Un modelo TRI de múltiples facetas para la evaluación del desempeño docente en el aula

Karen Rosana Córdoba Perozo
Estadística

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Enero de 2020

Un modelo TRI de múltiples facetas para la evaluación del desempeño docente en el aula

Karen Rosana Córdoba Perozo
Estadística

Disertación presentada para optar al título de
Magister en Ciencias Estadística

Director
Álvaro Mauricio Montenegro Díaz, Ph.D.
Doctor en Estadística

Línea de investigación
Teoría de respuesta al ítem

Grupo de investigación
SICS Research Group

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Enero de 2020

Título en español

Un modelo TRI de múltiples facetas para la evaluación del desempeño docente en el aula

Title in English

A multi-faceted TRI model for the evaluation of teacher performance in the classroom

Resumen: La evaluación del desempeño docente en la educación superior ha sido un tema controversial y de frecuente uso en este campo para la toma de decisiones. En la mayoría de casos, se basan en el supuesto de que los estudiantes aprenden más de profesores altamente calificados y su validez se sustenta en el hecho de que los estudiantes observan el desempeño de los docentes en el salón de clase. Por lo tanto, deberían ser ellos los evaluadores bajo el supuesto que responderán sinceramente cuando sean preguntados por el desempeño del docente. Sin embargo, muchos estudios ponen en duda las metodologías usadas para dichas mediciones, en general, porque los promedios de respuestas categóricas tienen poco sentido estadístico. En este documento, se propone la medición del desempeño docente a través de un modelo de TRI de múltiples facetas que tiene en cuenta parámetros asociados a la severidad del evaluador y un parámetro adicional que está relacionado con el efecto del curso evaluado. Para la estimación del modelo se hizo uso de técnicas de inferencia bayesiana debido a la gran cantidad de parámetros a estimar. La propuesta se aplicó a un conjunto de datos obtenido de una encuesta de percepción de desempeño docente realizada por la Facultad de Ciencias de la Universidad Nacional de Colombia a estudiantes de la misma durante el año 2015. El modelo propuesto se evaluó con estadísticas de bondad de ajuste y se comparó con un modelo que no contempla el parámetro adicional. Los resultados obtenidos indican que el modelo propuesto presentó mejor ajuste y los criterios de selección indican que el modelo propuesto es más apropiado para la medición. Adicionalmente, haciendo uso de variables auxiliares se observaron diferencias entre las calificaciones otorgadas por estudiantes mujeres y estudiantes hombres y por estudiantes pertenecientes a programas de posgrado y de pregrado, además, se observaron diferencias entre las dificultades promedio estimadas de los cursos según el área de estudio

Abstract: Teacher evaluation in higher education has been a controversial topic and its use is frequent in this field for taking decision. In most of cases, the evaluation is based on the assumption that students learn more from highly qualified teachers and its validity is based on the fact that students observe the teachers performances in the classroom. Therefore, the students should be the evaluators under the assumption that they will respond sincerely when asked about the teacher performance. However, many studies question about the methodologies used for getting such measurements, in general, because the averages by categorical responses have little statistical sense. In this document, the measurement of teaching performance is proposed through a multi-faceted TRI model that takes into account parameters associated with the severity of the evaluator and an additional parameter that is related to the effect of the evaluated course. For the model estimation, Bayesian inference techniques were used due to the large number of parameters to be estimated. The proposal was applied to a data set obtained from a survey of perception of teaching performance conducted by the Faculty of Science of the National University of Colombia to students of the same faculty. The proposed model was evaluated with goodness of fit statistics and compared with a model that did not

include the additional parameter. The results obtained indicate that the proposed model had a better fit and the selection criteria indicate that the proposed model was more appropriate for the measurement. Additionally, using auxiliary variables, differences were observed between the qualifications given by female students and male students and the qualifications by postgraduate students and undergraduate students, in addition, there were differences between the estimated average difficulties of the courses according to the study area that offers.

Palabras clave: Modelos de TRI de múltiples facetas, desempeño docente, inferencia bayesiana

Keywords: Multi-faceted TRI model, Teacher performance, Bayesian inference

Nota de aceptación

Trabajo de tesis
Aprobado

"Mención Meritoria o Laureada"

Jurado

Jurado

Jurado

Director
Álvaro Mauricio Montenegro Díaz

Bogotá, D.C., Enero de 2020

Dedicado a

A mi padre y a mi madre, por que son la luz que guía mi vida.

A mis hermanos Michel y Jesús complementos de mi vida en todo el tiempo de nuestra existencia.

A mis abuelas, ejemplos de vida que me acompañan desde el cielo.

Gracias por su cariño, por su amor, por su compañia y su confianza. La vida no sería la misma sin su amor. Los amo con mi vida.

Agradecimientos

Al profesor Álvaro, director de esta tesis, por su acompañamiento en todos estos años de estudio, por su ayuda y paciencia para ser guía de este trabajo.

A mi hermano y colega Michel por sus aportes, su guía y su acompañamiento.

A mis papás y hermano Jesús por estar siempre pendientes de cada paso en este importante proyecto.

A mi colega Cristian Montaño por darme siempre el impulso y su compañía para trabajar en este proyecto.

A mis amigos porque con su apoyo y aliento durante este tiempo logré mantener el impulso necesario para seguir trabajando.

Índice general

Índice de tablas

Índice de figuras

Introducción

En la actualidad, en diversos ámbitos se realiza medición de habilidades a través de pruebas o evaluaciones. Las pruebas pueden tener una variedad de funciones y, a menudo, se realiza una clasificación de las mismas. Por un lado, se consideran las pruebas cognitivas, las cuales miden habilidades o capacidades de los evaluados, y por otro lado, se consideran las pruebas no cognitivas que están diseñadas para medir constructos relacionados con actitudes y en general, con otros aspectos no cognitivos.

El uso de modelos de evaluación docente para la educación superior ha sido un tema controversial en el campo de la educación. Las evaluaciones de efectividad del desempeño docente SET (por sus siglas en inglés: Student evaluation of teaching) se basan en el supuesto de que los estudiantes aprenden más de profesores altamente calificados. Además, la validez de dichas evaluaciones se sustenta en el hecho de que los estudiantes observan el desempeño de los docentes en el salón de clase y por lo tanto responderán sinceramente cuando sean preguntados por dicho desempeño.

Sin embargo, las metodologías propuestas han sido puestas en duda por la comunidad académica, debido a la naturaleza de las mismas. Con frecuencia, se menciona el hecho de que existen factores adicionales sobre los que no se tiene control y que afectan de manera importante las evaluaciones del desempeño docente, ejemplos de dichos factores son: el campo de estudio, la motivación o interés del estudiante, el sexo del docente, el tamaño de la clase, etc. (Wachtel, 1998). Hay pruebas sólidas de que las respuestas de los estudiantes a las preguntas de "efectividad de la enseñanza"no miden este constructo. La comparación de promedios de respuestas categóricas, incluso si las categorías están representadas por números, tiene poco sentido estadístico(Stark & Freishtat, 2014). Por lo tanto, el uso en la práctica de promedios de los puntajes de las SET de los alumnos debe replantearse.

Por otra parte, los estudios de evaluación del desempeño docente se han convertido en un aspecto importante para la toma de decisiones administrativas en las instituciones de educación superior. Becker & Watts (1999) fueron los primeros en informar que las evaluaciones de los estudiantes en la enseñanza docente fueron el método más utilizado, y en general, el único, para evaluar la enseñanza en cursos de economía de pregrado. Además, concluyen que dichas evaluaciones se usan para tomar decisiones de promoción, en lugar de ser una parte regular de la evaluación de la enseñanza de un profesor.

En el presente documento se aborda la problemática de la validez de los instrumentos de evaluación docente a través de la implementación de modelos estadísticos. Los procedimientos usados para la generación de puntuaciones con dichos instrumentos son discutibles estadísticamente. La razón principal es que dichas puntuaciones se extraen de variables categóricas, donde existe un orden natural y por lo tanto, sus valores no deberían ser

promediados. En consecuencia, el objetivo de este trabajo es plantear el uso de un modelo de teoría de respuesta al ítem de múltiples facetas para la medición de la habilidad de la enseñanza docente a través de encuestas a los estudiantes. El modelo tiene en cuenta los factores adicionales que hacen parte del momento de evaluación a través de la estimación de un parámetro de severidad del evaluador, que contiene en sí mismo dichos factores adicionales. Adicionalmente, se trabajan los datos de la evaluación docente aplicada en la Facultad de Ciencias de la Universidad Nacional de Colombia en el año 2016.

A continuación, se describe el esquema del documento. En el capítulo 1 se presenta una introducción a la evaluación del desempeño docente mediante una revisión histórica y un abordaje de sus puntos a favor y en contra. En el capítulo 2 se ofrece una introducción a los modelos de teoría de respuesta al ítem de múltiples facetas. El capítulo 3 presenta una introducción al análisis de datos bayesiano. El capítulo 4 presenta la propuesta de este documento, que es el abordaje de evaluación del desempeño docente a través de modelos TRI de múltiples facetas. Finalmente, el capítulo 5 presenta una aplicación del modelo propuesto con datos recolectados en la Facultad de Ciencias de la Universidad Nacional de Colombia y una breve introducción al lenguaje de programación usado.

CAPÍTULO 1

Evaluación del desempeño docente

Las evaluaciones de efectividad del desempeño docente SET (por sus siglas en inglés: Student evaluation of teaching) son instrumentos con unas características especiales. Una de las características es que, en general, estas evaluaciones incluyen preguntas de tipo abierto y cerrado que indagan temas relacionados con el desempeño del docente en el curso y con el desarrollo del mismo. Ejemplos de estos tipos de preguntas pueden ser: ¿El docente incluye experiencias de aprendizaje en lugares diferentes al aula (talleres, laboratorios, etc.)? o ¿qué fortalezas destaca del docente? Otra característica importante de este tipo de evaluaciones es el uso de escalas consistentes a través de todo el cuestionario, además de que generalmente son evaluaciones cortas que no toman más de 10 o 15 minutos en ser respondidas por los estudiantes.

Sin embargo, es importante advertir que este tipo de evaluaciones deben tener un momento de aplicación efectivo, es decir, se debe informar a los estudiantes que las respuestas que ellos brindan son importantes y sirven de insumo para la toma de decisiones en la institución y que, además, se garantiza la confidencialidad de los datos.

En Bélanger & Longden (2009) se realiza una analogía de la relación entre un estudiante y un docente con una transacción convencional, en el sentido en que si se percibe que los docentes hacen un buen trabajo entonces las instituciones de educación superior aumentan su prestigio como proveedores de buena educación, y esto se logra en cierta manera a través del uso de los resultados de herramientas de evaluación de los docentes como las SET.

En el transcurso de los estudios universitarios, los estudiantes se exponen a diferentes estilos de enseñanza que provienen de diferentes profesores, por lo tanto, una hipótesis al respecto de las SET podría ser que los estudiantes dan mejores calificaciones a los docentes de los cuales consideran aprendieron más. En Bélanger & Longden (2009) se menciona que no es conveniente evaluar la eficacia de la enseñanza del docente sin tener en cuenta esos factores externos que provienen de los aprendizajes de los estudiantes.

Para el caso de Colombia es importante mencionar que en general los profesores son libres de desarrollar su estilo de enseñanza. Sin embargo, debería ser de interés conocer cuáles aspectos les gustan y cuáles no a los estudiantes sobre el estilo de enseñanza que reciben. Para lo anterior, Bélanger & Longden (2009) proponen tres conjuntos principales de características a conocer: la personalidad del profesor, el entorno del aula y el estilo de enseñanza.

Para medir las tres características anteriores, a través de las SET, se les pide a los estudiantes que califiquen sus percepciones de los docentes y de los cursos mediante formularios que presentan a menudo una escala Likert de 5 puntos que van desde *Muy en desacuerdo* hasta *Muy de acuerdo*. En general, estas evaluaciones se llevan a cabo en las últimas semanas de los cursos, antes de que se asignen las calificaciones finales.

En Marsh (2007) se menciona que la aplicación de las SET tiene algunos objetivos:

- Brindar un diagnóstico a la facultades sobre los docentes.
- Brindar medidas de la eficacia de la enseñanza para la toma de decisiones sobre el personal de la institución.
- Dar información a los estudiantes para la selección de cursos y de docentes.

En este capítulo se realiza una revisión histórica sobre el uso de la evaluación del desempeño docente, además, se hace una revisión sobre la discusión académica que existe alrededor del uso de estos instrumentos, y se presenta un análisis sobre los factores externos que influyen en los resultados de las SET.

1.1. Revisión histórica

Spencer & Flyr (1992) mencionan que la primera escala de calificación formal para una evaluación docente fue publicada en 1915, y Marsh (1987) señala que los procedimientos de evaluación para los docentes por parte de los estudiantes se introdujeron en varias de las principales universidades de Estados Unidos en la década de 1920.

En el transcurso de las décadas de los 70 y 80 Feldman publicó una serie de trabajos (Feldman, 1977, 1978, 1979, 1983, 1987, 1989) relacionados con los factores que inciden en los resultados de las evaluaciones del desempeño docente.

Por su parte, en los años 1980, se realizaron dos meta-análisis; Cohen (1981) y Feldman (1989) buscaron recopilar información relacionada con las correlaciones entre los resultados de las SET y el aprendizaje de los estudiantes.

En el año 1998 se realiza una revisión sobre las investigaciones asociadas a evaluaciones de desempeño docente de profesores universitarios. Allí se menciona que "La investigación sobre las evaluaciones de los estudiantes sobre la enseñanza y los factores que pueden afectarlos se remonta a la década de 1920 y al trabajo pionero de Remmers. En dichos trabajos, se abordaron algunos de los principales problemas en el área de investigación de evaluación de estudiantes, uno de los principales, se abordó la hipótesis de si las calificación de los estudiantes coinciden con las de ex alumnos". (Wachtel, 1998)

Para el 2017, se realiza un meta-análisis actualizado de dichas correlaciones. En el documento se expone que los meta-análisis de los años 80 tienen numerosos problemas relacionados con la localización de los estudios utilizados allí y que ninguno de los meta-análisis es replicable. Además, también indican que las correlaciones reportadas en dichos documentos presentan inconsistencias por efectos del tamaño del estudio. (Uttl et al., 2017)

1.2. Soporte y oposición al uso de evaluaciones SET

Como se mencionó anteriormente, el uso de las evaluaciones de desempeño docente SET ha generado controversia dentro de la comunidad académica. En general, existen dos posturas frente a su uso: los que lo defienden y los que no.

Aquellos que se encuentran a favor exponen diferentes razones para su uso: En Wachtel (1998) se refieren en primer lugar al tema económico, pues argumentan que este tipo de evaluaciones son baratas y fáciles de implementar, en segundo lugar, afirman que las SET dan valor a las opiniones de los estudiantes, y en Murray (2005) hacen énfasis en que los estudiantes son los únicos que pueden evaluar sus percepciones de los docentes en el salón de clase.

Por otra parte, quienes se oponen a su uso argumentan que las SET son evaluaciones que miden la satisfacción del estudiante y que, por lo tanto, la satisfacción del estudiante se ve influenciada por factores externos que no están relacionados con la efectividad de la enseñanza del docente. Por ejemplo, argumentan que si un estudiante obtuvo calificaciones diferentes a las esperadas es probable que su calificación en la SET sea baja, con relación a su baja satisfacción (Uttl et al., 2012).

1.3. Variables que pueden influir en la calificación de las SET

Los estudios relacionados con el campo de la enseñanza son extensos. Se incluyen temas como métodos de enseñanza, sesgo de género, rango académico y experiencia, dificultad del curso, rasgos de personalidad, reputación del docente, sentido del humor y el grado de indulgencia (Bélanger & Longden, 2009).

Se ha argumentado que variables externas pueden producir cambios significativos en las calificaciones dadas a los docentes por parte de los estudiantes. En esta sección, se abordan las características distintas a la efectividad de la enseñanza que pueden influir en los resultados de las evaluaciones de desempeño docente.

1.3.1. Características asociadas a la administración de evaluaciones de estudiantes

Es común que al momento de administrar una evaluación se hagan unas recomendaciones en cuanto a los procedimientos para la coordinación de las mismas. A continuación, se mencionan las características administrativas que pueden influir en las puntuaciones de la calificación de éstas.

- **Tiempo de evaluación:** Feldman (1979) argumenta que el tiempo en que son administradas las evaluaciones no tiene efecto, es decir, no influye si la evaluación es aplicada a la mitad del curso, al inicio, durante o después del examen final. Por otra parte, Aleamoni (1981) menciona que los resultados de la evaluación pueden distorsionarse si esta se administra antes o después de un examen.
- **Anonimidad de evaluaciones:** Feldman (1979) reporta que estudiantes tienden a mejorar las puntuaciones cuando ellos se identifican en el instrumento de evaluación que cuando no lo hacen, es decir, evaluaciones no anonimizadas presentan mejores

puntuaciones, por lo tanto, recomienda que las evaluaciones sean anónimas debido a temas de confidencialidad.

- **Presencia del docente en el salón de clase:** Feldman (1979) también reporta que las puntuaciones son más altas cuando el docente evaluado se encuentra presente en el salón de clase. Por lo tanto, hace la recomendación de que el docente esté ausente durante el proceso de evaluación.

- **Propósito de la evaluación:** Se menciona que las puntuaciones de los estudiantes tienden a ser más altas si ellos conocen de antemano el propósito de la evaluación, es decir, si saben que es una evaluación para promover o mantener a los docentes (Feldman, 1979).

1.3.2. Características del curso

Algunos autores hacen referencia a que las características del curso también tienen influencia al momento de proporcionar las calificaciones por parte de los estudiantes. A continuación, se presentan los hallazgos más importantes en la literatura.

- **Electividad del curso:** Se ha encontrado que docentes que dictan cursos electivos reciben puntuaciones mayores que aquellos docentes que dictan cursos obligatorios. Lo anterior, puede deberse a que los estudiantes muestran mayor interés en los cursos electivos que en los obligatorios. (Feldman, 1978).

- **Horario de la clase:** Algunos autores han encontrado que no existe relación entre las puntuaciones y la hora del día en la cual se dictan las clases (Centra, 1993), (Feldman, 1978). Sin embargo, Koushki & Kunh (1982) encontraron que clases muy temprano en la mañana y clases después de almuerzo reciben puntuaciones menores que las clases que se imparten en otras horas del día.

- **Nivel del curso:** Se comenta que cursos de niveles más avanzados tienden a recibir puntuaciones más altas (Feldman, 1978). Sin embargo, se menciona que en este tipo de relación puede ser relevante la influencia de la edad de los estudiantes que toman dichos cursos (Marsh, 1987).

- **Tamaño de la clase:** Esta es una característica bastante estudiada. Feldman (1978) menciona que clases pequeñas, es decir, con pocos estudiantes, reciben mejores calificaciones. Otra hipótesis es que la relación entre el tamaño de la clase y las puntuaciones dadas por los estudiantes no es una relación lineal, sino una relación de tipo U, donde los cursos de mayor y menor tamaño tienden a tener mejores calificaciones que aquellos cursos de tamaño mediano (Centra & Creech, 1976).

