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, we have found evidence of the positive impacts that an effective teacher can have, not only in the learning abilities but also in long-term outcomes, related to the development of the students. Specifically, Rockoff (2004); Rivkin et al. (2005); Duflo et al. (2009); Hanushek and Rivkin (2012); Muralidharan and Sundararaman (2013); Chetty et al. (2014a) and Araujo et al. (2016) showed that having an effective teacher significantly raise the student's reading and math test scores, hinting an improved knowledge of the subjects and a possible reduction of the learning gaps. Additionally, Araujo et al. (2016) found that better pedagogical practices are positively and significantly associated with the children's executive functions (working memory, cognitive flexibility, among others), while Chetty et al. (2014b) found that students assigned to effective teachers are more likely to attend college, earn higher salaries, and are less likely to have children as teenagers, among other long-lasting positive effects.

While it is appropriate to mention the positive side, it is also convenient to emphasize what happens when the students are exposed to ineffective teaching practices. Hanushek and Rivkin (2012) found, with no surprise, that having an effective teacher followed by an equally ineffective one, could cancel out the predicted lifetime earning gains of the students. Similarly, Ayala (2017) and Marotta (2019) showed that the detrimental effects of lower pedagogical practices have been more prevalent in temporary teachers, as they are usually associated with inferior attributes, compared to their contracted counterparts (Bertoni, Elacqua, Marotta, Martinez, Méndez, Montalva, Olsen, Santos and Soares, 2020), although these results have been highly dependent to the group's perceived incentives (Duflo et al., 2009; Muralidharan and Sundararaman, 2013; Chetty et al., 2015).

In summary, the literature has shown that effective teachers have multiple short and long term positive effects on students and also hinted that equally ineffective pedagogical practices can erase such benefits. This only provides more ground to the common perception that teachers are the main driver behind the students' learning process and achievement.

With this evidence, the design of an assessment system that can attract, select, develop, and retain the most effective teachers, should be the main agenda of any country's educational authority (Elacqua et al., 2018). For that purpose, it is paramount that such an organism first define what a "good" teacher should know and know how to do (Cruz-Aguayo et al., 2020), i.e. define a Standard of Teaching Performance (SoTP).

By the establishment of the SoTP, the educational authorities can set clear expectations about what is required from the teacher in the classroom. Cruz-Aguayo et al. (2020) already hinted that, beyond any specific, these requirements can be largely grouped into two: (a) to have the disciplinary knowledge and pedagogical practices, adequate to their classroom characteristics, context and teaching level, and (b) to display such knowledge and practices in the classroom, using the appropriate material and technological resources available. The evidence suggests that the teacher's knowledge (disciplinary and pedagogical) is a relevant observable factor that is part of the overall teacher quality and it is consistently associated with growth in student achievement (Metzler and Woessmann, 2012; Araujo et al., 2016).

But the questions on literature remains: how can we identify such qualities?. It is

clear that at the moment of design, that the educational authorities face a multiplicity of key decision, e.g. the selection of instruments and their weights in the final score, ways of implementation, among others.

Considerable uncertainty remains, however, concerning exactly which aspects of teachers are important, whether those aspects can be measured, and whether that effectiveness differs by type of student. (Clotfelter et al., 2006)

Without knowing what leads to better or worse performance, it is hard to know what should be done to train teachers. It is hard to know how to hire teachers who have no observed performance. And it is hard to decide on such issues as mentoring new teachers or providing professional development. (Hanushek and Rivkin, 2012)

- Paragraph's main point: How to identify good teachers

while teacher quality may be important, variation in teacher quality is driven by characteristics that are difficult or impossible to measure. (Rockoff, 2004)

A one-standard-deviation increase in teacher quality raises test scores by approximately 0.1 standard deviations in reading and math on nationally standardized distributions of achievement. (Rockoff, 2004)

- Paragraph's idea 1: multiple instruments are the key

Entonces, dada la complejidad del trabajo docente, es recomendable utilizar una multiplicidad de instrumentos para identificar y seleccionar a los mejores docentes. capacidad de impactar positivamente en sus estudiantes, la evaluación docente debe basarse en instrumentos que otorguen información válida 2 y confiable 3 cualidades son fundamentales para que la autoridad educativa pueda entender cuáles son las fortalezas y las debilidades de sus docentes, y pueda llevar adelante las acciones necesarias para potenciarlas o superarlas, respectivamente

