



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

ESTUDIO DE APLICABILIDAD DE MODELOS DE TEORÍA DE RESPUESTA AL
ÍTEM A LA PRUEBA DE SELECCIÓN UNIVERSITARIA DE MATEMÁTICAS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA
INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL MATEMÁTICA

JAVIERA FRANCISCA CASTILLO NAVARRO

PROFESOR GUÍA:
JAIME SAN MARTÍN ARISTEGUI

MIEMBROS DE LA COMISIÓN:
NANCY LACOURLY VENTRE
SALOMÉ MARTÍNEZ SALAZAR
MÓNICA SILVA RAVEAU
MARÍA LEONOR VARAS SCHEUCH

SANTIAGO DE CHILE
2017

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS
AL TÍTULO DE INGENIERA CIVIL MATEMÁTICA
POR: JAVIERA FRANCISCA CASTILLO NAVARRO
FECHA: 2017
PROF. GUÍA: SR. JAIME SAN MARTÍN ARISTEGUI

ESTUDIO DE APLICABILIDAD DE MODELOS DE TEORÍA DE RESPUESTA AL ÍTEM A LA PRUEBA DE SELECCIÓN UNIVERSITARIA DE MATEMÁTICAS

La construcción de preguntas para pruebas estandarizadas, como la Prueba de Selección Universitaria (PSU), supone un gran desafío y enfrentarlo requiere un análisis continuo por parte de instituciones especializadas, a través del uso de teorías de medición que constituyen el marco teórico para el diseño e implementación de estos instrumentos.

Actualmente, el análisis y desarrollo de la PSU se realiza a través de la denominada teoría clásica, con una creciente incorporación de la teoría de respuesta al ítem (IRT). Sin embargo, son conocidas las limitaciones de la teoría clásica y es deseable el estudio y la implementación de otros modelos que permitan efectuar análisis más completos de esta prueba. En particular, la implementación de un modelo avanzado permitiría obtener más información de las preguntas utilizadas y de las habilidades de los alumnos, además de generar preguntas enfocadas en habilidades específicas.

En este contexto se enmarca el presente trabajo de tesis, donde se realiza un estudio detallado y un análisis crítico de los modelos estándar de IRT que complementan los resultados de la teoría clásica. El objetivo es estudiar si se cumplen los supuestos para la aplicación de estos modelos.

Se aplican distintos modelos IRT unidimensionales a los datos de la PSU de matemática y se contrastan los resultados obtenidos. Además, se estudian extensiones multidimensionales de IRT, donde los supuestos necesarios son menos restrictivos. Los resultados obtenidos con estos modelos extendidos son comparados con los modelos originales.

Finalmente, en el caso de la PSU de matemática, la evidencia indica que no se cumplen los supuestos para poder aplicar los modelos IRT unidimensionales. Por otra parte, los modelos multidimensionales abordan aspectos que escapan del alcance de la teoría clásica y de los modelos IRT convencionales, por lo que su utilización podría entregar información significativa para comprender el comportamiento de grupos particulares de estudiantes o para la creación de preguntas que potencien habilidades específicas. Sin embargo, es necesario el estudio, comprensión e interpretación de los parámetros que definen a los modelos MIRT para complementar los análisis actuales de la PSU.

A mis padres.

Agradecimientos

Comienzo agradeciendo a mis papás, Gran Pa y Súper Mami, porque son una parte fundamental de mi vida. Soy muy afortunada de poder contar con ustedes siempre, incluso cuando nos encontrábamos a once mil kilómetros de distancia me hacían sentir como si estuviera en el segundo piso de la casa. Gracias por el amor, por los consejos, por las noches de tertulia, por la confianza, por transmitirme la curiosidad de aprender y por un sinfín de cosas que es imposible mencionar en un solo párrafo.

Gracias, Tito, por reírte de mis tonterías, por contarme chistes fomes y por ser el mejor hermano que podría haber tenido. Es gracias a ti (¿o por tu culpa?) que estoy aquí, porque tú me llevaste al lado oscuro ofreciéndome galletas. Gracias por la complicidad, por la confianza, por aconsejarme y permitirme aconsejarte. Y, sobretodo, gracias por extender nuestra familia con tres maravillosas personas.

Agradezco a la Emi, a la Romi, a Pesce, al Pato, a Rubb, a Povo y a Pelo, quienes fueron esenciales para sobrevivir a mis primeros años en la universidad. Gracias por convertirse en mi familia en Santiago, por el apoyo, las tardes de estudio, por todos los tacas que jugamos y los almuerzos que compartimos. Pese a que no nos vemos como antes, siguen ocupando el mismo lugar en mi pequeño corazón y se merecen mucho más que este lugar en mi tesis. Gracias también a la gente del dim, en especial a aquellos que hicieron que el regreso a Chile fuera menos traumático, gracias a Cabecitas, a Felipe Contreras, a Edgardo, al Seba por su amistad, compañía y apoyo, y, obviamente, a quienes han compartido conmigo todo este año y han sido los mejores compañeros de oficina, Garrido y Abner.

A Relaciones Internacionales Beauchef por darme la oportunidad de salir de mi burbuja y vivir dos años inolvidables. Agradezco también a la École Centrale Paris y Francia, por entregarme una formación completamente diferente. Bien sûr, gracias a los amigos que hicieron esta experiencia muchísimo más valiosa para mí: Chopán, Claudio, Seba, Joaquín, Kanokito, Karen, Gerardito, Steven, Sacha, Jean, Pierre y muchos otros. También gracias a la Vivi por el apoyo cuando estuve afuera y durante estos años de vuelta en la U.

Gracias al profesor Jaime San Martín, quien me ha guiado en el desarrollo de esta tesis. Gracias a Claudio Muñoz, por depositar en mí la confianza y permitirme realizar mis primeras clases auxiliares en otoño 2015. Gracias también a mi comisión, en especial a la profesora Nancy Lacourly, quien gentilmente ha respondido mis dudas durante el desarrollo de este trabajo y a la profesora Mónica Silva, por su dedicación y comentarios durante la revisión del trabajo. Gracias, en general, a los docentes y funcionarios del dim, que hacen del departamento una acogedora comunidad.

Y, por supuesto, gracias a Felipe, el Patato, parce que tu es le meilleur copain. Gracias por el cariño, por la paciencia, por el apoyo, por crecer juntos, por las aventuras compartidas

y las por compartir. Además, gracias a ti puedo contar con una nueva familia, gracias a tus papis, Héctor y Magaly, por recibirme con tanto cariño en su hogar.

Tabla de Contenido

Índice de Tablas	vii
Índice de Ilustraciones	viii
Introducción	1
1. Teoría de Respuesta al ítem.	4
1.1. Supuestos de IRT	5
1.2. Modelos IRT más usados	7
1.2.1. Modelo de Rasch o modelo logístico de un parámetro	7
1.2.2. Modelo logístico de dos parámetros	9
1.2.3. Modelo logístico de tres parámetros	10
1.3. Estimación de parámetros	11
1.3.1. Estimación de la habilidad, suponiendo conocidos los parámetros del ítem	12
1.3.2. Estimación conjunta de la habilidad y los parámetros del ítem	12
1.4. Software utilizado: El paquete ltm para R.	14
1.4.1. El Método de Máxima Verosimilitud Marginal	15
1.4.2. Estimación de las habilidades	16
1.5. Unidimensionalidad	17
1.5.1. Análisis Paralelo Modificado	18
2. Teoría de Respuesta al ítem multidimensional	22
2.1. Modelos MIRT	22
2.1.1. Modelos Compensatorios	23
2.1.2. Modelos no compensatorios.	26
2.2. Estimación de parámetros	27
2.2.1. Estimación del vector de habilidades θ suponiendo conocidos los parámetros del ítem	27
2.3. El paquete mirt para R.	28
3. Aplicación de la Teoría de Respuesta al Ítem a la PSU de matemática.	30
3.1. Muestra	30
3.2. Metodología	31
3.2.1. Primer enfoque	31
3.2.2. Segundo enfoque	31
3.3. Resultados: Análisis de los parámetros del Ítem.	32

3.3.1.	Aplicación del Modelo Rasch	32
3.3.2.	Aplicación del Modelo 2PL	35
3.3.3.	Comparación de resultados	37
3.4.	Resultados: Análisis de los parámetros de los individuos.	41
3.4.1.	Aplicación del Modelo Rasch	41
3.4.2.	Aplicación del Modelo 2PL	43
3.4.3.	Comparación de resultados	44
3.5.	Exploración de Unidimensionalidad	49
3.5.1.	Coefficiente Alfa de Cronbach	49
3.5.2.	Análisis de Componentes Principales	50
3.5.3.	Análisis Paralelo Modificado	53
3.6.	Aplicación de Modelos MIRT	56
3.7.	Comparación de Modelos	57
	Conclusión	60
	Bibliografía	64

Índice de Tablas

1.1. Valores referenciales de la capacidad de discriminación del ítem i , según a_i	10
1.2. Tabla de contingencia para x_i y x_j	20
3.1. Composición de la muestra fija de estudiantes, según clase.	32
3.2. Resumen de los resultados del análisis de los ítems a través del Modelo Rasch.	32
3.3. Dificultad promedio estimada por eje temático, modelo Rasch.	33
3.4. Discriminación estimada promedio, modelo Rasch.	35
3.5. Resumen de los resultados del análisis de la dificultad de los ítems a través del modelo 2PL.	35
3.6. Resumen de los resultados del análisis de la discriminación de los ítems a través del modelo 2PL.	36
3.7. Dificultad y discriminación promedio estimadas por eje temático, modelo 2PL.	36
3.8. Resumen de resultados del análisis de habilidades, modelo Rasch.	41
3.9. Resumen de resultados 2PL	43
3.10. Resumen Análisis de Componentes Principales.	50
3.11. Comparación de resultados por género, usando IRT y MIRT	56
3.12. Comparación de resultados por tipo de establecimiento educacional, usando IRT y MIRT	57

Índice de Ilustraciones

1.1. ICC del modelo Rasch para un ítem de dificultad $b_i = 0,3$	7
1.2. ICC del modelo Rasch para tres ítems con diferentes valores de dificultad b . .	8
1.3. ICC del modelo 2PL para tres ítems con diferentes valores de dificultad b y discriminación a	9
1.4. ICC del modelo 3PL para dos ítems con diferentes valores de dificultad b , discriminación a y guessing c	11
2.1. ICC de la extensión del modelo 2PL a dos dimensiones, donde se observa un ítem de características $a_1 = 0,5$, $a_2 = 1,5$ y $d = -0,7$	24
2.2. Curvas de nivel de la IRS para la extensión del modelo 2PL a dos dimensiones, donde se observa un ítem de características $a_1 = 0,5$, $a_2 = 1,5$ y $d = -0,7$	24
2.3. IRS de la extensión del modelo 3PL a dos dimensiones, donde se observa un ítem de características $a_1 = 0,5$, $a_2 = 1,5$, $c = 0,3$ y $d = -1$	26
3.1. Contraste dificultad y varianza, modelo Rasch.	34
3.2. Contraste dificultad y varianza, modelo 2PL.	37
3.3. Relación del plano dificultad-discriminación del modelo 2PL y la dificultad del modelo Rasch.	38
3.4. Relación entre las estimaciones de la dificultad: Rasch versus 2PL.	38
3.5. Diferencia entre estimaciones 2PL y Rasch para la dificultad.	39
3.6. Relación entre la proporción de respuestas correctas a los ítems y su dificultad.	40
3.7. Contraste Habilidad y varianza, modelo Rasch.	42
3.8. Contraste Habilidad y varianza, modelo 2PL.	44
3.9. Gráfico comparativo habilidad Rasch vs. habilidad 2PL.	45
3.10. Contraste entre estimaciones 2PL y Rasch para la habilidad.	46
3.11. Relación entre el número de respuestas correctas de cada alumno y la habilidad estimada, modelo Rasch.	47
3.12. Relación entre el número de respuestas correctas de cada alumno y la habilidad estimada, modelo 2PL.	47
3.13. Relación entre el puntaje PSU obtenido por cada alumno y la habilidad estimada, modelo Rasch.	48
3.14. Relación entre el puntaje PSU obtenido por cada alumno y la habilidad estimada, modelo 2PL.	49
3.15. ACP aplicado a la matriz de respuestas considerando todas las formas.	51
3.16. Resultados ACP sobre cada forma por separado.	52
3.17. Resultado de MPA, aplicado a cada forma de la prueba por separado.	55

Introducción

Las pruebas estandarizadas de medición en la educación, como la Prueba de Selección Universitaria (PSU) o el Sistema de Medición de la Calidad de la Educación (SIMCE) en Chile, requieren de profundos análisis y estudio continuo para su elaboración y para la construcción de repositorios de preguntas. Esta última labor ha de ser realizada a través de organizaciones especializadas en las teorías de medición, pues éstas constituyen el marco teórico para el diseño e implementación de este tipo de instrumentos, ya que permiten determinar propiedades de las preguntas que forman la evaluación, así como caracterizar a las personas que la rinden.

Es en este contexto que se desarrolla el presente trabajo de tesis, el cual corresponde a un análisis crítico de la aplicación de algunos modelos de teoría de medición a la PSU de matemática. Es por esto que para comprender su contenido es necesario introducir algunos conceptos comunes a toda teoría de medición. Cualquier instrumento de evaluación mide diferentes aspectos de los examinados: conocimientos en cierta área, razonamiento, análisis o comprensión de lectura, por ejemplo. En las teorías de medición educacional, se utiliza el término *habilidades* para identificar las características que son medidas en un test, así como cualidades de los individuos que lo rinden. Asimismo, existen parámetros que permiten caracterizar a las preguntas que forman la prueba. Entre los más utilizados, y recurrentes dentro de este trabajo, se encuentran la *dificultad* que suele medirse como la razón entre el número de personas que respondieron correctamente al ítem y el total de alumnos y que se interpreta como un indicador de qué tan complicada es la pregunta, y la *discriminación* que proporciona una medida de la capacidad de la pregunta para discernir entre los alumnos más y menos calificados.

