, and even direct statements warning against policy transfer" (Auld and Morris 2016, 213). At this point, one is left to wonder whether there is a contradiction between this caveat and the simultaneous declaration of grade repetition as, literally, the practice most strongly associated with low academic performance. While the statement is not literally causal, it is extremely forceful and, once again due to ordinalization, likely to make an impact.

Finally, why does PISA disclose the endogeneity issue, but not the comparison strategy?[32] In all fairness, neither does ERCE, and perhaps the subject is too technical for a results report. However, there is one final nuance to consider: while in ERCE's case, the grade-based comparison strategy may ameliorate the effects of the cross-sectional design to some extent; in PISA, the age-based comparison strategy surely accentuates them.

---

[32] This is surprising, as this was the topic of PISA's most cited external source, namely Allen et al. (2009), cited four times, as well as mentioned by Ikeda and Garcia (2014), published by OECD and based on PISA data, also cited four times.

3.6 Discussion

This study aimed to answer questions about three topics: the similarities and differences in ILSA discourses on grade repetition; their evolution over time; and the degree to which those similarities and differences can be linked to IOs' doctrines and ILSAs' designs and data. This section addresses those questions in view of the findings.

There are remarkable similarities in the structure and substance of the studied excerpts of ERCE and PISA reports, even across languages; e.g., the selective use of references when defending or critiquing grade retention. But most of all, the examined sections in these publications share "a common communicative purpose" (Auld and Morris 2016, 203), namely, identifying grade repetition as a policy strongly associated with low academic performance. This leads me to conclude that this is, indeed, a (sub)genre (Swales 1990). In addition to similarities in the rhetorical structure, there are also similarities in the substance of their messaging: crucially, their discourses on grade retention's effectiveness are similarly critical. Across ILSAs, this is the only mandatory step across all publications, boasting more references and charts than all other themes.

Simultaneously, describing this subgenre's conventions as a reference framework for comparison proved useful when contrasting ERCE and PISA reports. This helped uncover their differences: some quantitative; but mostly qualitative. As an example of a quantitative difference, the *Efficiency* theme is the norm for PISA (8 out of 9 publications), and the exception for ERCE (1 out of 4 publications). How to explain this difference? This is related to the education doctrines of the IOs behind these ILSAs, specifically, OECD's human capital theory focus as opposed to UNESCO's human rights-based approach (Cardoso 2020).

Similarly, definitions of repetition and description of operational procedures for its measurement are completely absent from ERCE's publications, while occasionally present in PISA's. How to explain this? There are at least two ways to explain this, although they may be complementary rather than mutually exclusive. PISA may be more "scientized", that is, more heavily influenced by a discourse that, among other things, emphasizes mechanical objectivity and its emphasis on "personal restraint" (Porter 1995, 4) and transparency in procedures. Or ERCE may be influenced by models in the social sciences as practiced in Latin America that, themselves, are not as closely aligned to those approaches (Abend 2006). This analysis can be expanded, e.g., by including PASEC reports.

Regarding qualitative differences, two stand out in the ILSAs' treatment of the *Effectiveness* issue. First, PISA uses ordinalization techniques (Fourcade 2016), i.e., rankings, in addressing this theme, while ERCE does not. This aligns with how these ILSAs present data regarding other topics, and links back to IOs' educational doctrines, as some scholars see connections between ordinalization and human capital theory (Sorensen and Robertson 2020). Second, while ERCE focuses on differences between groups within countries, PISA focuses on differences between countries. This difference may also partly stem from IOs' doctrines, but also from characteristics of ILSAs' data.

As for the role of data in shaping this discourse, the credibility and legitimacy of rankings also depends on the number of countries that participate in them (Sorensen and Robertson 2020): this enables rankings from an ILSA like PISA, but not from one like ERCE. Clearly, this characteristic of ILSAs data does not stem directly from their design, but from their uptake among countries. This also enables PISA to move on from the type of within-country analyses that ERCE prioritizes and instead analyze differences across systems, free from selection bias according to Hanushek and Woessman (2011) and, what is more, according to Strietholt et al. (2014), longitudinal, and therefore, an allegedly legitimate basis for causal inference. But while strength in numbers (of countries) enables these analyses, what motivates them?

There are both strategic and rhetorical factors at play: "As IOs are situated in a competitive field where multiple organisations participate and contribute their research knowledge, originality and quality in scientific research also becomes a crucial element in reputation-building and status battles among IOs." (Zapp 2020, 17). Strategically, there is an advantage for PISA in adopting an analytical approach -presented as the only valid one- that is beyond reach for other ILSAs. This effectively allows PISA to grant itself the monopoly of this analytical and policy niche, at least from a cross-nationally comparative perspective.

This also has important rhetorical consequences, as this analytical approach allows PISA to adopt the trappings of scientific discourse more fully, graduating from ranked bar charts, to scatterplots, to "scatter-less scatterplots". By literally erasing countries, these visual aids increasingly hint at the generality or, rather, generalizability of the postulated relationships between policy and outcomes; this stands in direct contrast to the claims of national or regional specificity implied by alphabetic bar charts. At the same time, erasing the scatter around the lines also adds to the appearance of certainty that some policymakers crave. This double rhetorical advantage is, ultimately, also strategic.

3.7 Conclusions

Perhaps the most surprising twist in this story is that, after embarking on diverging rhetorical journeys, ERCE and PISA converge in declaring grade repetition as the policy or practice most strongly associated with low academic performance. One may even feel tempted to call it a "worst practice". But while Auld and Morris (2016) emphasize the potential variability of secondary beliefs, or specific policy preferences, is this convergence really that surprising? I set out to describe the similarities and differences in discourse across

ILSAs, and to interpret them in the context of their design and data, as well as IOs' doctrines. Based on the differences in doctrine, design, and data, I expected discourse to be different as well. I found the expected differences regarding, for instance, *Efficiency*, but somewhat unexpected similarities regarding *Effectiveness*.

Prior research showed that, due to different designs, particularly their diverging comparison strategies, PISA and ERCE differ in their estimates of the (in)effectiveness of grade repetition, with PISA yielding harsher findings (Cardoso 2020). However, one must not overlook a crucial similarity in their design: ERCE and PISA are both cross-sectional, as all ILSAs currently are. This sets them apart from most studies discussed in the academic literature on grade repetition since at least the 1980s. In practice, this means that all ILSAs that address the issue of grade repetition's effectiveness will paint a rather bleak picture. Could this also be the reason why PIRLS and TIMSS choose not to collect information on this? Ultimately, in this case, the similarities in these two ILSAs' discourses must be interpreted in relation to the similarities in their designs, not in their differences there.

In summary, this study, while limited in scope, uncovered substantial similarities in the discourse of these two ILSAs on grade retention, proving that genre analysis can be used for this purpose. It also identified some important differences, regarding efficiency and evaluability, among others, that should be further explored. In PISA's case it was possible to chart some degree of evolution of discourse around these issues; for ERCE, whose cycles are, literally, few and far between -but lately accelerating-, a broader chronological scope might be needed. Finally, when trying to connect discursive features to doctrines, design and data, the focus shifted, perhaps unexpectedly, from the differences to the similarities. But methods were once more shown to be socially consequential, this time, in shaping discourse.

Chapter 4: Policymaking to the test? How international large-scale assessments (ILSAs) influence repetition rates

4.1 Introduction

Claims of ILSAs' effects on education policy come from both ILSAs' proponents (Hanushek & Woessmann, 2014; Schleicher, 1999; Wagemaker, 2013) and opponents (Meyer & Benavot, 2013; Smith, 2016). However, studies showing statistical associations between a country's participation in ILSAs and the adoption of specific education policies (e.g., Ramírez et al., 2018), are few and far between. Even if associations were found, a causal, independent effect of ILSAs on policy would be difficult to prove (Steiner-Khamsi, 2019). Finally, ILSAs' links with policy are deemed one reason, among many others, why countries opt into them (Addey et al., 2017). This study explores quantitative evidence of influence of ILSA participation on one education policy, namely grade retention.

Few education policies have caused as much worldwide debate as grade retention, since at least the 1970s (Jackson, 1975); accused of lacking effectiveness (Jimerson, 1999), efficiency (Haddad, 1979), equity (Schiefelbein, 1975) and empathy (Byrnes & Yamamoto, 1985). Critiques aside, grade repetition is pervasive in education systems around the world, with two nuances: rates differ substantially across regions, and have been decreasing worldwide for some time now (UIS, 2012). Therefore, while repetition is universal, it also varies substantially across geographies and cultures, as well as over time.

Recent sociology of quantification research shows that ILSAs with different designs, shaped by diverging IO doctrines, yield different data on grade retention's effectiveness (Cardoso 2020). This article expands that literature by showing how ILSA participation influences national policy decisions. It focuses on grade repetition, widespread in Latin

America (UIS, 2012), using data on participation in OECD's PISA (Programme for International Student Assessment) and UNESCO's ERCE (Comparative and Explanatory Regional Study). These assessments' designs differ in comparison strategy (Allen et al., 2009), grade- versus age-based; and implementation in primary versus secondary education. This is the research question: Can participation in ILSAs help explain why repetition rates vary over time? Specifically, can participation in different types of ILSAs, namely age- versus grade-based, help explain Latin America's evolving repetition rates?

Structure of the article

The *Theoretical framework(s)* are world society theory (WST), educational effectiveness research (EER), and the sociology of quantification; presented with three hypotheses: policy scripts, lessons, and incentives, respectively. The *Background* highlights a prior study on large-scale assessment and repetition rates (Ramírez et al., 2018), and the resulting debate (Smith, 2019). The *Data and methods* section specifies the analysed dataset (the World Development Indicators); variables (participation in ILSAs, repetition rates, and controls); scope (Latin America, 1992-2017); and analytical methods (fixed-effects panel regression). Subsequently, I present *Results* from models inspired by each hypothesis. The *Discussion* examines implications on the hypotheses; followed by the study's main *Limitations*. The *Conclusions* tie together IO doctrines, ILSA designs and data, as well as repetition rates, as outcomes of policy decisions.

4.2 Theoretical framework(s)

This section presents three theoretical frameworks, coupled with hypotheses about ILSA/policy links, namely: world society theory (WST), and policy scripts; educational effectiveness research (EER), and policy lessons; and finally, the sociology of quantification, and policy incentives. Each subsection briefly describes the theory; connects it to the ILSA debate; presents the hypothesis linking ILSAs and policy, generally, and focused on grade retention. A useful distinction should manage expectations, setting the scene for feasible empirical research: impact is ‘a difference in kind while influence is defined as a difference in degree’ (Rutkowski et al. 2020, p. 388). These scholars seek signs of influence, rather than impact, which is more elusive: ‘Policy influence can be viewed as the use of ILSA results to buttress or tweak policy settings that already exist’ (Rutkowski et al. 2020, p. 388). For instance, grade repetition is pervasive, but rates vary over time, thus illustrating Rutkowski and colleagues’ ‘tweaks’.

World society theory and the policy scripts hypothesis

According to WST, national policies are ‘depicted as if they were autonomous decisions because nation-states are defined as sovereign, responsible, and essentially autonomous actors. Taking into account the larger culture in which states are embedded, however, the policies look more like enactments of conventionalized scripts’ (Meyer et al, 1997, p. 159). WST highlights culture and explains the evolution of educational practices and policies over time, or their variations across geographic space, in terms of the norms and values deemed to underlie those practices. In this view, ‘the institutional system—the organizations and cultural meanings that write and rewrite the scripts—becomes central’ (Meyer, 2010, p. 4). For WST, education is a particularly important field of research, as it ‘is globally cast as the key to progress, or excellence, and justice, or equality’ (Ramírez et al., 2016, p. 43).

WST has linked participation in ILSAs to universalistic values that also underlie, or stand in contradiction with, other education policies: ILSAs ‘reflect a proactive commitment to global ideals of educational equality in world society’ (Furuta 2021, p. 87). This cultural tide change meant that, as allegedly egalitarian ILSAs expanded, other assessment practices dwindled, especially those traditionally ‘used to stratify and limit students’ opportunities’ (Furuta 2021, p. 87), e.g., primary school examinations, thus yielding a negative association between these two practices. It is thus crucial to determine which values (e.g., equity, efficiency, etc.) different practices represent, or which scripts countries perform when implementing certain policies.

The first hypothesis focuses on policy scripts: ILSA participation correlates positively with policies that reflect the same values attributed to ILSAs, and negatively with policies deemed lacking in them. Furuta (2021) sees participation in ILSAs as fuelled by a normative belief associated with universalistic and egalitarian conceptions of education. Crucially, WST scholars like Ramírez and colleagues (2018) and Furuta (2021) tend to treat ILSAs as homogeneous, a topic the Discussion will revisit. Meanwhile, grade repetition has long been deemed inequitable (Schiefelbein, 1975). From this perspective, if egalitarian values did drive the relationship, then participation in ILSAs should correlate with decreases in repetition rates. This is the policy-specific scripts hypothesis:

Hypothesis 1: Participation in ILSAs is negatively associated with repetition rates.

Educational effectiveness research (EER) and the policy lessons hypothesis

In the late 20th century, school effectiveness studies, aimed at explaining why some schools outperform others, expanded into educational effectiveness research (EER), covering a wider range of topics. They simultaneously moved from input-output studies, a.k.a. education production function studies (Kyriakides & Charalambous 2014), to inputs-process-output (Reynolds et al. 2014) and even context-inputs-process-output studies (Creemers & Scheerens, 1994). The 1990s saw an internationalization of EER studies (Reynolds et al.

2014) that foretold the advent of cyclical ILSAs. Finally, some EER scholars, joined by ILSA advocates, see the latter as a natural continuation of the former (Strietholt et al. 2014).

