KU LEUVEN

FACULTY OF SCIENCE

Generalized Linear Latent and Mixed Modeling:

method, estimation procedures, advantages, and applications to educational policy.

Jose Manuel Rivera Espejo

Supervisor: Prof. Geert Molenbegrhs
Affiliation *(optional)*

Co-supervisor: Prof. Wim Van den Noortgate *(optional)*
Affiliation *(optional)*

Proposal presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics and Data Science:
Social, Behavioral and Educational Sciences

Academic year 2020-2021

Dedication

To Manuel, for being my friend and father.
To Margarita, Susan and Ana, for their relentless encouragement.
To both of you, as you are always in my mind.
And to all that knowingly or not, help me to get here.
I am lucky due to all of you.
I hope I make you all proud.

A Manuel, por ser mi amigo y mi padre.
A Margarita, Susan y Ana, por su incansable aliento.
A ustedes dos, que siempre las tengo en mente.
Y a todos los que sabiendolo o no, me ayudaron a llegar aquí
Soy un suertudo gracias todos ustedes.
Espero llenarlos de orgullo.

Acknowledgment

(in the works)

Abstract

(in the works)
Keywords:

Contents

List of Figures

List of Tables

Abbreviations

GLLAMM	Generalized Linear Latent and Mixed Model.
SEM	Structural Equation Model.
GLM	Generalized Linear Model.
EFA	Exploratory Factor Analysis.
CFA	Confirmatory Factor Analysis.
IRT	Item Response Theory.

Symbols

J	total number of subjects.
j	index of specific subject.
I	total number of items.
i	index for the specific item.
$M_{(l)}$	total number of latent variables at level l.
m	index for the specific latent variable.
L	total number of levels (clusters).
l	index for the specific level (cluster).
$\mathbf{V}$	vector of the linear predictors.
$\boldsymbol{\beta}$	vector of fixed effects, for the I items.
$\mathbf{X}$	design matrix for the $\boldsymbol{\beta}$ parameters.
$\eta_{mj}^{(l)}$	latent variable, at the respective indices.
$\boldsymbol{\lambda}_m^{(l)}$	vector of loadings, at the respective indices.
$\mathbf{Z}_m^{(l)}$	design matrix for the $\boldsymbol{\lambda}_m^{(l)}$ parameters.

Chapter 1

Introduction

1.1 Preliminar considerations

The short and long term benefits of effective teaching practices can be observed throughout the literature: improvements in student achievements (Rockoff; 2004; Rivkin et al.; 2005; Duflo et al.; 2009; Hanushek and Rivkin; 2012; Muralidharan and Sundararaman; 2013; Chetty et al.; 2014a; Araujo et al.; 2016); development of executive functions (Araujo et al.; 2016), increased college attendance, higher salaries, lower possibility of premature parenthood (Chetty et al.; 2014b), among others. Similarly, the literature has shown most of the negative impacts resulting from the presence of teacher shortages[1] (Duflo et al.; 2009; Muralidharan and Sundararaman; 2013; Chetty et al.; 2015; Ayala; 2017; Marotta; 2019) or ineffective teaching practices (Hanushek and Rivkin; 2012).

However, while the evidence have a solid methodological support, Hanushek and Rivkin (2006) have indicated that some of the proxy variables used, are not consistently related to either teacher effectiveness or quality of instruction, examples of such are: out of field teaching[2] (Ingersoll; 1998; Dee and Cohodes; 2008; Bertoni et al.; 2020); teaching hours (Bruns et al.; 2015); years of experience or educational degree (Rockoff; 2004; Rivkin et al.; 2005; Clotfelter et al.; 2006, 2007; Hanushek and Rivkin; 2012); among others.

Given the lack of consistency of the effects that arises from using such proxies, Hanushek and Rivkin (2012) have pointed out that the analysis of teacher effectiveness has largely turned away, from attempts to identify the teacher's specific characteristics, to focus its attention into measuring the direct relationship between them and the student outcomes[3]. For that reason, considerable uncertainty is still present in the literature, regarding exactly which aspects of teachers are key for the student's learning and whether those qualities can be measured (Rockoff; 2004; Clotfelter et al.; 2006).