Históricamente, la teoría de medición más utilizada en el mundo dentro del ámbito educacional es la llamada *Teoría Clásica*. De hecho, es ésta la que se aplica actualmente en el análisis y corrección de la PSU. En la teoría clásica, el indicador de la habilidad de un examinado corresponde básicamente al número de respuestas correctas que haya respondido en el test. Por otro lado, la dificultad de un ítem se mide a través del inverso del número de personas que respondieron correctamente dicho ítem (es decir, si pocas personas lograron acertar a la pregunta, significa que ésta es más difícil y viceversa), mientras que la discriminación se calcula como la correlación entre la respuesta al ítem y el puntaje total de la prueba.

Esta manera de caracterizar la habilidad de una persona y la dificultad y discriminación de un ítem puede ser poco conveniente. Dado que el índice de habilidad de un individuo se basa en el número de respuestas correctas que respondió, es un indicador extremadamente dependiente del grupo de preguntas que forman el test. Si el conjunto de preguntas que forman

el test es fácil, un alumno tendrá un puntaje mayor y, por ende, reflejará una habilidad mayor que si la prueba es difícil. De la misma manera, la dificultad y la discriminación dependen del grupo de personas que responden a los ítems. Una pregunta puede ser clasificada como muy difícil si el grupo de personas que la responden es muy poco hábil o como muy fácil si el grupo de personas es excepcionalmente hábil. Por esta razón, no es posible contrastar resultados de grupos de alumnos que rinden pruebas diferentes, así como tampoco se puede comparar tests que han sido aplicados sobre distintos grupos de personas.

Esta dependencia entre la habilidad de una persona con respecto a la prueba rendida, así como entre las características del ítem respecto a los individuos que lo responden, se denomina *dependencia circular* y es el aspecto más criticado de la teoría clásica. Por este motivo, surge la inquietud por desarrollar una teoría que permita relacionar la habilidad de los examinados con una medida independiente de cuál sea el test que se rinda y que caracterice a los ítems sin depender de las personas que los respondan. Una teoría alternativa que resuelve este problema es la *teoría de respuesta al ítem* o IRT (por sus siglas en inglés *Item Response Theory*).

La teoría de respuesta al ítem tiene un enfoque totalmente diferente a la teoría clásica. En ella se postula que un individuo puede responder correctamente a un ítem con cierta probabilidad, la que depende del grado de habilidad del individuo y de los parámetros que definan al ítem. Existen diversos modelos IRT, más o menos complejos, que se diferencian por qué función utilizan para modelar la probabilidad de respuesta correcta a un ítem y por la cantidad de parámetros que usan para caracterizar a los ítems. Hay modelos en los cuales se usa sólo el grado de dificultad, otros en que se considera la dificultad junto con la discriminación, mientras que los más complejos agregan también un tercer parámetro que incorpora la posibilidad de responder bien por mero azar.

Este enfoque permitiría eliminar el problema de la dependencia circular que se observa en la teoría clásica y es la principal ventaja que tiene IRT por sobre ésta. En este caso, un estudiante debiese tener siempre la misma estimación para la habilidad, sin importar cuáles preguntas respondió. Asimismo, la estimación de parámetros de los ítems debería ser invariante ante el grupo de personas que haya rendido el test. Lamentablemente, esta propiedad de *invarianza* sólo puede ser alcanzada si se satisface los estrictos supuestos inherentes a los modelos IRT, entre ellos, la unidimensionalidad, que es el hecho de considerar que la prueba mide sólo una habilidad.

La aplicación de la teoría de respuesta al ítem a la PSU de matemática permitiría realizar análisis más completos de la prueba, como aconseja el informe realizado en 2005 por ETS [21]. Por ejemplo, se podría comparar de manera efectiva resultados de años distintos, lo que no es evidente usando la teoría clásica. De hecho, en la actualidad los puntajes PSU se usan como si fueran realmente equivalentes en años consecutivos sin que lo sean, según el análisis realizado por los expertos de Pearson Education [10]. Sin embargo, determinar si esta prueba de selección cumple los supuestos para la aplicación de modelos IRT no es un proceso sencillo y es lo que inspira este trabajo de tesis.

Objetivos

Objetivo General

El objetivo de este trabajo es determinar si los modelos IRT unidimensionales son aplicables a la Prueba de Selección Universitaria (PSU) de matemática. En particular, se desea estudiar si las habilidades que mide esta prueba pueden ser representados mediante una sola dimensión, es decir, si se satisface el supuesto de unidimensionalidad inherente a los modelos de esta teoría.

Objetivos Específicos

- (i) Aplicar los modelos unidimensionales a la prueba de selección universitaria de matemática y analizar el comportamiento de éstos, en base a los resultados obtenidos.
- (ii) Comparar los modelos unidimensionales con su extensión multidimensional.

Organización del documento

En el capítulo 1 se presenta la teoría de respuesta al ítem, explicando sus fundamentos y los supuestos que requiere. Se presentan los modelos más usados dentro de esta teoría, así como los parámetros que éstos involucran. Además, se realiza una revisión de las formas de estimación de parámetros más utilizadas y se presenta el software con el que se llevará a cabo el estudio.

El capítulo 2 describe los modelos de la teoría de respuesta al ítem multidimensional, también denominada MIRT por sus siglas en inglés. Se hace referencia a los métodos de estimación de parámetros y al software que será utilizado para la aplicación de estos modelos a los datos.

Finalmente, en el capítulo 3 se expone la metodología utilizada y se presentan los resultados obtenidos de la aplicación de los modelos unidimensionales a la PSU de matemática. Se comparan los resultados obtenidos con distintos modelos IRT y se estudia la unidimensionalidad de los datos. Por último, se contrastan resultados de modelos unidimensionales con resultados de modelos multidimensionales.

Capítulo 1

Teoría de Respuesta al ítem.

Durante la década de 1930 surge una nueva teoría de medición de tests, actualmente conocida como *Teoría de respuesta al ítem* o *Item Response Theory*, sin embargo, no fue hasta la década de 1970 en que se convirtió en el tema dominante de investigación entre los especialistas del área. Esta teoría es un conjunto de modelos que permiten establecer una relación entre el *desempeño observable* de un individuo en un test y las *habilidades* medidas por dicho test, definiendo parámetros inherentes a los ítems que componen el examen y parámetros que caracterizan a las personas sometidas al test (conocidos como habilidades o rasgos latentes).

Su principal virtud es que, si se cumplen los supuestos del modelo, las estimaciones de los parámetros que caracterizan a los ítems y a los individuos son invariantes, en el sentido de que no dependen de la muestra. Ésta es también su mayor ventaja respecto a la teoría clásica [12].

En términos matemáticos, los modelos IRT se basan en la idea de describir la probabilidad de obtener una respuesta correcta en cierto ítem como una función de una variable que caracteriza a una persona (habilidad) y parámetros que caracterizan al ítem. Los modelos IRT más utilizados son: el modelo logístico de un parámetro (1PL) o modelo Rasch, el modelo logístico a dos parámetros (2PL) y el modelo logístico a tres parámetros (3PL).

Si se considera la habilidad como una variable continua θ , característica de cada individuo, un modelo IRT propone que la probabilidad de que una persona con habilidad θ responda correctamente al ítem i es de la forma:

$$\mathbb{P}_i(\theta) = f(\theta, \boldsymbol{\eta}_i) \tag{1.1}$$

Donde $\boldsymbol{\eta}_i$ es un vector de parámetros que caracterizan al ítem i y que cuantifican distintas características, como la discriminación, la dificultad y el azar de la pregunta. Dependiendo del modelo que escogido se utiliza sólo uno, sólo dos o los tres parámetros anteriores.

Cabe destacar que la decisión sobre cuál modelo utilizar es compleja y tiene consecuencias importantes. Por ejemplo, utilizar el modelo 1PL implica la suposición de que el grado de

dificultad es la única característica que influye en el desempeño de quienes rinden el test. Al usar un solo parámetro se asume que todos los ítems son igualmente discriminantes y que la posibilidad de adivinar la respuesta correcta a una pregunta ítem es nula. La elección del modelo debe realizarse considerando tanto sus ventajas como sus limitaciones.

1.1. Supuestos de IRT

Todo modelo matemático involucra supuestos sobre los datos a los que es aplicado. Los principales supuestos de los modelos IRT son los siguientes:

1. Unidimensionalidad

En cualquier teoría general que busca describir el comportamiento humano a través de rasgos latentes, también llamados habilidades, se asume que un conjunto de k habilidades influyen en el desempeño de un individuo en un test. Estos k rasgos latentes o habilidades definen un *espacio latente*. En el caso de IRT, el primer supuesto corresponde a la unidimensionalidad del espacio latente, es decir, que existe una única variable que influye en el rendimiento de una persona y que es la habilidad θ medida por el test.

Cabe destacar que para que este supuesto sea satisfecho, es necesario que exista una habilidad que sea dominante.

2. Independencia Local

Este segundo punto establece que las respuestas de un individuo a distintos ítems en un test son estadísticamente independientes, cuando las habilidades que inciden en el rendimiento se mantienen constantes.

Sea X_i la respuesta de cierto individuo al ítem i , con $i = 1, \dots, n$ y θ su habilidad, la condición de independencia local se expresa matemáticamente según la ecuación (1.2)

$$\mathbb{P}(X_1, X_2, \dots, X_n | \theta) = \mathbb{P}(X_1 | \theta) \mathbb{P}(X_2 | \theta) \dots \mathbb{P}(X_n | \theta) \quad (1.2)$$

$$= \prod_{i=1}^n \mathbb{P}(X_i | \theta) \quad (1.3)$$

La importancia de este supuesto radica en que permite expresar analíticamente la verosimilitud como producto de probabilidades. Esta propiedad potencia la utilización del estimador de máxima verosimilitud para obtener los parámetros del modelo.

La noción de independencia local puede parecer contraintuitiva. ¿Cómo es posible que las respuestas de una persona a un conjunto de preguntas de un test no tengan correlación entre sí? Cuando se tienen variables correlacionadas, significa que éstas tienen rasgos en común, pero cuando estos rasgos se mantienen constantes, desaparece la correlación. Es

en este sentido que funciona la independencia local, es decir, este supuesto implica que los rasgos que tienen en común las preguntas de un test son las habilidades que se miden en él.

3. Naturaleza de la función característica

Como ya ha sido mencionado, todo modelo IRT se basa en el supuesto de que la probabilidad de que un individuo responda correctamente a un ítem es una función que depende de las habilidades del individuo y de ciertos parámetros que describen al ítem. A esta función de probabilidad se le llama *Curva Característica del Ítem* o ICC (por sus siglas en inglés). En general, se considera una función de forma logística como en la ecuación (1.4).

$$f(x) = \frac{e^x}{1 + e^x} \quad (1.4)$$

En otros modelos se considera una función ojiva normal, pero se ha demostrado que agregar un factor $D = 1,702$ en los modelos logísticos minimiza la diferencia máxima entre la curva logística y la curva ojiva normal [4].

4. Rapidez

Un supuesto implícito dentro de los modelos IRT es que el test no fue administrado contra el tiempo. Es decir, si un individuo responde incorrectamente fue exclusivamente por carencia de habilidad y no por falta de tiempo.

En estricto rigor, este supuesto está implícito en el de unidimensionalidad, pues si no, la rapidez sería otra habilidad a considerar en los modelos.

5. Invarianza de los parámetros

Si bien la invarianza de los parámetros es considerada, en general, como una característica de los modelos IRT, algunos autores como De Ayala [7] lo mencionan en la lista de supuestos.

Este punto hace referencia a que, a diferencia de la teoría clásica, en los modelos IRT no hay dependencia circular entre los parámetros. Las estimaciones de la dificultad, de la discriminación y de la pseudo-adivinanza son independientes de la muestra de estudiantes sobre la que el ítem es aplicado, mientras que la habilidad de un estudiante es independiente del test que rinda.

1.2. Modelos IRT más usados

1.2.1. Modelo de Rasch o modelo logístico de un parámetro

Éste es uno de los modelos más usados dentro de IRT y debe su nombre a su principal desarrollador, Georg Rasch. Se caracteriza porque sólo considera un parámetro para definir cierto ítem i : su dificultad, denotada por b_i .

Este modelo propone que la probabilidad de que una persona con habilidad θ responda correctamente un ítem de dificultad b_i está dada por:

$$\mathbb{P}_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \quad (1.5)$$

El parámetro b_i de un ítem corresponde al punto de la escala de habilidad donde la probabilidad de responder correctamente es de 0.5. Mientras mayor sea el valor de b_i , mayor será la habilidad requerida por un individuo para tener la respuesta correcta, es decir, el ítem es más difícil.

En la figura 1.1 se observa gráficamente la relación entre θ y la probabilidad de que una persona en ese nivel de habilidad responda correctamente a un ítem con parámetro $b = 0, 3$. Esta representación se denomina *curva característica del ítem* o ICC (por sus siglas en inglés *Item characteristic curve*).

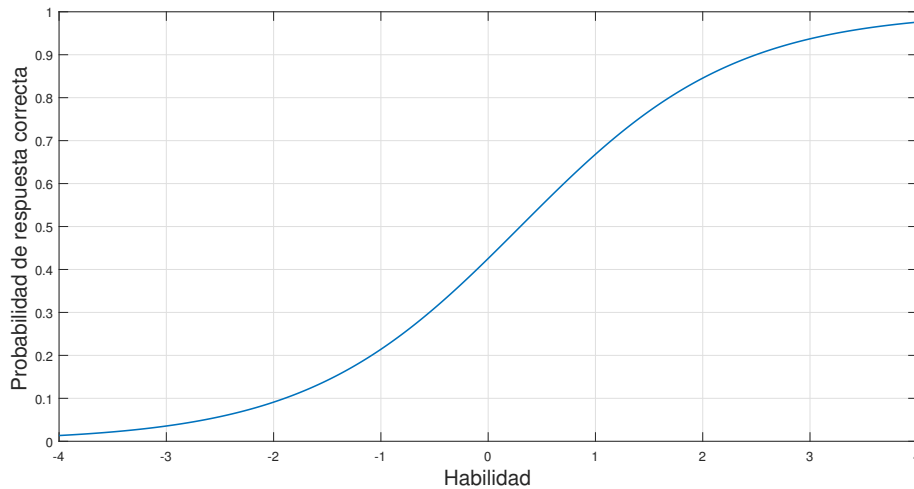


Figura 1.1: ICC del modelo Rasch para un ítem de dificultad $b_i = 0, 3$.