Therefore, ILSA 'data allow us to study the effectiveness of policy characteristics of educational systems by means of international comparisons' (Reynolds et al 2014, p. 209); a notion linked to their periodicity: 'as these data assessments take place regularly (…), it is possible to study changes over time of countries' educational effectiveness.' (Reynolds et al 2014, p. 210). Some ILSA advocates even claim that cyclical ILSAs 'have a longitudinal component at the system level. Unlike previous cross-sectional comparisons, such longitudinal designs allow researchers to estimate causal effects of changes in educational policies & other factors at the system level' (Strietholt et al 2014, p. 11), thus making ILSAs legitimate EER studies.[33] These lessons could prompt system-oriented changes (Pizmony-Levy 2018): if countries learn about policy from the ILSAs they join, the resulting empirical findings & policy directives could influence their policy decisions. If all ILSAs produce similar results about a policy's effectiveness, their influence should be similar. However, if their lessons differ, their policy influence could also diverge.

The second hypothesis focuses on policy lessons: an ILSA's influence on a policy depends on its findings on that policy's effectiveness -or other desirable properties. I assume that countries are more likely to learn lessons from the ILSAs they join, as opposed to from others. Therefore, participation in age-based ILSAs like PISA, harsher on grade retention, reduces repetition rates faster, yielding a negative association for both primary & secondary, regardless of in which level the ILSA is administered. Analogously, grade-based ILSAs like

[33] However, others claim that ILSAs' methodological limitations make them unfit for this purpose (Rutkowski & Rutkowski 2016).

ERCE & TIMSS[34], more lenient on grade retention, reduce repetition rates more slowly, thus yielding a positive association, once again, for both schooling levels. This is the policy-specific lessons hypothesis:

Hypothesis 2a: Participation in age-based ILSAs (e.g., PISA) is negatively associated with repetition rates.

Hypothesis 2b: Participation in grade-based ILSAs (e.g., ERCE) is positively associated with repetition rates.

The sociology of quantification & the policy incentives hypothesis

Linked to the sociologies of knowledge (Espeland & Stevens, 2008; Berman & Hirschman 2018) and expertise (Eyal 2019), it intersects substantially with politics and policy, since it mostly analyses the production of official statistics (Desrosières 2015). Here, the Devil is in the methodological details, but from a sociological, not prescriptive viewpoint. Methods are at once socially configured and socially consequential (Williamson, 2018); I will focus on the latter here, through the concepts of retroaction (Desrosières, 2015) and reactivity (Campbell, 1957). 'When a measure becomes a target, it ceases to be a good measure' (Strathern, 1996, p. 4); because it invites stakeholders to change behaviour in strategic ways that improve performance as measured, but not necessarily the underlying system in meaningful, sustainable ways. 'Actors may stick to an indicator's letter while ignoring its spirit, allowing its technicalities to conceal the underlying reality it was designed to reveal' (Cardoso 2020, p. 602). However, to game an ILSA, actors must first learn its rules.

[34] The international version of the TIMSS student questionnaire does not include questions on repetition. However, some countries, e.g., Chile in 2015, do include them, and prominently feature this topic in their national results report: http://www.revistadeeducacion.cl/wp-content/uploads/2016/11/TIMSS-PRESENTACION.pdf

EER argues that recurring ILSAs enable causal analyses (Strietholt et al, 2014), while the sociology of quantification reframes this regularity in terms of reactivity: while EER focuses on what we learn from an ILSA, the sociology of quantification focuses on what we learn about an ILSA. As stakeholders realize how a specific ILSA works, they acquire the means (i.e., the knowledge) and the motive (i.e., the ambition to improve the country's standing in league tables) to game it. This resembles the attempted reverse engineering of law school rankings (Espeland et al., 2016). ILSA-oriented changes are timebound and aim at improving a country's performance in the next cycle of a well-known ILSA (Pizmony-Levy, 2018), through either resource reallocation or student selection (Jennings, 2009).

This study focuses on selection practices that exclude low-performing students: 'Highlighted undesirable consequences associated with practices of the global testing culture include (…) shaping the testing pool by transferring students into non-tested student groups' (Smith, 2019, p. 123). If all ILSAs incentivized countries to select students in similar ways, their medium-term policy influence should be similar. For instance, ILSAs alike in comparison strategy, level and grade incentivize countries to use similar strategies to shape the testing pool. However, countries respond with different strategies to ILSAs whose designs provide different perverse incentives. Therefore, my third hypothesis focuses on policy incentives: in the medium term, an ILSA's influence on a policy depends on how the latter's implementation would affect the country's performance in that ILSA's upcoming cycle; consequently, this influence focuses on grades or ages preceding the assessment or, even more surgically, on the student cohort scheduled to participate.

An ILSA's comparison strategy (Allen et al., 2009) influences actions that shape the testing pool (Smith, 2019). To exclude low-performing students from ERCE in 6th grade, one can increase repetition rates in primary, before students are assessed; in secondary school, this is futile. ERCE provides a perverse incentive (Desrosières, 2015) to retain even more students, but only in grades preceding, not following, the assessment. With PISA, one cannot use this strategy to prevent low-performing students from being assessed, as they inevitable will be at 15, regardless of grade, unless they drop out.[35] This incentivizes promoting them to give them the opportunity to learn further content. As with ERCE, incentives only apply to grades preceding the assessment. But since PISA occur towards late lower secondary, it incentivizes reducing repetition in both levels. This is the policy-specific incentives hypothesis:

Hypothesis 3a: Participation in age-based ILSAs (e.g., PISA) is negatively associated with repetition rates in grades and levels preceding the assessment.

Hypothesis 3b: Participation in grade-based ILSAs (e.g., ERCE, TIMSS) is positively associated with repetition rates in grades and levels preceding the assessment.

Are these three explanations complementary or contradictory? The answer is rooted in the complex, multifarious role that ILSAs play in relation to policy. First, participating in ILSAs means implementing a policy that represents certain values, such as egalitarianism, according to Furuta (2021); this is how WST and the policy scripts hypothesis conceptualize the relationship. Second, aside from being policy, ILSAs generate knowledge -whether methodologically legitimate or not- about other policies and can therefore influence their adoption; this is EER's policy lessons hypothesis. Finally, as they become periodical, ILSAs monitor overall system performance, through specific mechanisms that are increasingly understood by interested actors, thus incentivizing policy tweaks geared towards so-called

[35]However, PISA's standard target population includes students in grade 7 or higher, which means that 15-year-old students in primary in Latin America are excluded. https://www.oecd.org/pisa/pisaproducts/MAIN-SURVEY-SCHOOL-SAMPLING-PREPARATION-MANUAL.pdf

quick wins in an upcoming ILSA cycle; this is a view informed by the sociology of quantification.

4.3 Background

This section revolves around an article (Ramírez et al., 2018) on the links between large-scale assessments and education policies, focusing on three of its findings and one recommendation. That article, like this one, does not analyse grade retention policies themselves, but repetition rates as their outcome and proxy. These are its findings: overall, repetition rates are negatively associated with ILSAs; repetition rates are not significantly associated with national assessments; and finally, the association between ILSAs and primary repetition rates is negative but only marginally significant. The section also explores the debates around these findings (Smith, 2019), and their implications from the perspectives of each of the three theoretical frameworks analysed here, including Ramírez and colleagues' WST.

Finding #1: ILSAs are negatively associated with repetition rates

Responding to critics of the global testing culture (Smith, 2016), these scholars hypothesized that 'primary and secondary repeat rates (…) could be driven up if the testing regime places pressures on nations to establish more stringent standards' (Ramírez et al, 2018, p. 354). However, their own findings supported the opposite conclusion: 'international testing appears to be associated with lower repeat rates at the primary and secondary levels' (Ramírez et al, 2018, p. 354). Why? The authors refrain from venturing a specific explanation, focusing instead on how the feared Draconian effects of the global testing regime have failed to materialize.

To explain this finding from the WST perspective, I explore how global scripts portray, on the one hand, ILSAs; and on the other hand, grade repetition. For instance, WST links ILSAs to universalistic, egalitarian views of education (Furuta 2021). Meanwhile, grade repetition has been accused of being inequitable (Cardoso 2020). Therefore, for WST, egalitarian global scripts would simultaneously lead to expanding ILSAs and decreasing repetition rates. This explains the negative association between ILSAs and repetition rates in a cross-national analysis (Ramírez et al, 2018).

Recommendation: pay attention to different types of assessments

While Ramírez et al. treat ILSAs as monolithic, they provide advice that this study takes to heart: 'pay more attention to how different types of tests are associated with various educational outcomes' (Ramírez et al, 2018, p. 358). This study addresses a distinction pondered by grade repetition experts (Allen et al, 2009) and ILSA practitioners (McGaw, 2008b; Wagemaker, 2008) alike. In grade retention research, the 'comparison strategy' distinguishes between studies that, post repetition, contrast repeaters' performance to two different comparison groups: either their new classmates, i.e., grade-based comparison; or their age-cohort peers, i.e., age-based comparison (Allen et al., 2009). The concept fuelled a rift between advocates of age-based PISA (McGaw, 2008a; McGaw, 2008b) and grade-based IEA studies (Wagemaker, 2008).

This deceivingly trivial methodological distinction matters to both EER and the Sociology of Quantification, but for different reasons. To EER, it matters simply because, as US studies on grade retention predict (Allen et al., 2009), lessons on grade retention effectiveness from each ILSA will depend on their respective comparison strategies, as shown by Cardoso (2020). Countries take these lessons at face value may implement system-oriented changes (Pizmony-Levy, 2018), to achieve lasting improvements in a country's education. Thus, as will be seen, countries participating in age-based ILSAs reduce their repetition rates faster than those participating in grade-based ILSAs.

Meanwhile, for the sociology of quantification, methods are socially configured and socially consequential (Williamson, 2018). ILSAs' comparison strategies are socially configured by the education doctrine of the IO behind it, and socially consequential because: (1) their findings regarding grade retention's effectiveness, even if only statistical artifacts, significantly diverge from each other; and (2) they reinforce its parent IO's prior bias about that policy, in a form of policy evidence by design (Cardoso 2020). Crucially, whether countries take these lessons at face value in implementing system-oriented changes, or instead react to the idiosyncrasies of the assessments' designs by implementing ILSA-oriented changes (Pizmony-Levy, 2018), the actual policy responses can be similar. Thus, for the sociology of quantification, the provisory expectation is, for different reasons, the same as for EER: countries in age-based ILSAs reduce repetition faster than those in grade-based ILSAs.

Finding #2: national assessments are not significantly associated with repetition rates

But does this diverging policy-relevant evidence lead to actual policy differences? Another finding proves crucial when examined through the lens of the comparison strategy (Allen et al., 2009). Ramírez and colleagues find that 'The coefficients for national testing are essentially zero' (Ramírez et al, 2018, p. 354). National and international assessment differ in many ways, but I will once again focus on the comparison strategy (Allen et al., 2009). The most pervasive ILSA, namely PISA, is age-based. Meanwhile, national assessments tend to be grade-based,[36] thus yielding less unfavourable lessons on repetition's effectiveness than international assessments. Therefore, from an EER perspective, it is plausible for national assessments to have no effect on repetition. I will revisit this finding from the perspective of the sociology of quantification after introducing the third and last finding.

Finding #3: ILSAs' negative association with primary repetition rates is marginally significant

Finally, the regression coefficients for repetition rates on ILSAS 'are substantively very small—and in the case of primary repeat rates, only marginally significant' (Ramírez et al, 2018, p. 354). Why this difference across schooling levels? I will once again explore the role of the comparison strategy (Allen et al, 2019); this time, coupled with the concepts of ILSA-oriented changes (Pizmony-Levy, 2018) and selection practices (Jennings, 2009). For critics of Ramírez et al. (2018), 'repetition rates are a potential variable of interest' (Smith, 2019, p. 123) when countries shape the ILSA testing pool. Age-based assessments may incentivize countries to further lower repetition prior to the assessment, while grade-based assessments provide the opposite incentive, if not to increase rates, to slow their decrease (Cardoso 2020).

How do primary and secondary ILSAs' comparison strategies differ? ILSAs in primary school are grade-based.[37] Therefore, they incentivize participating countries to retain

[36] An exception is PCAP (Pan-Canadian Assessment Program), which targets 13-year-olds two years before PISA. http://www.cmec.ca/Publications/Lists/Publications/Attachments/299/PCAP2007_Jurisdictional_Web_EN.pdf

[37] https://www.acer.org/gb/gem/gem-la

low-performing students before they make it to the assessment grade. This may lead countries to slow down the reduction of repetition rates in primary grades, but would not affect rates in secondary, which would continue to decrease rapidly. Since Ramírez et al. (2018) distinguish between repetition rates in primary and secondary but do not distinguish ILSAs by level, this relationship may be obscured.

Meanwhile, in lower secondary, there are two main ILSAs: PISA, age-based; and TIMSS 8th grade. Once again, following Cardoso's (2020) reasoning, PISA participation incentivizes countries to reduce repetition rates before students turn 15, when PISA assesses them. Conversely, TIMSS participation incentivizes countries to maintain or increase repetition rates in grades 1st to 7th. However, because PISA is more widespread and frequent, it drives the relationship between overall ILSA participation and repetition rates, making it negative. Because the lead-up to these assessments includes primary and secondary grades, both levels are affected. But primary-level, grade-based ILSAs counteract this influence over primary rates, perhaps rendering it only marginally significant overall. Once again, because Ramírez et al. (2018) do not distinguish ILSAs by level, this relationship may be obscured.