- **Área temática:** En general, se señala que no se ha encontrado efecto en las puntuaciones de los estudiantes según el área de conocimiento. Sin embargo, se ha observado que áreas relacionadas con matemáticas y ciencias tienen menores puntaciones (Feldman, 1978), (Centra & Creech, 1976).

Finalmente, en Cranton & Smith (1986) se menciona el efecto que tienen las características del curso en las puntuaciones de los SET. Sin embargo, los autores también encontraron que dichos efectos varían dentro de los departamentos de las universidades.

Por lo tanto, ellos concluyen que no se puede establecer una norma única para todo la universidad, y que en caso de extender dichos efectos a otros departamentos, esto debe hacerse con precaución.

1.3.3. Características del docente

Por otra parte, es importante revisar la influencia de las características de los docentes en los resultados de las calificaciones por parte de los evaluados. A continuación, se presentan los hallazgos más importantes.

- **Rango del docente y experiencia laboral:** En general, se ha observado que los docentes en propiedad tienen mejores puntuaciones que los asistentes docentes (Centra & Creech, 1976). Sin embargo, Feldman (1983) menciona que no se encontró relación entre los años de experiencia de los docentes y las puntuaciones dadas por los estudiantes.
- **Reputación:** Perry et al. (1974) encontraron que las expectativas a priori de los estudiantes relacionadas con el desempeño de los docentes basadas en su reputación influyen en las puntuaciones dadas. Además, ellos argumentan que estos resultados también influyen en la divulgación de información en la comunidad académica.
- **Productividad en investigación:** Una de las hipótesis que se maneja en cuanto a la eficacia de la enseñanza está relacionada con la productividad de la investigación docente. A menudo, se argumenta que dicha productividad mejora la efectividad porque permite que los docentes estén actualizados en sus campos de investigación. Sin embargo, los hallazgos han mostrado que no existe una relación entre la productividad de investigación y las puntuaciones (Feldman, 1987).
- **Género:** La discusión frente al efecto del género de los docentes es variada. Muchos autores sugieren que las puntuaciones de los estudiantes son sesgadas contra las mujeres docentes (Basow & Silberg, 1987), (Martin, 1984). Por otro lado, Feldman (1992) reporta en su meta-análisis que no existen diferencias significativas entre docentes hombres y mujeres.

1.3.4. Características de los estudiantes

Así como se revisan las características de los docentes, es importante revisar si las características de los estudiantes tienen influencia en los resultados de las calificaciones.

- **Personalidad:** Existe una hipótesis frente a las características de personalidad de los estudiantes en cuanto a las puntuaciones que otorgan a los docentes. Se cree que las opiniones, actitudes y preferencias tienen influencia en la forma como se califica a los docentes. Sin embargo, Abrami et al. (1982) reportan que no existe relación entre la personalidad de los estudiantes y las puntuaciones reportadas a los docentes.
- **Interés:** La evidencia sugiere que estudiantes que muestran más interés en el área de conocimiento del curso otorgan puntuaciones más altas que aquellos que no muestran interés (Feldman, 1977).

- **Género:** Muchos estudios han reportado que no hay diferencias entre el género del estudiante en las puntuaciones que se otorgan a los docentes. Sin embargo, Feldman (1977) reporta que estudiantes mujeres dan mejores puntuaciones que los hombres.
- **Estado emocional:** Small et al. (1982) mencionan que el estado emocional de los estudiantes a final del semestre o del curso está asociado con las calificaciones brindadas a los docentes. Específicamente, mencionan que estudiantes que mostraban mayores grados de ansiedad y depresión otorgaban menores puntuaciones. Por lo anterior, los autores creen que hay un serio problema de validez al administrar las evaluaciones finalizando los periodos académicos.
- **Expectativa e hipótesis de indulgencia:** Marsh (1987) menciona la hipótesis de indulgencia. Expone que docentes con más grado de indulgencia reciben en general calificaciones favorables. Lo anterior sugiere la existencia de un sesgo debido a que docentes que "brindan" calificaciones más altas a sus estudiantes posiblemente obtendrán mejores puntuaciones por parte de éstos, lo cual amenaza la validez de las calificaciones.

1.4. Perspectiva estadística de la evaluación docente

Como se ha mencionado, las puntuaciones de desempeño docente han sido debatidas por la comunidad académica. Una parte importante de dicho debate radica en la forma de construcción de dichas puntuaciones y en la forma como se divulgan estos resultados.

En la práctica, la generación de dichas puntuaciones se realiza usando promedios aritméticos de las calificaciones de todos los estudiantes. El procedimiento anterior es discutible estadísticamente desde dos puntos. En primer lugar, dichas puntuaciones provienen de variables de tipo categórico ordinal, por lo tanto, existe un orden natural, sin embargo, dichos números son en realidad etiquetas no valores reales, es decir, no deberían ser promediados. En segundo lugar, en una escala ordinal, no es posible asumir que una distancia de 1 es equivalente en todo el rango de la escala, es decir, no se puede asumir que una diferencia entre 4 y 5 es lo mismo que una diferencia entre 1 y 2, por lo tanto, la realización de comparaciones entre dichas escalas no tiene mucho sentido.

Suponga que, en una universidad, por ejemplo, se les pide a los estudiantes que califiquen 5 ítems relacionados con la forma en que un profesor desarrolla su clase. La escala de calificación es una escala de 5 puntos, que se describe de la siguiente manera: el valor de 1 significa *Muy en desacuerdo*, el valor de 2 es *En desacuerdo*, el valor de 3 es *Ni de acuerdo ni en desacuerdo*, el valor de 4 es *De acuerdo*, y el valor de 5 es *Muy de acuerdo*. A continuación se presentan las calificaciones que otorgaron dos estudiantes al desarrollo de dicha clase:

Ítem	1	2	3	4	5	Promedio
Estudiante 1	1	3	2	4	4	2.8
Estudiante 2	2	5	4	3	4	3.6

TABLA 1.1. Ejemplo de calificación

Según los procedimientos regulares de la institución, la puntuación final del desarrollo de la clase del profesor estaría establecida por el promedio de las puntuaciones de todos

sus estudiantes. Para el caso expuesto anteriormente, se observa que la nota promedio del estudiante 1 fue de 2.8 y la nota promedio del estudiante 2 fue de 3.6. Sin embargo, en la escala de calificación descrita anteriormente, no existe ninguna etiqueta para dichos valores, y por la naturaleza de la escala no es posible asignarle ningún valor en la misma. De igual manera, no es posible asumir que la diferencia entre 1 y 2, que para el caso descrito sería la diferencia entre *Muy en desacuerdo*, y *En desacuerdo*, es la misma diferencia que entre 4 y 5, que para el caso es la magnitud entre *De acuerdo*, y *Muy de acuerdo*. Por lo tanto, asumir que el promedio de valores pertenecientes a una escala ordinal es un valor es un error.

Stark & Freishtat (2014) hacen una revisión de las puntuaciones SET desde una perspectiva estadística. En el artículo mencionan puntos importantes, relacionados con la validez de usar las puntuaciones SET.

En primera instancia, hacen un análisis sobre la tasa de respuesta a las evaluaciones, comentan que algunos estudiantes no dan respuesta a las encuestas, por lo tanto las tasas de respuesta son menores a 100 %, lo que implica un problema de representatividad. Además, hacen énfasis en que los promedios de muestras pequeñas son más susceptibles que los promedios de muestras más grandes.

También mencionan que no tiene sentido hacer comparaciones, debido a que no se puede asumir que las diferencias en la escala son igual de interpretables en todos los puntos de la misma, pues la naturaleza de la escala es de tipo ordinal. Incluso afirman que el hecho de comparar el promedio de un docente con el promedio de un curso en general no tiene sentido, porque es poco informativo. Recomiendan el uso de la distribución de las puntuaciones tanto para los docentes como para los cursos.

En segunda instancia, comentan sobre la naturaleza de dichas investigaciones relacionadas con las SET. Mencionan que dichos estudios se basan en métodos observacionales, y por lo tanto, si se quisiera realizar inferencia sería necesario desarrollar un experimento controlado y aleatorio, donde los individuos sean asignados al azar a los cursos.

Por otra parte, Braga et al. (2014) evalúan el contenido de las evaluaciones de los estudiantes en contraste con medidas de la efectividad del docente. Para lo anterior, los autores usan el rendimiento de los estudiantes para estimar medidas de la efectividad. La metodología propuesta es similar a la de valor agregado, ellos proponen comparaciones de resultados de cursos más avanzados de estudiantes que asistieron a cursos obligatorios con diferentes docentes, partiendo de la idea de que los estudiantes que fueron enseñados por mejores docentes obtuvieron mejores resultados más adelante.

En primera medida, los autores estiman la media condicional de las calificaciones futuras de los estudiantes en cada clase, mediante una regresión que tiene en cuenta un vector de características a nivel de estudiante; los parámetros de interés son los interceptos que miden las medias condicionales de las calificaciones futuras de los estudiantes en clase. En segunda medida, los autores eliminan los coeficientes estimados del efecto de otras características de clase que podrían afectar el rendimiento de los estudiantes en cursos posteriores pero que no son atribuibles a los docentes. Las medidas finales de la efectividad de la enseñanza docente son los residuales de la regresión de los interceptos estimados sobre todas las variables observables.

Respecto a los resultados encontrados, los autores mencionan que los mismos arrojan dudas sobre la validez de las evaluaciones SET como medidas de calidad, por lo cual proponen algunas soluciones. En primer lugar, sugieren que dado que las evaluaciones

de los mejores estudiantes están más alineadas con la efectividad real de los docentes, las opiniones de los mejores estudiantes podrían tener un peso mayor en la medición del desempeño docente, pero para lo anterior debe perderse cierto grado de anonimato de las evaluaciones. En segundo lugar, proponen que los cuestionarios sean administrados en un momento posterior a la terminación de los cursos para dar a los estudiantes el tiempo para apreciar el valor real de la enseñanza. Sin embargo, ésta propuesta puede tener problemas de sesgo.

Como se describió anteriormente, la validez de las evaluaciones SET se ve comprometida. Por otra parte, propuestas como la de Braga et al. (2014) pueden ser de mucha utilidad debido a que abordan el problema desde otro enfoque. No obstante, su implementación es difícil debido a que se requieren mayores esfuerzos en cuanto a temas logísticos y de disponibilidad de información.

CAPÍTULO 2

Modelos de TRI de múltiples facetas

En ciertos campos del conocimiento se requiere hacer mediciones sobre algunos conceptos en los cuales no es posible hacer mediciones directas. El concepto de desempeño docente es una variable latente, lo que implica que es necesario evaluarlo a través de variables observables. Una forma de evaluarlo es a través de las SET, que como se ha mencionado en el capítulo anterior, son instrumentos compuestos de una serie de ítems que indagan sobre el desempeño del docente y el desarrollo de los cursos por parte del mismo. Para los modelos de Teoría de Respuesta al Ítem, en adelante TRI, estas respuestas no son mediciones directas, pero proporcionan los datos a partir de los cuales se pueden inferir las mediciones (Van Der Linden & Hambleton, 1997).

En la evaluación del desempeño docente frecuentemente se tiene más de una calificación por docente, ya que los cursos en general están compuestos por varios estudiantes. Por lo tanto, se tiene información de varias fuentes, es decir, en términos matemáticos, se cuenta con una matriz de respuestas para cada uno de los evaluados. Es necesario mencionar que cada ítem o pregunta relacionada en las evaluaciones SET tiene propiedades diferentes, debido a que por su naturaleza tienen un objetivo de medición distinto entre sí, es decir, que dos evaluaciones SET diferentes probablemente no miden la misma variable latente.

Según lo anterior, la mejor forma de relacionar dichas respuestas con el desempeño del docente es a través de un ajuste de un modelo, que relacione los valores del desempeño docente (variable latente) y de las propiedades de los ítems con los datos de respuesta recopilados por los estudiantes. Si el modelo ajusta los datos, se puede usar para producir medidas de habilidad que son independientes de las propiedades de los ítems en la prueba (Van Der Linden & Hambleton, 1997). En el presente capítulo se realiza una introducción a los modelos de TRI de múltiples facetas que generalizan los modelos de TRI.

2.1. Modelos de TRI

Con los modelos de TRI se intenta representar la probabilidad de que un individuo proporcione cierta respuesta a un ítem, como función de los parámetros del ítem y de la habilidad del individuo (de Andrade et al., 2000). Hambleton et al. (1991) mencionan que los modelos de TRI, se basan en dos postulados básicos: (a) el rendimiento de un evaluado en un ítem se puede explicar mediante un conjunto de factores llamados rasgos latentes

(o habilidades); y (b) la relación entre el desempeño del evaluado en el ítem y el conjunto de rasgos que subyacen al desempeño del ítem, puede describirse mediante una función monótona creciente llamada Curva Característica del Ítem (CCI).

Los modelos de TRI dependen de tres factores importantes: la naturaleza del ítem (dicotómico o politómico), el número de grupos poblacionales evaluados y la cantidad de variables latentes medidas (modelo unidimensional o multidimensional). Independientemente de los factores que incidan en el estudio, el objetivo de estos modelos es crear una escala para medir variables latentes. (de Andrade et al., 2000)

2.1.1. Modelos dicotómicos

Generalmente, los ítems utilizados en contextos de pruebas cognitivas o de conocimientos tienen un formato de tipo dicotómico. La asignación de dichos valores comúnmente se realiza en una escala de 0 y 1, donde se asigna el valor de 1 al ítem que el evaluado contestó correctamente y 0 al ítem que el evaluado contestó incorrectamente.

Los modelos dicotómicos son los más utilizados en la práctica, entre estos los más empleados son los que se diferencian por el número de parámetros que describen la probabilidad de contestar correctamente el ítem. Estos modelos son conocidos como modelos logísticos de 1, 2 y 3 parámetros, que consideran respectivamente:

1. Solamente la dificultad del ítem.
2. La dificultad y la discriminación del ítem.
3. La dificultad, la discriminación y la probabilidad de responder al azar y de manera correcta cuando el rasgo latente es bajo.

La forma matemática del modelo logístico de tres parámetros está dada por la probabilidad condicional de responder correctamente a un ítem, expresada como:

$$P(u_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + \exp(-a_i(\theta_j - b_i))} \tag{2.1}$$

donde U_{ij} es la variable dicotómica de respuesta del individuo j en el ítem i; θ_j es el rasgo latente del j-ésimo individuo; b_i es el parámetro de dificultad del ítem i; a_i es el parámetro de discriminación del ítem i y c_i es el parámetro de pseudo-azar del ítem i.

Cuando el parámetro de pseudo-azar c_i se hace igual a 0 entonces el modelo es un modelo logístico de dos parámetros, y cuando además de esto el parámetro de discriminación a_i se hace igual a uno entonces el modelo es un modelo logístico de un parámetro.

En el ámbito del modelo de un parámetro, existen dos corrientes para hacer el ajuste: Los modelos de Rasch y los modelos de TRI. La diferencia más general es que en un modelo de Rasch se busca un grupo de ítems que satisfaga las condiciones del modelo mientras que la otra alternativa busca un modelo de TRI que ajuste mejor al conjunto de ítems. Dado que la expresión matemática de los dos modelos es idéntica, a menudo se usan sin distinción lo cual es un error, pues los dos modelos son realmente diferentes.

2.1.2. Modelos politómicos

Los ítems utilizados en contextos de pruebas psicológicas, entre otras, tienen una forma estructural similar en todos los ítems del instrumento. Por lo general, el formato de la escala de calificación es de tipo Likert; un ejemplo de esta es una encuesta que sugiere una opinión a una situación usando una escala de 5 puntos que va desde *totalmente en desacuerdo* hasta *totalmente de acuerdo*.

Los modelos politómicos añaden parámetros a los modelos dicotómicos. Estos parámetros permiten describir el funcionamiento completo de la escala que induce todas las opciones de respuesta. La diferencia entre los diversos modelos politómicos se describe en términos de las expresiones usadas para representar el parámetro de localización b (Ostini & Nering, 2006).

El primer modelo a considerar es el modelo de *rating scale* (Andrich, 1978); este es adecuado para un instrumento donde se consideren ítems que tienen exactamente el mismo número de categorías de respuesta. La expresión matemática del modelo está dada por:

$$P_{i,k}(\theta_j) = \frac{1}{1+\exp^{-a_i(\theta_j - b_i + d_k)}} - \frac{1}{1+\exp^{-a_i(\theta_j - b_i + d_{k+1})}} \tag{2.2}$$

donde $P_{i,k}(\theta_j)$ es la probabilidad de que un individuo con habilidad θ_j escoja la categoría k dentro de las $(m+1)$ categorías del ítem i, a_i el parámetro de discriminación del ítem i, b_i el parámetro de localización del ítem i y d_k el parámetro de la categoría.

El segundo modelo a considerar es el modelo de crédito parcial (Masters, 1982). Es una generalización del modelo de *rating* ya que se elimina la restricción del mismo número de categorías de respuesta para todos los ítems, es decir, que permite que los instrumentos tengan en sí mismos diferentes escalas de calificación. Este modelo se define por la siguiente expresión:

$$P_{i,k}(\theta_j) = \frac{\exp \sum_{u=0}^{k} [\theta_j - b_{i,u}]}{\sum_{u=0}^{m_i} \exp \sum_{v=0}^{u} [\theta_j - b_{i,v}]} \tag{2.3}$$

donde $P_{i,k}(\theta_j)$ es la probabilidad de que un individuo con habilidad θ_j escoja la categoría k dentro de las $(m+1)$ categorías del ítem i, b_{ik} representa un parámetro de localización b para cada categoría k y cada ítem i.

2.2. Modelos de TRI de múltiples facetas

La evaluación se caracteriza por tener múltiples factores que inciden en los resultados de la medición del trazo latente. Por lo tanto, la construcción de medidas confiables y válidas de la habilidad de los evaluados depende crucialmente de la implementación de métodos bien diseñados para tratar con las múltiples fuentes de variabilidad de la situación.

Se usa el término "facetas"para referirse a cualquier componente del contexto de medición que contribuye con un error de medición sistemático (Wolfe & Dobria, 2008). La idea básica de un modelo de este tipo es principalmente incluir todas las facetas necesarias

para obtener mejores estimaciones de las habilidades de cada individuo involucrado en el contexto de la evaluación.

La definición anterior incluye facetas que son de interés sustancial, por ejemplo, los evaluados, así como facetas que están asociadas a un error de medición sistemático, por ejemplo, los evaluadores, los ítems, los criterios de evaluación, los tiempos de aplicación de la prueba, etc. (Eckes, 2011). Muchas facetas distintas a las asociadas con el constructo que se está midiendo pueden tener un impacto no despreciable en los resultados de la evaluación.

2.2.1. Elementos de la medición

Los modelos de múltiples facetas son una extensión de los modelos clásicos de medición. Dichos modelos incorporan más facetas (evaluador, criterio, tiempo, etc.) que las dos que involucran a los modelos clásicos (ítems y evaluados). Originalmente se presentaron en el marco de los modelos de Rasch, lo que no limita su implementación como modelos más generales de TRI, e incluyen los modelos de respuesta politómica como el modelo de *rating scale* (RSM) y el modelo de crédito parcial (PCM), descritos en la sección anterior.