El gráfico anterior evidencia que la probabilidad de una respuesta correcta es una función estrictamente creciente de la habilidad. También se observa que el modelo tiene una cota inferior en 0 y una cota superior en 1.

Por otro lado, el parámetro b entrega información sobre qué personas tendrán mayor probabilidad de responder correctamente un ítem. Si $\theta > b$ la probabilidad de una respuesta correcta es mayor que 0,5, por el contrario, si $\theta < b$ entonces la probabilidad es menor que 0,5.

Se observa, además que en el valor $\theta = b$ (0,3 en este caso) corresponde al punto donde la derivada de la ICC alcanza su máximo. De esta manera, el parámetro b indica en qué punto de la escala de habilidad el ítem es más discriminador, en el sentido de que cerca de $\theta = b$ una pequeña diferencia de habilidad implica una mayor diferencia en la probabilidad de responder correctamente al ítem.

A continuación, en la figura 1.2, se presentan las curvas características de tres ítems diferentes y, por ende, con distintos valores de dificultad. En ella se observa que, al variar el valor de b , la curva simplemente se desplaza a la izquierda o a la derecha, sin variar su forma.

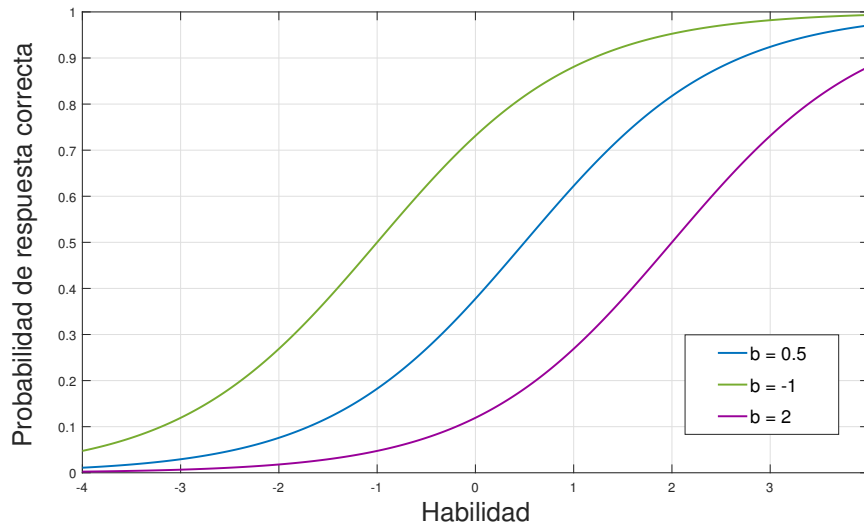


Figura 1.2: ICC del modelo Rasch para tres ítems con diferentes valores de dificultad b .

Una característica importante del modelo Rasch que se evidencia en el gráfico precedente (figura 1.2) es que, si bien el máximo de la derivada de la curva se alcanza en $\theta = b_i$ y b_i es diferente para cada ítem, el valor de este máximo es común para todos los ítems. Por esto, se dice que el modelo tiene el supuesto implícito de que todos los ítems tienen la misma discriminación.

Una propiedad útil del modelo de Rasch, es que el puntaje total definido por $r = \sum_{i=1}^n x_i$, con n el número de preguntas en el test, es un estadístico suficiente para la habilidad θ [16]. Esto significa que el mismo puntaje total representa el mismo nivel de habilidad, lo que permitiría simplificar las estimaciones de las habilidades [11].

1.2.2. Modelo logístico de dos parámetros

Planteado en 1968 por Allan Birnbaum [2], este modelo propone que la probabilidad de que una persona con habilidad θ responda correctamente al ítem i está dada por la ecuación:

$$\mathbb{P}_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \quad (1.6)$$

Donde b_i representa la dificultad del ítem i y a_i es el parámetro de *discriminación* del ítem. Este último parámetro es proporcional a la pendiente de la curva característica del ítem en el punto b_i de la escala de habilidad. Mientras que D es una constante introducida en el modelo para obtener valores semejantes a los de otros modelos que utilizan la función ojiva normal.

En la figura 1.3 se muestra gráficamente la curva característica de tres ítems con distintos parámetros para este modelo.

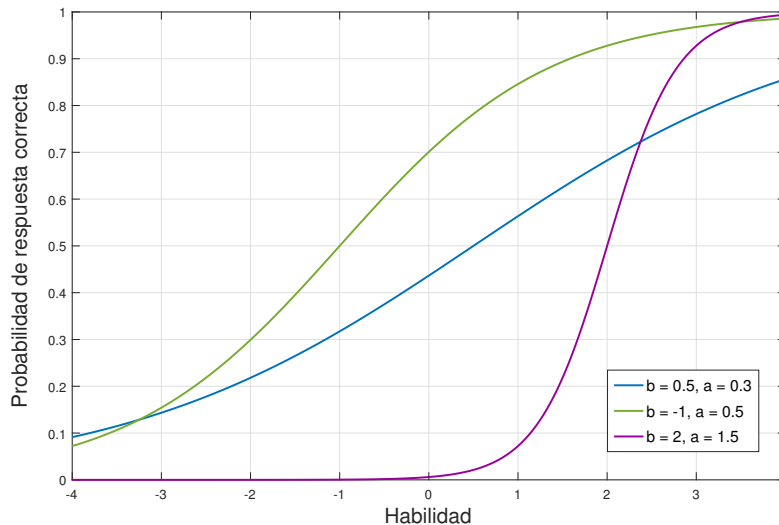


Figura 1.3: ICC del modelo 2PL para tres ítems con diferentes valores de dificultad b y discriminación a .

Al igual que en el modelo Rasch, el parámetro b permite identificar qué valores de habilidad tienen alta probabilidad de responder correctamente al ítem. Si $\theta > b$, entonces la probabilidad de una respuesta correcta es mayor que 0,5 y viceversa.

Además, el punto $\theta = b$ sigue siendo el valor donde la derivada de la curva característica alcanza su valor máximo, la diferencia es que este valor ya no es constante para todos los ítems, sino que depende del parámetro a de cada ítem. Así, mientras mayor es el valor de a_i , más pronunciada es la pendiente de la curva característica en el punto $\theta = b_i$ y, por lo tanto, el ítem es más discriminador.

Todo lo anterior se observa en la figura 1.3 donde, a diferencia del caso anterior, las ICC no sólo se desplazan horizontalmente, también se intersectan y varían su forma gracias a la

introducción del parámetro de discriminación. Sin embargo, las asíntotas horizontales de la curva continúan siendo las mismas: 1 hacia $+\infty$ y 0 hacia $-\infty$.

Cabe destacar que es deseable que un ítem tenga un alto poder de discriminación, pues permite distinguir mejor a los individuos que poseen el nivel de habilidad suficiente para responder correctamente de los que no. En la tabla 1.1 se exhibe los valores referenciales de discriminación propuestos por Baker en [1].

Tabla 1.1: Valores referenciales de la capacidad de discriminación del ítem i , según a_i .

Capacidad de discriminación	Rango de valores de a_i
Ninguna	0
Muy baja	0,01 - 0,34
Baja	0,35 - 0,64
Moderada	0,65 - 1,34
Alta	1,35 - 1,69
Muy alta	$>1,70$
Perfecta	$+\infty$

1.2.3. Modelo logístico de tres parámetros

Similar a los dos modelos anteriores, e introducida por Frederic M. Lord en 1980, la forma matemática de la curva logística de tres parámetros está dada por:

$$\mathbb{P}_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (1.7)$$

Donde a_i , b_i y D representan la discriminación, la dificultad y el factor de escalamiento, al igual que en el modelo a dos parámetros, y tienen las mismas interpretaciones que en éste.

En este caso se agrega un nuevo parámetro para describir al ítem: el *pseudo-azar*, también conocido como guessing o adivinanza. Con él se pretende integrar un aspecto no menor (y altamente observado) en las interacciones entre ítems y personas: el hecho de que un individuo pueda inferir o incluso seleccionar al azar la respuesta correcta a un ítem, sobretodo en preguntas de selección múltiple. Esto explicaría que una persona con baja habilidad pueda responder a un ítem correctamente, aun si éste posee una dificultad muy alta.

En la figura 1.4 se observa las curvas características de dos ítems con distintos parámetros para este modelo.

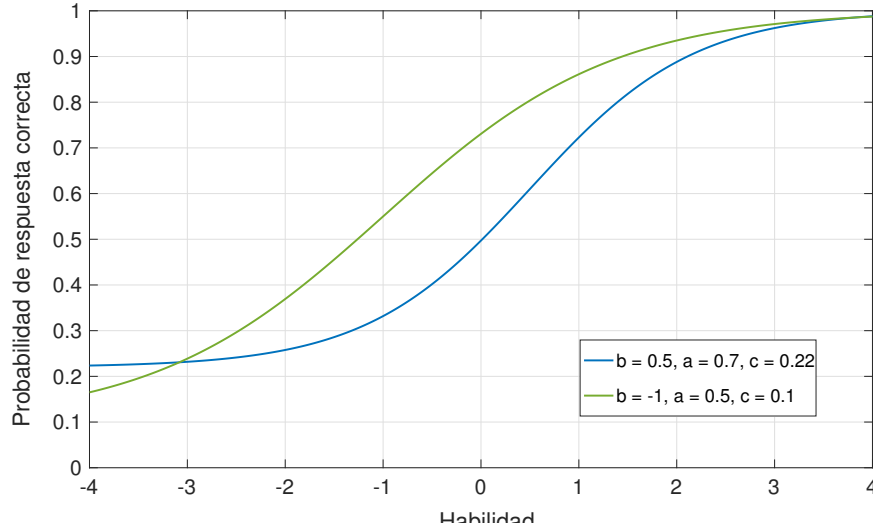


Figura 1.4: ICC del modelo 3PL para dos ítems con diferentes valores de dificultad b , discriminación a y guessing c .

Como es posible observar en el gráfico precedente (figura 1.4), la diferencia entre el modelo de tres parámetros y los dos anteriores es que, con la integración del pseudo-azar, la asíntota horizontal hacia $-\infty$ es justamente el valor c_i . De esta manera, c_i corresponde a la probabilidad con que una persona con habilidad infinitamente baja responda correctamente al ítem i .

1.3. Estimación de parámetros

Uno de los pasos más importantes, y a la vez complejos, en la aplicación de IRT a los datos de un test es la estimación de los parámetros del modelo.

La probabilidad de una respuesta correcta a cierto ítem depende de la habilidad θ de la persona sometida al test y de los parámetros que caracterizan al ítem. Sin embargo, tanto la habilidad como los parámetros del ítem son desconocidos, por lo que es necesario determinarlos a partir de la información disponible: las respuestas de todos los individuos que rindieron el test. De esta manera, determinar los parámetros de los modelos es un problema de estimación estadística.

Uno de los métodos más usados se basa en el estimador de máxima verosimilitud [11], sin embargo, a veces son realizadas estimaciones bayesianas.

Si el vector de respuestas de cierto individuo a las n preguntas de un test está dado por $u = (u_1, u_2, \dots, u_n)$, donde $u_i \in \{0, 1\}$, $\forall i = 1, \dots, n$ ($u_i = 1$ representa una respuesta correcta al ítem i y $u_i = 0$ una respuesta errónea). Gracias a la independencia local, la probabilidad de obtener este vector de respuestas se expresa como

$$\mathbb{P}(U = u|\theta) = \prod_{i=1}^n \mathbb{P}_i(\theta)^{u_i} (1 - \mathbb{P}_i(\theta))^{1-u_i}. \quad (1.8)$$

Asimismo, si se conocen los vectores de respuestas de N individuos a las n preguntas de un test, la probabilidad de obtener estos vectores de respuestas está dada por

$$\mathbb{P}(u_1, u_2, \dots, u_N | \theta_1, \theta_2, \dots, \theta_N) = \prod_{j=1}^N \prod_{i=1}^n \mathbb{P}_i(\theta_j)^{u_{ij}} (1 - \mathbb{P}_i(\theta_j))^{1-u_{ij}}. \quad (1.9)$$

La probabilidad expresada en la ecuación (1.9) se conoce como *función de verosimilitud* y se denota por

$$L(\theta_1, \dots, \theta_N, \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n) = \mathbb{P}(u_1, \dots, u_N | \theta_1, \dots, \theta_N). \quad (1.10)$$

Donde $\boldsymbol{\eta}_i$ representa los parámetros que describen al ítem i .

Los estimadores de máxima verosimilitud de los parámetros se obtienen de maximizar la función $L(\theta_1, \dots, \theta_N, \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n)$ en (1.10).

1.3.1. Estimación de la habilidad, suponiendo conocidos los parámetros del ítem

Si los parámetros del ítem son conocidos, la estimación de la habilidad se convierte en un problema relativamente sencillo. El método más utilizado se basa en el de máxima verosimilitud (MLE), es decir, se maximiza la función de verosimilitud en (1.10). Sin embargo, como se conocen los parámetros de los ítems, la función de verosimilitud depende sólo de $\theta_1, \dots, \theta_N$, es decir:

$$L(\theta_1, \dots, \theta_N, \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n) = L(\theta_1, \dots, \theta_N). \quad (1.11)$$

La ventaja de usar MLE es que los estimadores de habilidad obtenidos de esta forma son consistentes, eficientes y asintóticamente distribuyen como una variable normal [11].