If all the above hypotheses held, what would Ramírez et al.'s (2018) results look like? In primary, IEA studies and regional assessments -all grade-based-, would drive repetition up, while PISA drives it down, resulting in a muddled picture: a marginally significant negative relationship between ILSA participation and primary repetition rates. Primary level ILSAs would not influence lower secondary, which leaves PISA and TIMSS 8th grade to arm-wrestle each other, perhaps resulting in a statistically significant negative association between ILSA participation and secondary repetition rates, but with a regression coefficient that is not substantial.

In summary, these hypotheses explain Ramírez et al.'s (2018) findings. To advance this research agenda, the present study will make more specific predictions by introducing two new distinctions regarding ILSAs: by comparison strategy, i.e., grade- and age-based ILSAs (Allen et al., 2009); and by level, i.e., ILSAs administered in primary or secondary. I therefore distinguish three types: (1) grade-based ILSAs in primary (e.g., ERCE, PIRLS, or TIMSS 4^{th} grade); (2) grade-based ILSAs in secondary (i.e., TIMSS 8^{th} grade); and (3) age-based ILSAs in secondary (i.e., PISA). A fourth type is theoretically possible, but no age-based assessment targets mostly primary grades.

4.4 Data and methods

Following Ramírez et al. (2018), this study does not examine the grade retention policies in the books, but instead focuses on repetition rates as an outcome of, or a proxy for, those policies. This study analyses publicly available data from 18 Latin American countries on repetition rates by ISCED[38] level (see Figure 1 and 2) between 1992 and 2017 according to participation in ILSAs:

- Between 1995 and 2016, six countries participated in IEA studies PIRLS (mostly in 4^{th} grade) and TIMSS (mostly in 4^{th} or 8^{th} grade).
- Between 1997 and 2013, UNESCO implemented three rounds of ERCE, covering 18 countries between them. ERCE targets mostly 3^{rd} and 6^{th} grade.
- Between 2000 and 2015, OECD implemented six rounds of PISA, covering 10 countries between them. PISA targets 15-year-old students.

Geographic scope

[38] International Standardized Classification of Education.

Latin America stands out due to its relationship with repetition. First, it is second only to sub-Saharan Africa when it comes to repetition rates (UIS, 2012). Second, at the same time, from 2000, when it had the world's highest rates, to, 2010, it experienced the sharpest reduction in repetition rates of any region (UIS, 2012). Finally, grade repetition has been abundantly debated in Latin America for nearly half a century (Schiefelbein, 1975). Additionally, Latin America has high ILSA participation and overall data availability for educational and economic indicators, as compared to the rest of the developing world.

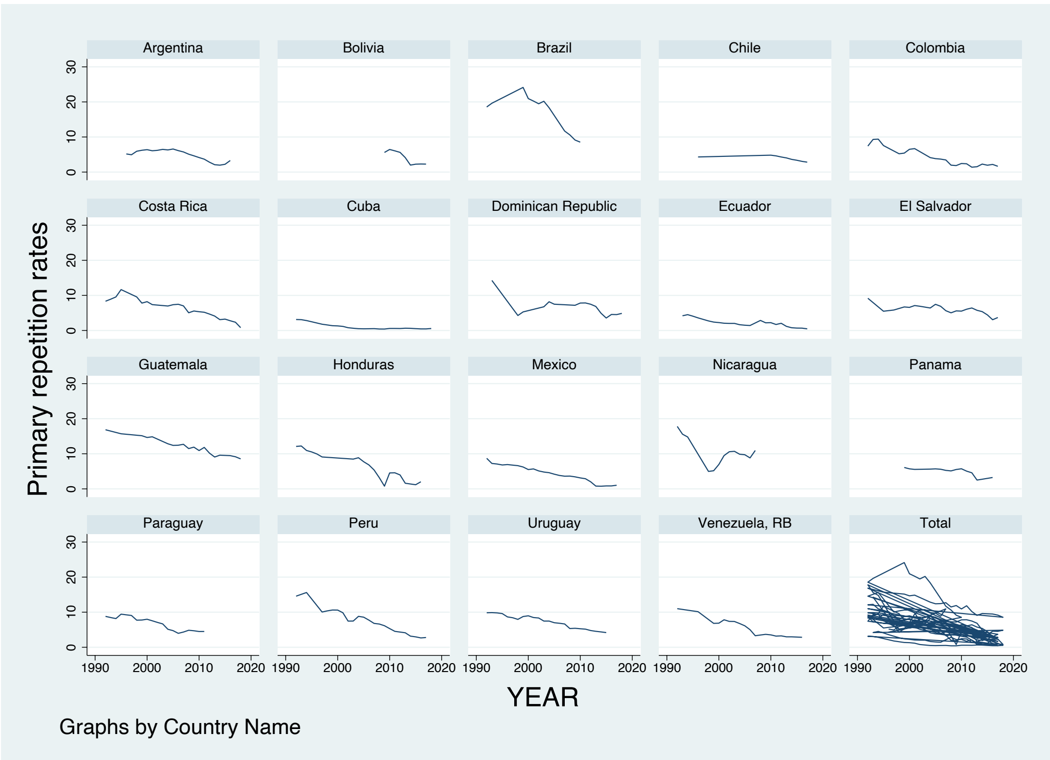


Figure 1. Primary repetition rates by year. Latin America, 1992-2018.

Dataset and variables

The dataset is the World Development Indicators (WDI) database, hosted by the World Bank; most of its education indicators are originally compiled by the UNESCO

Institute for Statistics (UIS) directly from Ministries of Education. The outcomes are repetition rates in ISCED levels that precede PISA: WDI series *Repetition rate in primary education (all grades), both sexes (%)* and *Repetition rate in lower secondary general education (all grades), both sexes (%)*. Both these measures were logged to reduce skewness. Secondary repetition rates became widely available in 1999, after ISCED 1997 split secondary into two levels. For primary, since ILSAs in Latin America started in 1995, I chose 1992, a high data availability year (13 countries), as baseline. Both series end in 2017, the last widely available year.

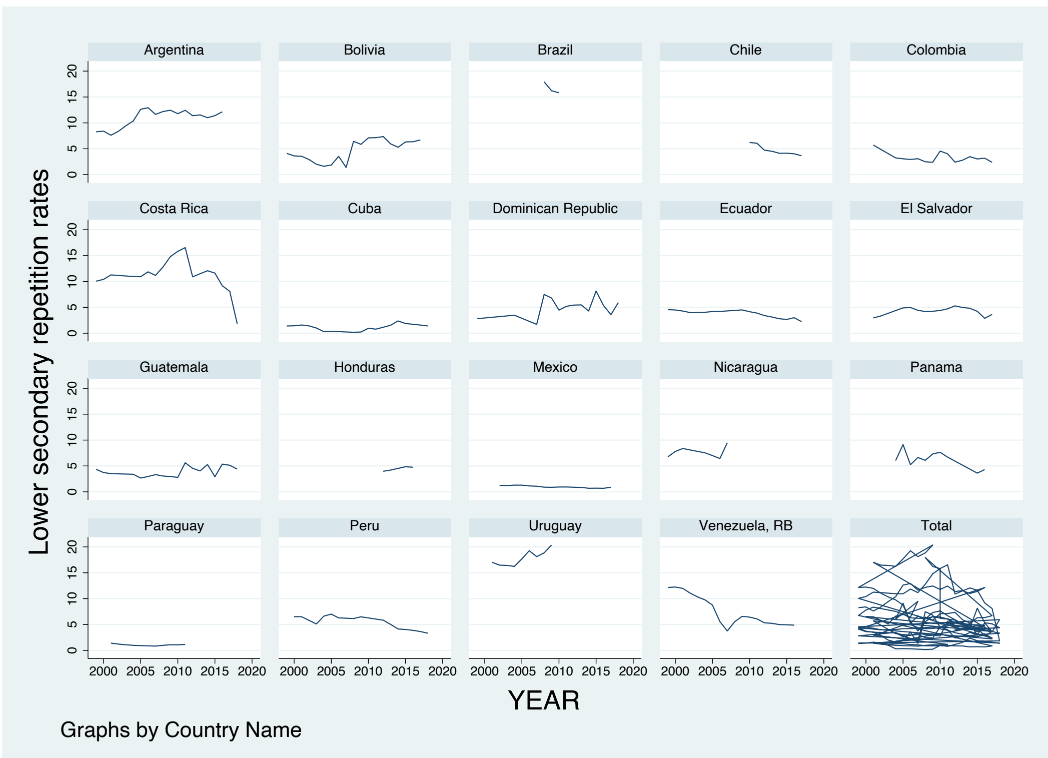


Figure 2. Lower secondary repetition rates by year. Latin America, 1998-2018.

All independent variables (see Table 1) indicate participation in ILSAs: PISA, ERCE, PIRLS and TIMSS (4th and 8th grade). Instances of participation in the previous ten years were summed for each ILSA to account for their lingering effects, following Ramírez et al.

(2018). To replicate their analyses, an overall composite variable was created, namely ILSAs, informed by WST. The two additional composites are informed by the remaining theoretical frameworks: Grade-based ILSAs in primary and secondary should similarly influence primary repetition rates according to both EER and the sociology of quantification; while Grade-based ILSAs in primary create incentives for countries to reduce their repetition rates in primary, but not in secondary, according to the sociology of quantification.

TABLE 1
INSTANCES OF PARTICIPATION IN SELECTED ILSAS.
LATIN AMERICA, 1992-2017.

	ERCE	PIRLS	PISA	TIMSS 4th	TIMSS 8th	Country totals
Argentina	3	1	5	0	2	11
Bolivia	1	0	0	0	0	1
Brazil	3	0	6	0	0	9
Chile	3	1	5	2	4	15
Colombia	3	2	4	1	2	12
Costa Rica	3	0	3	0	0	6
Cuba	2	0	0	0	0	2
Dominican Republic	3	0	1	0	0	4
Ecuador	2	0	0	0	0	2
El Salvador	1	0	0	1	1	3
Guatemala	2	0	0	0	0	2
Honduras	2	0	0	1	1	4
Mexico	3	0	6	1	1	11
Nicaragua	2	0	0	0	0	2
Panama	2	0	1	0	0	3
Paraguay	3	0	0	0	0	3
Peru	3	0	4	0	0	7
Uruguay	2	0	5	0	0	7
Venezuela	0	0	0	0	0	0
ILSA totals	43	4	40	6	11	

.

Three control variables were included.[39] First, since primary repetition rates trend downward in Latin America (see Figure 1), Year accounts for the passage of time. In the next two controls, I follow both Kijima (2010) and, again, Ramírez et al. (2018). Gross domestic product (GDP) is a frequent control in cross-national education analyses; here, it is expressed in purchasing power parity terms and logged to reduce skewness. Secondary Gross Enrolment Ratio (GER) is also controlled for; admittedly, this indicator is problematic, especially here, since repetition itself can inflate it. However, the better indicator, Net Enrolment Rate (NER), is rarely available.

Analytical approach

This study uses fixed effects regression models (*xtreg, fe* in Stata), estimated separately for each of the two ISCED levels under analysis, namely primary and lower secondary. The base model, upon which variations are built, adopts this form:

$$RepRate_L = \beta_0 + \beta_1 ILSAs + \beta_2 GDPPPP(log) + \beta_3 SGER + \beta_4 YEAR + \varepsilon \quad (1)$$

where sub-index L denotes ISCED Level (either ISCED 1=primary, or 2=lower secondary). Missing data were deleted listwise by country/years, thus excluding Cuba, without GDP data.[40]

[39] Some WST studies (see Kijima 2010) use embeddedness in IOs to predict education outcomes. However, Ramírez et al. (2018) do not, since their specific focus, like mine, is on ILSAs as predictors of education policies. And while including embeddedness in IOs might in theory explain away the overall association between ILSAs and repetition that Ramírez et al. postulate and find, my focus is even narrower, as I delve into different types of ILSAs.

[40] Outputs with Cuba, without GDP, with similar findings, are available upon request.

4.5 Results

This section presents results from fixed-effects panel regression models for primary (Table 2) and lower secondary (Table 3) repetition rates on participation in ILSAs. Each Table presents three models, analogously built: Model A replicates Ramírez et al.'s (2018) analyses,[41] combining all ILSAs; Model B introduces PISA individually, alongside the grade-based ILSAs composite; and finally, Model C introduces PISA and TIMSS 8th grade individually, alongside the composite Grade-based ILSAs in primary. As will be seen, Model A is inspired by WST and the policy scripts hypothesis, while Models B and C, jointly, test the policy lessons and incentives hypotheses, respectively inspired by EER and the sociology of quantification.

In Table 2, Model A shows no link between ILSAs and primary repetition rates, for which Ramírez et al. (2018) found a marginally significant negative association. Across the two remaining models, however, PISA does have a negative, highly significant (α=.01) association with the outcome. In Model B, Grade-based ILSAs in primary and secondary has a positive, statistically significant association (α=.05) with the outcome, contrasting with PISA. Model C introduces a final composite, Grade-based ILSAs in primary, alongside TIMSS 8th grade; both coefficients are positive, as expected, but not significant. The Discussion addresses how Model B results fit both the policy lessons and incentives hypotheses, respectively informed by EER, and the sociology of quantification.

[41] However, there are important differences: their geographic and chronological scope is broader; my ILSAs dataset more focused on primary and lower secondary.

TABLE 2

FIXED EFFECTS PANEL REGRESSION OF PRIMARY REPETITION RATES (LOGGED) ON PARTICIPATION IN SELECTED ILSAS IN PAST 10 YEARS. LATIN AMERICA, 1992-2017 (18 COUNTRIES, 297 OBSERVATIONS).