Para ejemplificar lo anterior, suponga la siguiente situación de evaluación: Un investigador le hace una prueba de escritura a un grupo de estudiantes para medir su nivel de desarrollo en esta competencia. La prueba le pide a los estudiantes elaborar una carta con unas condiciones específicas. Sin embargo, para evaluar el nivel de la competencia en escritura de los estudiantes no hay forma de hacerlo sin recurrir a un evaluador de los escritos. Por lo tanto, la institución decide contratar unos expertos en evaluar dichas competencias comunicativas con el objetivo de que ellos brinden unas calificaciones del desempeño de los estudiantes. En esta situación, se observan tres facetas. La primera, relacionada con el estudiante evaluado, debido a que la habilidad del estudiante tiene un componente de variabilidad. La segunda, vinculada con el ítem que respondió el estudiante, esto debido a que los diferentes ítems proporcionan diversa variabilidad al modelo. Y la tercera, la participación del experto en el tema para obtener las calificaciones finales, puesto que esto induce una variabilidad adicional que debe ser considerada en el modelo de calificación. Por lo tanto, el investigador se enfrenta a una situación de tres facetas: evaluados, ítems y evaluadores o calificadores.

Suponiendo que se han identificado las facetas relevantes (es decir, evaluados, ítems y evaluadores), un modelo adecuado de medición de tres facetas se puede expresar formalmente así:

$$\log \frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \beta_i - \gamma_j - \tau_k \tag{2.4}$$

donde P_{nijk} es la probabilidad de que el evaluado n reciba una calificación k o menos del evaluador j en el ítem i; θ_n es el trazo latente del evaluado n; β_i es la dificultad del ítem i; γ_j es la severidad del evaluador j y τ_k es la dificultad de recibir una calificación de k relativa a $k-1$. Este modelo es esencialmente un modelo aditivo lineal basado en una transformación logística.

CAPÍTULO 3

Análisis de datos bayesiano

Gelman et al. (2013) define la inferencia bayesiana como el proceso de ajustar un modelo de probabilidad a un conjunto de datos y resumir el resultado mediante una distribución de probabilidad en los parámetros del modelo y cantidades observadas. Además describen el proceso de análisis en los siguientes pasos:

1. **Establecer un modelo de probabilidad completo:** Es decir, una distribución de probabilidad conjunta para las cantidades observables y no observables.
2. **Condicionamiento de los datos observados:** Es decir, debe haber un cálculo e interpretación de la distribución posterior o condicional.
3. **Evaluación del ajuste del modelo**: Es necesario revisar si el modelo se ajusta a los datos y si son razonables las conclusiones.

3.1. Inferencia bayesiana

En inferencia bayesiana las conclusiones sobre un parámetro θ o sobre datos no observados $\tilde{y}$ se hacen en términos de probabilidad. Estas están condicionadas al valor observado de y, y se notan como $p(\theta|y)$. La función de densidad conjunta puede escribirse como:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

donde $p(\theta)$ es conocida como la *distribución a priori* y $p(y|\theta)$ es conocida como la *distribución muestral*. Cuando se condiciona sobre los valores conocidos de los datos y, usando la propiedad de la probabilidad condicional conocida como la regla de Bayes, la densidad posterior está dada por:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{3.1}$$

donde $p(y) = \sum_\theta p(\theta)p(y|\theta)$ (o $p(y) = \int p(\theta)p(y|\theta)$ en el caso continuo), la suma es sobre todos los posibles valores del parámetro θ. La distribución de los valores y desconocidos pero observables se define como:

$$p(y) = \int p(y, \theta)d\theta = \int p(\theta)p(y|\theta)d\theta \tag{3.2}$$

$p(y)$ Es conocida como distribución marginal de y o distribución a priori. Cuando los datos ya han sido observados, se puede predecir una variable observable desconocida $\tilde{y}$. La distribución de $\tilde{y}$ es llamada *distribución predictiva*:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta = \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta = \int p(\tilde{y}|\theta)p(\theta|y)d\theta \tag{3.3}$$

3.2. Métodos computacionales

Fox (2010) menciona que la introducción de poderosos métodos de simulación hizo posible el modelado bayesiano en varios campos de investigación. Los métodos de simulación posterior hacen que las distribuciones posteriores sean accesibles; es decir, los algoritmos para la simulación posterior se pueden usar para obtener aproximaciones de los momentos de las distribuciones posteriores.

Una clase de métodos de simulación lo constituyen los métodos de Monte Carlo basados en cadenas de Markov, mejor conocidos como MCMC (Monte Carlo Markov Chain) construyen secuencias que convergen en distribución a la distribución posterior. Luego, se calculan los promedios de dichas muestras para estimar las esperanzas posteriores. Para una descripción más detallada ver (Chen et al., 2012).

3.2.1. Muestreador de Gibbs

El método MCMC más popular es el muestreador de Gibbs, cuyo nombre se origina en una clase de distribuciones de probabilidad para modelar interacciones espaciales y procesos estocásticos espaciales (Geman & Geman, 1984).

En Casella & George (1992) se realiza la explicación detallada del método. Este comienza con la partición de parámetros o vectores aleatorios de interés en subvectores $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_Q)$. La densidad posterior conjunta del vector aleatorio es igual a $p(\theta|y)$, y esta también es la densidad objetivo. Un proceso de transición de $\theta^{(m)}$ a $\theta^{(m+1)}$ se define haciendo extracciones en la iteración $m+1$ de la función de densidad condicional de cada subvector,

$$\begin{aligned}
\theta_1^{(m+1)} &\sim p(\theta_1|\theta_2^{(m)}, \cdots, \theta_Q^{(m)}, y) \\
\theta_2^{(m+1)} &\sim p(\theta_2|\theta_1^{(m+1)}, \theta_3^{(m)}, \cdots, \theta_Q^{(m)}, y) \\
&\vdots \\
\theta_Q^{(m+1)} &\sim p(\theta_Q|\theta_1^{(m+1)}, \cdots, \theta_{Q-1}^{(m+1)}, y)
\end{aligned}$$

El muestreador de Gibbs gestiona el proceso de transición, la forma de las densidades condicionales y la elección del bloqueo caracteriza a cada muestra. Bajo algunas condiciones de regularidad la cadena MCMC tiene una densidad estacionaria igual a $p(\theta|y)$. Esto significa que $\boldsymbol{\theta}^{(m)}$ converge a $\boldsymbol{\theta}$ en distribución cuando $m \to \infty$. Roberts (1996) detalla las condiciones de regularidad.

3.2.2. Algoritmo Metropolis-Hastings

El algoritmo Metropolis-Hastings (M-H) fue desarrollado por Metropolis et al. (1953) y generalizado por Hastings (1970). Generaliza el muestreador de Gibbs ya que da una solución al problema del muestreo a partir de una distribución condicional, de la cual es difícil muestrear directamente. Es un procedimiento en dos pasos.

- En el primer paso, un candidato se extrae de una densidad propuesta que puede ser elegida como una aproximación a la densidad posterior deseada.
- En el segundo paso, el candidato extraído se acepta o se rechaza según un criterio de aceptación especificado. Se construye una probabilidad de aceptación a partir de la densidad posterior para evaluar al candidato.

La probabilidad de aceptación asegura que el algoritmo genere muestras a partir de la densidad objetivo. Se permite que la densidad propuesta dependa del estado actual $\boldsymbol{\theta}^{(m)}$.

Técnicamente, en la iteración m, un candidato, $\boldsymbol{\theta}_q^*$ es extraído de una densidad propuesta, $q(\boldsymbol{\theta}_q|\boldsymbol{\theta}_q^{(m)})$, y la transición de $\boldsymbol{\theta}_q^{(m)}$ a $\boldsymbol{\theta}_q^{(m+1)}$ se hace sí

$$u^{(m)} \leq \frac{p(\boldsymbol{\theta}_q^*|y)/q(\boldsymbol{\theta}_q^*|\boldsymbol{\theta}_q^{(m)})}{p(\boldsymbol{\theta}_q^{(m)}|y)/q(\boldsymbol{\theta}_q^{(m)}|\boldsymbol{\theta}_q^*)} \tag{3.4}$$

donde $u^{(m)}$ es la m-ésima observación de la variable aleatoria U que distribuye uniformemente en el intervalo $[0, 1]$. Si no se acepta el valor propuesto, la cadena permanece en su estado actual. El algoritmo es muy simple y la forma de la densidad propuesta no tiene restricción.

Es eficiente tener una relación de aceptación cercana a uno, pero los valores propuestos también deben cubrir todo el rango de valores probables bajo la distribución objetivo. Es decir, las propuestas deben extraerse de toda la región donde $p(\theta|y)$ es apreciable.

3.2.3. Método Monte Carlo Hamiltoniano

El método de Monte Carlo Hamiltoniano (HMC, por sus siglas en inglés) (Neal et al., 2011) es un algoritmo MCMC que evita el comportamiento de caminata aleatoria y la sensibilidad a los parámetros correlacionados. Para lo anterior, toma una serie de pasos informados por el gradiente de información de primer orden. Esto le permite converger a distribuciones objetivo de grandes dimensiones mucho más rápidamente que métodos como el muestreador de Gibbs o Metropolis-Hastings.

Para modelos con muchos parámetros, los métodos simples como el de Metropolis (Metropolis et al., 1953) y el muestreador de Gibbs (Geman & Geman, 1984) pueden

requerir un tiempo muy largo para converger a la distribución objetivo. Neal (1993) menciona que esto se debe en gran parte a la tendencia de dichos métodos a explorar el espacio de parámetros a través de caminatas aleatorias ineficientes.

Neal et al. (2011) muestra que cuando los parámetros del modelo son continuos el método HMC puede suprimir dicho comportamiento de caminata aleatoria mediante un inteligente esquema de variable auxiliar que transforma el problema del muestreo de una distribución objetivo en el problema de simular la dinámica Hamiltoniana. El método tiene dos requerimientos: el gradiente log-posterior y que el usuario especifique al menos dos parámetros: un tamaño de paso y una cantidad de pasos L para ejecutar un sistema Hamiltoniano simulado. No cumplir alguno de los requerimientos dará como resultado una caída en la eficiencia del método.

En HMC se introduce una variable auxiliar r_d para cada variable de modelo θ_d. Estas variables se extraen independientemente de la distribución normal estándar, produciendo la densidad conjunta

$$p(\theta, r) \exp\{\mathcal{L}(\theta) - \frac{1}{2} r \cdot r\}$$

donde $\mathcal{L}$ es el logaritmo de la densidad conjunta de las variables de interés θ y $x \cdot y$ denota el producto interno.

Para cada muestra m, primero volvemos a muestrear las variables de momento de una normal multivariada estándar. Luego aplicamos L actualizaciones a las variables de posición y momento θ y r, generando una propuesta del par de posición-momento $(\tilde{\theta}, \tilde{r})$. Luego se establece $\theta^m = \tilde{\theta}$ y $r^m = -\tilde{r}$, y se acepta o rechaza esta propuesta de acuerdo con el algoritmo de Metropolis. Neal et al. (2011) desarrollan los detalles del método.

3.2.4. Muestreador No-U-Turn

El muestreador *No-U-Turn*, NUTS por sus siglas en inglés, es un algoritmo MCMC derivado de HMC, sin embargo, este elimina la necesidad de elegir el parámetro problemático de número de pasos L (Hoffman & Gelman, 2014). Para esto, es necesario un criterio que señale cuándo se ha simulado la dinámica durante "tiempo suficiente", esto es, cuando ejecutar la simulación para más pasos ya no aumentaría la distancia entre la propuesta $\tilde{\theta}$ y el valor inicial de θ. Se utiliza un criterio conveniente basado en el producto punto entre $\tilde{r}$ (el momento actual) y $\tilde{\theta} - \theta$ (el vector desde la posición inicial a la posición actual), que es la derivada con respecto al tiempo (en el sistema Hamiltoniano) de la mitad de la distancia al cuadrado entre la posición inicial θ y la posición actual $\tilde{\theta}$:

$$\frac{d}{dt} \frac{(\tilde{\theta} - \theta) \cdot (\tilde{\theta} - \theta)}{2} = (\tilde{\theta} - \theta) \cdot \frac{d}{dt}(\tilde{\theta} - \theta) = (\tilde{\theta} - \theta) \cdot \tilde{r} \tag{3.5}$$

Esto sugiere un algoritmo en el que se ejecutan pasos de salto hasta que la cantidad en la ecuación 3.5 sea menor que 0; dicho enfoque simularía la dinámica del sistema hasta que la ubicación de la propuesta $\tilde{\theta}$ comenzara a retroceder hacia θ. NUTS comienza introduciendo una variable de corte u con distribución condicional $p(u|\theta, r) = U(u; [0, \exp \mathcal{L}(\theta) - \frac{1}{2} r \cdot r])$. Para ver los demás detalles del desarrollo del método remitirse a Hoffman & Gelman (2014).

CAPÍTULO 4

Un modelo TRI bayesiano de múltiples facetas para la evaluación del desempeño docente

En la Universidad Nacional de Colombia y en general, en la mayoría de universidades del país al finalizar cada semestre se realiza una encuesta a los estudiantes que está relacionada con el desempeño del docente en el transcurso del semestre. Por lo general, los docentes imparten más de un curso en un mismo semestre y además, estos cursos son impartidos a estudiantes que poseen diversas características sociodemográficas y de su vida universitaria, razón por la cual, es importante tener en cuenta dichas características al momento de realizar algún tipo de análisis sobre los datos recolectados con estas encuestas.

Como se ha mencionado en capítulos anteriores, en general, la estimación del desempeño tiene dos problemas desde el punto de vista estadístico. Se mencionó que no es correcto hacer uso de promedios de variables categóricas de naturaleza ordinal y que además no es posible cuantificar las distancias entre la escala, razón por la cual no es posible realizar comparaciones entre estas.

En este capítulo se propone un modelo de TRI Bayesiano de múltiples facetas para el tratamiento de datos de tipo SET. El modelo propuesto está formulado para aplicarse a datos en los que se tiene información asociada a un docente, un curso específico y un estudiante evaluador de dicho curso. Para abordarlo se hace una introducción a partir de la distribución logística ordenada, seguido de la descripción de la función de probabilidad asociada al modelo, así como la especificación de las distribuciones a priori y los resultados a posterior.

4.1. Distribución logística ordenada

La función logística inversa se define como:

$$\frac{1}{\text{logit}} = \frac{1}{1 + \exp^{-x}}. \tag{4.1}$$

La expresión anterior es la definición de la función de distribución acumulativa de la distribución logística.

Sea $\beta' = (\beta_1, \beta_1, \cdots, \beta_{K+1})$ un vector con valores en los reales, con $\beta_{k-1} < \beta_k$, y sea $\eta \in \mathbb{R}$. Se define la función de masa de probabilidad de la distribución logística ordenada como

$$f(k|\eta, \beta) = \begin{cases} 1 - \text{logit}^{-1}(\eta - \beta_1) & \text{si } k = 1 \\ \text{logit}^{-1}(\eta - \beta_{k-1}) - \text{logit}^{-1}(\eta - \beta_k) & \text{si } 1 < k < K \\ \text{logit}^{-1}(\eta - \beta_{K-1}) & \text{si } k = K \end{cases} \tag{4.2}$$

Como los modelos de múltiples facetas son modelos de TRI, en este caso las η son variables latentes que deben predecirse, y los valores β_k son los valores de los parámetros de los ítems. Es importante tener en cuenta que las variables η y β_k se encuentran en el mismo espacio vectorial. Además, para un valor fijo de η, si $\beta_k < \eta < \beta_{k+1}$, entonces, la categoría $k + 1$ generalmente tiene mayor probabilidad de ser respondida.

4.2. Función de probabilidad

El desempeño del docente se asume como una variable latente θ. Desde una perspectiva general, las variables latentes pueden considerarse efectos aleatorios (Bartholomew et al., 2011). Por otro lado, siguiendo la línea de la explicación del modelo de tres facetas mencionado en el capítulo anterior, para este se debe tener en cuenta la severidad γ del estudiante al evaluar un docente, y esta variable se mide también por una nueva variable latente.

Finalmente, en la situación de evaluación, un estudiante califica varios aspectos del desempeño docente a través de p ítems presentados en el cuestionario, donde cada ítem se basa en una escala Likert con categorías k_m con m el número de categorías del ítem k.

Sea Y_{ijsc} la variable aleatoria que representa la puntuación asignada por el estudiante s al docente i en el ítem j en el curso c, θ_i es el trazo latente o desempeño del docente i, η_c es un efecto aleatorio del curso evaluado c, β_j la dificultad del ítem j y γ_{is} la severidad del estudiante s al docente i. El modelo de probabilidad condicional se define como:

$$Pr\left[Y_{ijsc} = k|\theta_{ic}, \gamma_{is}, \beta_j\right] = \text{logit}^{-1}(\theta_i - \eta_c - \gamma_{is} - \beta_{j(k-1)}) - \text{logit}^{-1}(\theta_i - \eta_c - \gamma_{is} - \beta_{jk}) \tag{4.3}$$

Donde $i = 1, \cdots, N$; $c = 1, 2, \cdots, C$; $s = 1, 2, \cdots, S$; $j = 1, 2, \cdots, P$; $k = 1, \cdots, m$.

4.3. Distribuciones a priori y distribución a posterior

El modelo definido en la ecuación (4.3) es no identificable, como es común en los modelos de TRI. Para establecer una escala, se asume que $\theta_i \sim N(0, 1)$. Además, se asignan las siguientes distribuciones a priori para los demás parámetros:

$$\begin{gathered} \eta_c \sim N(0, \sigma_\eta^2) \\ \gamma_{is} \sim N(0, \sigma_\gamma^2) \\ \beta_{jk} \sim N(0, \sigma_\beta^2) \end{gathered}$$

$$\sigma_\eta \sim Cauchy(0,2)$$
$$\sigma_\gamma \sim Cauchy(0,2)$$
$$\sigma_\beta \sim Cauchy(0,2)$$

Para tener la distribución posterior del modelo se requieren dos supuestos: en primer lugar, las respuestas de los estudiantes son independientes. En segundo lugar, las respuestas de un estudiante a un docente son independientes. Estas suposiciones, especialmente la segunda, pueden ser controvertidas. La primera suposición puede ser violada si un estudiante responde a más de un cuestionario para el mismo docente. Esto puede ocurrir si el estudiante está tomando dos cursos con el mismo profesor. Sin embargo, este problema disminuye porque los cursos son diferentes. Además, no hay forma de saber cuándo ocurre esta situación porque las respuestas de los estudiantes son anónimas. Por otro lado, cada pregunta en el cuestionario está diseñada para medir un aspecto diferente del desempeño de los profesores en el aula. Por lo tanto, se espera que cada una de las preguntas sea respondida independientemente por el alumno.