Cabe destacar que cuando existen puntajes perfectos, puntajes nulos o puntajes patológicos (por ejemplo, si una persona responde ítems relativamente difíciles y discriminantes correctamente y falla los ítems más fáciles) el estimador de máxima verosimilitud falla. Esto puede evitarse removiendo los casos conflictivos de la muestra antes de realizar la estimación, sin embargo, no se sabe cuáles son las consecuencias que esto tiene sobre las propiedades del estimador. Cuando esto ocurre, suele utilizarse la estimación bayesiana.

1.3.2. Estimación conjunta de la habilidad y los parámetros del ítem

Lo más común es que al aplicar algún modelo IRT para analizar los datos de un test tanto los parámetros de los ítems como las habilidades de los individuos sean desconocidos.

Cuando éstos deben ser estimados simultáneamente surge un problema difícil y los estimadores, en general, no tienen buenas propiedades. A pesar de esto, existen métodos que permiten realizar estas estimaciones.

Antes de revisar brevemente los métodos más usados, es conveniente efectuar una pequeña reflexión sobre la cantidad de datos que deberán ser estimados: cuando N individuos se someten a un test de n ítems el número de habilidades θ_a que deben ser estimados es N , uno por cada sujeto. Mientras que el número de parámetros de los ítems depende del modelo IRT elegido y pueden ser n , $2n$ ó $3n$.

El problema es que esta gran cantidad de parámetros ($N + n$, $N + 2n$ y $N + 3n$ según el modelo) son todos inobservables, lo que conlleva cierto grado de indeterminación del modelo (también conocido como el problema de identificación).

Por ejemplo, en el modelo Rasch, la transformación $\theta^* = \theta + k$ y $b_i^* = b_i + k$, deja la función característica invariante, es decir:

$$\mathbb{P}_i(\theta^* | b_i^*) = \mathbb{P}_i(\theta | b_i)$$

Así, el origen está indeterminado. Para evitarlo, es necesario escalar los θ (o los parámetros b) para que su media esté fija (convenientemente, en cero). Teniendo fijo el origen, quedan $N + n - 1$ parámetros por determinar.

Similarmente, una transformación lineal de los parámetros en 2PL y 3PL deja las respectivas funciones características invariantes. Para evitar la indeterminación, lo usual es fijar las habilidades θ de manera tal que su media sea 0 y su varianza sea 1. Con estas restricciones, restan respectivamente $N + 2n - 2$ y $N + 3n - 2$ parámetros a determinar.

Los métodos más usados para la estimación de estos parámetros son:

1. Estimación conjunta de máxima verosimilitud

Es el método más utilizado de todos y está bien implementado computacionalmente.

Para realizar esta estimación se considera un método iterativo: se realiza una estimación inicial para los parámetros de los ítems y luego se estima la habilidad por máxima verosimilitud, suponiendo estos parámetros como conocidos. Luego, con esta primera estimación de las habilidades, se estiman los parámetros de los ítems por máxima verosimilitud y con estos resultados se vuelven a estimar las habilidades. El proceso continúa hasta que los valores convergen.

Debido a que se estiman los dos tipos de parámetros de manera conjunta, no se tienen las propiedades usuales de máxima verosimilitud, sin embargo, existe evidencia empírica de que los estimadores son consistentes.

2. Estimación de máxima verosimilitud condicional

Este método sólo es válido para el modelo Rasch (1PL), pues aprovecha la existencia de un estadístico suficiente. Como el puntaje observado, r , es estadístico suficiente para el parámetro θ , es posible expresar la función de verosimilitud $L(u|\theta_a, b_i)$ en términos de r en lugar de θ . Así, se puede obtener el estimador de máxima verosimilitud de la dificultad del ítem, b_i , sin necesidad de los parámetros de habilidad.

Este estimador tiene las propiedades deseadas de un estimador de máxima verosimilitud (listadas anteriormente), sin embargo, su implementación numérica parece ser compleja.

3. Estimación de máxima verosimilitud marginal

Estimar los parámetros estructurales (inherentes al ítem) puede realizarse si la función de verosimilitud no depende de la habilidad. Esto puede lograrse al integrar dicha función con respecto a la habilidad, si se supone ésta como un parámetro continuo, o sumando, si se considera como un parámetro discreto.

Los detalles de este método se encuentran en la sección 1.4.1.

4. Estimación bayesiana

Una solución alternativa a la estimación por verosimilitud es la estimación bayesiana.

Una de sus ventajas es que la estimación es directa y no se necesita imponer restricciones a los parámetros, ya que la información a priori controla la indeterminación antes explicada.

5. Procesos de estimación aproximada

Sobretudo para el modelo 3PL, donde los métodos antes expuestos pueden ser realmente costosos y lentos, es posible obtener estimadores que aproximen los estimadores de máxima verosimilitud.

A pesar de que no conserven las buenas propiedades de los estimadores de máxima verosimilitud, tienen grandes ventajas computacionales.

1.4. Software utilizado: El paquete `ltm` para R.

R es un lenguaje de programación para estadística computacional y gráficos. Posee ya implementada una gran variedad de técnicas estadísticas (modelos lineales y no lineales, tests clásicos, análisis de series de tiempo, clasificación, entre otros) y gráficos; pero lo que lo hace aún más interesante, es la posibilidad de extender las aplicaciones de R a través de paquetes.

Esto último hace de R un lenguaje muy versátil, adaptable a las necesidades de los usuarios: ellos mismos pueden crear sus propios paquetes (o librerías), los que quedarán almacenados en el repositorio de CRAN. CRAN es una red de ftp y servidores web en todo el mundo que recopilan versiones actualizadas de código y documentación para R.

Existen diversas librerías para el análisis de datos a través del enfoque de IRT, que difieren entre sí por los modelos que poseen implementados, así como en los métodos que utilizan para realizar las estimaciones.

En particular, este documento presentará el paquete `ltm` de R, cuyo objetivo es el análisis de datos dicotómicos y politómicos a través de IRT. Entre los modelos implementados incluye:

el modelo Rasch, el modelo logístico de dos parámetros, el modelo de Birnbaum de tres parámetros y modelos de respuesta graduada [20].

Para realizar un análisis usando IRT, es necesario estimar los parámetros propios del modelo utilizado, así como las habilidades de los individuos que se someten al test. Entre los métodos por Máxima Verosimilitud más utilizados para realizar estas estimaciones están: máxima verosimilitud conjunta, máxima verosimilitud condicional y máxima verosimilitud marginal.

El paquete ltm realiza las estimaciones de los parámetros utilizando este último (MMLE por sus siglas en inglés *Marginal Maximum Likelihood Estimation*).

A continuación se revisan los detalles de este método.

1.4.1. El Método de Máxima Verosimilitud Marginal

Como se mencionó anteriormente, cuando N individuos se someten a un test de n ítems el número de habilidades θ_j que deben ser estimadas es N , uno por cada sujeto. Mientras que el número de parámetros de los ítems depende del modelo IRT elegido y pueden ser n , $2n$ ó $3n$. Así el número total de parámetros a estimar es $N + n$, $N + 2n$ ó $N + 3n$ según el modelo con el que se esté trabajando. Además, si $N \gg n$ (es decir, si el número de personas que rinden el test es muy grande), el número de parámetros a estimar es gigante.

Por esto, resulta muy útil utilizar algún método que permita realizar las estimaciones en dos etapas: (i) la estimación de los parámetros de los ítems, a través de una función de verosimilitud modificada que considere las habilidades de manera implícita y (ii) la estimación de las habilidades [15].

Una manera de lograrlo es la que propone el Método de Máxima Verosimilitud Marginal.

Esta técnica asume que las habilidades son muestras de una variable aleatoria con función de distribución $F(\cdot)$ (la que podría ser discreta o continua) .

Bajo el supuesto de independencia condicional, la probabilidad conjunta del vector de respuestas \mathbf{x}_j del individuo j condicional a su habilidad θ_j es

$$L(\theta_j|\mathbf{x}_j, \boldsymbol{\phi}) := \mathbb{P}(\mathbf{x}_j|\theta_j, \boldsymbol{\phi}) = \prod_{i=1}^n \mathbb{P}(X_{ij} = x_{ij}|\theta_j, \boldsymbol{\phi}_i) \quad (1.12)$$

donde $\boldsymbol{\phi}_i$ es el vector de los parámetros del ítem i y $x_{ij} \in \{0, 1\}$ (respuesta incorrecta o correcta).

Integrando sobre la distribución de las habilidades, se tiene que

$$\mathbb{P}(\mathbf{x}_j|\boldsymbol{\phi}) = \int_{\Theta} L(\theta_j|\mathbf{x}_j, \boldsymbol{\phi}) dF(\theta_j) \quad (1.13)$$

Y, finalmente, la función de Verosimilitud Marginal para el vector de parámetros de los ítems ϕ , $L(\phi|\mathbf{X})$, se obtiene como el producto de las probabilidades de cada individuo

$$L(\phi|\mathbf{X}) = \prod_{j=1}^N \mathbb{P}(\mathbf{x}_j|\phi) \quad (1.14)$$

$$= \prod_{j=1}^N \int_{\Theta} \prod_{i=1}^n \mathbb{P}(X_{ij} = x_{ij}|\theta_j, \phi_i) dF(\theta_j) \quad (1.15)$$

Maximizando esta función sobre los parámetros de los ítems, ϕ se puede obtener las estimaciones de éstos. Sin embargo, típicamente se prefiere maximizar la función $\log L(\phi|\mathbf{X})$ dada por:

$$\log L(\phi|\mathbf{X}) = \sum_{j=1}^N \log \left(\int_{\Theta} \prod_{i=1}^n \mathbb{P}(X_{ij} = x_{ij}|\theta_j, \phi_i) dF(\theta_j) \right) \quad (1.16)$$

Observaciones:

1. La función de distribución de las habilidades F ahora forma parte del modelo IRT elegido y hay que ser cuidadoso al momento de elegir su forma paramétrica. Comúnmente se asume que F es una distribución normal con media 0 y varianza 1, pero ésta podría no ser adecuada para todas las aplicaciones.
2. Para resolver el problema de identificación, es posible poner restricciones sobre la misma función F , en lugar de hacerlo sobre los parámetros de los ítems.
3. Un inconveniente de este procedimiento es que además de necesitar métodos numéricos para realizar la optimización de la verosimilitud $L(\phi|\mathbf{X})$, se requiere también de técnicas de integración numéricas para calcular la integral de la ecuación (1.13).

1.4.2. Estimación de las habilidades

Una vez que se han calculado los estimadores de los parámetros de los ítems, la función `factor.scores` ofrece dos posibles métodos para estimar las habilidades:

1. *Maximum a posteriori*: Se calcula θ tal que se maximiza la función

$$p(\theta_j|\mathbf{x}_j) = p(\theta_j|\mathbf{x}_j, \phi_j). \quad (1.17)$$

2. *Expected a posteriori*: Se calcula θ que maximiza

$$\int \theta_j p(\theta_j|\mathbf{x}_j, \phi_j) d\theta. \quad (1.18)$$

1.5. Unidimensionalidad

Antes de continuar, se definen algunos conceptos que serán utilizados en esta sección.

Definición 1.1 (Unidimensionalidad) *La unidimensionalidad se define como la existencia de un rasgo latente subyacente a todos los ítems.*

Definición 1.2 (Homogeneidad) *La homogeneidad se refiere a la similitud de las correlaciones entre ítems. Un test perfectamente homogéneo es aquel donde todos sus ítems se inter-correlacionan de igual manera, es decir, todos los ítems miden los constructos igualmente.*

Cabe destacar que algunos autores, como Lord, Novick y McDonald, han usado el término homogeneidad como sinónimo de unidimensionalidad.

Definición 1.3 (Confiabilidad) *La confiabilidad de un conjunto de ítems se define como el cociente entre la varianza del puntaje real y la varianza del puntaje observado. En esencia, la confiabilidad de un instrumento es el grado de congruencia con la que mide la habilidad que se desea evaluar.*

Uno de los supuestos más críticos de la teoría de respuesta al ítem es el de la unidimensionalidad. Sin embargo, se trata de una condición difícil de satisfacer, sobretudo en pruebas con gran cantidad de preguntas, como es el caso de la PSU, pues *todos* los ítems deben medir la *misma* habilidad, aptitud u otra variable psicológica.

Por ejemplo, en una prueba donde se desea evaluar los conocimientos matemáticos de un grupo de personas, es deseable que el nivel de comprensión lectora de los individuos sometidos al test no afecte su rendimiento. Si una persona extremadamente hábil en matemática no logra comprender el enunciado de una pregunta y falla, significa que la pregunta no sólo evaluaba conocimientos matemáticos, también requería cierto grado de otra habilidad: la comprensión de lectura.

A pesar de su importancia, aún no se ha definido un índice efectivo que mida la unidimensionalidad de un test y que sea globalmente aceptado. Es más, durante las décadas de 1940 y 1950, el término unidimensionalidad era usado indistintamente de otras características de un test como la homogeneidad y la confiabilidad, por lo que algunos de los índices más conocidos y utilizados miden en realidad alguna otra característica y no necesariamente unidimensionalidad.

Hattie [13] hace una compilación de los índices más utilizados y el razonamiento detrás de ellos. En su trabajo, clasifica los distintos índices en cinco categorías: (1) índices basados en el patrón de respuestas, (2) índices basados en la confiabilidad, (3) índices basados en componentes principales, (4) índices basados en análisis factorial y (5) índices basados en modelos IRT. No obstante, concluye que la mayoría de los índices existentes no tienen una base teórica que los legitime, pues existen muy pocos estudios que comparen los distintos índices y se han realizado muy pocas pruebas de los índices usando datos sintéticos cuya dimensión sea conocida, además sólo cuatro de los más de ochenta índices estudiados logran distinguir consistentemente un conjunto de ítems unidimensional de uno con más de una

dimensión.