Models	A			B			C		
	Coeff.		S.E.	Coeff.		S.E.	Coeff.		S.E.
International assessments (ILSAs)	0.00		0.01						
PISA				-0.03	**	0.01	-0.03	**	0.01
Grade-based ILSAs in primary and secondary				0.02	*	0.01			
Grade-based ILSAs in primary							0.01		0.01
TIMSS 8th grade							0.03		0.03
Secondary Gross Enrolment Ratio	0.00	*	0.00	0.00		0.00	0.00		0.00
GDP (PPP, log)	0.10		0.19	0.13		0.19	0.14		0.19
Year	-0.03	***	0.00	-0.03	***	0.00	-0.03	***	0.00
Constant	59.36	***	5.95	53.90	***	5.98	53.84	***	5.99
Within R^2	0.676			0.692			0.693		
Between R^2	0.037			0.019			0.015		
Overall R^2	0.267			0.286			0.280		

NOTE: + P < .10. * P < .05. ** P < .01. *** P < .001.

Finally, Table 3 presents analogous models for secondary repetition rates. Model A's results approach Ramírez et al.'s (2018): ILSA participation is linked to decreasing repetition, although here the association is only marginally significant. In Model B, PISA's coefficient is negative and statistically significant (α=.05), but in Model C it is only marginally significant. Similarly to Table 2, grade-based ILSAs are introduced through an overall composite in Model B; and through composite Grade-based ILSAs in primary alongside TIMSS 8th grade in Model C. None of their coefficients is statistically significant. However, these results still offer food for thought: the coefficients for the grade-based composite and the primary grade-based composite, flip from positive on Table 2 (primary) to negative on Table 3 (secondary). Meanwhile, TIMSS 8th grade's coefficients remain positive throughout. Although these results lack statistical significance, they fit the policy incentives hypothesis, and call for replicating these analyses elsewhere.

TABLE 3
FIXED EFFECTS PANEL REGRESSION OF LOWER SECONDARY REPETITION RATES (LOGGED) ON PARTICIPATION IN SELECTED ILSAS IN PAST 10 YEARS.
LATIN AMERICA, 1999-2017 (18 COUNTRIES, 227 OBSERVATIONS).

Models	A			B			C		
	Coeff.		S.E.	Coeff.		S.E.	Coeff.		S.E.
International assessments (ILSAs)	-0.02	+	0.01						
PISA				-0.04	*	0.02	-0.03	+	0.02
Grade-based ILSAs in primary and secondary				-0.01		0.01			
Grade-based ILSAs in primary							-0.01		0.02
TIMSS 8th grade							0.02		0.03
Secondary Gross Enrolment Ratio	-0.01	***	0.00	-0.01	***	0.00	-0.01	***	0.00
GDP (PPP, log)	0.24		0.27	0.25		0.27	0.25		0.27
Year	0.01	+	0.00	0.01	*	0.00	0.01	*	0.00
Constant	-13.54	+	7.45	-17.15	*	7.78	-16.87	*	7.79
Within R^2	0.089			0.100			0.103		
Between R^2	0.069			0.089			0.088		
Overall R^2	0.006			0.007			0.005		

NOTE: + $P < .10$. * $P < .05$. ** $P < .01$. *** $P < .001$.

4.6 Discussion

This section revisits the findings and explores how each hypothesis conceptualizes the issue of possible differences between ILSAs in their ideology, methodology, and empirical results. In Table 3, Model A's results fit Furuta's (2021) characterization of ILSAs as egalitarian, which explains their negative association with repetition, deemed inequitable (Schiefelbein, 1975). This supports WST's interpretations focused on culture and values. In theory, if different ILSAs represented different values, their relationship with policy could diverge. But WST scholars explicitly dismiss tensions between values in global scripts, as they 'see education, simultaneously human capital and a human right, is constructed as a means to fulfil both visions of development' (Chabott, 2003, Kindle Location 451). In those scripts, 'education is globally cast as the key to progress, or excellence, and justice, or equality' (Ramírez et al., 2016, p. 43). From this perspective, ILSAs can be treated as monolithic.

Meanwhile, EER scholars assert ILSAs' neutrality: 'International achievement data (…) were not collected to support any specific economic research agenda' (Hanushek & Woessmann 2014, p. 147). Neither do these scholars dwell on methodological differences between ILSAs. However, they do not explicitly rule out the possibility that different ILSAs may yield different lessons. Therefore, Models B and C explore ILSAs' diversity. In Table 2, focused on primary, Model B challenges the policy scripts interpretation. As previously seen, ILSAs with different comparison strategies yield different empirical findings regarding grade repetition (Cardoso 2020).

Joining ILSAs that teach diverging lessons about a policy's effectiveness may differently influence how countries tweak existing policy settings (Rutkowski et al. 2020). Operating rationally as per EER, after countries join ILSAs that yield harsh findings about grade repetition, they reduce rates faster. Surprisingly, but analogously, participating in ILSAs that yield more lenient findings about this policy may increase its prevalence, all else equal (overall, repetition rates decrease in most countries, but at different paces). In summary, Model B supports the notion that countries 'tweak' their policies as they learn policy lessons from ILSAs.

From an EER perspective, Table 3, focused on secondary, should present similar results. However, that is not the case. PISA participation remains negatively associated with repetition rates, as per the policy lessons hypothesis. *Grade-based ILSAs in primary* (Model C) should have a positive coefficient; instead, it is negative, albeit not statistically significant. Moreover, in Model C, *TIMSS 8th grade* has positive, although not significant association. The coefficient sign for *Grade-based ILSAs in primary* changes from primary to secondary; not so for *TIMSS 8th grade*. Why? Secondary follows ERCE, PIRLS and TIMSS 4th grade, rather than precede them; therefore, no ILSA-oriented changes (Pizmony-Levy, 2018) are in order for future cycles. Conversely, PISA and TIMMS 8th grade keep pushing, but in different directions. These findings fit the policy incentives hypothesis.

This preliminary evidence points towards the sociology of quantification and its notion of retroaction (Desrosières, 2015). Repeated participation in a recurring ILSA teaches countries the peculiarities of its design, inciting Espeland and colleagues' (2016) reverse engineering. Retroaction should be understood within the broader framework of the sociology of quantification. Not only do ILSAs differ from each other in their design, the data they produce, and the influence they exert on decisions regarding policies like grade retention. They differ from each other in all those ways because, like all methods, they are socially configured and socially consequential (Williamson, 2018). Here, ILSAs are configured by the IOs behind them, with different views of education as either a human right or the engine of human capital. How are they consequential? One way is retroaction, where official statistics pose perverse incentives that prompt actors to strategically exploit their knowledge of superficial attributes of monitoring indicators. Consequently, policy, if still deserving of that name, becomes increasingly timebound and instrumental, nearly surgical.

4.7 Limitations

This section reviews four limitations, the first three focused on the outcome, i.e., repetition rates, and the last one, on the independent variable, namely ILSAs. These are the first three, from general to particular: the focus on policy outcomes, as opposed to processes; the use of repetition rates; and their use by level, rather than by grade. This article focuses on repetition rates as a policy outcome, instead of examining national policy processes. While these matter, they lie beyond this article's scope. However, this calls for further research, since during this period several Latin American countries introduced policy changes: 'In Argentina, Brazil, Chile, Mexico, and Peru, first grade has automatic promotion (…). In practice, this is not fully achieved, since repetition persists due to the resistance of some teachers and management teams' (Rivas, 2016, p. 103; my own translation). Peru introduced that change in 1994; 1st grade repetition rates dropped from 24% in 1992 to just 5% in 1998, while other grades remained unaffected. Policies in the books matter, but practices on the ground are not entirely consistent with them: 1st grade repetition endured. This topic deserves focused study.

Admittedly, repetition rates are imperfect. Stakeholders from national governments to classroom teachers have incentives to inflate or deflate them. IOs like the World Bank (Haddad, 1975) and OECD (Ward, 2018) encourage countries to reduce them by, in theory, retaining fewer students, but in practice may incentivize creative accounting on statistics (Jennings, 2009) without changing practices on the ground. Moreover, do repetition rates, or repetition itself, have the same meaning across countries? Probably not, as practices vary. This case exemplifies a general principle: 'educational systems are not comparable per se, but they are made comparable by using the same measurement' (Cardoso & Steiner-Khamsi, 2017, p. 401).

That said, those concerns apply to all studies involving repetition rates, including Ramírez et al. (2018). However, this also illustrates a strength of approaches focused on ILSAs' performativity. Paradoxically, indirect incentives from age-based ILSAs, not focused on repetition itself, may prompt deeper behavioural change in school systems. While creative accounting on repetition rates cannot change rankings, countries could reverse engineer PISA (Espeland et al., 2016) by reducing actual repetition. Ultimately, 'Seeing like PISA', is characterised by (…) the viewing of education as a global race with winners and losers' (Gorur, 2016, p. 608).

Critics of Ramírez and colleagues (2018) argue that students 'are increasingly held back in the year prior to the test (…). Therefore, without specific information on the tested grade, this analysis (…) does not validly test the critique' (Smith, 2019, p. 123). This also applies here: grade-based ILSAs like ERCE incentivize countries to increase repetition rates in the grades and years preceding the assessment. But in the assessment grade, countries are incentivized to promote low-performing students, especially in the year immediately preceding the assessment, thus reducing that grade's rates. Future studies should examine the ILSAs' influence on repetition rates by grade.

As for the independent variable, we should not forget that "in practice ILSAs are also diverse, global projects, and their characteristics, meanings and effects have to be considered in relation to the different locations, institutions, countries and contexts in which they take place" (Maddox 2018, Kindle Edition). While this study treats each country's participation in each ILSA each year as a dichotomous yes/no variable, there is clearly room for nuance. For instance, when Chile participated in TIMSS in 2015, it collected data on students' repetition history[42] by using questions that were not featured in the global questionnaire. Thus, attempts to understand the methods used by ILSAs "should be organised within an overarching review of the institutional and political arrangements within which the methods are mobilised" (Guadalupe 2017, 335).

4.8 Conclusions

There is no global conspiracy here. ILSAs are not designed with a specific policy question in mind; and optimizing their design to obtain the desired answer to every conceivable policy question would be

[42] http://archivos.agenciaeducacion.cl/presentacion_nacional_de_resultados_TIMSS_2015.pdf

an impossible task, to say nothing of the questionable ethics of such an enterprise. But a conspiracy need not happen for actors to react strategically to a changing environment. National actors may endorse an ILSA if the values it represents, the lessons it yields, or the incentives it provides are aligned with their own agendas. This does not necessarily mean that ILSAs have an independent effect on policy (Steiner-Khamsi, 2019), partly because 'numbers may simply justify decisions that would have been made anyway' (Berman & Hirschman, 2018, p. 258).

In connecting all empirical findings, the emerging picture points to the policy incentives hypothesis, informed by the notion of retroaction (Desrosières, 2015). But these alleged attempts at reverse engineering may not produce the outcomes that actors expect. Are decreasing repetition rates associated with improved performance in future PISA cycles? This is banal: decreasing repetition rates are normally attributed to improved academic achievement. The flipside seems more interesting: Do increasing or more slowly decreasing repetition rates in grades preceding ERCE or TIMSS lead to improved performance in its next cycle? This would be a stunning finding.

Regardless of that cliff-hanger's eventual resolution, there is potential for applying the sociology of quantification to the relationship between ILSAs and education policies. Paradoxically, ILSA proponents (Schleicher, 1999) and opponents (Smith, 2016) agree that ILSAs influence policy; therefore, the debate often focuses on whether that is good or bad, not on whether it happens. Simultaneously, valuable big-picture perspectives like WST may miss nuanced connections between ILSAs and policies. Against this backdrop, the sociology of quantification scrutinizes methods to shed light on actors using their emergent knowledge of those very details for their own purposes. As ILSAs become increasingly visible, and as we learn to see the world the way they do, are we also learning to see right through them, or just policymaking to the test?

Conclusion

Going back to the beginning…

"Follow your passion!" Large-scale assessment (LSA) has been my work for years. "But don't forget your pet peeves…", I might add. Reading about repetition in reports and articles with LSA data, I was often frustrated by simplistic treatments. Then, at a PISA 2015 presentation, I saw a scatterplot on the link between percentage of repeaters and mean science score by country. I immediately tried to explain my concerns to colleagues, but I lacked the vocabulary.

Other LSA findings bothered me before, but this was different, and when I tried to explain it, people looked at me with a mix of puzzlement and indifference. That is what I set out to do: I wanted to explain why that scatterplot bothered me. But first I would have to understand why it did… I first thought it seemed methodologically dubious. But then it dawned on me that what truly bothered me was that this chart was just one way to look at it, no better or worse than other ways that would lead to different policy outcomes. And that was deeply unsettling.

Around the same time, an anonymous reviewer introduced me to Alain Desrosières and the sociology of quantification. He was a sociologist and statistician who, while working for France's national statistical office, had the courage and intellectual honesty -and perhaps the institutional space- to inspect quantification practices through an interpretive and critical lens. His perspective made me realize that large-scale assessment, like other quantification processes, can at once reflect reality and create it. And like any other methods, they can be examined both methodologically -that is, normatively- or sociologically.

The problem with the scatterplot, and with the ILSA behind it, was not necessarily that there was something methodologically wrong, but that they were the results of choices that stemmed from ideological inclinations; and had an impact on the perception of the effectiveness of policies. And the same could be said of a different ILSA or a different chart that were the result of different ideological roots, and different methodological choices, and that potentially might have a different impact on discourse and even policy. The novelty, for me, was this relativistic view of positivistic science. However, this relativism was not absolute (pun intended). The likes of Desrosières (2002) and Porter (1996) find value in scientific and, particularly, statistical endeavors, but warn us that these are, inescapably, social processes.