[1] Bertoni et al. (2020) defined it as the context in which the teacher's supply, i.e. the number of available teachers in the system, is less than its demand. The authors further elaborate that one of the causes of these shortages is related to the applicants' lower quality or due to their faulty initial training, implying that the shortage can also be conceived as the lack of good quality teachers. In this sense, the evidence of such shortage has been more prevalent, but not decisive, with temporary teachers, as they are usually associated with inferior attributes, compared to their contracted counterparts

[2] Medeiros et al. (2018) defines it as teachers teaching a subject in which they are not specialized or do not have the appropriate certificate.

[3] The method is known as value-added analysis, and it is based on the perspective that a good teacher is one who consistently gets higher achievement from students after other determinants of such are controlled for. For a more detailed explanation of the method refer to Scherrer (2011).

However, because the evidence still largely supports the perception that teachers are the main driver behind the student's learning processes, one of the main points in the agenda of any educational authority should be the design of an assessment system that can attract, select, develop, and retain the most effective ones (Elacqua et al.; 2018).

In that sense, an Educational Performance Standard (EPS) that best agrees with the country's context have to be defined. With the EPS establishment, the authorities can set clear expectations about what a "good" teacher should know and know to do (Cruz-Aguayo et al.; 2020). While the specific requirements are not easy to define, Cruz-Aguayo et al. have hinted that most of them can be largely grouped into two: (i) to have the disciplinary knowledge and pedagogical practices adequate to the classroom characteristics, context and teaching level, and (ii) to display such knowledge and practices in the classroom, using the appropriate material and technological resources available.

As one can infer from the previous general conditions, and the slew evidence, the disciplinary knowledge is a relevant observable factor that it is consistently associated with teacher effectiveness and growth in the student's achievement (Santibañez; 2006; Clotfelter et al.; 2006, 2007; Hanushek and Rivkin; 2006; Marshall; 2009; Rockoff et al.; 2011; Kane et al.; 2010; Kane and Staiger; 2012; Ome; 2012; Metzler and Woessmann; 2012; Kane et al.; 2013; Araujo et al.; 2016; Bietenbeck et al.; 2018; Estrada; 2019); and in that sense, its measurement should be of interest for any educational authority.

The measurement of knowledge has a myriad of available tools, however, given that any educational department are bounded by budgetary constraints, valid[4] and reliable[5] standardized tests[6] stand out not only for its cost-effectiveness, and a much simpler implementation (Cruz-Aguayo et al.; 2020), but also because, they are one of tools with less subjective scoring processes and interpretations, compared to other instruments.

However, as no instrument is perfect, the teacher's subject knowledge scores will likely reflect measurement error (Metzler and Woessmann; 2012). As established by Angrist and Krueger (1999), measurement error in the explanatory variable could bias the estimated coefficients, which implies that evidence based on test scores could be an attenuated reflection of the true effects. On the other hand, the use of one composite value, i.e. the score, does not allow to test which specific factors -if any- leads to better or worse teacher performance, making also difficult to know which teachers should be hired or what should be done to train them (Hanushek and Rivkin; 2012).

But beyond the use of test results as explanatory variables in modeling processes, there is one more pressing argument on why the issue of measurement error should be addressed: approximately 60% of the Caribbean and Latin American countries use standardized test scores as part of or as a main teacher selection tool (Cruz-Aguayo et al.; 2020). In this setting, devoting effort to assess the issues related to measurements errors, could help the educational authorities to understand, for example: if the scores thresholds used for the selection processes are appropriately set, or ultimately, to know the characteristics of the teachers that are being integrated into the public teaching staff.

[4]the extend to which a measurement tool is well-founded and accurately corresponds to the real measure (Kelley; 1927)

[5]the overall consistency of a measure under consistent conditions.

[6]Assessment instrument in which the implementation, questions, scoring processes, and interpretations are consistent with a predetermined or typified way. The instrument is usually composed of questions or items that fulfill three conditions: (i) they are polytomous, i.e. they have multiple choices, (ii) the choice categories are nominal, i.e. do not present any specific order, and (iii) there is only one "correct" category or answer (Rivera; 2019)

In summary, teachers are one of the main drivers behind the student achievements. However, some of the evidence supporting this claim has been based on proxy variables that are not consistently related to the quality of instruction, or methods that are not concerned with the outline of the teaching factors responsible for the student's learning. Nevertheless, while the literature still reflects considerable uncertainty on what are the "ingredients for a good teacher", a good amount of evidence has supported the disciplinary and pedagogical knowledge as relevant components of the teacher effectiveness. Finally, the literature has shown that valid and reliable standardized tests are among the best tools to assess such factors, but also have emphasized that such scores could reflect the teacher's abilities with considerable noise.