Pese a su cuestionada efectividad, ya que en realidad es un indicador de la confiabilidad de un test, el índice históricamente más utilizado para verificar la unidimensionalidad de un conjunto de ítems es el *alfa de Cronbach* [6], definido como:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_{Y_i}^2}{\sigma_X^2} \right) \quad (1.19)$$

donde k es el número de ítems que componen el test, σ_X^2 es la varianza del puntaje total y $\sigma_{Y_i}^2$ es la varianza del i -ésimo ítem.

La lógica detrás de este indicador es que para que un test sea fiable, sus ítems deben tener alta correlación entre sí, pues deberían medir los mismos constructos. El nivel máximo de correlación se alcanza cuando todos los ítems son iguales, en cuyo caso $\alpha = 1$. Por el contrario, si todos los ítems son independientes $\alpha = 0$.

Otro método utilizado para estudiar la unidimensionalidad de una prueba es la realización de un análisis de componentes principales (ACP) [14], de ésta. Este método reposa fuertemente en el criterio de la persona que realiza el ACP, puesto que la prueba será considerada suficientemente unidimensional si los resultados indican que un porcentaje considerable de la varianza está explicado por la primera componente principal, en comparación con las demás.

El paquete `ltm` tiene implementada una función `unidimtest()` que determina, a partir de la matriz de respuestas y el análisis IRT de ella, si el test es *suficientemente unidimensional* para que los resultados obtenidos del análisis sean significativos. El método se basa en una modificación del análisis paralelo (*modified parallel analysis* en inglés), presentada por Drasgow y Lissak en 1983 [8], cuyo objetivo es determinar cuándo las violaciones al supuesto de unidimensionalidad son tan severas que no permiten estimar los parámetros satisfactoriamente.

1.5.1. Análisis Paralelo Modificado

En 1968, Lord y Novick [17] mostraron que, bajo ciertas condiciones, la unidimensionalidad de un conjunto de ítems es equivalente a que la matriz de correlaciones tetracóricas (ver definición 1.4) de éstos tenga un solo factor común, que corresponde a θ . Sin embargo, al trabajar con datos reales, difícilmente es posible encontrar sólo un factor común al realizar análisis de correlaciones de los ítems. Afortunadamente, Drasgow y Parsons (1983) [9] descubrieron, a través de diversas simulaciones, que no es necesario un espacio latente estrictamente unidimensional para una estimación satisfactoria de los parámetros de los modelos IRT, siempre y cuando exista un solo factor dominante y la dificultad de un ítem no sea confundida con dimensionalidad.

El análisis paralelo modificado (MPA) es un método para examinar la estructura del espacio latente de un conjunto de ítems dicotómicos. De esta manera, proporciona una medida

de qué tan poco unidimensionales son los datos para determinar si las estimaciones proporcionadas por el modelo IRT elegido son apropiadas.

La metodología desarrollada por Drasgow y Lissak tiene su base en el análisis factorial y consiste en una modificación del análisis paralelo (en inglés, *parallel analysis*), técnica utilizada para determinar el número de componentes a conservar al realizar un análisis de componentes principales o para determinar la cantidad de factores a retener luego de aplicar un análisis factorial.

El método consta de las siguientes etapas:

- (i) Se cuenta con la matriz de respuestas a ítems dicotómicos, de tamaño $N \times n$ y a valores en $\{0, 1\}$, donde 0 representa una respuesta incorrecta y 1, una respuesta correcta. N corresponde a la cantidad de personas que respondieron el test, mientras que n es el número de ítems que lo componen.
- (ii) Se calculan las correlaciones tetracóricas entre los ítems, generando una matriz de $n \times n$.
- (iii) Se realiza un análisis factorial sobre la matriz de correlaciones y se calculan sus valores propios.
- (iv) Se generan datos sintéticos unidimensionales (el conjunto de datos paralelo). Para ello, se selecciona un modelo IRT y se aplica a los datos reales, obteniendo así las estimaciones de las habilidades de los individuos y de los parámetros de los ítems. Utilizando los parámetros, las habilidades estimadas y un generador de números aleatorios se simulan respuestas dicotómicas que son unidimensionales.
- (v) Se repite los pasos (i), (ii) y (iii) sobre los datos sintéticos generados en (iv).
- (vi) Finalmente, se comparan los valores propios obtenidos a partir de los datos reales y de los datos sintéticos. Se considera que existe multidimensionalidad cuando el segundo valor propio obtenido de los datos reales difiere considerablemente del segundo valor propio obtenido a partir de los datos sintéticos.

Junto con la presentación de este procedimiento, Drasgow y Lissak, en su publicación [8], presentan cinco estudios empíricos donde se observa los buenos resultados del método.

No obstante, esta técnica no es completamente rigurosa. Sólo compara el segundo par de valores propios entre las dos matrices, ignorando completamente otros valores. La elección de este estadístico no tiene justificación teórica ni empírica según Budescu [3]. En su trabajo, Budescu critica también otros aspectos de MPA y propone otro método, denominado Análisis Paralelo Modificado Revisado (RMPA por sus siglas en inglés, *Revised Modified Parallel Analysis*), que solucionaría los aspectos técnicos criticados de MPA y agrega una segunda etapa que permite extraer un subconjunto de ítems que sí resulta ser unidimensional, de un conjunto multidimensional más grande.

Si bien RMPA es una técnica más avanzada que MPA, no será utilizada en este trabajo, ya que el objetivo no es la obtención de un subconjunto de preguntas de la PSU que sea unidimensional, sino simplemente determinar si la prueba es suficientemente unidimensional para

la aplicación de los modelos IRT convencionales. En este sentido y pese a sus defectos, MPA es una buena herramienta para distinguir entre tests unidimensionales y multidimensionales.

Definición 1.4 (Correlaciones Tetracóricas) *La correlación tetracórica entre dos variables dicotómicas, x_i y x_j , es la estimación de la correlación entre dos variables continuas, x_i^* y x_j^* , que han sido dicotomizadas según la ecuación (1.20)*

$$x_k = \begin{cases} 1, & \text{si } x_k^* > t_k, \\ 0, & \text{si } x_k^* \leq t_k. \end{cases} \quad \text{con } k = i, j. \quad (1.20)$$

Para estimar estas correlaciones se supone que la distribución conjunta de x_i^* y x_j^* es normal bivariada.

Consideremos la tabla de contingencia en 1.2. De ella se tiene que la frecuencia relativa observada de que un par de realizaciones de las variables x_i y x_j caiga en el casillero kl , con $k, l \in \{0, 1\}$, está dada por $P_{kl} = \frac{n_{kl}}{N}$, donde N es corresponde al total de datos.

Se denota por π_{kl} a las frecuencias relativas esperadas, dadas por la distribución de (x_i^*, x_j^*) y los umbrales t_i y t_j .

Tabla 1.2: Tabla de contingencia para x_i y x_j .

x_j	1	0
x_i	1	0
1	n_{11}	n_{10}
0	n_{01}	n_{00}

El objetivo es encontrar una distribución normal multivariada para las variables (x_i^*, x_j^*) y los valores de t_i y t_j que definen a las variables dicotómicas.

La distribución y los umbrales se calculan de forma que las frecuencias esperadas π_{kl} se aproximen lo más posible a las frecuencias observadas P_{kl} . Además, se supone que la distribución conjunta de (x_i^*, x_j^*) está dada por

$$(x_i^*, x_j^*) \sim \mathcal{N} \left(\hat{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \quad (1.21)$$

Dado que la media y la varianza de esta distribución está fija, el único parámetro a estimar es ρ que corresponde precisamente a la correlación tetracórica entre x_i y x_j .

La densidad del vector (x_i^*, x_j^*) está dada por

$$\phi(x_i^*, x_j^*; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left(-\frac{1}{2(1-\rho^2)} (x_i^{*2} - 2\rho x_i^* x_j^* + x_j^{*2}) \right) \quad (1.22)$$

Luego, la verosimilitud del problema, es decir, la probabilidad de obtener las frecuencias observadas dados los umbrales y la correlación ρ , está dada por

$$L(n_{00}, n_{10}, n_{01}, n_{11}; t_i, t_j, \rho) = C \prod_{k=0}^1 \prod_{l=0}^1 \pi_{kl}^{n_{kl}} \quad (1.23)$$

Donde C es una constante. Por simplicidad, se trabaja con la log-verosimilitud, que se presenta en la siguiente ecuación

$$\ell = \ln(L) = \ln(C) \sum_{k=0}^1 \sum_{l=0}^1 n_{kl} \pi_{kl} \quad (1.24)$$

$$= \ln(C) + n_{00}\pi_{00} + n_{01}\pi_{01} + n_{10}\pi_{10} + n_{11}\pi_{11} \quad (1.25)$$

La expresión exacta de las frecuencias π_{kl} puede ser calculada, gracias al supuesto sobre la distribución de (x_i^*, x_j^*) , obteniendo

$$\begin{aligned} \pi_{00} &= \int_{-\infty}^{t_i} \int_{-\infty}^{t_j} \phi(x, y; \rho) dx dy & \pi_{11} &= \int_{t_i}^{\infty} \int_{t_j}^{\infty} \phi(x, y; \rho) dx dy \\ \pi_{10} &= \int_{t_i}^{\infty} \int_{-\infty}^{t_j} \phi(x, y; \rho) dx dy & \pi_{01} &= \int_{-\infty}^{t_i} \int_{t_j}^{\infty} \phi(x, y; \rho) dx dy \end{aligned}$$

Luego, para obtener la estimación de la correlación tetracórica ρ y los umbrales t_i y t_j basta calcular el estimador de máxima verosimilitud, es decir, los valores de ρ , t_i y t_j que maximicen ℓ . Se debe resolver el problema de optimización

$$\min_{t_i, t_j, 0 \leq \rho \leq 1} \ell(t_i, t_j; \rho) \quad (1.26)$$

Obteniendo finalmente la correlación tetracórica [18].

Capítulo 2

Teoría de Respuesta al ítem multidimensional

La teoría de respuesta al ítem multidimensional (MIRT) o *Multidimensional Item Response Theory* surge como un conjunto de ideas de áreas como la psicología, educación, desarrollo de tests, psicometría y estadística.

En particular, nace de la necesidad de un modelo que represente mejor la realidad, pues en IRT el supuesto de unidimensionalidad es muy restrictivo: ¿es posible crear un test con ítems que midan exclusivamente una (y sólo una) habilidad?. La respuesta lógica es no (o, al menos, es extremadamente difícil), sin embargo, las herramientas computacionales y de análisis estadístico no eran suficientemente eficaces para desarrollar estos modelos hasta hace unos años. Con el progreso de éstas, los modelos multidimensionales de teoría de respuesta al ítem han podido comenzar a estudiarse en profundidad.

2.1. Modelos MIRT

En general, existen dos tipos de modelos MIRT: los modelos compensatorios y los modelos no compensatorios. Su principal diferencia es la manera en que combinan la información del vector $\boldsymbol{\theta}$ de habilidades con las características del ítem para obtener la probabilidad de responder correctamente al ítem [19].

En el caso de los modelos compensatorios, se utiliza una combinación lineal de las coordenadas del vector $\boldsymbol{\theta}$, de manera tal que el mismo valor de la suma puede ser obtenido con distintos vectores $\boldsymbol{\theta}$. Así, si cierta coordenada θ_k tiene un valor muy bajo, esto puede ser *compensado* por otra coordenada θ_ℓ de valor más alto. Dicho de otra forma, permite que la alta habilidad en una dimensión compense la baja habilidad en otras dimensiones. De allí el nombre de estos modelos.

Los modelos no compensatorios, por otro lado, separan el ítem en partes, según la habilidad correspondiente en cada tarea cognitiva, y se aplica un modelo unidimensional en cada parte.

Así, la probabilidad de una respuesta correcta es el producto de las probabilidades de abordar correctamente cada parte. En este caso no existe compensación entre habilidades, por lo tanto, una alta probabilidad de respuesta correcta implica una alta habilidad en cada una de las dimensiones consideradas.

A continuación, se describen ambos tipos de modelos, sin embargo, en el trabajo posterior sólo se aplicará modelos de tipo compensatorio, puesto que éstos son consistentes con una visión más holística de la interacción entre personas y preguntas en un test. Cuando un individuo responde un test trata de utilizar todas sus habilidades y conocimientos disponibles para enfrentar todos los aspectos de cada pregunta, por lo que desde este punto de vista podría resultar imposible disgregar todo el conjunto de habilidades. Además, existen estudios que indican que los modelos compensatorios suelen ajustarse mejor a los datos que los modelos no compensatorios [19].

2.1.1. Modelos Compensatorios

1. Extensión multidimensional del modelo logístico de dos parámetros.

En el modelo logístico de dos parámetros (ver 1.2.2) la función logística está evaluada en un término de la forma $a(\theta - b)$, que se puede reescribir como $a\theta - ab$. Definiendo el término $d := -ab$, la extensión natural del modelo unidimensional resulta de evaluar la función logística en el término $\mathbf{a}^t\boldsymbol{\theta} + d$, donde $\mathbf{a} \in \mathbb{R}^m$ es el vector que contiene la información sobre la discriminación del ítem, $\boldsymbol{\theta} \in \mathbb{R}^m$ es el vector de habilidades de la persona, m es el número de dimensiones del espacio latente y $d \in \mathbb{R}$.