Revisiting the first article

Article 1, published in 2020 by *Comparative Education Review*, and recipient of awards from AERA and CIES, as well as an honorable mention from CER itself, has taken a life of its own. All this has given me an opportunity to hear scholars' views and reinterpretations of the article, which has been immensely enriching, flattering, unsettling… And more importantly, it has encouraged me to continue to think about the issues. That is why I would like to share some ideas about further research building on the findings from Article 1.

In addition to the comparison strategy (i.e., grade-based v. age-based), whose effect on estimates of the effectiveness of grade repetition is well documented (Alexander et al. 2003; Allen et al. 2009), the actual age difference (15 for PISA; versus a varying but typically younger age in ERCE), or grade difference (6th grade in ERCE; versus a varying but typically higher grade in PISA) may play a role. Crucially, these are also design choices. Moreover, they are consistent with a human-rights focus (e.g., universal primary education, the right to literacy) and a human capital theory approach (e.g., skills upon entering the working age population), respectively. Therefore, differences in repetition-related achievement gaps across ILSAs are still an artifact of ILSA design that can ultimately be traced back to IO doctrines.

How would these design choices to focus on specific ages or grades affect the observed achievement gap between repeaters and non-repeaters across ILSAs? I will consider one such mechanism: dropout. If repetition and dropout are associated (Andrew 2014), attrition could affect PISA even more than it affects ERCE. In theory, some repeaters may drop out after 6th grade, but before turning 15. This may affect the repetition-related achievement gap that is observed in schools participating in PISA, as compared to ERCE's: if lower-performing repeaters are more likely to drop out, or do it sooner (especially, before turning 15), this would narrow the gap in PISA. Therefore, this effect goes in the opposite direction as the comparison strategy hypothesis.

If this is true, then the effect of the comparison strategy is even stronger than previously thought. This can be empirically tested in at least two ways. First, results from PISA for Development (PISA-D), which include out-of-school populations, can be analyzed to explore the effect of attrition by dropout; the four Latin American countries in it also participate in ERCE. Second, one could compare results from ERCE 6th grade in Peru's Young Lives Study (a longitudinal survey with two cohorts, which included in- and out-of-school populations) at 12 years old. This would control for age, while letting comparison strategy vary. The downside: YLS is a convenience, "pro-poor" sample; but sample matching procedures could be used.

Connecting the dots

Article 3 was, chronologically, the second article to be written. It concludes that, in Latin America, there is an association between a country participating in PISA and having lower repetition rates in both primary and lower secondary. Meanwhile, that is not the case with participation in ERCE or other grade-based ILSAs. In fact, participation in grade-based ILSAs is associated with higher repetition rates in primary, albeit not in lower secondary. All of this is consistent with the policy incentives hypothesis, which outperforms the policy scripts and lessons hypotheses. Shifting the focus from policy lessons to policy incentives opens new avenues for research.

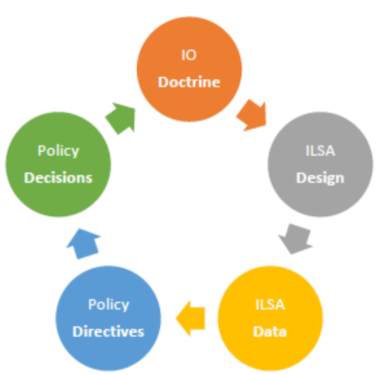
TIMSS and PIRLS do not include questions on grade repetition in their international questionnaires. Therefore, they cannot produce policy discourse, or lessons, on grade repetition based on their data. However, if the hypothesized association is with the incentives that ILSAs provide, rather than the lessons they yield, it would be possible to conduct research that explores the links between participation in TIMSS and PIRLS and tweaks to repetition rates, effectively replicating Article 3 in other regions where there is substantial overlap with PISA, or even globally.

Similarly, it would be possible to transcend Article 3, going one step further. This would entail exploring whether countries benefit from strategic tweaks to repetition rates with a view to improving performance in an upcoming ILSA. It should be possible to do this with all ILSAs, whether they collect data on grade repetition or not. In other words, Article 3 shows that policymaking to the test may actually happen, while an upcoming research project could show whether it works as an ILSA-oriented change (Pizmony-Levy 2018). That said, policymaking to the test does not equate to an independent effect of ILSAs on countries.

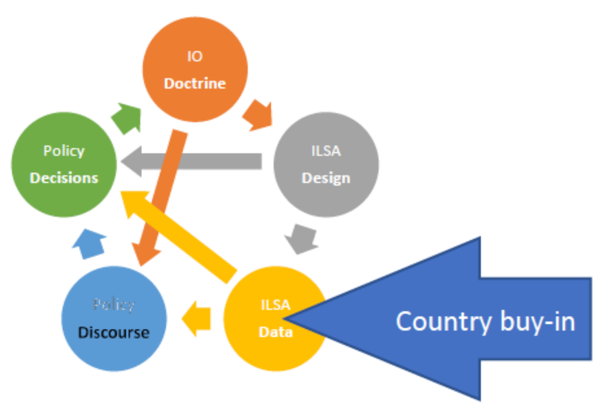
In fact, ILSA results, and incentives, can be used to trigger or accelerate changes or policy reforms that may have been in the making for years, independently of the ILSAs themselves. This is a legitimate empirical question, however challenging it may be to answer: If countries indeed adjust their behavior to improve their standing in an ILSA, one needs to question the notion that this is an asymmetry where countries are deprived of agency. Are they playing the ILSA game? Or is the ILSA a pawn in their own domestic game? These questions can be answered by exploring national policy processes, with a diversified methodological approach.

Speaking of which, article 2, the last to be written, is also the only one that used qualitative methods, specifically, document analysis. A perhaps obvious lesson is that, despite science supposedly being cumulative, the first rule of ILSAs is never to talk about other ILSAs. Therefore, it is up to third-party, independent researchers to connect the dots, bringing a critical, interpretive perspective. A second lesson derived from the process of writing Article 2 is that the process of qualitative inquiry is inherently non-linear and therefore open-ended, with all the risks that entails. Maintaining the focus, while being open to discovery can be a balancing act. Deciding when to stop was just as challenging. Finally, the inter-rater reliability analysis revealed the difficulty in distinguishing between steps in the *Critique* move and their equivalents in the *Results* move. What in a different approach would be dismissed as mere error or a reliability issue in the rubric becomes in this context a subject for future exploration. How does the ambiguity between presenting a statement as literature review or new evidence, or both and neither, play a role in the rhetoric?

Another lesson that I drew from this process is that I probably should have followed the logical order in writing the articles. Article 3 did not benefit from the conclusions presented in Article 2. Whether it would have looked any different is… well, a counterfactual. Looking at all three articles with the advantage of hindsight and going back to the diagram presented in the dissertation proposal, the connections seem now less linear. I would argue that Article 1 did what it set out to do: show the connection between IO doctrines and ILSA designs, and between ILSA design and data. But Article 2 showed that ILSA data and discourse (a broader concept than the old diagram's *directives*) are shaped by factors other than ILSA design and data. IOs' doctrines may influence discourse directly: e.g., OECD's focus on efficiency, for which ILSAs do not provide as much evidence, due to the paucity of expenditure data. Country buy-in appears to enable a different type of discourse on grade repetition, more ordinalized, scientized and allegedly generalizable, with ranked charts, correlation coefficients and scatterplots. Cultural differences regarding what qualifies as scientific discourse in different regions (Abend 2006) may play a part.



Before



After

Article 3, written prior to Article 2, could not incorporate discursive inputs which, admittedly, would have been extremely challenging to integrate into a quantitative approach. It concluded that the policy incentives hypothesis, which connects ILSA designs and policy decisions, received the most empirical support. ILSA data can be hypothesized to mediate in this relationship between ILSA design and policy directives. But it may be possible for ILSA designs to influence policy decisions on grade repetition without being mediated by ILSA data, and this could be empirically tested by replicating Article 3 with TIMSS and PIRLS in other regions, as mentioned above.

If this does not feel like a conclusion, it may be because this process has opened multiple avenues for further exploration. The most immediate items on the agenda would be replications of these findings in other regions, followed by explorations of the effect of repetition rates "tweaks" on countries' performance in ILSAs, as well as studies of the policy processes, intertwined with the ILSA processes, at the country level.

References

Introduction

Addey, C. (2020). The appeal of PISA for Development in Ecuador and Paraguay: theorising and applying the global ritual of belonging. *Compare: A Journal of Comparative and International Education*, *50*(8), 1159-1174.

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, *47*(3), 434-452.

Agasisti, T., & Cordero, J. M. (2017). The determinants of repetition rates in Europe: Early skills or subsequent parents' help? *Journal of Policy Modeling*, *39*(1), 129-146.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31*(4), 480-499.

Andrew, M. (2014). The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career. *Social Forces*, *93*(2), 653-685.

Barro, R., and Lee, J.-W. (1994). Sources of Economic Growth. *Carnegie-Rochester Conference Series on Public Policy*, 40: 1-46.

Bartlett, L., & Vavrus, F. (2017). Comparative case studies: An innovative approach. *Nordic Journal of Comparative and International Education (NJCIE)*, *1*(1).

Byrnes, D. A., & Yamamoto, K. (1985). Academic retention of elementary pupils: An inside look. *Education*, *106*(2).

Cardoso, M. E. (2020). Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition. *Comparative Education Review*, *64*(4), 000-000.

Cordero, J. M., Cristobal, V., & Santín, D. (2018). Causal inference on education policies: a survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys*, *32*(3), 878-915.

Desrosières, A. (2002). *The politics of large numbers: A history of statistical reasoning.* Harvard University Press.

Desrosières, A. (2015). Retroaction: How indicators feed back onto quantified actors. *The world of indicators. The making of governmental knowledge through quantification*, ed. R. Rottenburg, S. E. Merry, S. J. Park, and J. Mugler. Cambridge University Press.

Eide, E. R., & Goldhaber, D. D. (2005). Grade retention: What are the costs and benefits? *Journal of Education Finance*, 195-214.

Espeland, W. N., Sauder, M., & Espeland, W. (2016). *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401-436.

Eyal, G. (2019). *The crisis of expertise*. John Wiley & Sons.

Fourcade, M. (2016). Ordinalization: Lewis A. Coser memorial award for theoretical agenda setting 2014. *Sociological Theory*, *34*(3), 175-195.

Gómez Vera, G. (2013). Los efectos de la repitencia en tanto que política pública en cuatro países del cono sur: Argentina, Brasil, Chile y Uruguay. Un análisis en base a PISA 2009. *Revista Latinoamericana de Educación Comparada* 4 (4): 59-70.

Gorur, R. (2016). Seeing like PISA: A cautionary tale about the performativity of international assessments. *European Educational Research Journal*, *15*(5), 598-616.

Hanushek, E. A., & Woessmann, L. (2014). Institutional structures of the education system and student achievement: A review of cross-country economic research. *Educational policy evaluation through international comparative assessments*, *145*.

Ikeda, M., & García, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. *OECD Journal: Economic Studies*, *2013*(1), 269-315.

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of educational research*, *45*(4), 613-635.

Jimerson, S. R. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of school psychology*, *37*(3), 243-272.

Kamens, D. H., & Benavot, A. (2011). National, regional and international learning assessments: Trends among developing countries, 1960–2009. *Globalisation, societies and education*, *9*(2), 285-300.

Kijima, R. (2010). Why participate? Cross-national assessments and foreign aid to education. In *The impact of international achievement studies on national education policymaking*. Emerald Group Publishing Limited.

Kijima, R., & Lipscy, P. Y. (2017). The politics of international testing. *Available at SSRN 2995226*.

Kijima, R., & Lipscy, P. Y. (2020). International Assessments and Education Policy. *The Power of Global Performance Indicators*, 174.

Komatsu, H., & Rappleye, J. (2017). A new global policy regime founded on invalid statistics? Hanushek, Woessmann, PISA, and economic growth. *Comparative Education*, *53*(2), 166-191.

Maddox, B. (Ed.). (2018). *International large-scale assessments in education: Insider research perspectives*. Bloomsbury Publishing.

Maddox, B., Keslair, F., & Javrh, P. (2018). Investigating testing situations. *International Large-Scale Assessments in Education: Insider Research Perspectives*, 157.

Meyer, H. D., & Benavot, A. (Eds.). (2013, May). PISA, power, and policy: The emergence of global educational governance. Symposium Books Ltd.

Mølstad, C. E., Pettersson, D., & Forsberg, E. (2017). A game of thrones: Organising and legitimising knowledge through PISA research. *European Educational Research Journal*, *16*(6), 869-884.

Pizarro, A., Bilagher, M., Copetta, M., Ortiz, M., Viteri, A., & Riveros, R. (2016). Third Regional Comparative and Explanatory Study (TERCE): Learning Achievements'. *Latin American Laboratory for Assessment of the Quality of Education, UNESCO*.

Pizmony-Levy, O. (2018). Compare globally, interpret locally: international assessments and news media in Israel. *Globalisation, Societies and Education*, *16*(5), 577-595.

Pizmony-Levy, O., & Bjorklund Jr, P. (2018). International assessments of student achievement and public confidence in education: evidence from a cross-national study. *Oxford Review of Education*, *44*(2), 239-257.

Porter, T. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Ramírez, F. O., Schofer, E., & Meyer, J. W. (2018). International tests, national assessments, and educational development (1970–2012). *Comparative Education Review*, *62*(3), 344-364.

Reimers, F. (Ed.). (2000). *Unequal schools, unequal chances: The challenges to equal opportunity in the Americas* (Vol. 5). Cambridge, MA: Harvard University Press.

Rivas, A. (2015). América Latina después de PISA: lecciones aprendidas de la educación en siete países (2000-2015). Fundación CIPPEC.