Sea $p_{ijsck} = Pr\left[Y_{ijsc} = k|\theta_{ic}, \gamma_{is}, \beta_j\right]$. La distribución posterior para el modelo está dada por

$$\begin{aligned} L[\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\boldsymbol{y}] \propto \prod_{i=1}^{N}\prod_{j=1}^{P}\prod_{s=1}^{S}\prod_{c=1}^{C}\prod_{k=1}^{K_j}[p_{ijsck}1_{k(y_{ijs})}]\times \\ p(\theta_i)p(\eta_c|\sigma_\eta)p(\gamma_{is}|\sigma_\gamma)p(\beta_{jk}|\sigma_\beta)p(\sigma_\beta)p(\sigma_\gamma)p(\sigma_\eta) \end{aligned} \tag{4.4}$$

Donde $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_i, \cdots, \theta_N)'$, $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_c, \cdots, \eta_C)'$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_l, \cdots, \beta_j)'$ y $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_s, \cdots, \gamma_S)'$ son los vectores de parámetros, $\boldsymbol{y}$ es el vector de respuestas observadas y $1_{k(y_{ijs})}$ es una variable indicadora definida como:

$$1_{k(y_{ijs})} = \begin{cases} 1 & \text{Si } y_{ijsc} = k \\ 0 & \text{En otro caso} \end{cases} \tag{4.5}$$

4.4. Estadísticas de bondad de ajuste y criterios de selección de modelos

4.4.1. Estadísticas de bondad de ajuste

Para facilitar, en esta sección, se supone que el parámetro vectorial completo es $\boldsymbol{\theta}$, y los datos observados son $\mathbf{y}$. La probabilidad de las observaciones es $f(\mathbf{y}|\boldsymbol{\theta})$ y la densidad a priori de $\boldsymbol{\theta}$ es $\pi(\boldsymbol{\theta})$, por lo tanto, la especificación del modelo es $f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. La distribución predictiva de valores no observados $\boldsymbol{\omega}$ se denota $f(\boldsymbol{\omega}|\mathbf{y})$ y se define como:

$$f(\omega|\mathbf{y}) = \int f(\omega|\theta)\pi(\theta|y)d\theta \tag{4.6}$$

En los procedimientos de bondad de ajuste los datos ω son datos replicados. Sea $\mathbf{W}$ la variable aleatoria con densidad $f(\boldsymbol{\omega}|\mathbf{y})$. Para evaluar el ajuste de un modelo a los datos

observados, Box (1980) propuso utilizar la distribución predictiva $f(\omega|y)$. En particular, Box propuso calcular la esperanza $E[g(\mathbf{W},\boldsymbol{\theta})|y]$ de alguna estadística relevante $g(\mathbf{W},\boldsymbol{\theta})$.

Gelfand et al. (1992) propusieron algunas funciones de verificación de ajuste basadas en medidas de discrepancia, y Gelman et al. (1996) estudian a profundidad la idea de utilizar dichas medidas.

En este documento se propone el uso de la estadística $D(\mathbf{W},\boldsymbol{\theta}) = -2\log f(\mathbf{W}|\boldsymbol{\theta})$, lo anterior con el objetivo de tener una estadística similar a las utilizadas comúnmente en estadística frecuentista.

De acuerdo a la notación usada por Box (1980) y Gelfand et al. (1992) y usando la estadística de discrepancia D propuesta por Gelman et al. (1996), se construye la estadística de bondad de ajuste como sigue. Sea $B_{\omega,\theta} = \{(\omega,\theta) : D(\omega,\theta) \leq D(y,\theta)\}$. La estadística de verificación se define como $g(\mathbf{W},\boldsymbol{\theta}) = I_{B_{\omega,\theta}}(\mathbf{W},\boldsymbol{\theta})$. La decisión es basada en el valor $d_{\omega,\theta}$ definido por

$$\begin{aligned} d_{\omega,\theta} &= P(B_{\omega,\theta}) \\ &= E[g(\mathbf{W},\boldsymbol{\theta})|y] \\ &= \int\int g(\omega,\theta) f(\omega|\theta)\pi(\theta|y) d\theta d\omega \end{aligned}$$

Valores de $d_{\omega,\theta}$ cercanos a 0.5 indican que el modelo ajusta bien los datos.

En Gelman et al. (1996) se menciona que los cálculos se pueden realizar analíticamente para algunos problemas simples pero en modelos complicados, se logra más fácilmente a través de simulación de Monte Carlo. Ellos describen el procedimiento para el cálculo de la evaluación del modelo. Considere el cálculo requerido para comparar la discrepancia realizada $D(y,\theta)$. Dado un conjunto de t muestras $\theta^{(t)}$, $t = 1, ..., T$, solo se necesita realizar los siguientes dos pasos para cada t:

1. Dado θ^j, extraiga un conjunto de datos replicados simulados, y^{repj}, de la distribución muestral, $P_A(y^{rep}|H,\theta^j)$.
2. Calcule $D(y;\theta^j)$ y $D(y^{rep}j;\theta^j)$.
3. Habiendo obtenido lo anterior, calcule la proporción de los pares J para los cuales $D(y^{rep};\theta^j)$ excede $D(y;\theta^j)$.

4.4.1.1. Estadísticas de bondad de ajuste para modelos de múltiples facetas

En los modelos de respuesta a ítems, es común evaluar el ajuste del modelo a los datos completos y el ajuste del modelo a ítems y personas por separado. En consecuencia, se evalúa la adecuación de los modelos para los datos completos, los datos de cada uno de los profesores evaluados y cada uno de los ítems. Sea $y_{i..}$ el vector completo de respuestas al profesor i; $y_{.j.}$ el vector completo de respuestas al ítem j. Los datos de réplica correspondientes se denotarán $\omega_{i..}$ y $\omega_{.j.}$ respectivamente. Para evaluar la bondad del ajuste del modelo a los datos del i-ésimo profesor, se utiliza la densidad predictiva, dada por

$$f(\omega_{i..}|y_{i..}) = \int f(\omega_{i..}|\theta_i,\gamma_i,\eta_i,\beta)\pi(\theta_i,\gamma_i,\eta_i,\beta|y_{i..}) d\theta_i d\gamma_i d\eta_i d\beta \tag{4.7}$$

A continuación, se describe el procedimiento de calculo de la estadística de bondad de ajuste. Sea $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\eta}^{(t)})$ la t-ésima muestra completa de la distribución posterior $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}|\mathbf{y})$, con $t = 1, ..., T$. La estadística para los datos del i-ésimo docente se calcula como sigue:

1. Obtener un vector de datos replicado $\{\omega_{ijsc}\}$ de la distribución logística ordenada dada por

$$g_{ijsc}(k|\theta_i^{(t)}, \eta_c^{(t)}, \gamma_{is}^{(t)}, \beta_j^{(t)}) = \text{logit}^{-1}(\theta_i^{(t)} - \eta_c^{(t)} - \gamma_{is}^{(t)} - \beta_{j(k-1)}^{(t)}) - \text{logit}^{-1}(\theta_i^{(t)} - \eta_c^{(t)} - \gamma_{is}^{(t)} - \beta_{jk}^{(t)})$$

2. La medida de discrepancia para $\boldsymbol{\omega}_{i...}$ está dada por

$$D_i(\omega_{i..}|\theta_i^{(t)}, \eta_c^{(t)}, \gamma_i^{(t)}, \beta_j^{(t)}) = -2 \sum_{j=1}^{p} \sum_{s=1}^{n_i} \sum_{c=1}^{C} \sum_{k=1}^{K_j} 1_k(\omega_{ijsc}) \log g_{ijsc}(k|\theta_i^{(t)}, \gamma_{is}^{(t)}, \eta_c^{(t)}, \beta_j^{(t)})$$

3. La medida de discrepancia para $\boldsymbol{y}_{i...}$ está dada por

$$D_i(y_{i..}|\theta_i^{(t)}, \eta_c^{(t)}, \gamma_i^{(t)}, \beta_j^{(t)}) = -2 \sum_{j=1}^{p} \sum_{s=1}^{n_i} \sum_{c=1}^{C} \sum_{k=1}^{K_j} 1_k(y_{ijsc}) \log g_{ijsc}(k|\theta_i^{(t)}, \gamma_{is}^{(t)}, \eta_c^{(t)}, \beta_j^{(t)})$$

4. La estadística de bondad de ajuste está dada por

$$p_i = \frac{1}{T} \sum_{t=1}^{T} 1_{(D_i(\omega_{i..}|\theta_i^{(t)}, \eta_c^{(t)}, \gamma_i^{(t)}, \beta_j^{(t)})) - D_i(y_{i..}|\theta_i^{(t)}, \eta_c^{(t)}, \gamma_i^{(t)}, \beta_j^{(t)}) > 0}$$

4.4.2. Criterios de selección de modelos

Un procedimiento de selección de modelos es necesario para encontrar un modelo adecuado que explique los datos actuales y los datos futuros. Existen una variedad de procedimientos para selección de modelos en el marco de la inferencia bayesiana. A continuación se presentan los de uso más frecuente.

4.4.2.1. DIC

En Spiegelhalter et al. (2002) se propone un criterio de información DIC, el cual es una analogía con los resultados clásicos AIC y BIC. La desviación bayesiana se define como:

$$D(\theta) = -2 \log\{p(y|\theta)\} + 2 \log\{f(y)\} \tag{4.8}$$

donde $p(y|\theta)$ es la función posterior y $f(y)$ como un término de estandarización completamente especificado que es solo una función de los datos. Se define p_D como:

$$p_D = \bar{D(\theta)} - D(\bar{\theta}) \tag{4.9}$$

Finalmente, el criterio de información DIC, se define como una estimación clásica de ajuste, así:

$$DIC = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \tag{4.10}$$

La ecuación muestra que el DIC puede considerarse como una medida Bayesiana de ajuste penalizada por un término adicional p_D.

4.4.2.2. WAIC

El criterio de información ampliamente aplicable (WAIC, por sus siglas en inglés) fue introducido por Watanabe (2010). Gelman et al. (2014) y Ariyo et al. (2019) describen brevemente el desarrollo para la obtención de la estadística. Para una observación futura $\tilde{y}_i$, este criterio mide la precisión predictiva del modelo basado en la distribución predictiva log-posterior $\log p_{\theta|y}(\tilde{y}_i)$ del vector de parámetros $\boldsymbol{\theta}$. Dado que $\tilde{y}_i$ es desconocido, la precisión predictiva se define por la distribución log-predictiva esperada (elpd) como

$$elpd_i = E_f[\log p_{\theta|y}(\tilde{y}_i)] = \log p_{\theta|y}(\tilde{y}_i) f(\tilde{y}_i) d\tilde{y}_i$$

donde f es la distribución desconocida para el verdadero modelo. Para cada observación de un nuevo conjunto de datos, *elpd* es calculada para establecer la precisión predictiva de ese conjunto de datos. La distribución predictiva logarítmica puntual (*lppd*) está basada en los datos observados y se calcula de la siguiente manera:

$$lppd = \log \prod_{i=1}^{n} p_{\theta|y_i}(y_i) = \sum_{i=1}^{n} \log \int_{\theta} p(y_i|\theta) p(\theta|y) d\theta$$

En la práctica se puede estimar usando una muestra MCMC de la distribución posterior como sigue:

$$\widehat{lppd} = \sum_{i=1}^{n} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(y_i|\theta^k) \right] \tag{4.11}$$

Sea la medida p_{WAIC} una estimación del número efectivo de parámetros dada por

$$p_{WAIC} = 2 \sum_{i=1}^{n} \left[\log \left(\frac{1}{K} \sum_{k=1}^{K} p(y_i|\theta^k) \right) - \frac{1}{K} \sum_{k=1}^{K} \log p(y_i|\theta^k) \right] \tag{4.12}$$

El WAIC se puede expresar como

$$WAIC = -2\widehat{lppd} + 2p_{WAIC} \tag{4.13}$$

4.4.2.3. LOO

La validación cruzada es un enfoque para estimar la precisión predictiva fuera de la muestra utilizando ajustes dentro de la muestra. Requiere volver a ajustar el modelo con diferentes conjuntos de entrenamiento. La validación cruzada *Leave-one-out* (LOO) se puede calcular fácilmente usando muestreo de importancia (Gelfand et al., 1992).

Vehtari et al. (2017) muestran el desarrollo de la estadística. Consideran los cálculos utilizando la log-verosimilitud evaluada en las simulaciones posteriores habituales de los parámetros definida en la ecuación (4.11). La estimación bayesiana LOO del ajuste predictivo fuera de muestra es

$$elpd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i}) \tag{4.14}$$

donde

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta \tag{4.15}$$

es la densidad predictiva dada la información sin el i-ésimo punto de los datos.

4.5. Estimación de los parámetros del modelo

El modelo propuesto en el presente documento se implementa a través de la interfaz de R para Stan, cuyo nombre es Rstan. La implementación a través de Stan se realiza usando vectorización, por lo tanto, es necesario preparar el conjunto de datos para especificarlos en el modelo. Para familiarizar al lector con la estructura de datos a continuación, se muestra un ejemplo.

La tabla 4.1 presenta una estructura de datos original en forma matricial, donde cada fila es un vector de respuestas de un estudiante dado a la encuesta sobre el desempeño docente, además, se cuenta con la información asociada a la identificación del docente y del curso.

Estudiante	**Curso**	**Profesor**	**Item 1**	**Item 2**	$\cdots$	**Item 14**
1	1	1	4	4	$\cdots$	5
2	1	1	5	5	$\cdots$	5
3	1	1	4	5	$\cdots$	3
4	1	1	4	5	$\cdots$	5
5	1	1	3	4	$\cdots$	5

TABLA 4.1. Ejemplo de estructura de datos original

En resumen, los datos originales están estructurados como una matriz de dimensiones $M \times (p+3)$, donde M es el número de encuestas y p es el número de ítems en la encuesta. La columna 1 es la identificación del estudiante, la columna 2 es la identificación del curso y la columna 3 es la identificación del docente. Las columnas 4 a $p+3$ contienen las

puntuaciones que el estudiante asigna al profesor para cada ítem. Esas puntuaciones están entre 1 y K_j, para cada columna.

El proceso de vectorización consiste en generar un vector columna que contenga todas las respuestas del estudiante a todos los ítems. La tabla 4.2 presenta una estructura de datos vectorizada, allí se observa el reordenamiento de los datos.

Estudiante	**Curso**	**Profesor**	**Item**	**Y**
1	1	1	1	4
1	1	1	2	4
⋮	⋮	⋮	⋮	⋮
1	1	1	14	5
2	1	1	1	5
⋮	⋮	⋮	⋮	⋮
5	1	1	1	3
5	1	1	2	4
⋮	⋮	⋮	⋮	⋮
5	1	1	14	5

TABLA 4.2. Ejemplo de estructura de datos vectorizada

Los datos reestructurados son una matriz $N \times 5$, donde N es el número total de respuestas. La columna 1 es el identificador del estudiante, la columna 2 corresponde al identificador del curso, la columna 3 corresponde al identificador del docente, la columna 4 es el identificador del ítem y la columna 5 corresponde al valor de la respuesta dada por el estudiante. El código implementado en el el software R para la vectorización de los datos se encuentra en la sección C.1 del apéndice C.

Luego de la preparación de los datos se procede a ejecutar el programa Stan que contiene la estructura del modelo propuesto. Stan hace uso del muestreador NUTS descrito en el capitulo anterior. La estructura del código se encuentra en la sección C.2 del apéndice C. Las líneas de código 1 a 13 hacen referencia a la definición del bloque de datos donde se debe definir la estructura de datos vectorizada descrita anteriormente.

En cada paso del algoritmo implementado en Stan se obtiene una muestra de todos los parámetros. Las líneas de código 14 a 22 definen la estructura de los parámetros, esto es, si los parámetros son en forma vectorial o simplemente parámetros reales, y sirven para definir la estructura de extracción de las muestras en el primer paso del algoritmo. Las líneas 24 a 32 definen las distribuciones a priori de los parámetros para la extracción de las muestras. A partir de las muestras de parámetros se genera una muestra de la distribución posterior, esta se especifica en el bloque del modelo, en las líneas 34 a 35.

En situaciones ideales cuando no se requiere aproximación de la distribución posterior se obtienen muestras directamente de la misma. Sin embargo, como se deben hacer aproximaciones debido a que no se tienen las ecuaciones exactas, es necesaria la implementación de un esquema de aceptación de muestras similar al implementado en el algoritmo de Metropolis para verificar que los parámetros se encuentren en el espacio de la distribución objetivo.

Finalmente es necesario extraer la información de la log-verosimilitud del modelo, por lo tanto esta se define en el bloque de cantidades generadas en las líneas 39 a 41.

El ajuste del modelo a través de RStan se presenta en la sección C.3 del apéndice C, allí se muestran las líneas de código (líneas 1 a 19) necesarias para hacer la ejecución del modelo a través de R, así como las líneas de código necesarias para el calculo de los criterios de selección de modelos (líneas 26 a 30).

4.5.1. Diagnóstico

Cuando se genera una muestra de la distribución posterior se busca que los valores simulados sean representativos de la misma, para eso se deben tener suficientes estimaciones buscando que estas sean precisas y estables. Los métodos MCMC garantizan que cadenas largas (con tendencia hacia el infinito) logran representaciones perfectas, sin embargo, se debe tener un criterio para cortar la cadena y evaluar la calidad de las muestras simuladas.

Para determinar la convergencia de la cadena es conveniente realizar más de una cadena, con esto se busca ver si ha olvidado el estado inicial y revisar que algunas cadenas no hayan quedado atoradas en regiones inusuales del espacio de parámetros.

Para el ajuste del modelo se simulan cuatro cadenas, cada una tiene 3000 iteraciones de las cuales 1500 hace parte del periodo de calentamiento. En el gráfico 4.1 se representan cuatro cadenas. La primera cadena está relacionada con el parámetro de habilidad θ_1, la segunda con el parámetro de severidad γ_1, la tercera con el parámetro de curso η_1 y la cuarta está relacionada con el parámetro de ítem $\beta_{1,1}$. La primera parte de cada uno de los gráficos se encuentra sombreada, esto indica el punto de corte de las iteraciones del periodo de quemado. Esta gráfica ayuda a determinar si se eligió un periodo de calentamiento adecuado o si alguna cadena se encuentra alejada de las otras.

Además de la evaluación gráfica se puede usar la medida$\hat{R}$ que es una estadística de diagnostico de convergencia propuesta por Gelman et al. (1992). Se conoce como el *factor de reducción potencial de escala*. Se define como la posible reducción en la longitud de un intervalo de confianza si las simulaciones continuaran de manera infinita. Su expresión matemática está dada por:

$$\hat{R} = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\frac{\hat{V}}{\hat{W}}} \tag{4.16}$$

donde $\hat{W}$ es la varianza dentro de las cadenas, $\hat{V}$ es una estimación del varianza posterior del parámetro y B es la varianza entre las cadenas,

$$\begin{aligned}
\hat{W} &= \frac{1}{M}\sum_m \hat{\sigma}_m^2 \\
B &= \frac{N}{M-1}\sum_m (\hat{\theta}_m - \hat{\theta})^2 \\
\hat{V} &= \frac{N-1}{N}\hat{W} + \frac{M+1}{MN}B
\end{aligned}$$

Si $\hat{R}$ es mayor a 1,1 indica que las cadenas no se han mezclado bien.