2 Cuando los resultados que surgen de su implementación permiten identificar, mediante investigaciones rigurosas, docentes altamente efectivos o que tienen un impacto en el aprendizaje de sus estudiantes 3 Cuando los resultados que se obtienen de cada docente reflejan su desempeño típico en clase y no dependen del día en particular en que la información fue recolectada o de la persona que estuvo a cargo de esa recolección. (Bertoni, Elacqua, Méndez, Montalva, Munevar, Olsen and Román, 2020)

Un instrumento (o un conjunto de instrumentos) de evaluación docente otorga información válida si los resultados que surgen de su implementación permiten identificar, mediante investigaciones rigurosas, docentes altamente efectivos o que tienen un impacto en el aprendizaje de sus estudiantes. Un instrumento (o un conjunto de instrumentos) otorga información confiable cuando los resultados que se obtienen de cada docente reflejan su desempeño típico en clase y no dependen del día en particular en que la información fue recolectada o de la persona que estuvo a cargo de esa recolección. Estas dos cualidades son fundamentales para que la autoridad educativa pueda entender cuáles son las fortalezas y las debilidades de sus docentes y pueda llevar adelante las acciones necesarias para potenciarlas o superarlas, respectivamente. Si el diseño y la implementación de las evaluaciones docentes no aseguran que los instrumentos y la información que surja de ellas sean válidos y confiables, es poco probable que los recursos invertidos en ellas tengan frutos verdaderos. Incluso si esta primera condición se cumple es importante también el uso que se les dé a los resultados que arrojen las evaluaciones. (Cruz-Aguayo et al., 2020)

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 2: standardize MCQ evaluations are good enough (at least for the

purpose)

Bertoni et al (2020) - Concursos Docentes en Latinoamérica Existe amplia evidencia de que las pruebas de conocimiento y las observaciones estandarizadas de aula son instrumentos relacionados con una mayor efectividad docente (Bruno y Strunk, 2019; Kane et al., 2011; Kane y Staiger, 2012), así como también las entrevistas por parte del director u otro funcionario (Harris y Sass, 2014; Jacob y Lefgren, 2008).

estudios encuentran que los puntajes obtenidos por los docentes en pruebas de conocimientos están asociados a mayores aprendizajes de los estudiantes (Bietenbeck et al., 2018; Clotfelter et al., 2006, 2007)

Hincapie et al (2020) - Profesores a prueba Pruebas estandarizadas a los docentes Probablemente, la mayor ventaja de este instrumento es que, una vez su diseño asegure que las preguntas incluidas efectivamente evalúan si el docente cumple o no con los estándares de desempeño requeridos, su implementación es mucho más sencilla y menos costosa que la de los otros instrumentos arriba referenciados. Además, es el único instrumento para el cual hay evidencia causal positiva de la región que muestra cómo su implementación puede efectivamente mejorar el desempeño estudiantil (Ome, 2012; Brutti y Sánchez, 2017; Estrada, 2019).

Higher licensure test scores are associated with higher-test scores. Students assigned to teachers with higher licensure test scores apparently do better in math, but the effect is relatively modest. A one-standard-deviation increase in teacher test score implies at most a 0.017 standard deviation increase in average student math test scores and a somewhat smaller increase in reading scores. (Clotfelter et al., 2006)

higher average test scores are associated with higher math and reading achievement, with far larger effects for math than for reading. (Clotfelter et al., 2007)

The empirical analysis draws on data from the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ). After measurement-error correction, a one-standard-deviation increase in teacher subject knowledge raises student performance by 4% of a standard deviation. Results are robust to adding teacher fixed effects and are not driven by student or teacher sorting. Furthermore, teacher knowledge and school resources appear to be complements in student learning. This implies that teacher subject knowledge explains about 25% of the variation in teachers' overall effectiveness. We find that such resource based policies may be more effective in the presence of highly knowledgeable teachers. Interestingly, the coefficient on teacher subject knowledge changes only little when teacher characteristics, such as educational attainment and experience, are also controlled for. (Bietenbeck et al., 2018)

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)

there is a failure in the quantifiable characteristics to explain much of the variation in teacher effectiveness, possibly putting a missguided emphasis in assessments to enter the teaching career (Hanushek and Rivkin, 2012)

- Paragraph's idea 3: For now the evidence has been based on proxies, but more can be done