Rutkowski, L. (2016). *A Look at The Most Pressing Design Issues in International Large-Scale Assessments: A Paper Commissioned by The US National Academy of Education.* Washington, DC: NAED

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252-257.

Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the Policy Influence of International Large-Scale Assessments in Education. *Reliability and Validity of International Large-Scale Assessment*, 261.

Schiefelbein, E. (1975). Repeating: an overlooked problem of Latin American Education. *Comparative Education Review*, *19*(3), 468-487.

Schleicher, A. (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Organisation for Economic Co-Operation and Development, Paris, France

Smith, W. C., & Edwards, D. B. (Eds.). (2016). *The global testing culture: Shaping education policy, perceptions, and practice*. Oxford: Symposium Books.

Sorensen, T. B., & Robertson, S. L. (2020). Ordinalization and the OECD's Governance of Teachers. *Comparative Education Review*, *64*(1), 21-45.

Steiner-Khamsi, G. (2019). Conclusions: What Policy-Makers Do with PISA. *Understanding PISA's attractiveness: Critical analyses in comparative policy studies*, 233.

Strietholt, R., Bos, W., Gustafsson, J. E., & Rosén, M. (Eds.). (2014). *Educational policy evaluation through international comparative assessments*. Waxmann Verlag.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

UIS. (2012). *Global Education Digest 2012. Opportunities Lost: The Impact of Grade Repetition and Early School Leaving*. Quebec: UNESCO Institute for Statistics.

Wagemaker, H. (2013). International Large-Scale Assessments: From Research to Policy. In *Handbook of International Largescale Assessment: Background, Technical Issues, and Methods of Data Analysis,* ed. L. Rutkowski, M. von Davier, M., & D. Rutkowski. Boca Raton, FL: CRC Press.

Williamson, B. (2019). Assessment imaginaries: Methodological challenges of future assessment machines. *International large-scale assessments in education Insider research perspectives*, 3151.

Willms, J. D., & Somer, M. A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *School effectiveness and school improvement*, *12*(4), 409-445.

Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, *47*(3), 434-452.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31*(4), 480-499.

Barro, R., and Lee, J.-W. (1994). Sources of Economic Growth. *Carnegie-Rochester Conference Series on Public Policy*, 40: 1-46.

Bieber, T. (2016). *Soft governance, international organizations and education policy convergence: comparing PISA and the Bologna and Copenhagen processes*. Springer.

Brimer, M., and Pauli, L. (1971). *Wastage in Education, a World Problem: a Study Prepared for the International Bureau of Education*. Paris: UNESCO.

Demanet, J., & Van Houtte, M. (2016). Are flunkers social outcasts? A multilevel study of grade retention effects on same-grade friendships. *American Educational Research Journal*, *53*(3), 745-780.

Desrosières, A. (2002). *The politics of large numbers: A history of statistical reasoning.* Harvard University Press.

Desrosières, A. (2015). Retroaction: How indicators feed back onto quantified actors. *The world of indicators. The making of governmental knowledge through quantification*, ed. R. Rottenburg, S. E. Merry, S. J. Park, and J. Mugler. Cambridge University Press.

Emmerij, L. (1967). Economic Objectives of Education: Reflections on The OECD Experience. *The OECD Observer* 27: 19.

Garfinkle, A. (2015). *Politics and society in modern Israel: Myths and realities*. Routledge.

Gómez Vera, G. (2013). Los efectos de la repitencia en tanto que política pública en cuatro países del cono sur: Argentina, Brasil, Chile y Uruguay. Un análisis en base a PISA 2009. [Effects of repetition as public policy in four South Cone countries: Argentina, Brazil, Chile and Uruguay: An analysis based on PISA 2009]. *Revista Latinoamericana de Educación Comparada* 4 (4): 59-70.

Goos, M., Schreier, B. M., Knipprath, H. M. E., De Fraine, B., Van Damme, J., & Trautwein, U. (2013). How can cross-country differences in the practice of grade retention be explained? A closer look at national educational policy factors. *Comparative Education Review*, *57*(1), 54-84.

Gorur, R. (2016). Seeing like PISA: A cautionary tale about the performativity of international assessments. *European Educational Research Journal*, *15*(5), 598-616.

Green Saraisky, N. (2015). *The politics of international large-scale assessment: The Programme for International Student Assessment (PISA) and American education discourse, 2000-2012* (Doctoral dissertation, Columbia University).

Hanushek, E. A., & Wößmann, L. (2007). The role of education quality for economic growth. *World Bank policy research working paper*, (4122).

Ikeda, M., & García, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. *OECD Journal: Economic Studies*, *2013*(1), 269-315.

Jimerson, S., Carlson, E., Rotert, M., Egeland, B., & Sroufe, L. A. (1997). A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of school psychology*, *35*(1), 3-25.

Jones, P. W. (1990). UNESCO and the politics of global literacy. *Comparative Education Review*, *34*(1), 41-60.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative education review*, *54*(1), 5-25.

Lingard, B., & Sellar, S. (2013). 'Catalyst data': Perverse systemic effects of audit and accountability in Australian schooling. *Journal of Education Policy*, *28*(5), 634-656.

Lintott, H. J. (1949). Machinery for Rebuilding the European Economy: II. The Organization for European Economic Cooperation. *International Organization*, *3*(2), 269-277.

Lorence, J., & Dworkin, A. G. (2006). Elementary grade retention in Texas and reading achievement among racial groups: 1994–2002. *Review of Policy Research*, *23*(5), 999-1033.

Marmot, M. G. (2004). Evidence-Based Policy or Policy-Based Evidence? Willingness to Take Action Influences the View of the Evidence—Look at Alcohol. *BMJ: British Medical Journal* 328 (7445): 906–7.

McGaw, B. (2008a). Further reflections. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 279-282.

McGaw, B. (2008b). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 223-243.

Pizmony-Levy, O. (2013). Testing for all: The emergence and development of international assessment of student achievement, 1958–2012.

Pizmony-Levy, O. (2014). Back to the Future on International Assessments. *Quality Assurance in Education* 22 (4): 321–22.

Porter, T. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Psacharopoulos, G., & Nguyen, N. (1987). The age-efficiency of education systems: an international comparison. In *Discussion Paper: Education and Training Series, No. EDT55*(p. 31). World Bank.

Rose, N. (1999). *Powers of Freedom: Reframing Political Thought*. Cambridge: Cambridge University Press.

Rutkowski, L. (2016). *A Look at The Most Pressing Design Issues in International Large-Scale Assessments: A Paper Commissioned by The US National Academy of Education.* Washington, DC: NAED

Sandoval Hernández, A., Rutkowski, D., Matta, T., and Miranda, D. (2019). Pensémoslo de nuevo: ¿Podemos comparar las escalas de antecedentes socioeconómicos? [Back to the drawing board: Can we compare socioeconomic background scales?]. *Revista de Educación* 383:37–62.

Schiefelbein, E. (1975). Repeating: an overlooked problem of Latin American Education. *Comparative Education Review*, *19*(3), 468-487.

Schlaufer, C. (2018). The Narrative Uses of Evidence. *Policy Studies Journal* 46 (1): 90–118.

Schleicher, A. (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Organisation for Economic Co-Operation and Development, Paris, France.

Sellar, S., and Lingard, B. (2013). PISA and the Expanding Role of The OECD In Global Educational Governance. In *PISA, Power, and Policy: The Emergence of Global Educational Governance*, ed. H.-D. Meyer and A. Benavot. Oxford: Symposium.

Sellar, S., and Lingard, B. (2014). The OECD and the Expansion of PISA: New Global Modes of Governance in Education. *British Educational Research Journal* 40 (6): 917–36.

Steiner-Khamsi, G., & Waldow, F. (Eds.). (2012). *World yearbook of education 2012: Policy borrowing and lending in education*. Routledge.

Tyack, D., and Tobin, W. (1994). The 'Grammar' of Schooling: Why Has It Been So Hard to Change? *American Educational Research Journal* 31 (3): 453–79.

UNESCO. (1990). *Basic Texts*. Paris: UNESCO.

UNESCO Institute for Statistics. (2012). *Global Education Digest 2012. Opportunities Lost: The Impact of Grade Repetition and Early School Leaving*. Quebec: UNESCO Institute for Statistics.

UN (United Nations). (1948). *Universal Declaration of Human Rights*. New York: United Nations.

Wagemaker, H. (2008). Choices and trade-offs: reply to McGaw. *Assessment in Education: principles, policy & practice*, *15*(3), 267-278.

Wagemaker, H. (2013). International Large-Scale Assessments: From Research to Policy. In *Handbook of International Largescale Assessment: Background, Technical Issues, and Methods of Data Analysis,* ed. L. Rutkowski, M. von Davier, M., & D. Rutkowski. Boca Raton, FL: CRC Press.

Willms, J. D., & Somer, M. A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *School effectiveness and school improvement*, *12*(4), 409-445.

Ydesen, Christian, and Grek, S. (2019). Securing Organisational Survival: A Historical Inquiry into the OECD's Work in Education during the 1960s. *Paedagogica Historica* 56 (3), 1–16.

Yore, L. D., Anderson, J.O., and Chiu, M.-H. (2010). Moving PISA Results into the Policy Arena: Perspectives on Knowledge Transfer for Future Considerations and Preparations. *International Journal of Science and Mathematics Education* 8 (3): 593–609.

When ILSAs Met Policy: Evolving Discourses on Grade Repetition

Abend, G. (2006). Styles of sociological thought: Sociologies, epistemologies, and the Mexican and US quests for truth. *Sociological Theory*, *24*(1), 1-41.

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, *47*(3), 434-452.

Agasisti, T., & Cordero, J. M. (2017). The determinants of repetition rates in Europe: Early skills or subsequent parents' help? *Journal of Policy Modeling*, *39*(1), 129-146.

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary school grades*. Cambridge University Press.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates' effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31*(4), 480-499.

Andrew, M. (2014). The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career. *Social Forces*, *93*(2), 653-685.

Auld, E., & Morris, P. (2016). PISA, policy and persuasion: Translating complex conditions into education 'best practice'. *Comparative Education*, *52*(2), 202-229.

Baek, C. (2020). *Knowledge Utilization in Education Policymaking in the United States, South Korea, and Norway: A Bibliometric Network Analysis* (Doctoral dissertation, Teachers College, Columbia University).

Baroutsis, A., & Lingard, B. (2021). A methodological approach to the analysis of PISA microblogs: social media during the release of the PISA 2015 results. *Journal of Education Policy*, 1-21.

Bartlett, L., & Vavrus, F. (2017). Comparative case studies: An innovative approach. *Nordic Journal of Comparative and International Education (NJCIE)*, *1*(1).

Brophy, J. (2006). *Grade repetition*. UNESCO.

Byrnes, D. A., & Yamamoto, K. (1986). Views on grade repetition. *Journal of Research & Development in Education*.

Byrnes, D. A., & Yamamoto, K. (1985). Academic retention of elementary pupils: An inside look. *Education*, *106*(2).

Cardoso, M. E. (2020). Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition. *Comparative Education Review*, *64*(4), 598-618.

Cardoso, M., & Steiner-Khamsi, G. (2017). The making of comparability: Education indicator research from Jullien de Paris to the 2030 sustainable development goals. *Compare: A journal of comparative and international education*, *47*(3), 388-405.

Chabbott, C. (2013). *Constructing education for development: International organizations and education for all*. Routledge.

Coleman, D., & Jones, P. W. (2004). *The United Nations and education: Multilateralism, development and globalisation*. Routledge.

Desrosières, A. (2010). *La politique des grands nombres : Histoire de la raison statistiques.* La Découverte/Poche.

Desrosières, A. (2015). Retroaction: How indicators feed back onto quantified actors. *The world of indicators. The making of governmental knowledge through quantification*, ed. R. Rottenburg, S. E. Merry, S. J. Park, and J. Mugler. Cambridge University Press.

Eide, E. R., & Goldhaber, D. D. (2005). Grade retention: What are the costs and benefits? *Journal of Education Finance*, 195-214.

Emmerij, L. (1967). Economic Objectives of Education: Reflections on The OECD Experience. *The OECD Observer* 27: 19.

Espeland, W. N., Sauder, M., & Espeland, W. (2016). *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401-436.

Eyal, G. (2019). *The crisis of expertise*. John Wiley & Sons.

Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. Psychology Press.

Fourcade, M. (2016). Ordinalization: Lewis A. Coser memorial award for theoretical agenda setting 2014. *Sociological Theory*, *34*(3), 175-195.

Gómez Vera, G. (2013). Los efectos de la repitencia en tanto que política pública en cuatro países del cono sur: Argentina, Brasil, Chile y Uruguay. Un análisis en base a PISA 2009. *Revista Latinoamericana de Educación Comparada* 4 (4): 59-70.

Gorur, R. (2017). Towards productive critique of large-scale comparisons in education. *Critical Studies in Education*, *58*(3), 341-355.

Gorur, R. (2019). In search of patterns: with enough data, do the numbers speak for themselves? *World Yearbook of Education 2019: Comparative Methodology in the Era of Big Data and Global Networks*, 53.

Haddad, W. D. (1979). *Educational and Economic Effects of Promotion and Repetition Practices*. Staff Working Paper No. 319.

Hanushek, E. A., & Woessmann, L. (2011). Sample selectivity and the validity of international student achievement tests in economic research. *Economics Letters*, *110*(2), 79-82.

Hanushek, E. A., & Woessmann, L. (2014). Institutional structures of the education system and student achievement: A review of cross-country economic research. *Educational policy evaluation through international comparative assessments*, *145*.

Ikeda, M., & García, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. *OECD Journal: Economic Studies*, *2013*(1), 269-315.

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of educational research*, *45*(4), 613-635.

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33-58.

