KU LEUVEN

FACULTY OF SCIENCE

Proposal:

Item Response Theory and Confirmatory Factor Analysis equivalence: Application on a teacher evaluation process in Peru

Jose Manuel RIVERA ESPEJO

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

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

Mentor: *(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

1 Topic justification and context

Throughout the literature, the benefits of effective teaching practices spans from short-term outcomes, like 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); to long-term effects, like the 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] or ineffective teaching practices (Duflo et al., 2009; Hanushek and Rivkin, 2012; Muralidharan and Sundararaman, 2013; Chetty et al., 2015; Ayala, 2017; Marotta, 2019).

However, while the evidence have a solid methodological support, some of it have been based on proxy variables that are not consistently related to either teacher effectiveness or quality of instruction (Hanushek and Rivkin, 2006). Examples of such variables are, out of field teaching[2] (Ingersoll, 1998; Dee and Cohodes, 2008; Bertoni, Elacqua, Marotta, Martinez, Méndez, Montalva, Olsen, Santos and Soares, 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); while Sutcher et al. (2016) details other proxies used to measure teacher shortage.

Given the consistency issue of proxy variables, Hanushek and Rivkin (2012) pointed out that the analysis of teacher effectiveness has largely turned away from attempts to identify specific characteristics of teachers, and instead has focused its attention to measure the direct relationship between teachers and student outcomes through a value-added analysis[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).

Because the evidence significantly support the perception that teachers are the main driver behind the student's learning processes and outcomes, the main agenda of any educational authority should be the design of an assessment system that can attract, select, develop, and retain the most effective teachers (Elacqua et al., 2018). But first, it would be necessary to define the Educational Performance Standards (EPS) that best agrees with the country's context. With the EPS establishment, the educational authorities can set clear expectations about what a "good" teacher should know and know to do (Cruz-Aguayo et al., 2020), and eventually, assess the teacher shortage.

While the specific requirements for a "good "teacher are not as easy to define, Cruz-

[1]Bertoni, Elacqua, Marotta, Martinez, Méndez, Montalva, Olsen, Santos and Soares (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 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).

Aguayo et al. (2020) has 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, knowledge is a relevant observable factor that it is consistently associated with teacher effectiveness and growth in 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); in that sense, its measurement should of interest for any educational authority.

While Bertoni, Elacqua, Méndez, Montalva, Munevar, Olsen and Román (2020) had advocated for use of multiple instruments, and in fact, the measurement of knowledge has a myriad of available tools, we have to keep in mind that the educational authorities are bound by budgetary constraints. Is in this setting that, compared to other instruments, valid[4] and reliable[5] standardized tests[6] shines because not only are cost-effective and much simpler to implement (Cruz-Aguayo et al., 2020), but they are among the most subjective-free tools available.

However, as no instrument is perfect, the teacher's subject knowledge scores likely reflect considerable noise or measurement error (Metzler and Woessmann, 2012). As established by Angrist and Krueger (1999), measurement error in the explanatory variable may lead to a downward bias in the estimated coefficient, meaning that, the previous evidence on teacher's knowledge could be an attenuated reflection of the true effect. 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's performance, making also difficult to know which teacher 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 measurement error issue should be addressed: approximately 60% of the Caribbean and Latin American countries use standardized test scores as part or as a main teacher's selection tool (Cruz-Aguayo et al., 2020). For this reason, the educational authorities could benefit from testing if the scores thresholds used for the selection are appropriately set, and what kind of teacher are they letting in to the system, as a result.

In summary,

- Paragraph's conclusion:

Second, it would be valuable to develop richer measures of teacher quality which go beyond the mean test score impacts that we analyzed here. (Chetty et al., 2014a)

(Cruz-Aguayo et al., 2020) Las falencias en la calidad de la formación inicial y las características de los interesados en ingresar a la profesión docente en la región (reseñadas

[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)

por (Elacqua et al., 2018)), sumadas a la dificultad que tienen los ministerios de educación para remover del cargo a docentes con bajos niveles de desempeño, convierten a las evaluaciones de ingreso a la carrera en un elemento esencial para identificar mejor las características y capacidades de los futuros docentes. En ese sentido, continuar con los procesos de mejora y de implementación adecuada de esta evaluación podría traer beneficios en la calidad de la fuerza docente en la región. Aunque la evidencia aún es escasa, estudios para Colombia (Ome, 2012; Brutti y Sánchez, 2017) y México (Estrada, 2019) sugieren que estos sistemas de selección (con evaluaciones para ingresar a la carrera) están teniendo algunos impactos positivos en la calidad educativa de los países que los implementan.