1.2 Objectives

This research will have two main goals. First, to describe the method, estimation procedures, and advantages of the Generalized Linear Latent and Mixed Modeling framework (GLLAMM, Rabe-Hesketh et al.; 2004a,b; Skrondal and Rabe-Hesketh; 2004; Rabe-Hesketh et al.; 2012). Second, to test the real implications of the method, in a data composed of large repeated Teacher's standardized educational assessments from Peru.

Specifically, for the first objective of the research, the author expects to appraise:

1. If the method can provide a general framework that could serve multiple psychometric purposes, e.g. to analyze the quality of the items, to obtain a dynamical noise-free "score" for the disciplinary abilities of the teachers, among others; and

2. What are the advantages or disadvantages of such models, specially compared to factor, item-response theory and multilevel models.

For the second objective, the author expects to shed some lights about some key policy decisions related to those large evaluation processes, to mention a few:

- Are the educational authorities screening the teachers with higher disciplinary knowledge?, and in that sense, what differentiate a contract teacher from a temporary one?,

- What are the general characteristics of the teaching-career applicants?, What is the level of their disciplinary knowledge, and how it evolves?,

- Do the initial training or socioeconomic status help to explain the disciplinary knowledge profile of the applicants?

- What specific factor of the disciplinary knowledge is consistently related to a good performance in the classroom?

- Do the instruments guarantee a fair assessment of minority groups with different abilities?

In this sense, the researcher believes the master's thesis contributes to the literature in two aspects:

1. In a the theoretical and methodological sense, as the research is focused on offering an exhaustive description and analysis of the GLLAMM framework; and

2. In a more practical sense, as it helps to provide evidence on some of key policy decisions that Latin America countries are currently facing.

Finally, it is important to mention, that the computational implementation of the methods will be developed in `R` (R Core Team; 2015) and `WinBUGS` (Lunn et al.; 2000).

1.3 Organization

Chapter 2, The Generalized Linear Latent and Mixed Model, will describe the model, its components, characteristics, assumptions and properties, to finally assess its benefits against factor (EFA and CFA), IRT and Multilevel models.

Chapter 3, Estimation, will describe two of the methods that can be used to fit such models: Likelihood and Bayesian methods. The chapter will also present the computational implementation of the model.

Chapter 4, Application, will describe the instruments and the "dimensions" under analysis. Additionally, it will describe briefly the data collection process, the sample design, and the results of the analysis under the GLLAMM framework.

Finally, **Chapter 5, Conclusions**, will discuss the conclusion for the method, its estimation process, the policy implications derived from the implementation of the model in a large teacher's assessment process, and the path of future research that can be derived from the present effort.

Chapter 2

The Generalized Linear Latent and Mixed Model

The Generalized Linear Latent and Mixed Model (GLLAMM) is a framework that unifies a wide range of latent variable models. Developed by Rabe-Hesketh et al. (2004a); **?**, the method was motivated by the need of a multilevel Structural Equation Models (SEM) that accommodates for unbalanced data, noncontinuous responses and the use of cross-level effects among latent variables.

This chapter will present the definition of such model, its characteristics, assumptions and properties.

2.1 Definition

Following Rabe-Hesketh et al. (2004a), we depart from the traditional multivariate framework for formulating factor and structural models, i.e. a "wide" format, and adopt a univariate approach, i.e. "long" format. In that sense, all the response variables for each unit are "stacked" in a single response vector with different variables distinguished from each other by a design matrix.

With the aforementioned structure, we proceed to outline the three parts of the framework: (i) the response model, (ii) the structural latent variable model, and (iii) the distribution of the latent variables. For a detailed description of some of the special cases of multilevel SEM that can be derived with this framework, refer to Appendix A.

2.1.1 Response model

As outlined in Rabe-Hesketh et al. (2004a, 2012), conditional on the latent variables, the response model is a generalized linear model (GLM) defined by a systematic and a distributional part. For the systematic part, a linear predictor and a link function are selected. Finally, for the distributional part, a distribution from the exponential family is also selected.