De acuerdo con (Sutcher et al., 2016), los principales indicadores que han sido utilizados

en la literatura son: 1) porcentaje de vacantes con dificultades para ser llenadas en los concursos docentes; 2) tamaños de clase, ya que la mayoría de los sistemas tienen límites máximos del número de estudiantes por profesor; 3) Porcentaje de docentes sin la preparación necesaria, de acuerdo a los estándares de formación inicial establecidos en la legislación; 4) Porcentaje de docentes sin experiencia; 5) Porcentaje de docentes con contratos temporales, sustitutos o con certificaciones ad-hoc para ejercer la docencia; 6) Profesores que enseñan una materia o asignatura 1 distinta a aquella en la que se especializaron o en la que obtuvieron su licencia (out of field teachers); 7) Docentes representativos de minorías étnicas, indígenas o de necesidades especiales. Cada uno de estos indicadores tiene ventajas y desventajas para aproximarse al concepto de escasez de docentes.

(Bertoni, Elacqua, Marotta, Martinez, Méndez, Montalva, Olsen, Santos and Soares, 2020) En segundo lugar, para Brasil, Chile y Ecuador, se consideran indicadores de idoneidad de los profesores. Según Medeiros et al., (2018), la falta de idoneidad (out-of-field-teaching, en inglés), ocurre cuando el docente enseña una materia en la que no tiene la especialidad y/o el certificado correspondiente

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.

Sin embargo se puede tener una medicion de idoneidad mas precisa, si se conocen los instrumentos. Solo usan el puntaje para medir un proxy de idoneidad

últimas décadas del siglo XX. Además, las habilidades pedagógicas y los conocimientos disciplinares de los docentes en la región están por debajo de lo que sugieren los estándares internacionales (Cruz-Aguayo et al., 2020). Por ejemplo, los docentes de la región dedican 20% menos del tiempo efectivo en clase de lo que estos recomiendan. Es decir, en América Latina, las diferencias en tiempos de enseñanza efectiva de los docentes implican que los estudiantes de la región reciben en promedio un día menos de clase a la semana (Bruns et al., 2015). 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).

First, neither a graduate degree nor additional years of experience past the initial year or two translate into significantly higher instructional effectiveness. Second, descriptions of unequal access to quality teachers as measured by experience, education, or other quantifiable characteristics fail to portray accurately any actual differences in the quality of instruction by student demographics, community characteristics, and specific schools (Hanushek and Rivkin, 2012)

The analysis of teacher effectiveness has largely turned away from attempts to identify specific characteristics of teachers. Instead attention has focused directly on the relationship

between teachers and student outcomes. This outcome-based perspective, now commonly called value-added analysis, takes the perspective that a good teacher is simply one who consistently gets higher achievement from students (after controlling for other determinants of student achievement such as family influences or prior teachers). (Hanushek and Rivkin, 2012)

we find little evidence that any observable teacher characteristic, save experience, explains any of this variation. (Clotfelter et al., 2006)

Ultimately, two characteristics – teacher experience and licensure test scores – emerge as robust determinants of test scores for fifth grade students. Compared to students assigned to teachers with no prior experience, students assigned to highly experienced teachers attain standardized reading and math test scores roughly one-tenth of a standard deviation higher in math and slightly less than a tenth of a standard deviation in reading. About half of this gain occurs for the first one or two years of teaching. Students assigned to teachers with higher licensure test scores apparently do better in math, but the effect is relatively modest. A one-standard-deviation increase in teacher test score implies at most a 0.017 standard deviation increase in average student math test scores and a somewhat smaller increase in reading scores. (Clotfelter et al., 2006)

The most surprising result is the consistently negative effect of a master's degree on student achievement. The coefficients suggest that, all else constant, teachers with master's degrees are less effective than those without. (Clotfelter et al., 2006)

We conclude that a teacher's experience, test scores and regular licensure all have positive effects on student achievement, with larger effects for math than for reading. Taken together the various teacher credentials exhibit quite large effects on math achievement, whether compared to the effects of changes in class size or to the socio-economic characteristics of students. (Clotfelter et al., 2007)

As expected, we find clear evidence that teachers with more experience are more effective in raising student achievement than those with less experience. Though the positive results by years of teacher experience are clear and robust to various model specifications, the thorny issue remains of whether the rising returns to experience reflect improvement with experience or differentially higher attrition of the less effective teachers (Rockoff, 2004).