De manera similar, en lo que respecta a los temas disciplinares, pruebas en Perú, Chile y México y estudios internacionales aplicados a los propios docentes indican que sus conocimientos de matemáticas son insatisfactorios (Elacqua et al., 2018).

- Paragraph's main point: How the results are used

Todas estas políticas deben tener como objetivo desarrollar y potenciar los conocimientos disciplinares y las habilidades pedagógicas de los docentes. Las evaluaciones docentes pueden ayudar a identificar las diferencias de desempeño entre los profesores. Además, el uso adecuado de sus resultados puede otorgar la información necesaria para aprovechar al máximo sus fortalezas, buscar superar las falencias, y potenciar la excelencia en la profesión. (Cruz-Aguayo et al., 2020)

La segunda característica necesaria dentro de la teoría de cambio (reflejada en el segundo punto del primer círculo de la cadena) es el uso que se les da a los resultados de la evaluación. (Cruz-Aguayo et al., 2020)

- Paragraph's idea 1: how they are used

la Organización para la Cooperación y el Desarrollo Económicos (OCDE) este uso puede tener dos objetivos: i) mejorar las prácticas y las habilidades pedagógicas y/o disciplinares a partir del diagnóstico y la vinculación a programas de desarrollo profesional diseñados para superar los resultados; ii) mejorar la composición y la motivación de la fuerza docente por medio del otorgamiento de bonificaciones, reconocimientos especiales o ascensos para aquellos docentes con resultados excelentes o excluir al docente del sistema —o al menos retirarlo de las aulas— en los casos en que muestre de manera consistente que no cumple con las condiciones requeridas por la profesión (OCDE, 2009 y 2013). El problema que surge en relación con estos dos objetivos (plasmados en el segundo eslabón del gráfico 2.1) es que son difíciles de lograr con una única herramienta de evaluación. Para poder detectar los aspectos por mejorar de las prácticas y el conocimiento pedagógico y disciplinar, y reconocer la excelencia docente, es necesario que los docentes estén completamente abiertos a revelar sus prácticas y logros y dispuestos a compartirlos con las autoridades (Cruz-Aguayo et al., 2020)

Igualmente, una vez que los procesos finalizan, debe decidirse cómo entregar la información recolectada a los docentes y, lo que es más importante aún, cómo utilizarla para asegurar la mejora de la labor pedagógica. (Cruz-Aguayo et al., 2020)

- Paragraph's idea 2: evidence about impacts on the use of the results
- Paragraph's conclusion: results can have serious impacts into multiple facets
- Closing thoughts

en este documento tratamos de responder a dos preguntas fundamentales: ¿cómo identificamos y seleccionamos a los mejores docentes? y ¿cómo los asignamos a las escuelas de una manera eficiente y equitativa? (Bertoni, Elacqua, Méndez, Montalva, Munevar,

Olsen and Román, 2020)

2 Methods

Four measurements issues receive considerable attention in the research literature: (a) random measurement error, (b) the focus of test on particular portions of the achievement distribution, (c) cardinal versus ordinal comparisons of test scores, and (d) the multidimensionaly of educational outcomes. Not only do the test measurements issues introduce noise into the estimates of the teacher effectiveness, but they also bias upwards estimates of the variance in teacher quality (Hanushek and Rivkin, 2012). While this was mentioned for the value-added measures it is equally valid for the standardized evaluation of teachers.

We address measurement error by correcting the estimated coefficients using a reliability ratio estimated on the basis of answers to all items on the teacher tests (see Section 5.3).

- Paragraph's main point: what method are you using?

one can improve the value-added measures if we incorporate other measures of teacher quality, such as teacher characteristics (Chetty et al., 2014a)

- Paragraph's idea 1: IRT and the focus on items

De esta forma, mientras que los modelos para respuestas dicot?micas, tales como Rasch (**?**), de uno, dos, tres par?metros (**?**) y cuatro par?metros (**?**), expresan la probabilidad de elegir la alternativa correcta en funci?n de la "habilidad" del individuo; el **Modelo de Respuesta Nominal (NRM)** y todas sus extensiones (**?** y **?**, cap?tulo 2), expresa la probabilidad de elegir cada alternativa de la pregunta en funci?n de la misma "habilidad".