Así, la probabilidad de que la persona j responda al ítem i correctamente, según este modelo, está dada por:

$$\mathbb{P}(X_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{e^{\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i}}. \quad (2.1)$$

La expansión del exponente de e en la ecuación (2.1) refleja la manera en que los elementos de \mathbf{a} y los del vector $\boldsymbol{\theta}$ interactúan:

$$\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \dots + a_{im}\theta_{jm} + d_i \quad (2.2)$$

$$= \sum_{\ell=1}^m a_{i\ell}\theta_{j\ell} + d_i. \quad (2.3)$$

Este exponente es una función lineal de los elementos de $\boldsymbol{\theta}$ con el parámetro d como término de intersección y los elementos del vector \mathbf{a} como los parámetros de “pendiente”. La expresión en la ecuación (2.2) define una línea en un espacio de dimensión m y es una interesante propiedad de este modelo. Si el exponente se fija a un valor constante k , los vectores $\boldsymbol{\theta}$ que satisfacen la relación $k = \mathbf{a}_i^t \boldsymbol{\theta}_j + d_i$ forman una línea recta y tienen la misma probabilidad de respuesta correcta.

En las siguientes imágenes se presenta la forma gráfica de este modelo. En la figura 2.1 se representa la probabilidad de respuesta correcta al ítem como la altura sobre el plano (θ_1, θ_2) . De esta manera se obtiene la *Superficie de respuesta al ítem* o IRS, por sus siglas en inglés, el equivalente en varias dimensiones a la ICC.

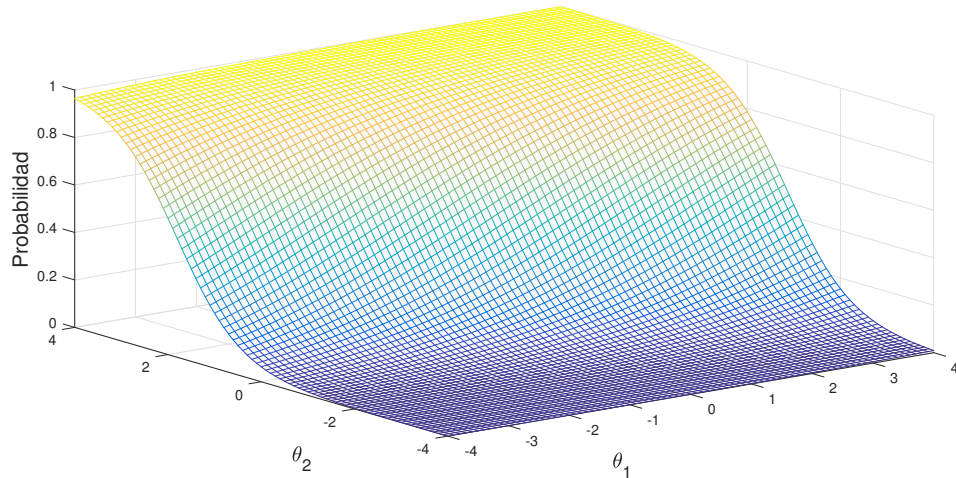


Figura 2.1: ICC de la extensión del modelo 2PL a dos dimensiones, donde se observa un ítem de características $a_1 = 0,5$, $a_2 = 1,5$ y $d=-0.7$.

Por otro lado, en la figura 2.2 se observa la probabilidad como curvas de nivel de la superficie anterior (figura 2.1). Cada recta representa todas las combinaciones de (θ_1, θ_2) que tendrán la misma probabilidad de respuesta correcta al ítem.

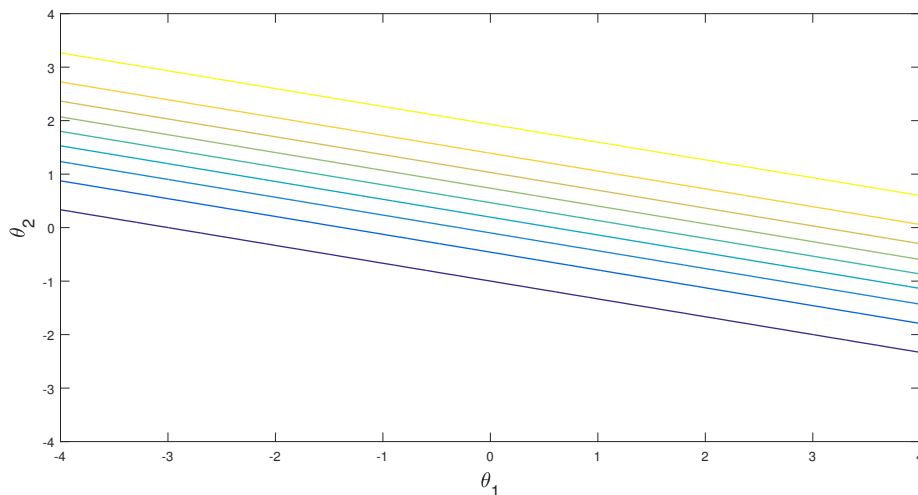


Figura 2.2: Curvas de nivel de la IRS para la extensión del modelo 2PL a dos dimensiones, donde se observa un ítem de características $a_1 = 0,5$, $a_2 = 1,5$ y $d=-0.7$.

2. Extensión multidimensional del modelo Rasch.

Para el caso del modelo Rasch, parece lógico realizar una extensión análoga a la del modelo anterior, pero considerando un vector \mathbf{a} tal que todas sus coordenadas sean iguales. Sin embargo, esto no resulta útil, pues al escribir el exponente queda

$$\mathbf{a}^t \boldsymbol{\theta} + d = \sum_{k=1}^m a_k \theta_k + d \quad (2.4)$$

Dado que se tiene $a_k = a^*$, $\forall k \in \{1, \dots, m\}$ lo anterior se escribe como

$$a^* \left(\sum_{k=1}^m \theta_k \right) + d \quad (2.5)$$

Definiendo $\theta^* := \sum_{k=1}^m \theta_k$, se observa que esta propuesta de modelo multidimensional se reduce simplemente al modelo Rasch unidimensional, que considera la habilidad como la suma de las habilidades en las distintas dimensiones.

Para realizar una extensión útil del modelo Rasch, se considera que la probabilidad de que la persona j responda correctamente al ítem i está modelada por

$$\mathbb{P}(X_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{e^{\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i}}. \quad (2.6)$$

La diferencia con el modelo multidimensional de dos parámetros en (2.1) es que en éste el vector de características del ítem, \mathbf{a}_i , es estimado a partir de los datos; mientras que en la extensión del modelo Rasch, \mathbf{a}_i debe ser fijado por el desarrollador del test y es un vector constante para todos los ítems.

De esta manera, los parámetros a estimar a partir de los datos son: el vector de habilidades $\boldsymbol{\theta}_j$ de cada individuo j y el coeficiente d_i para cada ítem i . Una ventaja de extender el modelo de esta forma es que se heredan propiedades del modelo Rasch unidimensional, como la existencia de un estadístico suficiente para el parámetro d_i .

Un inconveniente de este modelo es que la bondad de ajuste del modelo a los datos dependerá de la elección de los valores del vector \mathbf{a} .

3. Extensión multidimensional del modelo logístico de tres parámetros.

De manera análoga a la extensión de 2PL, se extiende el modelo 3PL (ver 1.2.3) a su versión multidimensional, que permite tener una asíntota inferior distinta de cero para el modelo. Esto permite modelar la posibilidad de que un alumno con bajas habilidades pueda dar una respuesta correcta al ítem. Así, este modelo se escribe

$$\mathbb{P}(X_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, c_i, d_i) = c_i + (1 - c_i) \frac{e^{\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i^t \boldsymbol{\theta}_j + d_i}}, \quad (2.7)$$

c_i es entonces el parámetro de pseudo-azar, que especifica la probabilidad de respuesta correcta para alumnos con baja habilidad.

A continuación, en la figura 2.3 se presenta la IRS de un ítem, según este modelo. Al igual que en los modelos unidimensionales, la diferencia respecto a la IRS de la extensión 2PL (figura 2.1) es que la cota inferior para la probabilidad de respuesta correcta al ítem i corresponde al valor de c_i en lugar de ser 0.

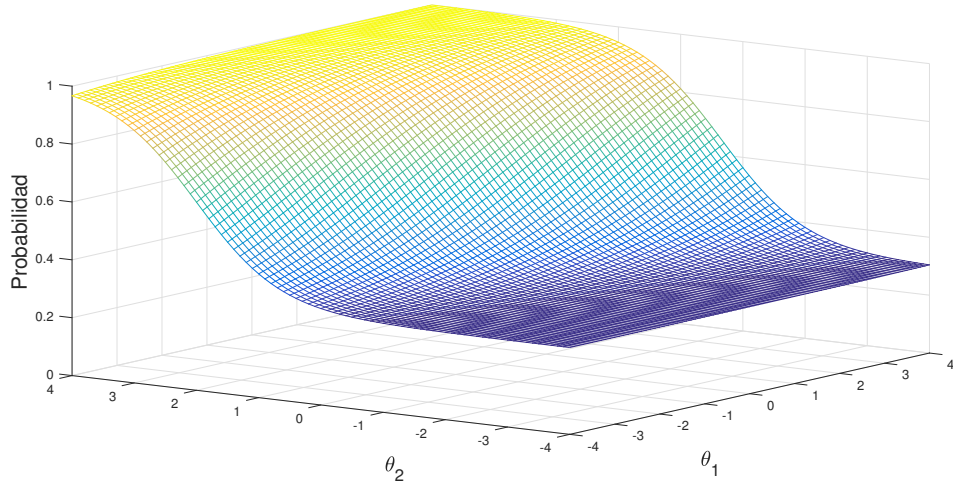


Figura 2.3: IRS de la extensión del modelo 3PL a dos dimensiones, donde se observa un ítem de características $a_1 = 0,5$, $a_2 = 1,5$, $c = 0,3$ y $d = -1$.

2.1.2. Modelos no compensatorios.

Estos modelos surgen para representar situaciones donde los modelos compensatorios no son suficientes. Por ejemplo, se cuenta con una pregunta de matemáticas que requiere fuertemente dos habilidades: comprensión de lectura (pues el enunciado debe ser representado matemáticamente) y habilidades aritméticas (para poder resolver y encontrar el resultado). Si un alumno tiene extraordinarias habilidades aritméticas, pero no logra traducir el texto a la buena expresión matemática a resolver, no podrá responder correctamente al ítem. Su habilidad aritmética no puede compensar su falta de comprensión de lectura.

En este caso, un modelo compensatorio no modelará adecuadamente la situación, pero un modelo que no considere compensaciones tan fuertes entre las habilidades sí puede resultar útil.

La forma general en que estos modelos representan la probabilidad de un individuo de responder correctamente cierto ítem está dada por

$$\mathbb{P}(X_{ij} = 1 | \theta_j, \phi_i) = \prod_{k=1}^m p_k \quad (2.8)$$

donde p_k es la probabilidad de que el individuo j aborde correctamente la dimensión k -ésima del ítem i .

A modo de ejemplo, se presenta a continuación la extensión del modelo Rasch.

1. Extensión no compensatoria del modelo Rasch.

$$\mathbb{P}(X_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{b}_i) = \prod_{k=1}^m \frac{e^{(\theta_{jk} - b_{ik})}}{1 + e^{(\theta_{jk} - b_{ik})}} \quad (2.9)$$

donde θ_{jk} representa la habilidad del individuo j en la dimensión k y b_{ik} corresponde a la dificultad del ítem i según la dimensión k .

2.2. Estimación de parámetros

Al igual que en el capítulo anterior con los modelos unidimensionales, es necesario estimar los parámetros de los ítems y las habilidades de las personas, siendo difícil determinarlos independientemente.

Por otro lado, la cantidad de parámetros a estimar es de orden superior. A modo de ejemplo, para el caso del modelo compensatorio de dos parámetros se requiere estimar $n(m + 1) + Nm$ parámetros, donde m son las dimensiones del espacio latente considerado, N es el número de personas sometidas al test y n es el número de ítems en el test.

Asimismo, existen indeterminaciones en los modelos (similares a las mencionadas en el caso unidimensional) en cuanto al origen del espacio, las unidades de medida de las coordenadas, entre otras cosas que deben solucionarse al momento de implementar la estimación numéricamente.

2.2.1. Estimación del vector de habilidades $\boldsymbol{\theta}$ suponiendo conocidos los parámetros del ítem

Si los parámetros que caracterizan a los ítems ya han sido determinados, resta sólo estimar los vectores de habilidades de las personas que rinden el test. Principalmente son tres los métodos más utilizados para realizar esta estimación:

1. Máxima Verosimilitud

Análogo al caso unidimensional. Para cada individuo j , el estimador máximo verosímil para el vector $\boldsymbol{\theta}_j$ es el que maximiza la probabilidad del vector de respuestas del individuo.

2. Estimación Bayesiana

Este estimador encuentra su base en el teorema de Bayes, de donde se obtiene que

$$h(\boldsymbol{\theta}|\mathbf{X}_j) = \frac{L(\mathbf{X}_j|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int_{\Theta} L(\mathbf{X}_j|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \quad (2.10)$$

donde $f(\boldsymbol{\theta})$ es la densidad de probabilidad a priori de $\boldsymbol{\theta}$, \mathbf{X}_j es el vector de respuestas del individuo j , $L(\mathbf{X}_j|\boldsymbol{\theta})$ es la probabilidad del vector de respuestas, dada la habilidad (corresponde a la función de verosimilitud que se maximiza en el método anterior) y $h(\boldsymbol{\theta}|\mathbf{X}_j)$ es la densidad de probabilidad a posteriori de $\boldsymbol{\theta}$ dado el vector de respuestas.

Como el denominador de la expresión anterior es constante para $\boldsymbol{\theta}$, se tiene que

$$h(\boldsymbol{\theta}|\mathbf{X}_j) \propto L(\mathbf{X}_j|\boldsymbol{\theta})f(\boldsymbol{\theta}) \quad (2.11)$$

Con esto, los métodos de estimación bayesiana maximizan el término a la derecha de la expresión en (2.11). Cabe destacar que este enfoque requiere especificar la distribución $f(\cdot)$ antes de realizar la estimación.

Este método es especialmente útil cuando el estimador máximo verosímil no es finito, pues la hipótesis sobre la distribución a priori permite regular esto.