Having a graduate degree exerts no statistically significant effect on student achievement and in some cases the coefficient is negative. Thus, the higher pay for graduate degrees would appear to be money that is not well spent, except to the extent that the option of getting a master's degree keeps effective experienced teachers in the profession. (Clotfelter et al., 2007)

teachers have powerful effects on reading and mathematics achievement, though little of the variation in teacher quality is explained by observable characteristics such as education or experience. The results suggest that the effects of a costly ten student reduction in class size are smaller than the benefit of moving one standard deviation up the teacher quality distribution, highlighting the importance of teacher effectiveness in the determination of school quality. (Rivkin et al., 2005)

Consistent with prior findings, there is no evidence that a master's degree raises teacher effectiveness. In addition, experience is not significantly related to achievement following the initial years in the profession. (Rivkin et al., 2005)

I also find evidence that teaching experience significantly raises student test scores, particularly in reading subject areas. Reading test scores differ by approximately 0.17

standard deviations on average between beginning teachers and teachers with ten or more years of experience. Evidence of gains from experience for math subjects is weaker. The first two years of teaching experience appear to raise scores significantly in math computation (about 0.1 standard deviations). However, subsequent years of experience appear to lower test scores, though standard errors are too large to conclude anything definitive about this latter trend. (Rockoff, 2004)

easily-observed teacher characteristics, such as education, gender, and teaching experience (except for the first few years), are not consistently related to teacher effectiveness (Hanushek and Rivkin, 2006).

At least within our sample of Peruvian teachers, there is no indication that the effect of teacher subject knowledge levels out at higher knowledge levels. (Metzler and Woessmann, 2012)

- Paragraph's idea 4: Selection has been based on scores

En el contexto de un proceso de evaluaci?n, muchas veces el individuo se ve enfrentado a una prueba "estandarizada"; es decir, una evaluaci?n dise?ada de tal manera que las preguntas, las condiciones para ser administrada, los procedimientos de calificaci?n e interpretaciones son consistentes con una manera predeterminada o tipificada. En este contexto, el individuo es expuesto a un instrumento de evaluaci?n cuyos ?tems usualmente cumplen con las siguientes tres caracter?sticas: (i) preguntas de opci?n m?ltiple o polit?micas, (ii) preguntas que exhiben categor?as nominales, sin un ordenamiento espec?fico y (iii) una respuesta "correcta"; tal y como se observa en la **Figura ??**

(measurement error problem)

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.

Other analyses have emphasized the importance of measurement error in using test outcome data (e.g., Kane Staiger 2002, McCaffrey et al. 2009).

Like all test scores, teacher subject knowledge in our data is likely measured with error. Measurement error in the explanatory variable might lead to an attenuation bias, which is aggravated in the student fixed-effects model ((Angrist and Krueger, 1999), Section 4). 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).

As is well known, measurement error in the explanatory variable may lead to a downward bias in the estimated coefficient, and this bias may be aggravated in the student fixed-effects models (Angrist and Krueger, 1999)

The only teacher trait consistently associated with gains in student performance is teacher cognitive skills as measured by achievement tests ((Hanushek and Rivkin, 2006; Rockoff et al., 2011)). In the context of developing countries, several studies have found positive correlations between teacher test scores and student achievement; see, for example, (**?**) for Mexico, (Marshall, 2009) for Guatemala, and Behrman et al. (2008) for Pakistan.

We find that teacher subject knowledge exerts a statistically and quantitatively significant impact on student achievement. After measurement-error correction, one standard deviation

in subject-specific teacher achievement increases student achievement by about 9% of a standard deviation in math. Effects in reading are significantly smaller and mostly not significantly (Metzler and Woessmann, 2012)

The available measure may proxy only poorly for the concept of teachers' subject knowledge, as in most existing studies, the examined skill is not subject-specific. Furthermore, any specific test will measure teacher subject knowledge only with considerable noise. (Metzler and Woessmann, 2012)

- 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)

standardized evaluation help to know the knowledge of the teacher

(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 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.

Without knowing what leads to better or worse performance, it is hard to know what should be done to train teachers or how to / which to hire them (Hanushek and Rivkin, 2012)

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

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

- Paragraph's main point: what method are you using?
- 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
- 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[1][2][3][4], 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 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.

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

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

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

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

3 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.

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.

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*

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*

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*

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*

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

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