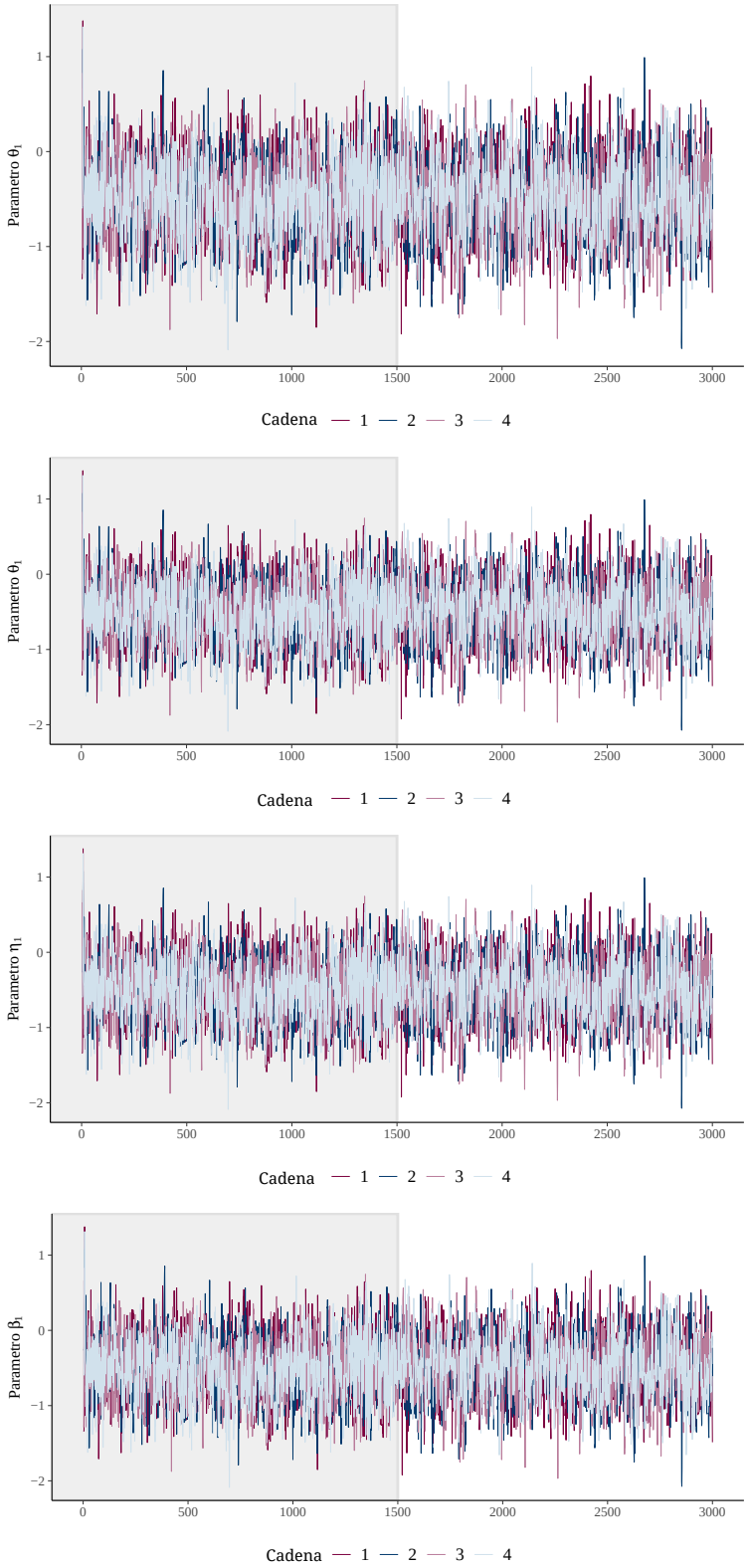

FIGURA 4.1. Convergencia de cadenas para algunos parámetros del modelo

Por otra parte, una vez se tiene una muestra de la distribución posterior, es necesario asegurarse de que la muestra es lo suficientemente grande para producir estimaciones

estables y precisas. Una estadística que ayuda a verificar esto es el n_eff el cual es el tamaño efectivo de muestra, dado que las simulaciones en general están correlacionadas, el tamaño efectivo indica que tamaño de muestra de observaciones independientes daría la misma información que las simulaciones de la cadena. La expresión matemática está dada por:

$$n\_eff = \frac{N}{1 + 2\sum_{k=1}^{\infty} ACF(k)} \tag{4.17}$$

Se esperaría tener un tamaño efectivo de al menos 100 para cada uno de los parámetros.

A continuación se muestra un ejemplo de salida de Rstan, obtenida después del ajuste del modelo.

```
    Inference for Stan model:
4 chains, each with iter=3000; warmup=1500; thin=1;
post-warmup draws per chain=1500, total post-warmup draws=6000.

          mean se_mean  sd    2.5%    25%   50%  75%   97.5%  n_eff Rhat
theta[1]  -0.5       0  0.4   -1.3   -0.8  -0.5  -0.2   0.3   1717    1
theta[2]  -0.1       0  0.3   -0.7   -0.3  -0.1   0.1   0.5   1134    1
theta[3]   0.6       0  0.3   -0.1    0.4   0.6   0.8   1.3   1398    1
theta[4]   1.8       0  0.4    0.9    1.5   1.8   2.1   2.6   1794    1
theta[5]  -0.2       0  0.4   -0.9   -0.5  -0.2   0.0   0.5   1637    1
```

En esta se observa un resumen de estadísticas para cada uno de los parámetros estimados. Entre las estadísticas obtenidas se destacan promedio, desviación estándar, cuartiles para los intervalos de credibilidad. Además, incluye dos resultados adicionales, un valor n_eff que se interpreta como el número efectivo de muestras para la distribución posterior sobre la cual se realiza la inferencia y el valor $Rhat$ que es la estadística de diagnostico de convergencia descrita anteriormente, una buena convergencia de las cadenas se identifica con valores cercanos a 1.

CAPÍTULO 5

Aplicación: Evaluación del desempeño docente en la Universidad Nacional de Colombia

Este capítulo aborda la aplicación de la propuesta diseñada en el capítulo anterior relacionada con un modelo de múltiples facetas para la evaluación del desempeño docente. Los datos provienen de una encuesta de estudiantes que se realiza cada semestre en la Facultad de Ciencias de la Universidad Nacional de Colombia. La encuesta está diseñada para medir la percepción del rendimiento de los docentes en el aula, la relevancia del curso y la eficiencia de los recursos disponibles en el campus.

En este estudio se utilizaron los datos del segundo semestre de 2015. El cuerpo de datos inicial contenía respuestas asociadas a 14703 registros. Esta contiene información relacionada con los identificadores de docente, del curso y del estudiante. Además de esto, se dispone de los datos relacionados con el departamento que ofrece la asignatura y el sexo de los docentes y de los estudiantes.

De acuerdo al modelo propuesto en este documento, donde se quiere capturar variabilidad del efecto de curso, se hace una depuración de los datos, con el fin de obtener docentes que ofrecen al menos dos cursos en ese semestre. Teniendo en cuenta lo anterior, se verifica que 203 docentes presentan dicha condición. Para efectos prácticos en términos computacionales, también para llevar a cabo el análisis de los datos y con el fin de probar el modelo propuesto, se selecciona una muestra de 50 docentes. En dicha muestra, quedan representadas un total de 2505 respuestas de estudiantes, que están relacionadas con 94 cursos. De la muestra de docentes seleccionados, 36 de ellos imparten dos cursos y los 14 restantes imparten tres cursos. Por otra parte, de los 94 cursos que son seleccionados en la muestra, 79 son impartidos por un único docente, mientras que 10 por 2 docentes, 3 por 3 docentes y 1 impartido por 4 docentes.

La encuesta se realiza semestralmente y contiene 25 preguntas que indagan al estudiante por su percepción sobre el desempeño docente en el aula. La encuesta se compone de 8 secciones, así:

1. **Datos generales del estudiante:** Indaga sobre las características sociodemográficas de los estudiantes sin solicitar información personal explícita debido a que las respuestas de la encuesta son de carácter confidencial.

2. **Gestión del curso:** Indaga sobre la gestión del docente en la asignatura, se abordan temas relacionados con el cumplimiento del programa, asistencia regular a las clases por parte del docente, y sobre si se propicia el trabajo en grupo y se reconocen los logros de los estudiantes por parte del docente.

3. **Impacto del curso:** Indaga sobre el impacto de la asignatura en la formación profesional; se abordan temas relacionados directamente con los contenidos.

4. **Ambientes de aprendizaje:** Indaga sobre los ambientes de aprendizaje, se abordan temas relacionados con el uso de otros espacios para el desarrollo de la asignatura y finalmente se indaga sobre la promoción de autodidactismo por parte del docente.

5. **Uso de la tecnología:** Indaga sobre el uso de tecnologías y herramientas digitales para el desarrollo de competencias por parte de los estudiantes.

6. **Planta física e implementos:** Indaga sobre las condiciones de infraestuctura y la calidad de los implementos utilizados para el desarrollo de la asignatura.

7. **Monitores:** Indaga sobre la necesidad y la importancia del apoyo de los monitores en la asignatura.

8. **En general:** Indaga sobre temas relacionados con fortalezas y debilidades del desempeño general del docente y del contenido de la asignatura.

Las respuestas a las preguntas vienen presentadas en dos escalas tipo Likert de 5 opciones, las cuales dependen del tipo pregunta que se realiza. La primera escala viene dada por las opciones *Deficiente*, *Malo*, *Regular*, *Bueno* y *Excelente* y está dirigida a preguntas relacionadas con la percepción del desempeño global del propio estudiante, la percepción del desempeño del docente y la percepción del contenido de la asignatura. La segunda escala viene dada por las opciones *Nunca*, *Casi Nunca*, *A veces*, *Casi Siempre* y *Siempre* y está dirigida a preguntas específicas sobre la percepción del desempeño del docente y el desarrollo de la asignatura.

Teniendo en cuenta lo anterior, se realiza una inspección inicial de los datos y se observa que las categorías más bajas en las dos escalas: *Deficiente* o *Nunca*, según sea el caso tienen pocas respuestas, razón por la cual, las dos primeras categorías fueron colapsadas y los ítems por lo tanto son recodificados en cuatro categorías ordinales de respuesta.

Como se mencionó la encuesta está dividida en 8 secciones. Sin embargo, cada una de las secciones indaga por un aspecto en particular relacionado con la percepción del estudiante sobre el curso y sobre el desempeño del docente. Dado que en este estudio nos enfocamos en el aspecto del desempeño docente, no se tuvo en cuenta las secciones: Impacto del curso y planta física e implementos. En el apéndice A se encuentra la encuesta final depurada que se utilizó para la aplicación expuesta en este documento.

5.1. Análisis de unidimensionalidad

El primer paso para definir una escala de medida es verificar que los datos cumplan el supuesto de unidimensionalidad, esto es, que midan un único constructo. Existen varias herramientas para evaluar la unidimensionalidad de un conjunto de datos. En este documento se abordan dos de ellas: el análisis de componentes principales (Jolliffe, 2002) y el

análisis de confiabilidad a través del coeficiente Alfa de Cronbach (Cronbach, 1951) con el uso de la curva de Cronbach-Mesbah (Cameletti & Caviezel, 2012).

La herramienta para verificar la unidimensionalidad a través del análisis de componentes principales se define a partir del uso del número de dimensiones halladas por medio del gráfico de valores propios. En general, si el primer valor propio es sustancialmente más grande comparado con los otros valores se sugiere unidimensionalidad en los datos.

El coeficiente Alfa de Cronbach es un índice entre 0 y 1 que representa la fuerza de la relación entre los ítems; se construye a partir de la varianza de la puntuación total y la varianza de la puntuación de cada ítem. En Lord & Novick (2008) se menciona que valores más grandes que 0,75 pueden estar asociados con la medición de un único constructo.

En la figura 5.1 el panel izquierdo presenta el gráfico de barras del porcentaje de varianza explicada por los valores propios obtenidos a través del análisis de componentes principales, y el panel derecho presenta la curva de Cronbach-Mesbah.

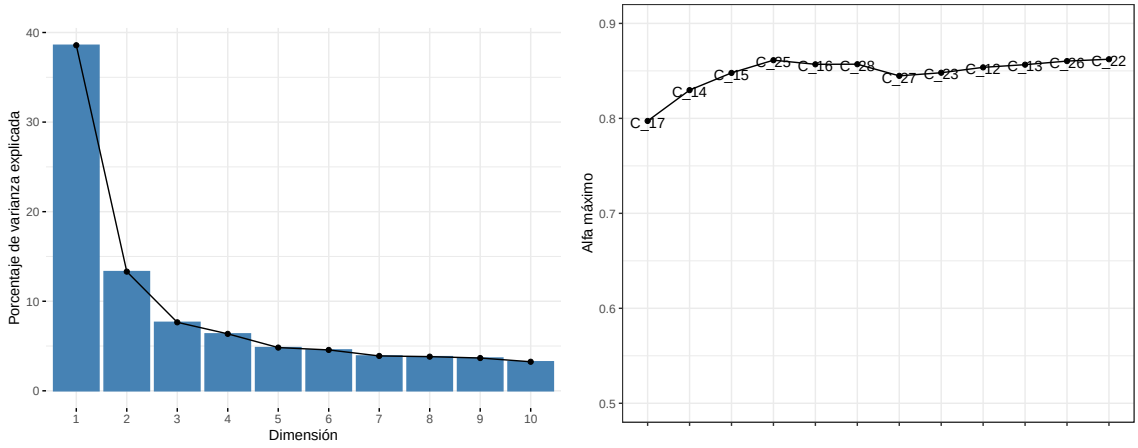

FIGURA 5.1. Gráficos para verificación de unidimensionalidad del instrumento

De la figura anterior, el panel izquierdo sugiere visualmente que los datos son unidimensionales ya que el porcentaje de varianza explicada por el primer valor propio es sustancialmente mayor al porcentaje explicado por el segundo valor. Por otra parte, en el panel derecho, la curva de Cronbach-Mesbah muestra el coeficiente de Cronbach después de que cada ítem que retirado maximiza la confiabilidad de los datos es removido.

En Cameletti & Caviezel (2012) se menciona que si los datos son unidimensionales, entonces la curva es monótonamente creciente. Sin embargo, aunque la curva no tiene ese comportamiento, se observa que el cambio en la confiabilidad al extraer dos ítems es despreciable en términos de la escala. Asimismo, se mencionó que valores mayores a 0,75 del coeficiente están asociados con la medición de un único constructo, se observa que los valores estimados del Alfa de Cronbach se encuentran entre 0,79 y 0,87. Dado lo anterior, se concluye que el instrumento cumple con el supuesto de unidimensionalidad.

Además, el valor estimado del Alfa de Cronbach para la muestra es de 0.86, por lo tanto, según lo mencionado en Lord & Novick (2008) se presume que el instrumento está midiendo un único constructo.

5.2. Estimación de parámetros

El modelo propuesto en el presente documento se implementa a través de la interfaz de R para Stan, cuyo nombre es Rstan. Para hacer uso de Rstan se debe escribir un programa *.stan* que calcula directamente la densidad log-posterior. Luego este código se compila y se ejecuta con el conjunto de datos, a través de la interfaz de R. Estos dos códigos son complementarios. Para ampliar información sobre el software utilizado remítase al apéndice C. En esta sección se describen los resultados más relevantes relacionados con la estimación de los parámetros.

5.2.1. Estimación de la severidad del evaluador

La inclusión de un parámetro de severidad en el modelo permite capturar la variabilidad asociada a las respuestas dadas por los estudiantes, determinando para cada uno de ellos el grado de severidad o indulgencia con que perciben el desempeño de cada docente que evalúa. Valores negativos estimados del parámetro implican que el evaluador tiene una tendencia a dar calificaciones más altas que el promedio en los aspectos evaluados, es decir, que es indulgente en la calificación del desempeño del docente mientras que valores positivos estimados implican que el evaluador presenta una tendencia a dar calificaciones bajas en los aspectos, lo que implica una severidad alta.

En la figura 5.2 se presenta la información principal de la estimación del parámetro de severidad. En la parte izquierda se observa la distribución del parámetro de severidad γ para los 2505 evaluadores y en la parte derecha las bandas de credibilidad ordenadas para algunos de ellos.

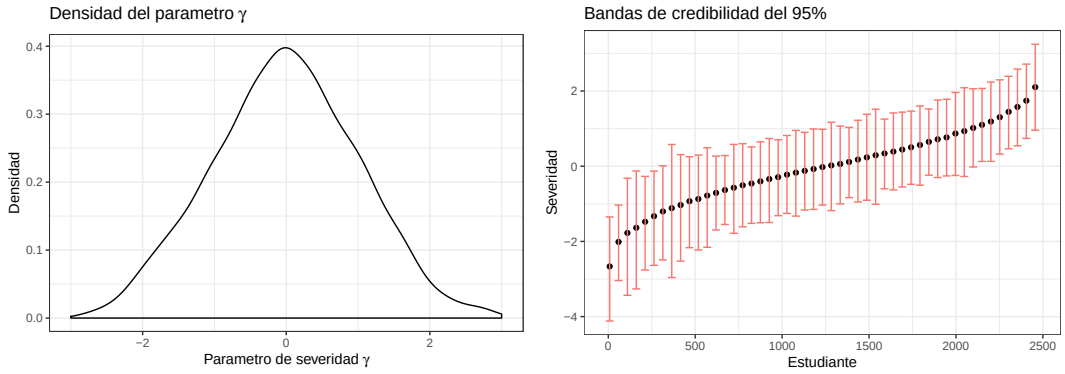

FIGURA 5.2. Resultados de la estimación del parámetro de severidad γ

La distribución de los parámetros de severidad es aproximadamente simétrica; los valores estimados se encuentran en un rango de -3,564 a 4,983, con un valor promedio estimado de -0,007 y una desviación estándar de 1,031. El modelo sugiere que en la muestra hay evaluadores indulgentes y con un alto grado de severidad en la misma proporción.

Adicionalmente se posee información auxiliar sobre el género de los estudiantes y sobre el tipo de vinculación que tienen con la universidad, es decir, si hacen parte de un programa de pregrado o de uno de posgrado. En la figura 5.3 se presentan los gráficos de caja que muestran las distribuciones de los parámetros de severidad estimados según las clasificaciones descritas.