3. Mínimos Cuadrados

Los métodos mencionados anteriormente son similares en el sentido de que ambos maximizan cierta probabilidad. El método de mínimos cuadrados, por su parte, minimiza la diferencia al cuadrado del valor observado y el valor esperado. Notando que para el caso dicotómico $\mathbb{E}(X|\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{P}(x = 1|\boldsymbol{\theta}, \boldsymbol{\phi})$, la expresión a minimizar para obtener este estimador es

$$SS_{\boldsymbol{\theta}} = \sum_{i=1}^n (x_i - \mathbb{P}(x_i = 1|\boldsymbol{\theta}, \boldsymbol{\phi}))^2, \quad (2.12)$$

donde $SS_{\boldsymbol{\theta}}$ es la suma de las diferencias al cuadrado para cierto $\boldsymbol{\theta}$ y x_i es la respuesta al ítem i . Cabe mencionar que éste es el criterio menos utilizado de los tres.

2.3. El paquete mirt para R.

Al igual que para la teoría de respuesta al ítem unidimensional, existen diversos softwares que permiten analizar datos desde la perspectiva de la teoría de respuesta al ítem multidimensional, entre ellos, TESTFACT, NOHARM, ConQuest, BMIRT y paquetes de R como MCMCpack y mirt. La diferencia entre ellos radica en el método empleado para realizar las estimaciones, así como en la implementación de éstos. En particular, en este trabajo se utilizará el paquete `mirt` de R que permite el análisis de datos, tanto dicotómicos como politómicos, usando IRT unidimensional y multidimensional.

Este paquete permite un análisis de acuerdo a dos enfoques: el enfoque exploratorio y el enfoque confirmatorio. El análisis exploratorio se utiliza cuando no hay una hipótesis clara sobre la estructura de los datos. El análisis confirmatorio, por otro lado, requiere una hipótesis precisa para la estructura de los datos, es decir, requiere hipótesis sobre el número de dimensiones en el espacio latente.

Las estimaciones son realizadas mediante métodos de cuadratura, como el algoritmo de *Expectation Maximization*) o métodos estocásticos, como el algoritmo *Metropolis-Hastings Robbins-Monro*) [5].

Capítulo 3

Aplicación de la Teoría de Respuesta al Ítem a la PSU de matemática.

3.1. Muestra

La Prueba de Selección Universitaria (PSU) es el instrumento de medición utilizado en Chile desde el año 2004 para la elección de candidatos a las universidades.

Se compone de cuatro pruebas: Lenguaje y Comunicación; Matemática; Ciencias e Historia, Geografía y Ciencias Sociales. Las dos primeras son de carácter obligatorio, mientras que es suficiente rendir una de las últimas dos pruebas para participar del proceso. Son exámenes de selección múltiple, con una sola opción de respuesta correcta, por lo que pueden ser tratadas como variables dicotómicas. Las preguntas han sido construidas por el Departamento de Evaluación, Medición y Registro Educativo (DEMRE), en base a los objetivos fundamentales y a los contenidos mínimos obligatorios de Enseñanza Media, declarados en la Actualización Curricular 2009.

Para la realización de este trabajo, se consideran los datos de la prueba de matemática del proceso de admisión 2016, rendida el día 30 de noviembre de 2015 y de la que participaron 252.745 estudiantes.

Los contenidos que se evalúan en este instrumento están agrupados en cuatro ejes temáticos: (i) Números, (ii) Álgebra, (iii) Geometría y (iv) Datos y azar. Asimismo, los diseñadores de las pruebas pretenden que éstas midan las habilidades cognitivas que los estudiantes han desarrollado durante su enseñanza básica y media, éstas se agrupan en tres: (i) Comprensión, (ii) Aplicación y (iii) Análisis, Síntesis y Evaluación.

Para la elaboración de este examen, se genera un banco de 120 preguntas en total, con éstas se constituye 4 formas de prueba diferentes, de 80 preguntas cada una. Además, cada una de estas preguntas puede ser clasificada según el eje temático que mide y la habilidad cognitiva que requiere.

Cabe mencionar que de las 80 preguntas que componen una forma de la prueba, 5 de ellas son preguntas piloto que no son consideradas al realizar el análisis de los resultados de los estudiantes. De esta manera, el puntaje máximo que puede obtener un alumno es de 75 respuestas correctas.

3.2. Metodología

El objetivo principal de este trabajo es analizar la aplicabilidad de los modelos IRT unidimensionales a la PSU de matemática, ante la posibilidad de que la prueba no satisfaga el supuesto de unidimensionalidad.

Para ello, se utilizan muestras de los datos de la PSU de matemáticas del proceso de admisión 2016, los cuales son analizados a través de los modelos Rasch y 2PL y, finalmente, se estudian los resultados obtenidos. Hay dos focos: analizar el comportamiento de los modelos al estimar los parámetros de los ítems y el análisis del comportamiento de los modelos para las estimaciones de habilidades.

3.2.1. Primer enfoque

La idea es estudiar el desempeño de los modelos IRT, en particular del modelo Rasch y 2PL, en la estimación de los parámetros que describen a los ítems. Para ello, del universo total de datos, se genera 1.000 muestras de manera aleatoria, cada una de 2.000 estudiantes, y con cada muestra se realiza un análisis IRT usando el modelo Rasch y el modelo 2PL.

De esta manera, es posible analizar la robustez de los modelos, en el sentido de observar la dependencia (o independenciam) de las estimaciones de éstos con respecto a la muestra de estudiantes, entre otras características.

3.2.2. Segundo enfoque

Similar al caso anterior, el objetivo es estudiar el desempeño de los modelos Rasch y 2PL, salvo que esta vez el énfasis es observar el comportamiento al estimar los parámetros que describen a los examinados, es decir, la estimación de las habilidades.

Para esto, se considera nuevamente 1.000 muestras de 2.000 observaciones. Sin embargo, esta vez hay 1.000 observaciones tomadas aleatoriamente en cada muestra y otras 1.000 que son fijas, comunes a todas las muestras.

Las observaciones comunes a todas las muestras se generan de manera de ser representativas de todos los tipos de alumnos que han rendido el test, en cuanto a resultados obtenidos. Para lograrlo, se agrupa a los estudiantes en clases, según la cantidad de respuestas correctas obtenidas en la prueba. En la tabla 3.1 se detalla la composición de la muestra fija de mil

alumnos.

Tabla 3.1: Composición de la muestra fija de estudiantes, según clase.

Nºrespuestas correctas	Cantidad de estudiantes para la muestra
Entre 0 y 15	184
Entre 16 y 30	529
Entre 31 y 45	172
Entre 46 y 60	81
Entre 61 y 75	34

La idea es estudiar la robustez de los modelos desde el punto de vista de la estimaciones de las habilidades de los individuos. Se desea observar la dependencia de las estimaciones con respecto a la muestra de estudiantes.

3.3. Resultados: Análisis de los parámetros del Item.

3.3.1. Aplicación del Modelo Rasch

En esta sección se exponen los resultados obtenidos al aplicar el modelo Rasch a los datos.

La tabla 3.2 a continuación presenta un resumen de este análisis. En ella se incluye los 5 ítems con menor dificultad promedio y los 5 ítems con mayor dificultad promedio, estimada con el modelo Rasch, la varianza de las estimaciones de la dificultad y el eje temático al que pertenecen dichas preguntas.

Tabla 3.2: Resumen de los resultados del análisis de los ítems a través del Modelo Rasch.

Item	Dificultad	σ^2	Eje Temático
75	-0,9420	0,0035	Datos y azar
19	-0,6518	0,0035	Álgebra
26	-0,5581	0,0036	Álgebra
33	-0,4871	0,0034	Geometría
1	-0,4840	0,0032	Números
118	2,1098	0,0152	Números
64	2,1775	0,0144	Datos y azar
13	2,1871	0,0093	Números
59	2,2592	0,0099	Datos y azar
74	2,3815	0,0105	Números

En la tabla 3.2 se observa que la varianza σ^2 es mayor en las preguntas de mayor dificultad que en las preguntas de dificultad menor, esto se muestra con mayor detalle en la figura 3.1b más adelante.

Un segundo aspecto llamativo de la tabla 3.2 es que las preguntas con menor dificultad sean mayoritariamente del eje temático “álgebra”, mientras las preguntas más difíciles son mayoritariamente del eje “datos y azar” y “números”. A continuación, en la tabla 3.3, se reporta la dificultad promedio por eje temático.

Tabla 3.3: Dificultad promedio estimada por eje temático, modelo Rasch.

Eje Temático	Dificultad Promedio
Álgebra	0,6380
Números	0,6688
Geometría	0,9434
Datos y Azar	1,0003

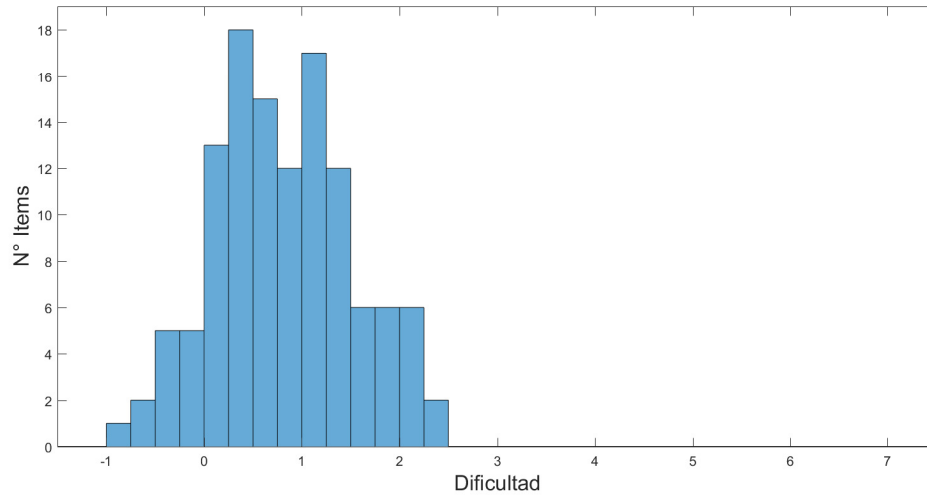
El resultado que arroja la tabla 3.3 lleva a reflexionar acerca de la existencia de una sola habilidad para responder a los ítems de la PSU de matemática. ¿Realmente las preguntas de “datos y azar” tienen, en promedio, dificultad más elevada que las de “álgebra” o tal vez cada eje temático mide habilidades diferentes e independientes entre sí?

En la figura 3.1a se presenta el histograma de la dificultad promedio estimada con el modelo Rasch, donde se observa que la distribución de la estimación de las dificultades con el modelo Rasch tiende a ser gaussiana.

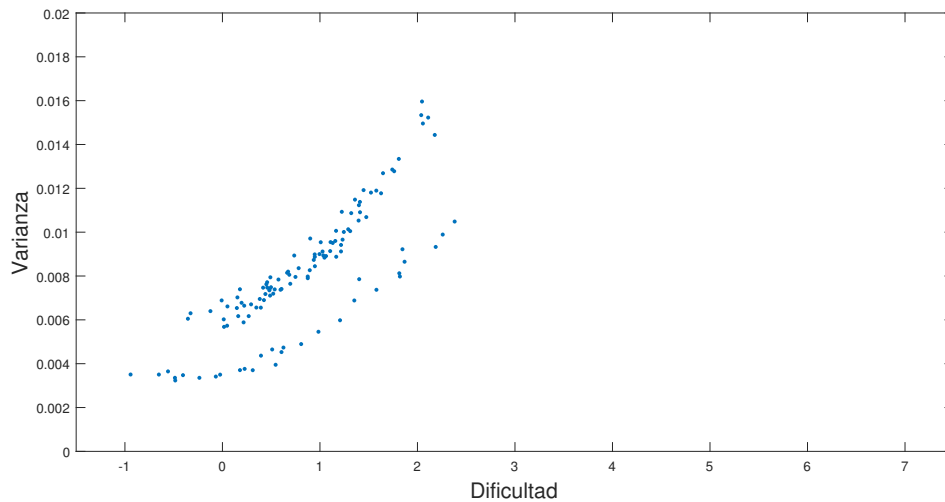
Por otro lado, en la figura 3.1b se puede ver que la varianza de la estimación de la dificultad crece al mismo tiempo que la dificultad. Esto implica que a medida que aumenta la dificultad la estimación del parámetro es menos certera.

Un aspecto curioso en esta imagen es que se pueden observar dos curvas de varianza que separan a las preguntas en dos grupos distintos. Al estudiar este fenómeno con detención, se advierte que el grupo de abajo en el gráfico, es decir, el que tiene menor varianza en función de la dificultad, corresponde a las 30 preguntas que son comunes a las cuatro formas de la prueba; mientras que el grupo de varianza mayor corresponde a las 90 preguntas que no están presentes en todas las formas.

Lo anterior implica que las estimaciones son más robustas cuando la información está completa. En consecuencia, debería ser más conveniente realizar un análisis IRT por cada forma por separado, en lugar de considerar la matriz completa con todas las formas.



(a) Histograma de la dificultad promedio estimada según el modelo Rasch.



(b) Gráfico dificultad estimada promedio vs. varianza muestral.

Figura 3.1: Contraste dificultad y varianza, modelo Rasch.

Finalmente, se ha estimado la discriminación promedio de la prueba según el modelo Rasch. Si bien este modelo en principio propone que la discriminación de todos los ítems es igual a 1, es posible hacer una estimación de la discriminación que sea igual para todos los ítems. En este caso, el valor obtenido para la discriminación estimada sirve como una medida de pertinencia del modelo. Una discriminación estimada de 1 es acorde a los supuestos del modelo Rasch, si se obtiene una discriminación estimada mayor a 1 significa que los ítems discriminan más de lo esperado por el modelo Rasch, según su dificultad, y una discriminación estimada menor a 1 significa que los ítems discriminan menos de lo esperado.

En la tabla 3.4 se presenta la discriminación estimada promedio para la PSU de matemática, según el modelo Rasch, y su respectiva varianza.