A diferencia de los modelos de respuesta graduada (**??** y **?**, cap?tulo 5), el NRM no se sustenta sobre el concepto de la dicotomizaci?n de las alternativas, que derivan en los umbrales por categor?as caracteristicos de los modelos mencionados; por el contrario, la probabilidad correspondiente a cada alternativa es modelada directamente, implementando una generalizaci?n multivariada del modelo de rasgos latentes log?stico (**?**, **?**).

- Paragraph's idea 2: SEM and the focus on abilities

- Paragraph's idea 3: IRT and SEM equivalence (evidence)

(Brown, 2015) The potential consequences of treating categorical variables as continuous variables in CFA are manifold: (1) They produce attenuated estimates of the relationships (correlations) among indicators, especially when there are floor or ceiling effects; (2) they lead to "pseudofactors" that are artifacts of item difficulty or extremeness; and (3) they produce incorrect test statistics and standard errors. ML can also produce incorrect parameter estimates, such as in cases where marked floor or ceiling effects exist in purportedly interval-level measurement scales (i.e., because the assumption of linear relationships does not hold).

Rasch Model with SEM

1. Requires to set the loadings = 1 in all items (there are no evidence that different items should load differently in all sub-factors, if that happen then we can say that an item does not behave good)

2. Thresholds can be transformed into difficulty parameters. They will be from the normal ogive model.

Evidence: It is well known that factor analysis with binary outcomes is equivalent to a two-parameter normal ogive IRT model (e.g., Ferrando Lorenza-Sevo, 2005; Glöckner-Rist Hoijtink, 2003; (Kamata and Bauer, 2008; Takane and de Leeuw, 1987).

Item difficulties have been alternatively referred to in the IRT literature as item threshold or item location parameters. In fact, item difficulty parameters are analogous to item thresholds (t) in CFA with categorical outcomes (Muthén et al., 1991).

Item discrimination parameters are analogous to factor loadings in CFA and EFA because they represent the relationship between the latent trait and the item responses.

Muthén (1988; Muthén et al., 1991) has shown that MIMIC models (see Chapter 7) with categorical indicators are equivalent to DIF analysis in the IRT framework (see also Meade Lautenschlager, 2004).

Muthén (1988; Muthén et al., 1991) notes that the MIMIC framework offers several potential advantages over IRT. These include the ability to (1) use either continuous covariates (e.g., age) or categorical background variables (e.g., gender); (2) model a direct effect of the covariate on the latent variable (in addition to direct effects of the covariate on test items); (3) readily evaluate multidimensional models (i.e., measurement models with more than one factor); and (4) incorporate an error theory (e.g., measurement error covariances). Indeed, a general advantage of the covariance structure analysis approach is that the IRT model can be embedded in a larger structural equation model (e.g., Lu, Thomas, Zumbo, 2005).

De esta forma, en el contexto de una evaluaci?n estandarizada, suponemos que n sujetos responden p ?tems de opci?n m?ltiple eligiendo **una sola** alternativa de m_j disponibles, las mismas que pueden variar de ?tem a ?tem y poseen un orden arbitrario. Entonces, el NRM define ***Funciones de Respuestas de las Categor?as del ?tem*** (ICRF, acorde con **?**) o Curvas Caracter?sticas de la Alternativas del ?tem (IOCC, acorde con **?**) de la siguiente manera:

$$P_{jk}(\theta_i) = \frac{e^{z_{jk}(\theta_i)}}{\sum_{h=1}^{m} e^{z_{jh}(\theta_i)}} \tag{1}$$

Donde:

$$z_{jk} = a_{jk}\theta_i + c_{jk} \quad \forall \quad i = 1, \ldots, n; \quad j = 1, \ldots, p; \quad k = 1, \ldots, m_j$$

El par?metro θ_i representa la "habilidad" del individuo i, a_{jk} corresponde al par?metro de discriminaci?n de la alternativa k del ?tem j y c_{jk} es proporcional a la "popularidad" de la alternativa k del ?tem j. El vector compuesto por los vectores $z_{j1}, z_{j2}, \ldots z_{jm_j}$ es usualmente definido como el vector *logit multinomial*. La presente parametrizaci?n del modelo es expresada en t?rminos del intercepto y la pendiente de las ICRFs; sin embargo, la literatura utiliza una parametrizaci?n que hace la estimaci?n computacionalmente m?s eficiente.