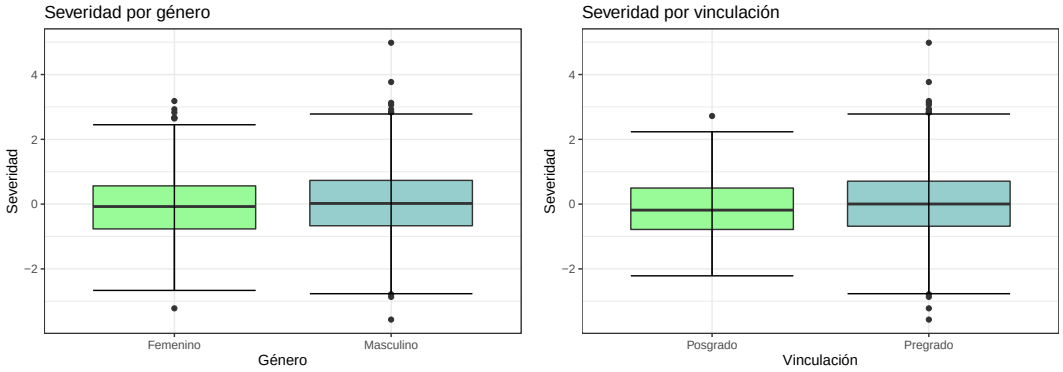

FIGURA 5.3. Resultados de parámetro de severidad según género y tipo de vinculación

De lo anterior, visualmente se observa que no hay diferencias sustanciales entre las calificaciones otorgadas por los estudiantes de genero femenino con los de masculino. Sin embargo, la severidad promedio de las mujeres es de -0,072 mientras que la de los hombres es de 0,027, lo que sugiere que los hombres son ligeramente menos indulgentes al momento de calificar el desempeño docente. Para ilustrar si hay diferencias entre estos promedios, se realiza una prueba de diferencia de medias y se obtiene un p-valor de 0,023, lo cual indica que si hay diferencia entre la severidad promedio de las mujeres y de los hombres, sin embargo, es importante tener en cuenta que los tamaños de muestra para realizar esta prueba son grandes (853 mujeres y 1652 hombres) y por lo tanto está prueba será significativa en favor de la diferencia.

Como se mencionó en el capítulo 1, Feldman (1977) menciona que estudiantes mujeres dan mejores puntuaciones que estudiantes hombres, para los datos de la Facultad de Ciencias se observa este comportamiento, debido a que la severidad promedio de las mujeres es menor que la de los hombres, lo que indica que en general las estudiantes mujeres otorgan calificaciones más altas a los docentes.

Para la variable tipo de vinculación se observan diferencias entre las dos distribuciones, sin embargo, para este caso, la distribución de las severidades de los estudiantes de pregrado tiene mayor variabilidad. La severidad promedio de los estudiantes de posgrado es de -0,164 mientras que la de los estudiantes de pregrado es de 0. Lo anterior sugiere, que los estudiantes de posgrados son en promedio más indulgentes que los estudiantes de pregrado. De igual forma, al realizar una prueba de diferencia de medas se obtiene un p-valor de 0,105, lo que indica que hay diferencias entre la severidad promedio de los estudiantes de pregrado y posgrado.

En el capítulo 1 también se menciona que cursos de niveles más avanzados tienden a recibir mayores puntuaciones, para el caso de esta aplicación, se observa que los estudiantes de posgrado otorgan en promedio calificaciones más altas debido a que su severidad promedio es menor que la de estudiantes de pregrado.

5.2.2. Estimación del parámetro de curso

El parámetro que captura la variabilidad de los cursos en el modelo, mejora el ajuste de la estimación del desempeño docente al controlar el efecto de las características del curso. Este parámetro se interpreta como una dificultad general del curso, por lo tanto, valores negativos estimados del parámetro indican que el curso tiene una percepción de

dificultad baja mientras que valores positivos indican que el curso tiene una percepción de dificultad alta.

En la figura 5.4 se presenta la información principal de la estimación del parámetro del curso. En la parte izquierda se observa la distribución del parámetro para los 97 cursos y en la parte derecha las bandas de credibilidad ordenadas.

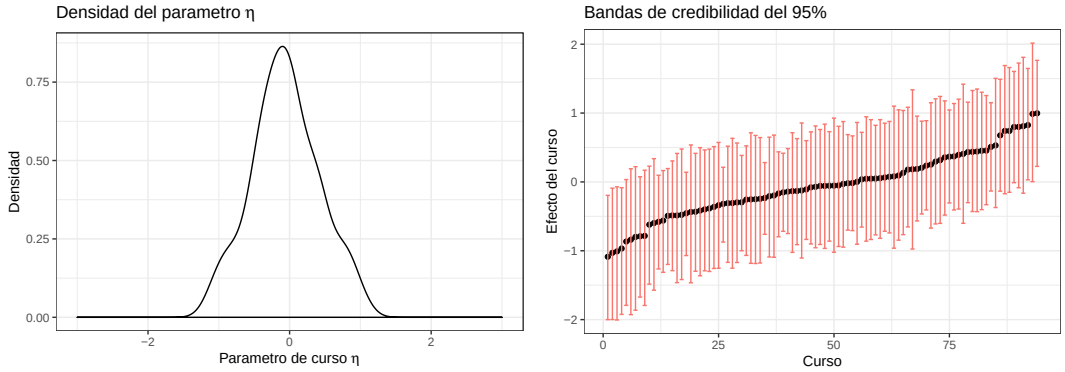

FIGURA 5.4. Resultados de la estimación del parámetro de curso η

Según se observa, la distribución de los parámetros del curso η tiene una distribución aproximadamente simétrica, los valores estimados se encuentran en un rango de -1,087 a 0,997, con un valor promedio estimado de -0,056 y un valor de desviación de 0,473. La distribución presenta una variabilidad baja, lo que indica que la dificultad estimada de los cursos se concentra alrededor de 0. Por la simetría evidenciada en la gráfica, el modelo sugiere que en la muestra hay cursos con percepción de dificultad baja y alta aproximadamente en la misma proporción.

Adicionalmente se cuenta con información sobre el departamento académico que imparte el curso en la universidad. En la tabla 5.1 se presenta el resumen de la estimación del parámetro para cada uno de los departamentos de la Facultad de Ciencias.

		Dificultad η	
Departamento	**n**	**Promedio**	**Desviación**
Biología	10	-0,137	0,545
Estadística	4	-0,074	0,777
Farmacia	14	-0,085	0,447
Física	17	-0,076	0,472
Geología	9	0,073	0,674
ICN	4	-0,022	0,725
Matemáticas	13	0,102	0,393
Química	26	-0,084	0,371

TABLA 5.1. Resultados de parámetro del curso η por departamento

Se observa que el parámetro promedio estimado más alto lo poseen los cursos pertenecientes al departamento de Matemáticas, además, tienen una variabilidad baja, lo que indica que en general estos cursos tienen una percepción de dificultad más alta para todos los estudiantes. Por otra parte, el departamento de Biología presenta una dificultad promedio más baja con una variabilidad media, lo que indica que los cursos en general tienen una percepción de dificultad más baja para los estudiantes evaluadores. Finalmen-

te, el departamento de Estadística presenta la mayor variabilidad en la percepción de la dificultad del curso en todos los departamentos de la Facultad de Ciencias.

5.2.3. Estimación del desempeño docente y parámetros de los ítems

El parámetro θ relacionado con el desempeño docente es quien juega uno de los papeles más importantes en la estructura del modelo, debido a que este representa la habilidad del evaluado, en el caso representada como la percepción del desempeño docente en el aula. Valores negativos estimados del parámetro indican una percepción de desempeño baja del docente según el criterio de los estudiantes mientras que valores positivos indican una percepción de desempeño alta.

En la parte izquierda de la figura 5.5 se observa la distribución del parámetro y en la parte derecha las bandas de credibilidad del 95 % ordenadas para cada uno de los 50 docentes de la muestra.

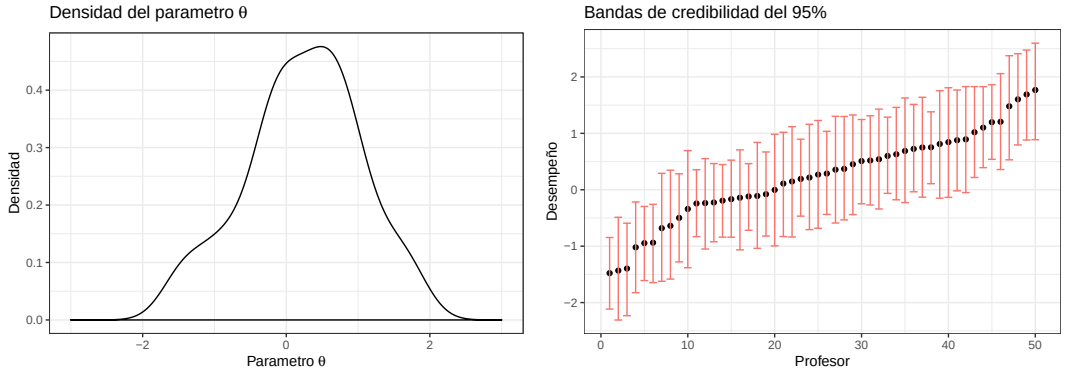

FIGURA 5.5. Resultados de la estimación del parámetro de habilidad θ

La distribución posterior tiene un valor promedio de 0,235 y una desviación estándar de 0,796, evidentemente es una distribución asimétrica negativa, es decir, presenta percepciones de desempeño docente negativas atípicas en comparación del conjunto de la muestra. En general, la percepción del desempeño docente en la muestra es buena, sin embargo, hay un pequeño conjunto de docentes con una percepción de desempeño muy baja en comparación al promedio. Lo que implica que la distribución sea sesgada negativamente.

En la figura 5.6 se presenta la distribución de las habilidades por género del docente.

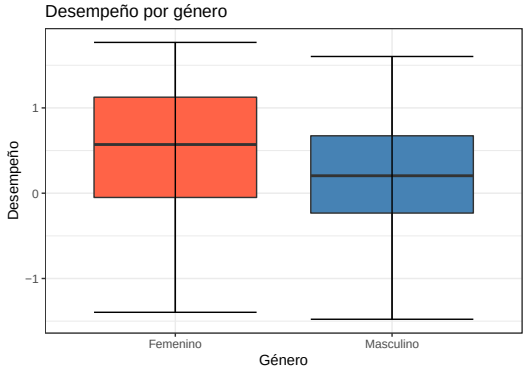

FIGURA 5.6. Distribución de habilidades estimadas por género

La distribución para el género femenino es más variable que la del género masculino. El desempeño promedio estimado en las docentes se percibe más alto que en los docentes, con valores de 0,501 y de 0,105, respectivamente. Al realizar una prueba de diferencia de medias entre el desempeño promedio de las docentes y de los docentes se obtiene un p-valor de 0,094, valor que indica que existe una diferencia entre los valores promedio estimados. En el capítulo 1 se menciona que las puntuaciones de los estudiantes son sesgadas para las mujeres docentes, los resultados anteriores sugieren que en la Facultad de Ciencias hay diferencias significativas entre la percepción del desempeño docente de las mujeres docentes y de los hombres, en favor de las mujeres, es decir, los estudiantes otorgan en general calificaciones más altas a las docentes mujeres.

La tabla 5.2 presenta un resumen de las habilidades estimadas por departamento académico del docente.

		Desempeño θ	
Departamento	**n**	**Promedio**	**Desviación**
Biología	4	0,593	0,363
Estadística	2	0,287	0,099
Farmacia	5	0,904	0,817
Física	9	0,315	0,637
Geología	4	-0,397	0,631
ICN	2	0,526	1,646
Matemáticas	8	-0,412	0,931
Química	16	0,328	0,667

TABLA 5.2. Resumen de habilidades estimadas por departamento

En la tabla se observa que los docentes de los departamentos de Geología y de Matemáticas son los que tienen menor percepción de desempeño promedio y el departamento de Farmacia es el que tiene mayor percepción de desempeño promedio de todos los departamentos de la Facultad de Ciencias.

La figura 5.7 muestra el comportamiento de los ítems en la prueba con respecto al parámetro β. Cada línea representa un ítem, y los puntos rojos son aquellos puntos de corte β_k estimados. Cada ítem tiene 3 valores de β_k, naturalmente debido a que cada uno tiene 4 opciones de respuesta.

Se observa que la distribución de los puntos de corte β_k de los ítems se encuentran concentrados en valores entre -5 y 2, esto es, valores sesgados a la izquierda. Teniendo en cuneta esto, únicamente dos ítems tienen su valor de β_k más alto llegando a 2. Esto es un indicador de que el test en general tiene un comportamiento ligeramente sesgado a la izquierda, es decir, que los ítems se perciben con una dificultad baja para la medición de la percepción del desempeño docente.

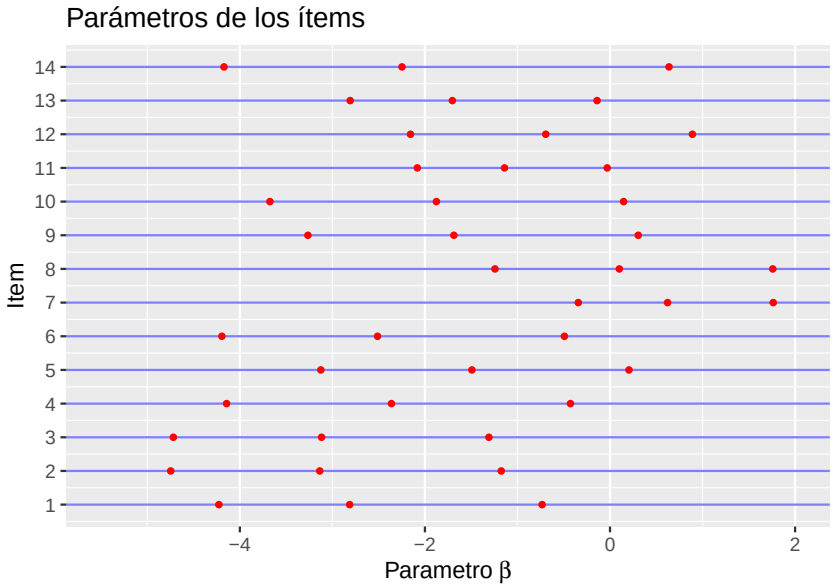

FIGURA 5.7. Parámetros estimados de los ítems

Los ítems 2 y 3 presentan la categoría de dificultad más baja; en estos dos los estudiantes tienden a dar a los docentes calificaciones muy altas. Las preguntas son:

- Pregunta 2: ¿El docente respetó las fechas acordadas para las actividades académicas, incluidas las evaluaciones y entrega de resultados?
- Pregunta 3: ¿El docente prepara con anterioridad cada una de las sesiones de la asignatura?.

Son preguntas de las que se espera en general unas buenas calificaciones para todos los docentes, debido a que en general los docentes cumplen con las fechas establecidas y preparan sus clases con anterioridad.

En contraste, los ítems 7 y 8 presentan la categoría de dificultad más alta del cuestionario ya que los estudiantes otorgaron calificaciones homogéneamente más bajas. Las preguntas son:

- Pregunta 7: ¿El docente incluye experiencias de aprendizaje en lugares diferentes al aula (talleres, laboratorios, empresa, comunidad, etc)?
- Pregunta 8: ¿El docente organiza actividades que me permiten ejercitar mi expresión oral y escrita?.

Los valores estimados de los parámetros de los ítems se pueden ver en el apéndice B.

5.2.4. Ajuste del modelo y criterios de selección

En esta sección, se realiza la revisión del ajuste del modelo y se presenta un modelo con menos parámetros que sirve de comparación para el modelo propuesto en este documento, para efectos prácticos el modelo de comparación se denomina *modelo cero*.

Tecnicamente, el modelo cero es aquel modelo que en su forma matemática no contempla el parámetro adicional η, es decir, este modelo de múltiples facetas no reconoce la variabilidad asociada al curso impartido por el docente. El modelo de probabilidad condicional se define como:

$$Pr\left[Y_{ijs} = k|\theta_i, \gamma_{is}, \beta_j\right] = \text{logit}^{-1}(\theta_i - \gamma_{is} - \beta_{j(k-1)}) - \text{logit}^{-1}(\theta_i - \gamma_{is} - \beta_{jk}) \tag{5.1}$$

Donde Y_{ijs} es la variable aleatoria que representa la puntuación asignada por el estudiante s al docente i en el ítem j, θ_{ic} es el trazo latente o desempeño del docente i, β_j la dificultad del ítem j y γ_{is} la severidad del estudiante s al docente i.

Dada la expresión anterior, y haciendo una revisión de la expresión (4.5) donde se define la probabilidad condicional del modelo propuesto, se puede deducir que el modelo propuesto es una generalización del modelo cero debido a que el objetivo de la inclusión del parámetro η es capturar variabilidad adicional que no debe influir en la estimación del parámetro de habilidad θ.

La figura 5.8 muestra las distribuciones de los parámetros que tienen en común los dos modelos, es importante mencionar que el ajuste del modelo cero se realizó con la misma muestra de 50 docentes con la que se ajustó el modelo propuesto. En el panel izquierdo de la figura se observan los resultados de la comparación de las distribuciones para el parámetro de habilidad (θ) y en el panel derecho los resultados para el parámetro de severidad ($gamma$). En color azul se encuentra la distribución obtenida con el modelo propuesto y en color rosa la obtenida con el modelo cero.

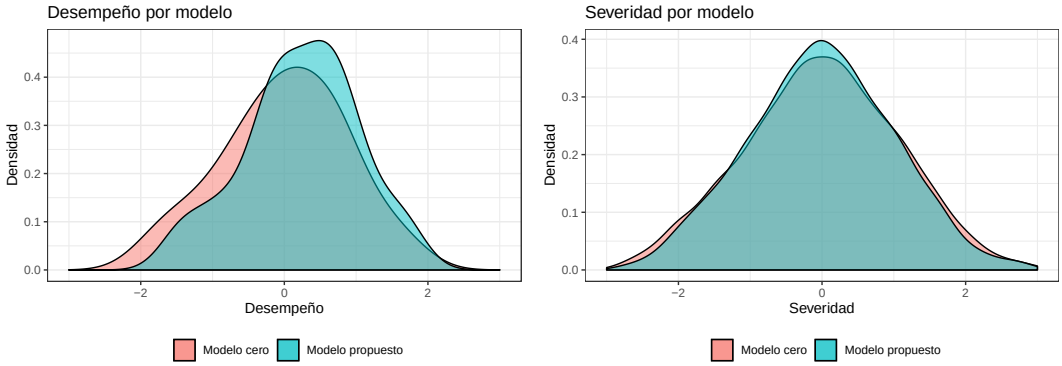

FIGURA 5.8. Comparación de las distribuciones de los parámetros θ y γ

Para el parámetro de habilidad θ la distribución obtenida con el modelo propuesto tiene un sesgo a la derecha comparada con la distribución obtenida en el modelo cero. Para el parámetro de severidad, las distribuciones estimadas para los dos modelos tienen un comportamiento similar, de hecho, las curvas se traslapan en casi todo el rango de medición. Lo anterior se explica en la inclusión del parámetro adicional η que se denominó efecto del curso, esto indica que la inclusión de dicho parámetro está afectando únicamente la estimación del parámetro de habilidad que mide la percepción del desempeño docente al mejorar las estimaciones de dichas habilidades dado que aísla el efecto de percepción de dificultad que tienen los evaluados sobre el curso en el momento de evaluación.

La figura 5.9 muestra la comparación de los valores estimados β_k para cada uno de los ítems, en los dos modelos. Los puntos de color rojo representan los valores para el modelo cero y los puntos de color café los del modelo propuesto.