In the following sections, we proceed to describe the linear predictor, the link function and the distributions accommodated by the framework.

Linear predictor

For a model with L levels and M_l latent variables at $l > 1$ levels, the linear predictor takes the following form:

$$v = \mathbf{X}\boldsymbol{\beta} + \sum_{l=2}^{L}\sum_{m=1}^{M_{(l)}} \eta_m^{(l)}\mathbf{Z}_m^{(l)}\boldsymbol{\lambda}_m^{(l)} \tag{2.1}$$

where $\mathbf{X}$ is a design matrix that maps the parameter vector $\boldsymbol{\beta}$ to the linear predictor, $\eta_m^{(l)}$ the mth latent variable at level l ($m = 1, \ldots, M_{(l)}$ and $l = 1, \ldots, L$), and $\mathbf{Z}_m^{(l)}$ a design matrix that maps the vector of loadings $\boldsymbol{\lambda}_m^{(l)}$ to the mth latent variable at level l.

Note that wo do not use subscripts for the units of observation at different levels. This decision was made with the purpose of avoiding the use of mathematical definitions with large number of subscripts. However, a careful reader should consider that equation 2.1 rest on the assumption that each unit is identified at their appropriate level. For special cases of multilevel SEM and their use of subscripts refer to Appendix A.

Links and Distributions

To make the mathematical explanation more tractable, the author first proceeds to define some nomenclature. First, $\boldsymbol{\eta}^{(l)} = [\eta_1^{(l)}, \ldots, \eta_{M_{(l)}}^{(l)}]^T$ denotes the vector of latent variables, and $\mathbf{Z}^{(l)} = [\mathbf{Z}_1^{(l)T}, \ldots, \mathbf{Z}_{M_{(l)}}^{(l)T}]^T$ the "stacked" design matrix of explanatory variables, at level l respectively. Finally, $\boldsymbol{\eta} = [\eta^{(2)T}, \ldots, \eta^{(L)T}]^T$ describes the "stacked" vector of latent variables, and $\mathbf{Z} = [\mathbf{Z}^{(2)T}, \ldots, \mathbf{Z}^{(L)T}]^T$ the "stacked" design matrices of explanatory variables, for all L levels respectively.

For the link functions, the model "links" the expectation of the conditional response to the linear predictor through a response function $h(\cdot)$, in the following form:

$$\mu = E[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = h(v) \tag{2.2}$$

where equation 2.2 can be re-written in terms of the link function $g(\cdot) = h^{-1}(\cdot)$:

$$g(\mu) = g(E[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}]) = v \tag{2.3}$$

Finally, the response model specification is complete when we select an appropriate distribution from the family of exponential distributions, for the conditional response. The types of responses that can be accommodated by the framework are the following:

1. **Continuous:**
   It results form selecting an identity link function for the mean scaled response,

   $$\mu^* = v^* \tag{2.4}$$

   On the other hand, the distributional part use an standard normal distribution,

   $$f(y^*|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}) = \phi(\mu^*)\sigma^{-1} = \phi(v^*)\sigma^{-1} \tag{2.5}$$

where $y^* = y\sigma^{-1}$, $v^* = v\sigma^{-1}$, and $\phi(x) = (2\pi)^{-1/2}exp(-x^2/2)$, describes the Standard Normal density distribution. Additionally, in the presence of heteroscedasticity, this can be modeled as:

$$log(\sigma) = \boldsymbol{\alpha}^T \mathbf{Z}^{(1)} \tag{2.6}$$

where σ is the standard deviation of the errors, and $\mathbf{Z}^{(1)}$ is the design matrix at level 1, that maps the regression parameters $\boldsymbol{\alpha}$. is this a random intercepts at level 1 definition?

2. **Dichotomous:**
It results from selecting an appropriate link function for the expected value of the response, which describe the probability of endorsing one of the two available categories,

$$\mu = E[y = 1|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = P[y = 1|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = \pi = h(v) \tag{2.7}$$

with a response function that can be defined in three ways:

$$h(x) = exp(x)[1 + exp(x)]^{-2} \tag{2.8}$$
$$h(x) = (2\pi)^{-1/2}exp(-x^2/2) \tag{2.9}$$
$$h(x) = exp(x - exp(x)) \tag{2.10}$$

which corresponds to the logistic, standard normal, and Gumbel density distributions, respectively. In turn, in terms of link functions, the distributions corresponds to the logit, probit and complementary log-log link functions, respectively.