- Paragraph's idea 4: what can be gain from this merge

De la parametrizaci?n anterior se espera que, al igual que los modelos para respuestas dicot?micas, la ICRF de la alternativa "correcta" sea monot?nicamente creciente respecto a la "habilidad", mientras que la forma de las ICRFs de los distractores depender? de como la alternativa sea percibida por el evaluado (**?**). De este modo, se plantea estudiar la formulaci?n, supuestos, caracter?sticas y propiedades del NRM.

De manera complementaria al estudio del modelo, el presente proyecto plantea la estimaci?n de los par?metros de inter?s a trav?s de simulaciones de **Cadenas de Markov de Montecarlo (MCMC)**, perteneciente a los m?todos de inferencia bayesiana. Se elige los m?todos bayesianos debido a que: (i) elimina los problemas de no convergencia

y estimaci?n impropia de los par?metros encontrados en los procedimientos de m?xima verosimilitud conjunta y/o marginal (**?**), (ii) bajo escenarios en los que la complejidad del modelo incrementa, el m?todo se vuelve m?s atractivo, pues usa simulaciones en vez de m?todos num?ricos; (iii) los modelos MCMC se vuelven particularmente ?tiles cuando los datos son dispersos o cuando es poco probable que la teor?a asint?tica se mantenga (**?**); (iv) la flexibilidad y escalabilidad de las soluciones implementadas y (v) una mayor capacidad de recuperaci?n de par?metros de inter?s, de los cuales existen muchos ejemplos (**?**, **?**, entre otros).

- Paragraph's idea 5: What are the difficulties
- Paragraph's conclusion: SEM/IRT merge provides multiple benefits

3 Data

Con respecto a los requisitos generales, para ser docente en Perú es necesario poseer el título de profesor o licenciado en educación, otorgado por una institución de formación docente acreditada en el país o en el exterior (en este último caso, el título debe ser revalidado en el Perú) 16 Además de los requisitos generales también se deben cumplir requisitos específicos, por ejemplo: a) manejar fluidamente la lengua materna de los estudiantes y conocer la cultura local para postular a vacantes de instituciones educativas pertenecientes a educación intercultural bilingüe (EIB); b) acreditar la especialización en la modalidad para postular a vacantes de instituciones educativas pertenecientes a educación básica especial (EBE); y c) se permite enseñar en inicial a los docentes con título de profesor o de licenciado en educación en la modalidad de educación básica regular (EBR) en el nivel primaria, con estudios concluidos de segunda especialidad en educación inicial y con experiencia mínima de dos (02) años lectivos en el nivel inicial.

- Paragraph's main point: What data do we have?

Finalmente, el modelo investigado ser? aplicado a un conjunto de datos reales pertenecientes al sector educativo.

- Paragraph's idea 1: Standardized MCQ in Peru for multiple purposes

En el actual escenario de la revalorizaci?n de la carrera magisterial[7][8][9][10], el Ministerio de Educaci?n del Per? (MINEDU) aprob? en el a?o 2012 e inici? la implementaci?n en el a?o 2014 las evaluaciones a docentes con el prop?sito de: (i) evaluar las capacidades y/o competencias de los docentes nombrados en las especialidades que corresponden a su ense?anza y (ii) revalorizar las escalas salariales de los docentes nombrados. En este contexto, en el a?o 2015, el ministerio aplic? la evaluaci?n de "Ingreso a la Carrera Publica Magisterial y Contratacion Docente" (en adelante **Nombramiento 2015**), la cual permiti? el ingreso de nuevos docentes a la primera de las siete escalas de la carrera magisterial.