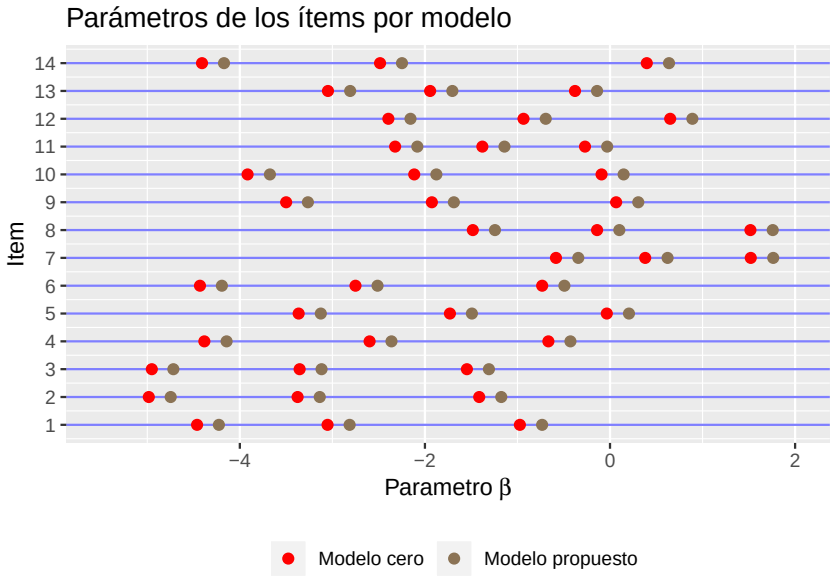

FIGURA 5.9. Parámetros estimados de los ítems por modelo

Se observa en la figura que los parámetros estimados para el modelo propuesto presentan una traslación hacia la derecha respecto a los parámetros estimados con el modelo cero. De igual manera que con el parámetro de habilidad, esto indica que la estimación de los parámetros de los ítems está aislando el efecto del parámetro η.

Ajuste de los modelos

En la sección 4.4.1.1 se describe una estadística para la evaluación de bondad de ajuste del modelo. En esta sección, Se revisa el ajuste para los parámetros estimados de habilidad θ y de los ítems β_k, así como la obtención de la estadística de bondad de ajuste a nivel global del modelo. El ajuste se verificó para el modelo propuesto y para el modelo cero.

El valor de la estadística de bondad de ajuste global en el modelo propuesto es de 0,41, mientras que para el modelo cero es de 0,40. Como se mencionó, valores cercanos a 0,5 indican buen ajuste del modelo. Por lo tanto, los dos modelos indican buen ajuste global, sin embargo, es importante mencionar que el valor más cercano a 0,5 es el del modelo propuesto.

El promedio de los valores de las estadísticas de bondad de ajuste de los ítems en el modelo propuesto es de 0,43, mientras que para el modelo cero es de 0,44. Por otra parte, el promedio de los valores de las estadísticas de bondad de ajuste para los habilidades de los evaluados en el modelo cero fue de 0,55 y para el modelo propuesto fue del mismo valor. Según el criterio establecido, estos valores indican buen ajuste de los modelos.

Criterios de selección de los modelos

Para cada uno de los modelos se realiza la estimación de las estadísticas de los criterios de selección. En la tabla 5.3 se presentan los resultados obtenidos para los criterios WAIC y LOO detallados en el capítulo 4 del documento, debido a que estos criterios son más robustos en general en modelos jerárquicos o con efectos aleatorios.

Modelo	**WAIC**	**LOO**
Modelo cero	64808,8	64827,7
Modelo propuesto	64788,4	64806,6

TABLA 5.3. Criterios para selección de modelos

Los valores estimados para los dos modelos son cercanos, sin embargo, es importante mencionar que la ganancia que se obtiene al incluir el parámetro adicional en términos de aislar la variabilidad asociada al curso es importante. Los valores de las estadísticas son menores para los dos criterios WAIC y LOO para el modelo propuesto. Dado esto, se infiere que según los criterios de selección el modelo que debe seleccionarse es el modelo propuesto debido a que presenta valores menores.

CAPÍTULO 6

Conclusiones

En el desarrollo de este documento se abordó el problema de la estimación del desempeño docente a partir de evaluaciones de percepción de los estudiantes. En la literatura consultada se encontró evidencia que en general las instituciones hacen un uso incorrecto de los resultados de las encuestas de percepción de desempeño docente, esto, debido a que realizan promedios de escalas ordinales asignándole un valor numérico sin ningún sustento teórico. Por lo tanto, el objetivo principal en el desarrollo de este documento es dar un enfoque estadístico al uso de las evaluaciones de percepción de desempeño docente a partir del uso de modelos de TRI, con el fin de obtener estimaciones de trazos latentes tanto para el evaluado como para el evaluador, esto teniendo en cuenta que la evaluación involucra aspectos que pueden influir los cuales no se pueden controlar.

La evidencia muestra que la implementación de un modelo estadístico es importante en dos sentidos: 1) Permite realizar inferencias sobre los parámetros estimados asociados al modelo además de tener asociada una medición del error y 2) Captura variabilidad asociada a las características de la evaluación que no se tienen en cuenta cuando se realizan otros procedimientos, esto es, el parámetro del curso es importante para aislar el efecto del curso que toman los estudiantes y que indirectamente están midiendo al evaluar al docente.

En este documento se abordó el problema con unos datos provenientes de una encuesta de estudiantes de percepción de desempeño docente de la Facultad de Ciencias de la Universidad Nacional de Colombia. El instrumento final después de la revisión contaba con 14 ítems enfocados en temas que indagaban sobre la percepción del desempeño del docente en el aula.

La aplicación presentada en el documento tuvo dos orientaciones, la primera enfocada en el ajuste del modelo y el análisis de los resultados, y la segunda, enfocada en el ajuste del modelo y en la comparación del modelo propuesto con un modelo llamado modelo cero que no consideraba en su expresión matemática el parámetro asociado al curso η.

En la parte aplicada se hicieron varios hallazgos: El instrumento de medición cumple con el supuesto de unidimensionalidad, razón por la cual, se pudo ajustar el modelo de TRI propuesto, esto es, todos los ítems están midiendo un constructo, que para el caso se puede denominar *Percepción del desempeño docente*.

En cuanto al parámetro de severidad del modelo se observa que tiene un comportamiento simétrico en su distribución, lo cual indica que hay un rango de severidades

apropiado para la medición de las habilidades de los docentes. La importancia de incluir este parámetro es que existen características asociadas a los estudiantes que influyen en la calificación que estos otorgan a los estudiantes, como por ejemplo, el género o el tipo de vinculación que tiene con la institución, en este caso, se obtuvo que en general las estudiantes mujeres son más indulgentes, así como, los estudiantes asociados a programas de posgrado también lo son respecto a los que están inscritos en programas de pregrado.

En cuanto al parámetro de curso también se observa una distribución simétrica con variabilidad baja, es decir, que la mayoría de parámetros se encuentran alrededor del valor 0. La inclusión de este parámetro en el modelo es de mucha importancia debido a que este parámetro captura el efecto asociado a la percepción que tiene el estudiante sobre el curso y aísla este efecto de la estimación del parámetro de habilidad del docente. Dentro de los resultados del análisis para este parámetro se encontró que los cursos adscritos al departamento de Matemáticas tienen en promedio una dificultad mayor a los cursos de otros departamentos, y que los cursos pertenecientes al departamento de Biología son los que tienen en promedio una menor dificultad estimada.

Por otra parte, el modelo fue evaluado por dos vías: la primera, asociada con estadísticas de ajuste del modelo y la segunda, a través de una revisión de criterios de selección de modelos en comparación con un modelo con menos parámetros llamado modelo cero. En este análisis se observo que la estadística de ajuste del modelo es cercana a 0,5 (con un valor de 0,41) lo cual indica que el modelo ajusta bien los datos. Al realizar la comparación de los resultados obtenidos mediante las estadísticas de los criterios de selección de modelos se obtuvo que el modelo propuesto tiene valores menores en las estadísticas de WAIC y LOO, lo cual es un indicador que la inclusión del parámetro adicional η relacionado con el efecto del curso en el modelo ajusta mejor y es apropiado para los datos.

Teniendo en cuenta lo anterior, se concluye que la implementación de la metodología propuesta en este documento es de mucha utilidad, porque aborda estadísticamente una problemática a la que se enfrentan muchas instituciones en el momento de evaluar a sus docentes, esto en el sentido de mejorar los instrumentos de evaluación y de mejorar en sí el proceso de evaluación de los docentes.

CAPÍTULO 7

Trabajo futuro

El trabajo anterior sirve de motivación para profundizar en esta línea de investigación, en particular, se puede ampliar el estudio mediante la inclusión de más variables explicativas del modelo que posiblemente tengan efecto en la percepción del desempeño docente, esto es, proponer y desarrollar algunos modelos con más facetas. Lo anterior, dado que hay evidencia a partir de otros estudios que existen otras variables de relevancia, como por ejemplo, el género del docente o el tipo de vinculación que tiene el docente con la institución.

Respecto al modelo desarrollado se puede diseñar un escenario en el cual se proponga un modelo de múltiples de facetas multidimensional, que intente capturar múltiples habilidades de un evaluado, para el caso, medir diferentes habilidades de un docente.

Finalmente, con base en las ideas planteadas en este documento, se pueden implementar modelos de múltiples facetas que midan otras variables latentes de interés y que involucren en el problema evaluadores con diferentes características, como por ejemplo, evaluar competencias en escritura a través de un instrumento que debe ser calificado por evaluadores con diferente formación profesional.

APÉNDICE A

Cuestionario para evaluación del desempeño de los docentes

A continuación se presenta el listado de los ítems que hacen parte del cuestionario para evaluación de docentes de la Facultad de Ciencias de la Universidad Nacional de Colombia.

1. El docente asiste a clases regular y puntualmente.
2. El docente respetó las fechas acordadas para las actividades académicas, incluidas las evaluaciones y entrega de resultados.
3. El docente prepara con anterioridad cada una de las sesiones de la asignatura.
4. El docente es accesible y está dispuesto a brindar ayuda académica.
5. El docente propicia el trabajo en grupo, reconociendo los éxitos y logros en las actividades de aprendizaje.
6. El docente muestra compromiso y entusiasmo en sus actividades docentes.
7. El docente incluye experiencias de aprendizaje en lugares diferentes al aula (talleres, laboratorios, empresa, comunidad, etc).
8. El docente organiza actividades que me permiten ejercitar mi expresión oral y escrita.
9. El docente desarrolla el contenido de la clase de una manera ordenada y entendible.
10. El docente promueve el autodidactismo y la investigación.
11. El docente emplea la tecnología (computador, *videobeam*, plataformas digitales, correo) como un medio que facilite el aprendizaje de los estudiantes.
12. El docente promueve el uso de diversas herramientas digitales para gestionar (recabar, procesar, evaluar y usar) información.
13. El docente promueve el uso seguro, legal y ético de la información digital.
14. En general el desempeño del docente fue:

APÉNDICE B

Estimaciones de los parámetros de los ítems

En la tabla B.1 se presentan los resultados obtenidos del proceso de estimación de los parámetros de los ítems. La tabla presenta para cada uno de los parámetros el promedio estimado, la desviación estándar, algunos cuartiles de la distribución obtenida, el tamaño de muestra efectivo obtenido del proceso de estimación y el valor de la estadística $\hat{R}$.

Parámetro	Promedio	Desviación	q2.5	q25	q50	q75	q97.5	n_eff	$\hat{R}$
$\beta_{1,1}$	-4,226	0,206	-4,624	-4,366	-4,226	-4,085	-3,829	657,183	1,006
$\beta_{1,2}$	-2,814	0,180	-3,158	-2,938	-2,814	-2,691	-2,461	526,274	1,008
$\beta_{1,3}$	-0,734	0,170	-1,066	-0,850	-0,734	-0,617	-0,400	442,490	1,009
$\beta_{2,1}$	-4,747	0,224	-5,189	-4,898	-4,743	-4,590	-4,319	720,166	1,005
$\beta_{2,2}$	-3,136	0,184	-3,498	-3,262	-3,137	-3,008	-2,785	519,682	1,006
$\beta_{2,3}$	-1,175	0,170	-1,506	-1,291	-1,175	-1,056	-0,844	465,720	1,009
$\beta_{3,1}$	-4,718	0,223	-5,155	-4,871	-4,718	-4,567	-4,287	715,350	1,004
$\beta_{3,2}$	-3,117	0,184	-3,478	-3,246	-3,119	-2,988	-2,758	520,174	1,005
$\beta_{3,3}$	-1,308	0,171	-1,639	-1,428	-1,307	-1,188	-0,979	474,061	1,007
$\beta_{4,1}$	-4,143	0,203	-4,540	-4,283	-4,142	-4,001	-3,759	538,254	1,007
$\beta_{4,2}$	-2,362	0,177	-2,704	-2,485	-2,360	-2,240	-2,019	477,773	1,007
$\beta_{4,3}$	-0,428	0,169	-0,750	-0,547	-0,429	-0,311	-0,099	452,423	1,008
$\beta_{5,1}$	-3,125	0,183	-3,480	-3,249	-3,125	-3,003	-2,764	477,284	1,008
$\beta_{5,2}$	-1,492	0,171	-1,826	-1,612	-1,490	-1,375	-1,160	447,191	1,008
$\beta_{5,3}$	0,205	0,169	-0,125	0,088	0,205	0,321	0,531	435,686	1,009
$\beta_{6,1}$	-4,194	0,203	-4,583	-4,333	-4,191	-4,055	-3,800	599,736	1,006
$\beta_{6,2}$	-2,512	0,177	-2,859	-2,635	-2,513	-2,390	-2,169	452,545	1,008
$\beta_{6,3}$	-0,493	0,170	-0,817	-0,611	-0,495	-0,373	-0,165	435,131	1,008
$\beta_{7,1}$	-0,343	0,169	-0,670	-0,459	-0,343	-0,228	-0,014	450,456	1,008
$\beta_{7,2}$	0,622	0,169	0,295	0,505	0,622	0,739	0,950	444,261	1,008
$\beta_{7,3}$	1,764	0,172	1,425	1,645	1,764	1,883	2,093	457,299	1,008
$\beta_{8,1}$	-1,243	0,171	-1,573	-1,364	-1,244	-1,127	-0,911	456,982	1,008
$\beta_{8,2}$	0,101	0,169	-0,226	-0,017	0,102	0,217	0,427	443,212	1,008
$\beta_{8,3}$	1,758	0,171	1,429	1,639	1,760	1,874	2,090	460,742	1,008
$\beta_{9,1}$	-3,264	0,186	-3,626	-3,392	-3,264	-3,143	-2,896	512,494	1,007
$\beta_{9,2}$	-1,687	0,174	-2,028	-1,806	-1,686	-1,568	-1,354	432,418	1,008
$\beta_{9,3}$	0,306	0,169	-0,023	0,191	0,303	0,422	0,633	434,226	1,008
$\beta_{10,1}$	-3,675	0,192	-4,043	-3,810	-3,675	-3,544	-3,300	537,177	1,007
$\beta_{10,2}$	-1,876	0,175	-2,211	-1,997	-1,875	-1,757	-1,536	426,882	1,009
$\beta_{10,3}$	0,147	0,170	-0,177	0,030	0,146	0,262	0,478	424,026	1,009
$\beta_{11,1}$	-2,082	0,175	-2,423	-2,203	-2,082	-1,962	-1,748	469,911	1,007
$\beta_{11,2}$	-1,140	0,171	-1,474	-1,256	-1,140	-1,020	-0,816	437,022	1,008
$\beta_{11,3}$	-0,031	0,169	-0,362	-0,145	-0,030	0,085	0,296	447,719	1,008
$\beta_{12,1}$	-2,156	0,174	-2,498	-2,277	-2,156	-2,034	-1,821	469,995	1,007
$\beta_{12,2}$	-0,694	0,169	-1,023	-0,812	-0,692	-0,578	-0,364	452,131	1,008
$\beta_{12,3}$	0,890	0,169	0,566	0,771	0,890	1,007	1,216	465,294	1,008
$\beta_{13,1}$	-2,808	0,177	-3,147	-2,931	-2,809	-2,684	-2,468	474,871	1,007
$\beta_{13,2}$	-1,703	0,172	-2,040	-1,825	-1,703	-1,586	-1,373	446,791	1,009
$\beta_{13,3}$	-0,140	0,169	-0,472	-0,257	-0,140	-0,023	0,191	453,576	1,008
$\beta_{14,1}$	-4,171	0,200	-4,560	-4,309	-4,170	-4,035	-3,784	573,597	1,007
$\beta_{14,2}$	-2,248	0,176	-2,592	-2,369	-2,246	-2,126	-1,900	449,512	1,008
$\beta_{14,3}$	0,638	0,169	0,317	0,521	0,637	0,756	0,962	439,464	1,009

Tabla B.1. Estimaciones de los parámetros de los ítems

APÉNDICE C

Código RStan para estimación del desempeño docente

Stan es un lenguaje de programación probabilístico para el modelado estadístico que tiene una serie de herramientas inferenciales para ajustar modelos que sean robustos, escalables y eficientes (Carpenter et al., 2017). Lleva el nombre de Stanislaw Ulam, un matemático que fue uno de los desarrolladores del método de Monte Carlo en la década de 1940 (Metropolis & Ulam, 1949).

A menudo, se piensa que es un lenguaje similar a BUGS o JAGS, debido a que permite escribir un modelo bayesiano en una forma similar a la notación estadística. Sin embargo, Stan utiliza el muestreador No-U-Turn (Hoffman & Gelman, 2014) o una versión del método Monte Carlo Hamiltoniano (Neal et al., 2011). Es considerado más eficiente que dichos software para estadística Bayesiana (Luo & Jiao, 2018).

Los modelos escritos en Stan son compilados a C++, lo que hace que sea más rápido y permite la traducción a otros lenguajes de programación. Otra de las ventajas de Stan es que cuenta con interfaces para los lenguajes del análisis de datos más comunes como por ejemplo R o Python. El código es abierto y está disponible en *http://mc-stan.org/*.

C.1. Preparación de los datos

Para ejecutar la estimación de un programa Stan es necesario preparar los datos adecuadamente. A continuación, se presenta el código de R que se implementa para la preparación de los datos.