Finally, the distributional part is defined by a Binomial distribution,

$$f[y = 1|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = \binom{n}{k}\mu^k(1-\mu)^{n-k} = \binom{n}{k}\pi^k(1-\pi)^{n-k} \tag{2.11}$$

where k denotes the number of successes in n independent Bernoulli trials.

3. **Polytomous:**
It results from selecting an appropriate link function for the expected value of the response, which in this case describe the probability of endorsing one of the S unordered available categories,

$$\mu_s = E[y = y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = P[y = y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = \pi_s = h(v_s) \tag{2.12}$$

with a response function that can be defined in two ways:

$$h(x) = exp(x)\left[\sum_{s=1}^{S} exp(x)\right]^{-1} \tag{2.13}$$
$$h(x) = exp(-x - exp(-x)) \tag{2.14}$$

where the former describes the generalized logistic density distribution, and the latter, the Gumbel (extreme value type I) distribution. It is important to note that under the Gumbel parametrization, the modelling process will corresponds with the "random utility model" parametrization, often used in econometrics.

make sure of the following part

Under this framework, we can model the category of interest against a reference category, in the following form:

$$\mu_{s-r} = E[y = y_{s-r}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = P[y = y_{s-r}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = \pi_{s-r} = h(v_s - v_r) \tag{2.15}$$

where r denotes the reference category, $y_{s-r} = y_s - y_r$, and $h(\cdot)$ is a logistic distribution like in 2.8. Alternatively, the framework can model the category of interest against the $S-1$ categories in the following form,

$$\mu_{s-S} = E[y = y_{s-r}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = P[y = y_{s-r}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = \pi_{s-r} = h(v_s - v_r) \tag{2.16}$$

Finally, the distributional part is defined by a Multinomial distribution,

$$f[y = \{y_1, \cdots, y_S\}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = \frac{n!}{k_1! \cdots k_S!} \prod_{s=1}^{S} \mu_s^{y_s} = \frac{n!}{k_1! \cdots k_S!} \prod_{s=1}^{S} \pi_s^{y_s} \tag{2.17}$$

where, in the case of the Gumbel parametrization, the product will span over $S-1$ set of alternatives, and the parameters will be defined as y_{s-r}, μ_{s-r} and π_{s-r}.

4. **Ordinal and discrete time duration:**

5. **Counts and continuous time duration:**
   It results from selecting a log link function for the expected value of the response,

$$ln(\mu) = ln(E[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}]) = ln(\lambda) = v \tag{2.18}$$

   and a Poisson distribution for the conditional distribution of the counts,

$$f[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] = exp(-\mu)\mu^y(y!)^{-1} = exp(-\lambda)\lambda^y(y!)^{-1} \tag{2.19}$$

6. **Rankings and pairwise comparisons:**
   Similar parametrization can be used as the building block for the conditiona distribution of rankings (Skrondal and Rabe-Hesketh; 2003).

7. **Mixed responses:**

2.1.2 Structural model for the latent variables

2.1.3 Distribution of the latent variables

2.1.4 Model identification

2.2 Relationship with other modeling schemes

Rabe-Hesketh et al. (2012)

The motivation for multilevel regression models is to handle hierarchical data where elementary units are nested in clusters, such as students in schools, which in turn may be nested in higher-level clusters (e.g., school districts or states). The latent variables, often called "random effects" in this context, can be interpreted as the effects of unobserved covariates at different levels that induce dependence among lower-level units. In contrast, the motivation for structural equation models is to handle variables that cannot be measured directly, and are hence latent, and to model their relationships with each other and with observed or manifest variables. The latent variables, often called "common factors" inthis context, are measured by manifest variables and induce dependence among them.