- Paragraph's idea 2: Definition of the sample and variables

El presente proyecto opt? por implementar el modelo investigado en 40 de los 90 ?tems disponibles de Nombramiento 2015, aplicados a 11826 docentes de la especialidad

[7]Ley N? 28044, Ley General de Educaci?n

[8]Ley N? 29944, Ley de Reforma Magisterial

[9]Decreto Supremo N? 011-2012-ED, que aprueba el Reglamento de La Ley de Educaci?n

[10]Decreto Supremo N? 004-2013-ED, que aprueba el Reglamento de la Ley de Reforma Magisterial, y sus modificaciones

de Matem?tica de la Modalidad de Educaci?n B?sica Regular Nivel Secundaria. El instrumento se encuentra dise?ado para medir un ***trazo latente unidimensional*** que corresponde a las ***competencias pedag?gicas y de especialidad*** que los docentes poseen. La elecci?n del modelo se sustent? en que este no solo provee informaci?n acerca de la alternativa elegida (presuntamente "correcta"), sino tambien, permite conocer la "popularidad" con la que el individuo percibe el resto de categor?as disponibles, informaci?n especialmente valiosa para el an?lisis de distractores y validez te?rica de constructo de los ?tems utilizados en el instrumentos de evaluaci?n.

- Paragraph's idea 3: Composition of the exam - Paragraph's idea 4: Selection of factors and why - Paragraph's conclusion: The process can be performed in this data

En conclusi?n, el presente proyecto de tesis estudiar? los supuestos, propiedades y caracter?sticas del Modelo de Respuesta Nominal (NRM) e implementar? la estimaci?n de sus par?metros desde el enfoque de la inferencia bayesiana. Entre los t?picos que adicionalmente ser?n presentados se encuentran: (i) estudios de simulaci?n que comparan la recuperaci?n de par?metros de inter?s entre el m?todo cl?sico de estimaci?n y el bayesiano y (ii) la aplicaci?n a un conjuntos de datos reales del sector educativo, acorde con lo detallado en parrafos previos.

4 Thesis objectives

El objetivo general de la tesis consiste en estudiar la formulaci?n, supuestos, caracter?sticas y propiedades del **Modelo de Respuesta Nominal (NRM)** en el contexto de la Teor?a de Respuesta al ?tem (IRT). Del mismo modo, se pretende realizar un estudio de simulaci?n que compare el m?todo cl?sico de estimaci?n del NRM frente a los m?todos bayesianos. Finalmente, se aplicar? el modelo descrito a un conjuntos de datos reales del sector educativo, desde el enfoque de la inferencia bayesiana. De manera espec?fica:

- Se realizar? una extensiva revisi?n de la literatura acerca del modelo de inter?s.
- Se estudiar?n los supuestos, caracter?sticas y propiedades del modelo, desde la perspectiva cl?sica y bayesiana.
- Se implementar?n m?todos de inferencia bayesiana para la estimaci?n de los par?metros de inter?s.
- Se realizar?n estudios de simulaci?n para comprobar la capacidad de recuperaci?n de los par?metros de inter?s por parte del m?todo cl?sico y bayesiano.
- Se aplicar? el modelo de inter?s a un conjunto de datos reales pertenecientes al sector educativo.

References

Angrist, J. and Krueger, A. (1999). Chapter 23 empirical strategies in labor economics, *in* O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol. 3, Elsevier, 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.

Bertoni, E., Elacqua, G., Méndez, C., Montalva, V., Munevar, I., Olsen, A. and Román, A. (2020). Concurso docentes en latinoamérica: Claves para mejorar calidad, eficiencia y equidad en educación, *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*

Brown, T. (2015). *Confirmatory Factor Analysis for Applied Research*, Methodology in the Social Sciences, The Guilford Press.

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). Chapter 18 teacher quality, *in* E. Hanushek and F. Welch (eds), *Handbook of the Economics of Education*, Vol. 2, Elsevier, 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

Kamata, A. and Bauer, D. (2008). A note on the relation between factor analytic and item response theory models, *Structural Equation Modeling: A Multidisciplinary Journal* **15**(1): 136–153.
URL: *https://doi.org/10.1080/10705510701758406*

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=METGatheringFee*

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.

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*

Muthén, B., Kao, C. and Burstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Applications of a new irt-based detection technique, *Journal of Educational Measurement* **28**(1): 1–22.
URL: *https://files.eric.ed.gov/fulltext/ED338678.pdf*

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

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*

Sutcher, L., Darling-Hammond, L., and Carver-Thomas, D. (2016). A coming crisis in teaching? teacher supply, demand, and shortages in the u.s., *Technical report*, Learning Policy Institute.

Takane, Y. and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables, *Psychometrika* **52**(3): 393–408.
URL: *https://doi.org/10.1007/BF02294363*

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