Jimerson, S. R. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of school psychology*, *37*(3), 243-272.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for specific purposes*, *24*(3), 269-292.

Kemple, T. M., & Mawani, R. (2009). The sociological imagination and its imperial shadows. *Theory, Culture & Society*, *26*(7-8), 228-249.

Lockheed, M., Prokic-Bruer, T., & Shadrova, A. (2015). The experience of middle-income countries participating in PISA 2000-2015 (PISA series). *Washington, DC & Paris: The World Bank & OECD Publishing. doi: http://dx. doi. org/10.1787/9789264246195-en*.

Lorence, J., & Dworkin, A. G. (2006). Elementary grade retention in Texas and reading achievement among racial groups: 1994–2002. *Review of Policy Research*, *23*(5), 999-1033.

McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of school psychology*, *37*(3), 273-298.

McDonald, T. J. (2007). *Romantic comedy: Boy meets girl meets genre*. Columbia University Press.

Mortimore, P. (2009). *Alternative Models for Analysing and Representing Countries' Performance in PISA*. Education International.

OECD/UNESCO Institute for Statistics (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264102873-en.

Pizmony-Levy, O. (2013). Testing for all: The emergence and development of international assessment of student achievement, 1958–2012.

Pizmony-Levy, O. (2014). Back to the Future on International Assessments. *Quality Assurance in Education* 22 (4): 321–22.

Pizmony-Levy, O. (2018). Compare globally, interpret locally: international assessments and news media in Israel. *Globalisation, Societies and Education*, *16*(5), 577-595.

Pizmony-Levy, O., & Torney-Purta, J. (2018). How journalists and researchers communicate results of international large-scale assessments. *How Journalists and Researchers Communicate Results of International Large-Scale Assessments*, 51-65.

Porter, T. M. (2010). *Karl Pearson*. Princeton University Press.

Range, B. G., Pijanowski, J., Holt, C. R., & Young, S. (2012). The Perceptions of Primary Grade Teachers and Elementary Principals about the Effectiveness of Grade-Level Retention. *Professional Educator*, *36*(1), n1.

Reimers, F. (Ed.). (2000). *Unequal schools, unequal chances: The challenges to equal opportunity in the Americas* (Vol. 5). Cambridge, MA: Harvard University Press.

Rutkowski, L. (2016). *A Look at The Most Pressing Design Issues in International Large-Scale Assessments: A Paper Commissioned by The US National Academy of Education.* Washington, DC: NAED

Sabatier, P. A., & Weible, C. M. (2007). "The advocacy coalition framework–innovations and clarifications" In *Theories of the Policy Process,* ed. P.A. Sabatier. 2nd. Edition. Boulder, CO: Westview.

Schiefelbein, E. (1975). Repeating: an overlooked problem of Latin American Education. *Comparative Education Review*, *19*(3), 468-487.

Schleicher, A. (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Organisation for Economic Co-Operation and Development, Paris, France

Smith, W. C., & Edwards, D. B. (Eds.). (2016). *The global testing culture: Shaping education policy, perceptions, and practice*. Oxford: Symposium Books.

Shepard, L. A., & Smith, M. L. (1990). Synthesis of research on grade retention. *Educational Leadership*, *47*(8), 84-88.

Sorensen, T. B., & Robertson, S. L. (2020). Ordinalization and the OECD's Governance of Teachers. *Comparative Education Review*, *64*(1), 21-45.

Steiner-Khamsi, G. (2021). What Is a Reference? Theoretically Understanding the Uses of Evidence in Education Policy. *Evidence and Expertise in Nordic Education Policies: A Comparative Network Analysis from the Nordic Region*.

Steiner-Khamsi, G., Karseth, B., & Baek, C. (2020). From science to politics: Commissioned reports and their political translation into White Papers. *Journal of Education Policy*, *35*(1), 119-144.

Stevens, S. S. (1946). On the theory of scales of measurement.

Strietholt, R., Bos, W., Gustafsson, J. E., & Rosén, M. (Eds.). (2014). *Educational policy evaluation through international comparative assessments*. Waxmann Verlag.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tomchin, E. M., & Impara, J. C. (1992). Unraveling teachers' beliefs about grade retention. *American Educational Research Journal*, *29*(1), 199-223.

UIS. (2012). *Global Education Digest 2012. Opportunities Lost: The Impact of Grade Repetition and Early School Leaving*. Quebec: UNESCO Institute for Statistics.

Wagemaker, Hans. 2013. "International Large-Scale Assessments: From Research to Policy." In *Handbook of International Largescale Assessment: Background, Technical Issues, and Methods of Data Analysis*, ed. L. Rutkowski, M. von Davier, M., and D. Rutkowski. Boca Raton, FL: CRC Press.

Williamson, B. (2019). Assessment imaginaries: Methodological challenges of future assessment machines. *International large-scale assessments in education Insider research perspectives*, 3151.

Willms, J. D., & Somer, M. A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *School effectiveness and school improvement*, *12*(4), 409-445.

Wholey, J. S. (1987). Evaluability Assessment: Developing Program Theory. *New directions for program evaluation*, *33*, 77-92.

Wong, Y. L. (2013). Why are in-grade retention rates so high in Macao? *Current Issues in Education*, *16*(3).

Ydesen, C. (Ed.). (2019). *The OECD's Historical Rise in Education: The Formation of a Global Governing Complex*. Springer Nature.

Zapp, M. (2017). The scientization of the world polity: International organizations and the production of scientific knowledge, 1950–2015. *International Sociology*, *33*(1), 3-26.

Zapp, M. (2020). The authority of science and the legitimacy of international organisations: OECD, UNESCO and World Bank in global education governance. *Compare: A Journal of Comparative and International Education*, 1-20.

Policymaking to the test? How international large-scale assessments (ILSAs) influence repetition rates

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, *47*(3), 434-452.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31*(4), 480-499.

Berman, E. P., & Hirschman, D. (2018). The sociology of quantification: Where are we now? *Contemporary Sociology*, *47*(3), 257-266.

Byrnes, D. A., & Yamamoto, K. (1985). Academic retention of elementary pupils: An inside look. *Education*, *106*(2).

Cadigan, D., Entwisle, D. R., Alexander, K. L., & Pallas, A. M. (1988). First-grade retention among low achieving students: A search for significant predictors. *Merrill-Palmer Quarterly (1982-)*, 71-88.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, *54*(4), 297.

Cardoso, M. E. (2020). Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition. *Comparative Education Review*, *64*(4), 000-000.

Cardoso, M., & Steiner-Khamsi, G. (2017). The making of comparability: Education indicator research from Jullien de Paris to the 2030 sustainable development goals. *Compare: A journal of comparative and international education*, *47*(3), 388-405.

Chabbott, C. (2003). *Constructing education for development: International organizations and education for all*. Psychology Press.

Creemers, B. P., & Scheerens, J. (1994). Developments in the educational effectiveness research programme. *International Journal of Educational Research*, *21*(2), 125-140.

Desrosières, A. (2015). Retroaction: How indicators feed back onto quantified actors. In R. Rottenburg, S. E. Merry, S. J. Park, and J. Mugler (Eds.), *The world of indicators. The making of governmental knowledge through quantification*. Cambridge University Press.

Espeland, W. N., Sauder, M., & Espeland, W. (2016). *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401-436.

Eyal, G. (2019). *The Crisis of Expertise*. Wiley.

Furuta, J. (2021). Western Colonialism and World Society in National Education Systems: Global Trends in the Use of High-Stakes Exams at Early Ages, 1960 to 2010. *Sociology of Education*, 0038040720957368.

Gorur, R. (2016). Seeing like PISA: A cautionary tale about the performativity of international assessments. *European Educational Research Journal*, *15*(5), 598-616.

Guadalupe, C. (2017). Standardisation and diversity in international assessments: barking up the wrong tree? *Critical Studies in Education*, *58*(3), 326-340.

Haddad, W. D. (1979). *Educational and Economic Effects of Promotion and Repetition Practices*. Staff Working Paper No. 319.

Hanushek, E. A., & Woessmann, L. (2014). Institutional structures of the education system and student achievement: A review of cross-country economic research. In R. Strietholt, W. Bos, J. E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments*, 145-176. Waxmann Verlag.

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of educational research*, *45*(4), 613-635.

Jennings, J. (2009). *Accountability and Inequality in American Education* (Unpublished doctoral dissertation, Columbia University).

Jimerson, S. R. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of school psychology*, *37*(3), 243-272.

Kijima, R. (2010). Why participate? Cross-national assessments and foreign aid to education. In A. W. Wiseman (Ed.), *The impact of international achievement studies on national education policymaking*, 35-61. Emerald Group Publishing Limited. (2010).

Kyriakides, L., & Charalambous, C. Y. (2014). Educational effectiveness research and international comparative studies: Looking back and looking forward. In R. Strietholt, W. Bos, J. E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments*, 33-50. Waxmann Verlag.

Maddox, B. (Ed.). (2018). *International large-scale assessments in education: Insider research perspectives*. Bloomsbury Publishing.

McGaw, B. (2008a). Further reflections. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 279-282.

McGaw, B. (2008b). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 223-243.

Meyer, H. D., & Benavot, A. (Eds.). (2013, May). PISA, power, and policy: The emergence of global educational governance. Symposium Books Ltd.

Meyer, J. W. (2010). World society, institutional theories, and the actor. *Annual review of sociology*, *36*, 1-20.

Meyer, J. W., Boli, J., Thomas, G. M., & Ramírez, F. O. (1997). World society and the nation-state. *American Journal of sociology*, *103*(1), 144-181.

Pizmony-Levy, O. (2018). Compare globally, interpret locally: international assessments and news media in Israel. *Globalisation, Societies and Education*, *16*(5), 577-595.

Ramírez, F. O., Meyer, J. W., & Lerch, J. (2016). World society and the globalization of educational policy. In, K. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy*, *2016*, 43-63. John Wiley & Sons.

Ramírez, F. O., Schofer, E., & Meyer, J. W. (2018). International tests, national assessments, and educational development (1970–2012). *Comparative Education Review*, *62*(3), 344-364.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School effectiveness and school improvement*, *25*(2), 197-230.

Rivas, A. (2015). *América Latina después de PISA: Lecciones aprendidas de la educación en siete países (2000-2015) [Latin America after PISA: Lessons learned from education in seven countries (2000-2015)]*. Fundación CIPPEC.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252-257.

Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the Policy Influence of International Large-Scale Assessments in Education. In H. Wagemaker (Ed.), *Reliability and Validity of International Large-Scale Assessment: Understanding IEA's Comparative Studies of Student Achievement*, 261. Springer Nature.

Schiefelbein, E. (1975). Repeating: an overlooked problem of Latin American Education. *Comparative Education Review*, *19*(3), 468-487.

Schleicher, A. (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Organisation for Economic Co-Operation and Development, Paris, France

Smith, W. C. (2019). Evaluating the Global Testing Culture: A Response to Ramírez and Colleagues. *Comparative Education Review*, *63*(1), 121-126.

Smith, W. C., & Edwards, D. B. (Eds.). (2016). *The global testing culture: Shaping education policy, perceptions, and practice*. Oxford: Symposium Books.

Steiner-Khamsi, G. (2019). Conclusions: What Policy-Makers Do with PISA. In F. Waldow, &, G. Steiner-Khamsi (Eds.), *Understanding PISA's attractiveness: Critical analyses in comparative policy studies*, 233. Bloomsbury Publishing.

Strathern, M. (1996). From improvement to enhancement: an anthropological comment on the audit culture. *Cambridge Anthropology*, 1-21.

Strietholt, R., Bos, W., Gustafsson, J. E., & Rosén, M. (Eds.). (2014). *Educational policy evaluation through international comparative assessments*. Waxmann Verlag.

UIS. (2012). *Global Education Digest 2012. Opportunities Lost: The Impact of Grade Repetition and Early School Leaving*. Quebec: UNESCO Institute for Statistics.

Wagemaker, H. (2008). Choices and trade-offs: reply to McGaw. *Assessment in Education: principles, policy & practice*, *15*(3), 267-278.

Wagemaker, H. (2013). International Large-Scale Assessments: From Research to Policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Largescale Assessment: Background, Technical Issues, and Methods of Data Analysis,* 11-36. Boca Raton, FL: CRC Press.

Ward, M. (2018). PISA for development: Results in focus.

Williamson, B. (2018). Assessment imaginaries: Methodological challenges of future assessment machines. In B. Maddox (Ed.), *International large-scale assessments in education Insider research perspectives*, 3151. Bloomsbury Publishing.

Conclusion

Abend, G. (2006). Styles of sociological thought: Sociologies, epistemologies, and the Mexican and US quests for truth. *Sociological Theory*, *24*(1), 1-41.

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary school grades*. Cambridge University Press.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates' effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, *31*(4), 480-499.

Andrew, M. (2014). The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career. *Social Forces*, *93*(2), 653-685.

Desrosières, A. (2002). *The politics of large numbers: A history of statistical reasoning.* Harvard University Press.

Pizmony-Levy, O. (2018). Compare globally, interpret locally: international assessments and news media in Israel. *Globalisation, Societies and Education*, *16*(5), 577-595.