2.2.1 Factor Models

2.2.2 Item Response Theory and Generalized Latent Models

2.2.3 Multilevel Models

2.3 Advantages and Disadvantages

Chapter 3

Estimation

3.1 Likelihood methods

3.1.1 Likelihood function

3.1.2 Adaptive Quadrature

3.2 Bayesian methods

3.2.1 Prior distributions

3.2.2 Initial start

3.2.3 Posterior distributions

Chapter 4

Application

4.1 Instruments

4.2 Data

4.2.1 Collection

4.2.2 Sample scheme

4.3 Results

4.3.1 Hypothesis 1:

4.3.2 Hypothesis 2:

4.3.3 Hypothesis 3:

Chapter 5

Conclusion and Discussion

5.1 Discussion

5.2 Conclusions

5.3 Future development

Appendix A

Additional Theory

A.1 Special cases for the GLAMM

Appendix B

Code

Bibliography

Angrist, J. and Krueger, A. (1999). Empirical strategies in labor economics, *in* O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol. 3, Elsevier, chapter 23, pp. 1277 – 1366.
URL: *http://www.sciencedirect.com/science/article/pii/S1573446399030047*

Araujo, M., Carneiro, P., Cruz-Aguayo, Y. and Schady, N. (2016). Teacher quality and learning outcomes in kindergarten, *The Quarterly Journal of Economics* **131**(3): 1415–1453.
URL: *https://publications.iadb.org/publications/english/document/Teacher-Quality-and-Learning-Outcomes-in-Kindergarten.pdf*

Ayala, M. (2017). *Efecto de los docentes provisionales sobre desempeño escolar - evidencia para la educación secundaria oficial en colombia*, Master's thesis, Universidad de los Andes.
URL: *http://biblioteca.uniandes.edu.co/acepto201699.php?id=11802.pdf*

Bertoni, E., Elacqua, G., Marotta, L., Martinez, M., Méndez, C., Montalva, V., Olsen, A., Santos, H. and Soares, S. (2020). Escasez de docentes en latinoamérica: ¿cómo se puede medir y que políticas están implementando los países para resolverlo?, *Technical report*, Banco Interamericano de Desarrollo.

Bietenbeck, J., Piopiunik, M. and Wiederhold, S. (2018). Africa's skill tragedy: Does teachers' lack of knowledge lead to low student performance?, *Comparative Education Review* **53**(3): 553–578.
URL: *http://jhr.uwpress.org/content/53/3/553.abstract*

Bruns, B., Luque, J., De Gregorio, S., Evans, D., Fernández, M., Moreno, M., Rodriguez, J. Toral, G. and Yarrow, N. (2015). Great teachers: How to raise student learning in latin america and the caribbean, *Technical report*, World Bank Group.

Chetty, R., Friedman, J. and Rockoff, J. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates, *American Economic Review* **104**(9): 2593–2632.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2593*

Chetty, R., Friedman, J. and Rockoff, J. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood, *American Economic Review* **104**(9): 2633–2679.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2593*

Chetty, R., Friedman, J. and Rockoff, J. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools, *Journal of Public Economics* **123**: 92–110.
URL: *http://www.sciencedirect.com/science/article/pii/S0047272714002412*

Clotfelter, C., Ladd, H. and Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness, *Working Paper 11936*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w11936*

Clotfelter, C., Ladd, H. and Vigdor, J. (2007). How and why do teacher credentials matter for student achievement?, *Working Paper 12828*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w12828*

Cruz-Aguayo, Y., Hincapié, D. and Rodríguez, C. (2020). Profesores a prueba: claves para una evaluación docente exitosa, *Technical report*, Banco Interamericano de Desarrollo.

Dee, T. and Cohodes, S. (2008). Out-of-field teachers and student achievement: Evidence from matched-pairs comparisons, *Public Finance Review* **36**(1): 7–32.
URL: *https://doi.org/10.1177/1091142106289330*

Duflo, E., Dupas, P. and Kremer, M. (2009). Additional resources versus organizational changes in education: Experimental evidence from kenya.

Elacqua, G., Hincapié, D., Vegas, E. and Alfonso, M. (2018). Profesión: profesor en américa latina ¿por qué se perdió el prestigio docente y cómo recuperarlo?, *Technical report*, Banco Interamericano de Desarrollo.