El primer paso es generar unas rutas donde se van a encontrar los archivos de interés y desde donde se leen los datos.

```
   # Ubicación de archivos
1  inPath <- file.path("..", "input")
2  outPath <- file.path("..", "output")
3  scrPath <- file.path("..", "scr")

   # Lectura de datos
```

```
4  datos <- read.csv(file.path(inPath, "D2_Sample_Data_Mayor2.txt"),
header=TRUE, sep=";")
```

El segundo paso es ordenar los datos y seleccionar las respuestas a los ítems de interés, los cuales están relacionados con el desempeño docente.

```
   # Ordenar los datos por Id de estudiante
5  data <- datos[order(datos[,1]),]

   # Extraer la matriz de respuestas a los ítems datos
6  Y <- as.matrix(data[,15:38]) # 1380 x 24

   # Selección de ítems relacionados con el desempeño docente
7  idx_prof <- c(1:7,11:17,23) # 14 ítems
```

El tercer paso es recategorizar las respuestas de los ítems, teniendo en cuenta que las categorías de menor frecuencia se transforman en una sola y seleccionar la matriz de datos con los ítems recategorizados.

```
   # Recategorización
8  for (i in idx_prof){
9    z <- Y[,i]
10   z <- ifelse(z==1 |z==2,1,z)
11   z <- ifelse(z==3,2,z)
12   z <- ifelse(z==4,3,z)
13   z <- ifelse(z==5,4,z)
14   Y[,i] <- z
15 }

   # Subconjunto de datos que contiene solo ítems de desempeño docente
16  y_prof <- Y[,idx_prof]
```

El cuarto paso es la generación de nuevos índices debido a que Stan maneja rangos de índices ordinales desde 1, esto es, que identificadores como el código del estudiante que se componen de seis dígitos, son necesarios recategorizarlos desde 1. Este paso incluye también la generación de nuevos identificadores para los docentes y los cursos.

```
    # Indices originales
17  Id_Stu_or <- sort(unique(data$Id_Student))
18  Id_Pro_or <- sort(unique(data$Id_Professor))
19  Id_Sub_or <- sort(unique(data$Id_Subject))
20  Id_Sel_or <- sort(unique(data$Self_Perfomance))

    # Longitudes de los identificadores
21  L1 <- length(Id_Stu_or)
22  L2 <- length(Id_Pro_or)
23  L3 <- length(Id_Sub_or)
24  L4 <- length(Id_Sel_or)
```

```
   #Generación de nuevos índices a partir de las longitudes
25 Id_Stu_new <- 1:L1
26 Id_Pro_new <- 1:L2
27 Id_Sub_new <- 1:L3
28 Id_Sel_new <- 1:L4
```

El quinto paso se realiza a partir de la generación de los nuevos índices, en este, se crean las nuevas estructuras de datos para incluir la información de estudiantes, docentes, y cursos y se realiza la asignación de los nuevos índices en las estructuras nuevas.

```
   # Generación de nuevos datos
29 N <- nrow(Y) #Número de respuestas
30 Student   <- c(Id_Stu_new)
31 Professor <- array(dim=N)
32 Subject   <- array(dim=N)

  # Asignación de nuevos índices para identificador de docente
33 for(i in 1:L2){
34   w = which(data$Id_Professor==Id_Pro_or[i])
35   Professor[w] = Id_Pro_new[i]
36 }

  # Asignación de nuevos índices para identificador de curso
37 for(i in 1:L3){
38   w = which(data$Id_Subject==Id_Sub_or[i])
39   Subject[w] = Id_Sub_new[i]
40 }
```

Finalmente, se seleccionan las variables auxiliares de interés, para el caso las que están relacionadas con género de estudiantes y docentes y el departamento asociado al curso, y se generan los nuevos datos y se guardan en la ruta deseada.

```
  # Variables auxiliares para los datos
41 varsAux <- c("Student_sex", "Carrer", "Professor_sex")


   # Generación de nueva matriz de datos re-ordenada
42 key_Student <- Id_Stu_or
43 complete_data <- cbind(key_Student,Student, Professor, Subject,
           Y, datos[,varsAux])

  # Guardar los nuevos datos recodificados
44 write.csv2(complete_data,file=file.path(outPath,
    "D3_Complete_Recode_Data_Sample_mayor2.txt"), row.names = FALSE)
```

C.2. Implementación del modelo en Stan

Un programa Stan se organiza en una secuencia de bloques, cuyos contenidos son declaraciones de variables. A continuación se describe cada componente del modelo para el código en Stan. (Luo & Jiao, 2018)

Bloque de datos

El bloque de datos es el primer componente necesario en el código, allí se especifica la información más relevante. Para este bloque, se debe reportar información relacionada con el número de respuestas, de docentes, de estudiantes, de ítems y de categorías para el ítem, y el número de cursos; asimismo, es necesario asignar los correspondientes identificadores de los datos. A continuación, se relaciona el código usado para el bloque de datos:

```
1  data{
2      int<lower=1>  N;                        // # de respuestas
3      int<lower=10> N_prof;                   // # de profesores
4      int<lower=10>  N_stud;                 //  # de estudiantes
5      int<lower=2>  N_item;                   // # de items
6      int<lower=2,upper=5> N_cat;             // # de categorias por item
7      int<lower=10> N_sub;                     // # de cursos
8      int<lower=1,upper=N_cat> y[N];          // Respuesta; y[n], n-ésima respuesta
9      int<lower=1,upper=N_prof> professor[N];// Profesor de la respuesta [n]
10     int<lower=1,upper=N_item> item[N];      // item de la respuesta [n]
11     int<lower=1,upper=N_stud> student[N];  // Estudiante de la respuesta [n]
12     int<lower=1,upper=N_sub> subject[N];  // Curso de la respuesta [n]
13     }
```

Bloque de parámetros

En este bloque se especifican los parámetros del modelo. Para el contexto de los modelos de TRI de múltiples facetas se deben especificar los parámetros de ítems, de habilidad y de severidad. Adicionalmente, para el modelo propuesto en el presente documento es necesario especificar el parámetro asociado al curso. Por otra parte, es necesario también especificar los hiper-parámetros de los parámetros descritos anteriormente. Las variables declaradas en este bloque corresponden a las variables que se deben muestrear. A continuación, se relaciona el código empleado:

```
14  parameters{
15    vector[N_prof] theta;          // Rasgo latente del profesor
16    vector[N_stud] gamma;          // Severidad del estudiante
17    vector[N_sub] eta;             // Parámetro del curso
18    ordered[N_cat-1] beta[N_item];// Parámetros de ítems
19    real<lower=0> sigma_eta;       // Desviación del parámetro eta
20    real<lower=0> sigma_gamma;     // Desviación del parámetro gamma
21    real<lower=0> sigma_beta;      // Desviación de los parámetros beta
22  }
```

Bloque de parámetros transformados

El bloque de parámetros transformados es opcional. Se usa cuando algunos parámetros especificados en el bloque de parámetros necesitan una transformación. En general, un parámetro transformado está limitado a ser función de algunos otros parámetros. En Stan es necesario separar los parámetros estimados libremente de los parámetros restringidos.

Bloque del modelo

Este bloque está relacionado con el modelo y es el más importante. Todos los parámetros relacionados ahí deben ser especificados en el bloque de parámetros. En este se define el modelo de probabilidad. A continuación se muestra el código implementado.

```
23  model{
24    theta ~ normal(0,1);               // Escala de trazos latentes
25    eta ~ normal(0,sigma_eta);         // Escala de parámetro del curso
26    gamma ~ normal(0,sigma_gamma);     // Escala de severidad de estudiantes
27    for(j in 1:N_item){
28      beta[j] ~ normal(0,sigma_beta);// Priori de los parámetros de los items
29    }
30    sigma_beta ~ cauchy(0,2);   // Hiper-prior para sigma_beta
31    sigma_gamma ~ cauchy(0,2);  // Hiper-prior para sigma_gamma
32    sigma_eta ~ cauchy(0,2);    // Hiper-prior para sigma_eta
33    for(n in 1:N){
34      y[n] ~ ordered_logistic(theta[professor[n]]-eta[subject[n]]
35             -gamma[student[n]], beta[item[n]]); //Función de probabilidad
36    }
37   }
```

Bloque de cantidades generadas

Este bloque es diferente a los demás bloques, pues este no afecta los valores de los parámetros muestreados. Si una cantidad no desempeña un papel en el modelo, debe definirse en el bloque de cantidades generadas. Entre los objetivos de este se encuentran: generar predicciones para nuevos datos, calcular las probabilidades de eventos posteriores y calcular las esperanzas posteriores. En este caso se calcula la log-verosimilitud asociada. A continuación se presenta el código implementado.

```
38  generated quantities{
39    vector[N] log_lik;  // Vector para extraer la log-verosimilitud
40      for(n in 1:N){
41        log_lik[n] = ordered_logistic_lpmf(y[n]|theta[professor[n]]
           -eta[subject[n]]-gamma[student[n]], beta[item[n]]);  // Función
42    }
43  }
```

C.3. Ajuste del modelo en RStan y cálculo de criterios de selección de modelos

Como se mencionó un programa Stan se puede ejecutar desde la interfaz de R mediante una librería se nombre RStan. A continuación, se presenta la implementación del código en R para realizar dicha ejecución.

El primer paso es realizar el cargue de las librerías necesarias para la correcta ejecución del programa, además de definir las rutas de trabajo. En este paso, es importante definir la opción de trabajo en paralelo a través de todos los núcleos del procesador del computador en uso,

```
   # Cargue de librerías
1  library(rstan)
2  rstan_options(auto_write = TRUE)
3  options(mc.cores = parallel::detectCores())
4  library(loo)
5  library(openxlsx)

   # Ubicación de archivos
6  inPath <- file.path("..", "input")
7  outPath <- file.path("..", "output")
8  scrPath <- file.path("..", "scr")
9  stanPath <- file.path("..", "scr", "Stan-codes")
```

El segundo paso es leer los datos en la estructura vectorizada, realizar la definición de los parámetros para la ejecución del modelo y definir una lista con esos parámetros como insumo para el objeto que requiere el modelo Stan.

```
   # Lectura de datos
10  professor_data <- read.csv(file.path(inPath,
"D11_professor_data_sample.txt"),header=TRUE,sep=";")

    # Definición de parámetros para modelo de Stan
11  NN <- nrow(professor_data) # de registros
12  NS <- length(unique(professor_data[,"Student"])) # de estudiantes
13  NP <- length(unique(professor_data[,"Professor"])) # de docentes
14  NI <- length(unique(professor_data[,"Item"])) # de ítems
15  NC <- max(professor_data[,"Y_prof"]) # de categorías de respuesta
16  NM <- length(unique(professor_data[,"Subject"])) # de cursos

    # Definición de lista para objeto Stan
17 dat <- list(student = professor_data[,"Student"],
           professor = professor_data[,"Professor"],
               y = professor_data[,"Y_prof"], item = professor_data[,"Item"],
               subject = professor_data[,"Subject"], N = NN, N_stud = NS,
```

```
            N_prof = NP, N_cat = NC, N_item = NI, N_sub = NM)
```

El cuarto paso es definir el objeto Stan para su compilación. Este se realiza a su vez en dos pasos: La primera compilación se realiza como prueba para verificar el ajuste del modelo y no es necesario ejecutar todas las iteraciones. La segunda compilación es en sí el ajuste del modelo y se realiza con los parámetros necesarios: límites de iteraciones y número de cadenas.

```
    # Primera compilación del modelo para verificar la escritura del modelo
18  prof_fit_p_ <- stan(file = file.path(stanPath,
            'S3_Multi_Faceted_Ciencias_professor_student_item.stan'),
                        data = dat, iter = 4, chains = 1,
                        control = list(adapt_delta = 0.99))

    # Ajuste del modelo usando el modelo compilado
19  prof_fit_2<- stan(fit = prof_fit_p_,  data = dat, iter = 3000,
                      chains = 4, control = list(max_treedepth = 12,
                      adapt_delta = 0.99))
```

Después del ajuste del modelo, se realiza la extracción de los parámetros estimados y se calculan los criterios LOO y WAIC.

```
    # Extracción de parámetros
20  prof_fit_summary1 <- summary(prof_fit_2)
21  prof_summary <- prof_fit_summary1$summary

    ## Guardar parámetros del modelo
22  params <- data.frame(prof_summary)
23  params[, "Parametro"] <- row.names(params)
24  row.names(params) <- NULL

    ## Escribir archivo de parámetros
25  write.xlsx(params, file.path(outPath, "resultadosParametros.xlsx"))

    ## Calculo de criterio LOO
26  log_lik <- extract_log_lik(prof_fit_2, merge_chains = FALSE)
27  rel_n_eff <- relative_eff(exp(log_lik))
28  loo1 <- loo(log_lik1, r_eff = rel_n_eff, cores = 2)

    ## Calculo de criterio WAIC
30   waic1 <- waic(loglik)
```

Bibliografía

Abrami, P. C., Perry, R. P. & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement., *Journal of Educational Psychology* **74**(1): 111.

Aleamoni, L. M. (1981). Student ratings of instruction, *in* J. Millman (ed.), *Handbook of Teacher Evaluation*, Beverly Hills, Calif.: Sage Publications, Beverly Hills, pp. 110–145.

Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika* **43**(4): 561–573.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G. & Lesaffre, E. (2019). Bayesian model selection in linear mixed models for longitudinal data, *Journal of Applied Statistics* pp. 1–24.

Bartholomew, D. J., Knott, M. & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, Vol. 904, John Wiley & Sons.

Basow, S. A. & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently?, *Journal of educational psychology* **79**(3): 308.

Becker, W. E. & Watts, M. (1999). How departments of economics evaluate teaching, *American Economic Review* **89**(2): 344–349.

Bélanger, C. H. & Longden, B. (2009). The effective teacher's characteristics as perceived by students, *Tertiary Education and Management* **15**(4): 323–340.

Box, G. E. (1980). Sampling and bayes' inference in scientific modelling and robustness, *Journal of the Royal Statistical Society: Series A (General)* **143**(4): 383–404.

Braga, M., Paccagnella, M. & Pellizzari, M. (2014). Evaluating students' evaluations of professors, *Economics of Education Review* **41**: 71–88.

Cameletti, M. & Caviezel, V. (2012). The cronbach-mesbah curve for assessing the unidimensionality of an item set: The r package cmc.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017). Stan: A probabilistic programming language, *Journal of statistical software* **76**(1).

Casella, G. & George, E. I. (1992). Explaining the gibbs sampler, *The American Statistician* **46**(3): 167–174.

Centra, J. A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness. The Jossey-Bass Higher and Adult Education Series.*, ERIC.

Centra, J. A. & Creech, F. R. (1976). The relationship between student teachers and course characteristics and student ratings of teacher effectieness, *Project Report*, Princeton, NJ, Educational Testing Service, pp. 76–1.

Chen, M.-H., Shao, Q.-M. & Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*, Springer Science & Business Media.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? a meta-analysis and review of the literature, *Journal of Marketing Education* **31**(1): 16–30.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies, *Review of educational Research* **51**(3): 281–309.

Cranton, P. A. & Smith, R. A. (1986). A new look at the effect of course characteristics on student ratings of instruction, *American Educational Research Journal* **23**(1): 117–128.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *psychometrika* **16**(3): 297–334.

de Andrade, D. F., Tavares, H. R. & da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações, *ABE, Sao Paulo* .

Eckes, T. (2011). Introduction to many-facet rasch measurement, *Frankfurt: Peter Lang* .

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis, *Research in Higher Education* **6**(3): 223–274.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't, *Research in Higher Education* **9**(3): 199–242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses, *Research in Higher Education* **10**(2): 149–172.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students, *Research in Higher Education* **18**(1): 3–124.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration, *Research in higher education* **26**(3): 227–298.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies, *Research in Higher education* **30**(6): 583–645.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part i—evidence from the social laboratory and experiments, *Research in Higher Education* **33**(3): 317–375.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*, Springer Science & Business Media.

Galbraith, C. S., Merrill, G. B. & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? a neural network and bayesian analyses, *Research in Higher Education* **53**(3): 353–374.

Gelfand, A. E., Dey, D. K. & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods, *Technical report*, STANFORD UNIV CA DEPT OF STATISTICS.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian data analysis*, Chapman and Hall/CRC.

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models, *Statistics and computing* **24**(6): 997–1016.

Gelman, A., Lee, D. & Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization, *Journal of Educational and Behavioral Statistics* **40**(5): 530–543.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statistica sinica* pp. 733–760.

Gelman, A., Rubin, D. B. et al. (1992). Inference from iterative simulation using multiple sequences, *Statistical science* **7**(4): 457–472.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on pattern analysis and machine intelligence* (6): 721–741.

Griewank, A. & Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*, Vol. 105, Siam.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*, Vol. 2, Sage.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hoffman, M. D. & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo., *Journal of Machine Learning Research* **15**(1): 1593–1623.

Jolliffe, I. T. (2002). Principal components in regression analysis, *Principal component analysis* pp. 167–198.

Koushki, P. A. & Kunh, H. A. J. (1982). How realiable are student evaluations of teachers?, *Engineering Education* **72**: 362–367.

Lord, F. M. & Novick, M. R. (2008). *Statistical theories of mental test scores*, IAP.

Luo, Y. & Jiao, H. (2018). Using the stan program for bayesian item response theory, *Educational and psychological measurement* **78**(3): 384–408.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research, *International journal of educational research* **11**(3): 253–388.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness, *The scholarship of teaching and learning in higher education: An evidence-based perspective*, Springer, pp. 319–383.

Martin, E. (1984). Power and authority in the classroom: Sexist stereotypes in teaching evaluations, *Signs: Journal of Women in Culture and Society* **9**(3): 482–492.

Masters, G. N. (1982). A rasch model for partial credit scoring, *Psychometrika* **47**(2): 149–174.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines, *The journal of chemical physics* **21**(6): 1087–1092.

Metropolis, N. & Ulam, S. (1949). The monte carlo method, *Journal of the American statistical association* **44**(247): 335–341.

Murray, H. G. (2005). Student evaluation of teaching: Has it made a difference, *Annual Meeting of the Society for Teaching and Learning in Higher Education. Charlottetown, Prince Edward Island*.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*, Department of Computer Science, University of Toronto Toronto, ON, Canada.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics, *Handbook of markov chain monte carlo* **2**(11): 2.

Ostini, R. & Nering, M. L. (2006). *Polytomous item response theory models*, number 144, Sage.

Perry, R. P., Niemi, R. R. & Jones, K. (1974). Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance., *Journal of Educational Psychology* **66**(6): 851.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms, *Markov chain Monte Carlo in practice* **57**: 45–58.

Small, A. C., Hollenbeck, A. R. & Haley, R. L. (1982). The effect of emotional state on student ratings of instructors, *Teaching of Psychology* **9**(4): 205–211.

Spencer, P. A. & Flyr, M. L. (1992). The formal evaluation as an impetus to classroom change: Myth or reality?.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the royal statistical society: Series b (statistical methodology)* **64**(4): 583–639.

Sta (2018). *Stan Modeling Language Users Guide and Reference Manual*. Version 2.18.0. **URL:** *http://mc-stan.org*

Stark, P. & Freishtat, R. (2014). An evaluation of course evaluations, *ScienceOpen Research* .

team, S. D. (n.d.). *Stan Modeling Language Users Guide and Reference Manual*, http://mc-stan.org.

Uttl, B., Eche, A., Fast, O., Mathison, B., Valladares Montemayor, H. & Raab, V. (2012). Student evaluation of instruction/teaching (sei/set) review, *Calgary, AB, Canada: Mount Royal Faculty Association Retrieved from: http://mrfa.net/files/MRFA_SEI_Review_v6. pdf* .

Uttl, B., White, C. A. & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related, *Studies in Educational Evaluation* **54**: 22–42.

Van Der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions, *Handbook of modern item response theory*, Springer, pp. 1–28.

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic, *Statistics and computing* **27**(5): 1413–1432.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review, *Assessment & Evaluation in Higher Education* **23**(2): 191–212.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**(Dec): 3571–3594.

Wolfe, E. W. & Dobria, L. (2008). Applications of the multifaceted rasch model, *Best practices in quantitative methods* pp. 71–85.