Porter, T. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Appendices

Appendix 2.A

Regression Coefficients and Standard Errors for OLS Models, PISA 2015, and TERCE 2013

	Mathematics				Reading				Science			
	PISA		TERCE		PISA		TERCE		PISA		TERCE	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Costa Rica:												
Nonrepeater	62.17	3.77	29.41	4.57	68.42	4.64	37.05	4.72	70.33	3.94	38.85	4.79
SES	27.12	2.44	33.16	2.43	31.99	2.33	34.49	2.55	33.00	2.45	34.08	2.90
Girl	-22.95	3.34	-15.68	3.75	14.78	2.86	7.19	3.84	-25.54	2.70	-11.85	3.98
Dominican Republic:												
Nonrepeater	37.29	2.40	19.63	3.18	59.77	3.17	26.52	4.51	48.98	3.01	16.83	3.80
SES	21.51	2.36	25.89	2.41	31.03	2.56	38.86	3.65	30.46	3.17	35.01	2.94
Girl	0.05	2.47	-10.91	2.56	22.81	2.64	11.72	4.58	-5.74	2.99	2.71	4.31
Mexico:												
Nonrepeater	82.07	6.54	49.64	7.37	79.85	6.53	47.13	6.17	73.37	5.09	36.88	7.03
SES	24.26	2.05	40.59	2.82	27.47	1.92	44.45	2.70	24.31	1.80	38.38	2.95
Girl	-12.63	3.15	-16.62	4.40	17.74	2.60	7.68	3.74	-12.25	2.69	-4.28	3.91
Peru:												
Nonrepeater	64.70	5.05	48.17	5.53	63.89	3.79	49.56	5.34	57.92	3.57	43.18	5.12
SES	37.70	2.23	45.71	2.90	41.54	1.97	48.77	2.35	36.15	1.87	42.36	2.35
Girl	-12.02	3.15	-22.13	4.24	8.93	3.26	1.45	3.92	-11.89	3.10	-8.31	3.10

Note.—Intercepts not presented.
Source.—For PISA, OECD; for TERCE, UNESCO.

<u>Appendix 3.A</u> Description of the corpus

Code	Author(s)	Year	Title	Repetition	Total pages	Examined pages
E2010	Treviño, E., Valdés, H., Castro, M., Costilla, R., Pardo, C., & Donoso Rivas, F.	2010	Factores asociados al logro cognitivo de los estudiantes de América Latina y el Caribe	24	159	3
E2015a	Treviño, E., Fraser, P., Meyer, A., Morawietz, L., Inostroza, P., & Naranjo, E.	2015	Informe de resultados TERCE: Factores asociados	35	184	3
E2015b	OREALC	2015	Is grade repetition effective? TERCE in sight	7	6	6
E2016	Treviño, E., Villalobos, C., & Baeza, A.	2016	Recomendaciones de Políticas Educativas en América Latina en base al TERCE.	44	168	5
P2007	OECD	2007	PISA 2006: Science competencies for tomorrow's world: Volume 1: Analysis.	9	390	2
P2010	OECD	2010	PISA 2009 Results: What makes a school successful? Resources, policies and practices. Vol. IV	32	308	5
P2011	OECD	2011	When Students Repeat Grades Or Are Transferred Out of School: What Does it Mean for Education Systems? PISA in Focus # 6	14	4	4
P2013	OECD	2013	PISA 2012 results: What makes schools successful? Resources, policies and practices (volume IV).	89	546	11
P2014	OECD	2014	Are disadvantaged students more likely to repeat grades? PISA in Focus # 43	16	4	4
P2016a	OECD	2016	Low-Performing Students: Why They Fall Behind and How to Help Them Succeed, PISA	55	212	8
P2016b	OECD	2016	Equations and Inequalities: Making Mathematics Accessible to All, PISA	20	224	3
P2016c	OECD	2016	PISA 2015 results: Excellence and equity in education	27	494	4
P2016d	OECD	2016	PISA 2015 Results (Volume II). Policies and Practices for Successful Schools	108	472	11

Appendix 3.B Codebook: Two-level (move and step) rhetorical structure, charts, and references

Name	Description	Files	References
1 Opening		11	27
1.1 Introduction	…situates the topic in the broader context of "education" or "school systems", without mentioning repetition.	7	9
1.2 Definition	…takes the form "repetition is…", "…is defined as", "…happens when…" or "a repeater/retainee is a student who…", etc.	3	4
1.3 Justification	…provides a rationale for using grade repetition, not necessarily espoused by the authors. It may take the form "repetition is used because…", "some teachers retain students in grade in order to…", etc.	9	14
2 Critique		12	59
Effectiveness (C)	…presents highlights of the literature on grade retention that address repetition's links with academic outcomes, achievement, performance, etc., without addressing financial costs.	9	16
Efficiency (C)	…presents highlights of the literature on grade retention focused on the financial resources repetition uses (costs), whether simultaneously addressing or not the academic outcomes it is expected to contribute to (benefits). Adjectives: inefficient, (not) cost-effective, wasteful, costly, etc.	7	10
Empathy (C)	…presents highlights of the literature on grade retention addressing repetition's links with socio-emotional variables. Terms: self-esteem, stigma, isolation, motivation, engagement, self-efficacy, etc.	7	15
Equity (C)	…presents highlights of the literature on grade retention statements addressing repetition's links with students' sociodemographic characteristics: SES, race/ethnicity, gender, home language, immigrant status, etc. Adjectives: unequal, inequitable, inegalitarian, unjust, unfair, biased, etc. (grade retention); disadvantaged, underprivileged, vulnerable, marginalized, etc. (groups).	5	8
Evaluability (C)	…presents highlights of the literature on grade retention addressing the methodological challenges found when evaluating grade retention's impact, mainly on	1	1

Name	Description	Files	References
	academic achievement.		
Expulsion (C)	…presents highlights of the literature on grade retention addressing repetition's link with dropout.	7	9
3 Quantification		11	53
3.1 Collection	…explains how the ILSA collects data from students on their repetition history, namely either repeater or non-repeater.	3	3
3.2 Level	…describes how prevalent repetition is across the entire set or subsets of participating countries; sometimes broadly, e.g., "…remains high", "…is still widely used".	11	34
3.3 Evolution	…describes changes in repetition level over time. They can be broad, e.g., "repetition has been decreasing" or precisely quantify change in all countries or in subsets.	5	16
4 Results		13	101
Effectiveness (R)	…presents the ILSA's own findings on repetition's links with academic outcomes, achievement, performance, etc., without addressing financial costs.	11	49
Efficiency (R)	…presents the ILSA's own findings on the financial resources repetition uses (costs), whether simultaneously addressing or not the academic outcomes it is expected to contribute to (benefits). Adjectives: inefficient, (not) cost-effective, wasteful, costly, etc.	4	9
Empathy (R)	…presents the ILSA's own findings on repetition's links with socio-emotional variables. Terms: self-esteem, stigma, isolation, motivation, engagement, self-efficacy, etc.	2	3
Equity (R)	…presents the ILSA's own findings on repetition's links with students' sociodemographic characteristics: SES, race/ethnicity, gender, home language, immigrant status, etc. Adjectives: unequal, inequitable, inegalitarian, unjust, unfair, biased, etc. (grade retention); disadvantaged, underprivileged, vulnerable, marginalized, etc. (groups).	9	34
Evaluability (R)	…presents the ILSA's own findings on the methodological challenges found when evaluating grade retention's impact, mainly on academic achievement.	3	6
Expulsion (R)	…presents the ILSA's own findings on repetition's link with dropout, in theory. But	0	0

Name	Description	Files	References
	because the ILSAs are school-based, this does not apply (N/A).		
5 Directives	…provide explicit guidance to countries on actions to take regarding the use of grade retention, or on practices that are deemed substitutes for repetition.	7	17
Chart		12	29
With country level information		12	23
Bar chart	A graph presenting country level data, “in which bars of varying height with spaces between them are used to display data for variables defined by qualities or categories.” https://dictionary.apa.org/bar-graph	11	20
Alphabetical bar chart	A bar chart in which countries are sorted in alphabetical order.	8	9
Ranked bar chart	A bar chart in which countries are sorted in ascending or descending order, according to the values of one of the variables being represented in the chart.	7	11
Ascending bar chart		4	5
Descending bar chart		5	6
Scatterplot	A chart presenting country-level data in “a graphical representation of the relationship between two continuously measured variables in which one variable is arrayed on each axis and a dot or other symbol is placed at each point where the values of the variables intersect.” https://dictionary.apa.org/scatterplot	3	3
Without country-level data		4	6
Reference token	An academic reference, usually enclosed in parentheses, including the author(s) name(s), and the year of publication. When several sources are included in the same parentheses and separated by semi-colons, they are coded separately. Each occurrence of the same source is coded separately, hence “token”.	12	82
Multiple references	A set of references enclosed in the same parentheses.	6	19
Multiple statements and references	A set of references enclosed in the same parentheses; preceded by a set of statements that could be attributed to that set of sources. The set of statements and the set of references. should be coded together.	5	6

<u>Appendix 3.C</u> Inter-rater reliability

A Research Assistant independently coded five randomly selected publications out of the 13 coded by me. Stratified by ILSA and cycle, these were 2 out of 4 ERCE publications; and 3 out of 9 PISA publications. I sorted the list in random order in Excel and selected the highest listed file from each cycle. This yielded 3 results reports, 1 monographic report, and 1 monographic leaflet, though not stratified by document type:

Code	**Title**	**Total pages**	**Examined pages**
E2010	Factores asociados al logro cognitivo de los estudiantes de América Latina y el Caribe	159	3
E2015b	Is grade repetition effective? TERCE in sight	6	6
P2007	PISA 2006: Science competencies for tomorrow's world: Volume 1: Analysis.	390	2
P2013	PISA 2012 results: What makes schools successful? Resources, policies and practices (volume IV).	546	11
P2016a	Low-Performing Students: Why They Fall Behind and How to Help Them Succeed, PISA	212	8

After randomly selecting those for double coding, I chose one remaining document (P2016d) as an example for Coder 2 to use as guidance. The consistency between the coders was high. Out of the initial discrepancies between coders, 15 were selected for referral back to Coder 2 for her revision, with the information that there was a discrepancy, but without disclosing Coder 1's code. Out of those 15 discrepancies were resolved, 11 were resolved and four remain.

Consistency was calculated at two levels: move/file and step/file.

The move/file level indicates the proportion of decisions that were consistent across coders regarding whether a given move (e.g., Opening) was present in a given file (e.g., E2010). In this case, because there were 5 double-coded cases and 5 moves, each coder had a total of 25 decisions to make. At this level, the two coders agreed on all the decisions; therefore, consistency was 100%.

The step/file level indicates the proportion of decisions that were consistent across coders regarding whether a given step (e.g., Justification) was present in a given file (e.g., E2010). In this case, because there were 5 double-coded cases and 13 moves, each coder had a total of 65 decisions to make. At this level, the two coders initially agreed on 55 out of 65 decisions; therefore, consistency was 85%. After Coder 2 independently revised the discrepancies (which had been identified by Coder 1), the number of remaining discrepancies dropped to 6 out of 65, raising consistency to 91%.

<u>Appendix 3.D</u> Moves and steps by coded publication

	E2010	E2015a	E2015b	E2016	P2007	P2010	P2011	P2013	P2014	P2016a	P2016b	P2016c	P2016d
1 Opening	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
1.1 Introduction	No	No	Yes	Yes	No	No	Yes	No	Yes	Yes	No	Yes	Yes
1.2 Definition	No	No	No	No	Yes	No	No	No	No	No	No	Yes	Yes
1.3 Justification	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
2 Critique	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
3 Quantification	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
3.1 Collection	No	No	No	No	Yes	No	No	Yes	No	No	No	No	Yes
3.2 Level	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
3.3 Evolution	No	No	Yes	No	No	No	No	Yes	Yes	Yes	No	No	Yes
4 Results	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Effectiveness (C,R)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Efficiency(C,R)	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Empathy (C,R)	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	No	Yes
Equity (C,R)	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Evaluability (C,R)	No	No	No	No	No	No	No	Yes	No	Yes	Yes	No	Yes
Expulsion(C)	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	No	Yes
5 Directives	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	No	Yes

Appendix 3.E
Examples of ambiguous attribution to sources
(Multiple statements attributed to multiple sources)

1. “some research suggests that repeating a grade generally does not yield improvements in learning outcomes and is associated with high economic and social costs (Alexander, Entwisle and Dauber, 2003; Hauser, 2004)” (P2010, 62)

2. “grade repetition may have adverse system-level effects as retained students are more likely to drop out, stay longer in the school system, or spend less time in the labour force (Rumberger, 2011; OECD, 2011b)” (P2013, 88)

3. “the probability that students drop out, stay longer in the school system, or spend less time in the labour force (Allen et al., 2009; Alexander, Entwisle and Dauber, 2003; Ikeda and García, 2014; Jacob and Lefgren, 2009; Manacorda, 2010)” (P2016b, 94).

4. “Previous PISA reports have suggested that grade repetition is a costly policy, that it is sometimes used as a form of punishment to sanction misbehaviour in school, and that it can reinforce inequalities in education because socio-economically disadvantaged students are more likely to repeat grades than advantaged students (OECD, 2013d; OECD, 2014b; OECD, 2015a).” (P2016a, 85).

5. “grade repetition in the earlier grades makes a student more likely to perform poorly in a later grade, because teachers have lower expectations for these students, because these students might have greater difficulties in integrating themselves into peer and school cultures, or for other reasons (Kaplan, Peck and Kaplan, 1997; Roderick, 1994)” (P2016a, 88).

6. “grade repetition is not necessarily beneficial for students. in fact, it may increase the probability that students drop out, stay longer in the school system, or spend less time in the labour force (Allen et al., 2009; alexander, Entwisle and dauber, 2003; Ikeda and García, 2014; Jacob and Lefgren, 2009; Manacorda, 2010)” (P2016b, 94)

7. “Students who have repeated a grade tend to display more negative behaviours and attitudes towards school, are more likely to drop out and may be stigmatised among their classmates (Ikeda and García, 2014; Rumberger and Lim, 2008; Thompson and Cunnigham, 2000; West, 2012)” (P2016d, 232)