Estrada, R. (2019). Rules versus discretion in public service: Teacher hiring in mexico, *Journal of Labor Economics* **37**(2): 545–579.
URL: *https://doi.org/10.1086/700192*

Hanushek, E. and Rivkin, S. (2006). Teacher quality, *in* E. Hanushek and F. Welch (eds), *Handbook of the Economics of Education*, Vol. 2, Elsevier, chapter 18, pp. 1051 – 1078.
URL: *http://www.sciencedirect.com/science/article/pii/S1574069206020186*

Hanushek, E. and Rivkin, S. (2012). The distribution of teacher quality and implications for policy, *Annual Review of Economics* **4**(1): 131–157.
URL: *https://doi.org/10.1146/annurev-economics-080511-111001*

Ingersoll, R. (1998). The problem of out-of-field teaching.
URL: *https://repository.upenn.edu/gse$_p$ubs*/137

Kane, T., McCaffrey, D., Miller, T. and Staiger, D. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment, *Research paper*, Bill Melinda Gates Foundation.
URL: *https://files.eric.ed.gov/fulltext/ED540959.pdf*

Kane, T. and Staiger, D. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains, *Research paper*, Bill Melinda Gates Foundation.
URL: *https://k12education.gatesfoundation.org/download/?Num=2678filename=MET_Gathering_Fee*

Kane, T., Taylor, E., Tyler, J. and Wooten, A. (2010). Identifying effective classroom practices using student achievement data, *Working Paper 15803*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w15803*

Kelley, T. (1927). *Interpretation of educational measurements*, Measurement and adjustment series, World Book Co.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility, *Statistics and Computing* (10): 325–337.

Marotta, L. (2019). Teachers' contractual ties and student achievement: The effect of temporary and multiple-school teachers in brazil, *Comparative Education Review* **63**(3): 356–376.

Marshall, J. (2009). School quality and learning gains in rural guatemala, *Economics of Education Review* **28**(2): 207–216.
URL: *http://www.sciencedirect.com/science/article/pii/S0272775708000745*

Medeiros, M., Gómez, C., Sánchez, M. and Orrego, V. (2018). Idoneidad disciplinar de los profesores y mercado de horas docentes en chile, *Calidad en la Educación* (48): 50–95.
URL: *https://doi.org/10.31619/caledu.n48.479*

Metzler, J. and Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation, *Journal of Development Economics* **99**(2): 486–496.
URL: *https://ideas.repec.org/a/eee/deveco/v99y2012i2p486-496.html*

Muralidharan, K. and Sundararaman, V. (2013). Contract teachers: Experimental evidence from india, *Working Paper 19440*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w19440*

Ome, A. (2012). The effects of meritocracy for teachers in colombia, *Research report*, Fedesarrollo.
URL: *https://ideas.repec.org/p/col/000124/010260.html*

R Core Team (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004a). Generalized multilevel structural equation modeling, *Psychometrika* **69**(2): 167–190.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004b). *GLLAMM Manual*, UC Berkeley Division of Biostatistics.
URL: *http://www.biostat.jhsph.edu/ fdominic/teaching/bio656/software-gllamm.manual.pdf*

Rabe-Hesketh, S., Skrondal, A. and Zheng, X. (2012). Multilevel structural equation modeling, *in* R. Hoyle (ed.), *Handbook of Structural Equation Modeling*, The Guilford Press, chapter 30, pp. 512–531.

Rivera, J. (2019). *El modelo de respuesta nominal: Aplicación a datos educacionales*, Master's thesis, Pontificia Universidad Católica del Peru.
URL: *http://hdl.handle.net/20.500.12404/14600*

Rivkin, S., Hanushek, E. and Kain, J. (2005). Teachers, schools, and academic achievement, *Econometrica* **73**(2): 417–458.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00584.x*

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data, *The American Economic Review* **94**(2): 247–252.
URL: *http://www.jstor.org/stable/3592891*

Rockoff, J., Jacob, B., Kane, T. and Staiger, D. (2011). Can you recognize an effective teacher when you recruit one?, *Education Finance and Policy* **6**(1): 43–74.
URL: *https://doi.org/10.1162/EDFP*$_{a0}$0022

Santibañez, L. (2006). Why we should care if teachers get a's: Teacher test scores and student achievement in mexico, *Economics of Education Review* **25**(5): 510–520.
URL: *http://www.sciencedirect.com/science/article/pii/S0272775705000804*

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea, *NASSP Bulletin* **95**(2): 122–140.
URL: *https://doi.org/10.1177/0192636511410052*

Skrondal, A. and Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings, *Psychometrika* **68**: 267–287.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, Chapman Hall/CRC Press.

AFDELING
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be