Tabla 3.4: Discriminación estimada promedio, modelo Rasch.

Discriminación	σ^2
0.9010	0,0004

Dado que se obtiene un valor inferior a 1, los ítems son menos discriminadores (entre las personas de alto y bajo rendimiento) que lo que se espera del modelo, esto significa que los datos no se ajustan completamente a éste.

3.3.2. Aplicación del Modelo 2PL

En esta sección se presenta los resultados obtenidos al aplicar el modelo 2PL a los mismos datos de la sección precedente (ver 3.3.1). Las tablas a continuación presentan el resumen de este análisis.

En la tabla 3.5 se incluyen los 5 ítems con menor dificultad promedio y los 5 ítems con mayor dificultad promedio, estimada con el modelo logístico de dos parámetros, la varianza de las estimaciones de la dificultad, la discriminación estimada promedio, la varianza de la discriminación y el eje temático al que pertenecen dichas preguntas.

Tabla 3.5: Resumen de los resultados del análisis de la dificultad de los ítems a través del modelo 2PL.

Item	Dificultad	σ_{dif}^2	Discriminación	σ_{dis}^2	Eje temático
75	-0,7735	0,0027	1,3621	0,0099	Datos y azar
114	-0,6521	0,0403	0,3897	0,0044	Datos y azar
19	-0,5098	0,0015	2,0035	0,0194	Álgebra
26	-0,4532	0,0014	2,1672	0,0221	Álgebra
33	-0,4314	0,0019	1,5412	0,0131	Geometría
13	4,0482	0,3128	0,4455	0,0036	Números
91	4,1445	1,5866	0,3063	0,0050	Álgebra
59	5,1242	0,8027	0,3623	0,0035	Datos y azar
22	5,2654	1,6938	0,2878	0,0033	Álgebra
74	7,0050	3,8852	0,2841	0,0037	Datos y azar

Al igual que en el caso anterior (ver sección 3.3.1), se observa que los ítems de mayor dificultad tienen una varianza mucho mayor que los de dificultad menor, esto se evidencia en la figura 3.2b. Además, en la estimación realizada con el modelo logístico de dos parámetros, el rango de dificultad es muy amplio, extendiéndose desde $-0,77$, para el ítem 75, hasta $7,00$ para el ítem 74. Con los 10 datos expuestos no es suficiente para determinar si las preguntas de cierto eje temático tienen una dificultad mayor que el resto, por lo que en la tabla 3.7 se estudia esto.

Por otro lado, en la tabla 3.6 se muestra los 5 ítems con menor discriminación promedio y los 5 ítems con mayor discriminación promedio, estimada con el modelo logístico de

dos parámetros, la varianza de las estimaciones de la discriminación, la dificultad estimada promedio, la varianza de la dificultad y el eje temático de cada pregunta.

Tabla 3.6: Resumen de los resultados del análisis de la discriminación de los ítems a través del modelo 2PL.

Item	Dificultad	σ_{dif}^2	Discriminación	σ_{dis}^2	Eje temático
74	7,0050	3,8852	0,2841	0,0037	Datos y azar
22	5,2654	1,6938	0,2878	0,0033	Álgebra
91	4,1445	1,5866	0,3063	0,0050	Álgebra
115	2,5098	0,4099	0,3448	0,0050	Datos y azar
59	5,1242	0,8027	0,3623	0,0035	Datos y azar
3	0,0321	0,0028	2,1588	0,0265	Números
26	-0,4532	0,0014	2,1672	0,0221	Álgebra
76	-0,0226	0,0026	2,1991	0,0345	Números
1	-0,4123	0,0013	2,2402	0,0224	Números
28	0,0488	0,0020	2,6245	0,0218	Álgebra

Examinando las tablas anteriores (3.5 y 3.6) se observa que las preguntas con mayor dificultad coinciden con las que tienen un nivel de discriminación *muy bajo* (de acuerdo al criterio de Baker en 1.1) y, a su vez, las preguntas de menor dificultad tienen discriminación *muy alta*.

A continuación, en la tabla 3.7, se muestra la dificultad promedio y la discriminación promedio, estimadas con el modelo 2PL, para cada eje temático.

Tabla 3.7: Dificultad y discriminación promedio estimadas por eje temático, modelo 2PL.

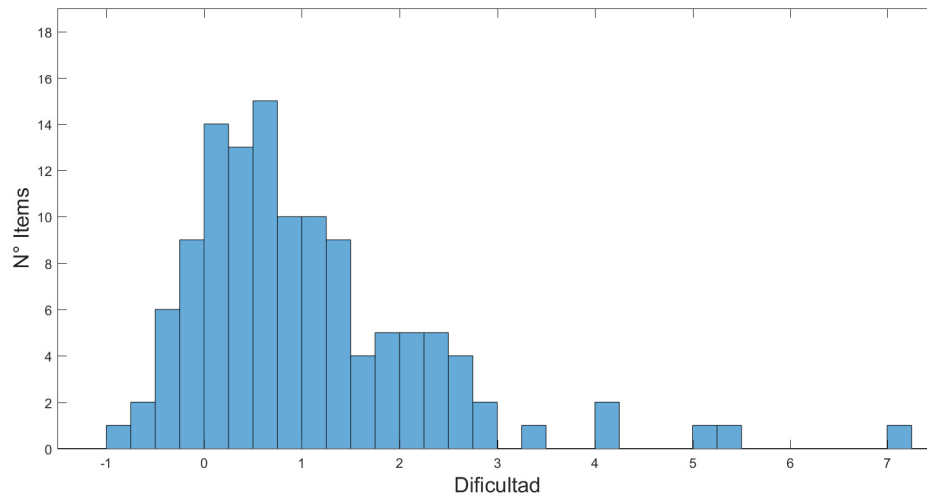
Eje Temático	Dificultad Promedio	Discriminación Promedio
Números	0,6917	1,1937
Álgebra	0,9195	1,1199
Geometría	1,0411	0,9012
Datos y azar	1,5557	0,7419

El orden de dificultad por eje no es exactamente el mismo que para el modelo Rasch (ver tabla 3.3). No obstante, las diferencias de dificultad y discriminación entre ejes temáticos es significativa y la pregunta de si esto podría implicar la existencia de más de una habilidad continúa abierta.

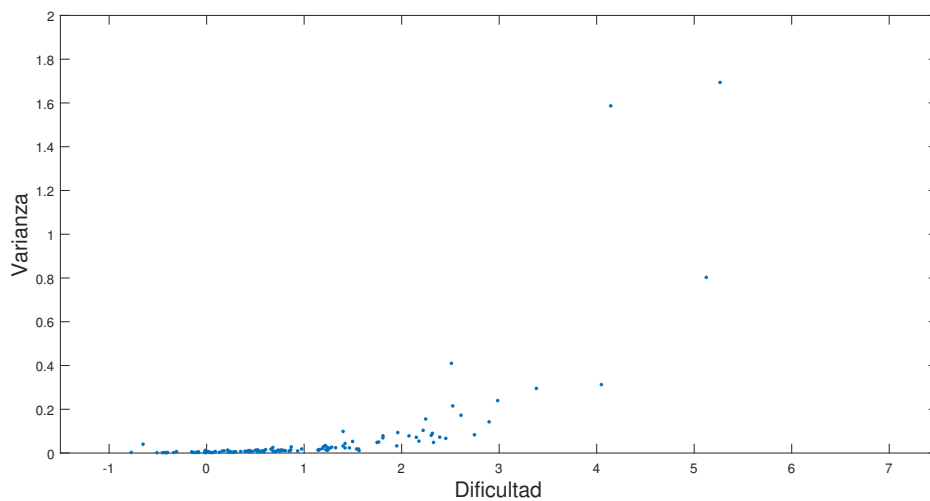
A continuación, en la figura 3.2a, se presenta el histograma de la dificultad promedio estimada con el modelo 2PL. Se observa que la distribución de la dificultad para esta estimación tiene un comportamiento diferente al del modelo Rasch (figura 3.1a), no gaussiano.

Por otro lado, en la figura 3.2b se muestra la relación entre dificultad estimada con este modelo y su varianza. Se puede ver que la varianza de la estimación y la dificultad crecen al mismo tiempo, al igual que en el análisis realizado para el modelo Rasch. Además, resulta

alarmante la alta varianza que tienen ciertas estimaciones. En ese sentido, las estimaciones del modelo 2PL resultan menos confiables que las del modelo Rasch.



(a) Histograma de la dificultad promedio estimada según el modelo 2PL.



(b) Gráfico Varianza Muestral vs. Dificultad estimada promedio.

Figura 3.2: Contraste dificultad y varianza, modelo 2PL.

3.3.3. Comparación de resultados

Luego de observar los resultados de la estimación de los parámetros del ítem para cada uno de los modelos, es natural preguntarse si existe algún tipo de relación entre ellos. La presente sección tiene por objetivo estudiar esto.

A continuación, en la figura 3.3 se grafica el plano dificultad 2PL - discriminación 2PL versus la dificultad Rasch.

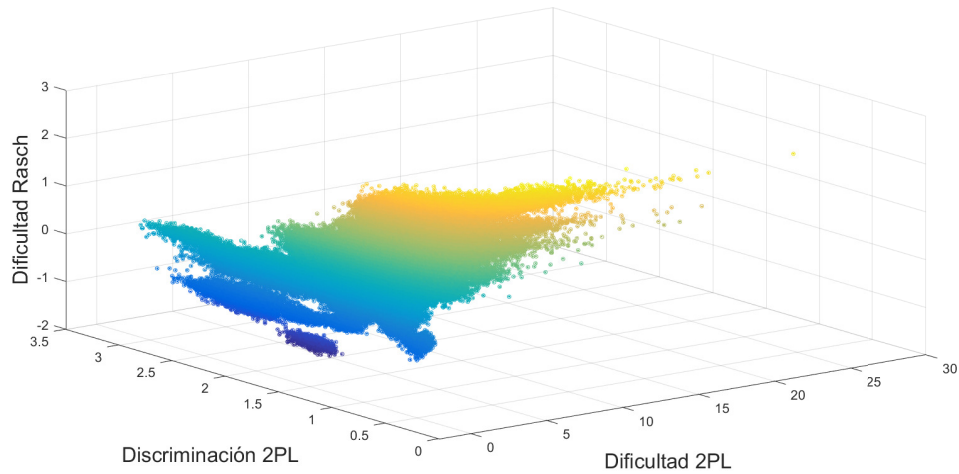


Figura 3.3: Relación del plano dificultad-discriminación del modelo 2PL y la dificultad del modelo Rasch.

Se observa que los resultados definen una superficie en tres dimensiones. Esto implicaría que efectivamente existe una función $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, tal que $g(b_{2pl}, a_{2pl}) = b_{Rasch}$. Es decir, las estimaciones del modelo 2PL podrían reducirse simplemente a Rasch.

Por otro lado, en la figura 3.4 se presenta el plano dificultad Rasch - dificultad 2PL.

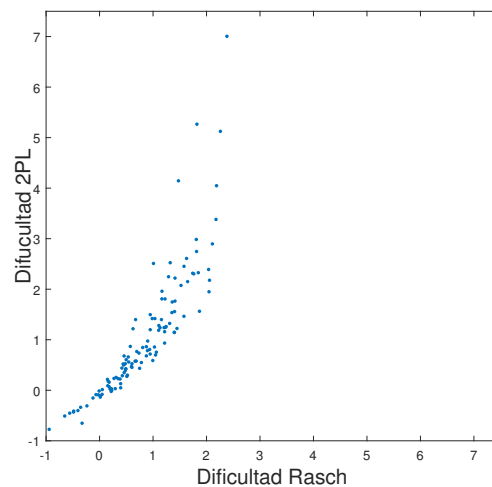


Figura 3.4: Relación entre las estimaciones de la dificultad: Rasch versus 2PL.

De la figura anterior (figura 3.4), se desprende que ambas estimaciones son coherentes entre sí, en el sentido de que las preguntas clasificadas como más difíciles según el modelo Rasch, también lo son según el modelo 2PL. Por otro lado, resulta interesante que la relación entre la dificultad Rasch y la 2PL es aparentemente cuadrática.

La figura 3.5 presenta específicamente la diferencia entre la estimación de la dificultad con

el modelo 2PL y la diferencia de la estimación con el modelo Rasch. Esto permite identificar cuáles son las preguntas cuya estimación es más disímil y si existe alguna característica en particular que pueda explicar este hecho.

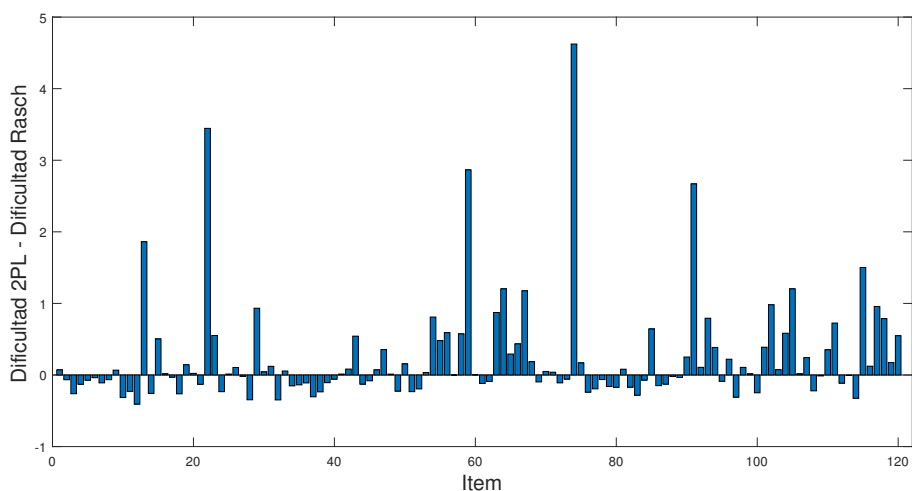
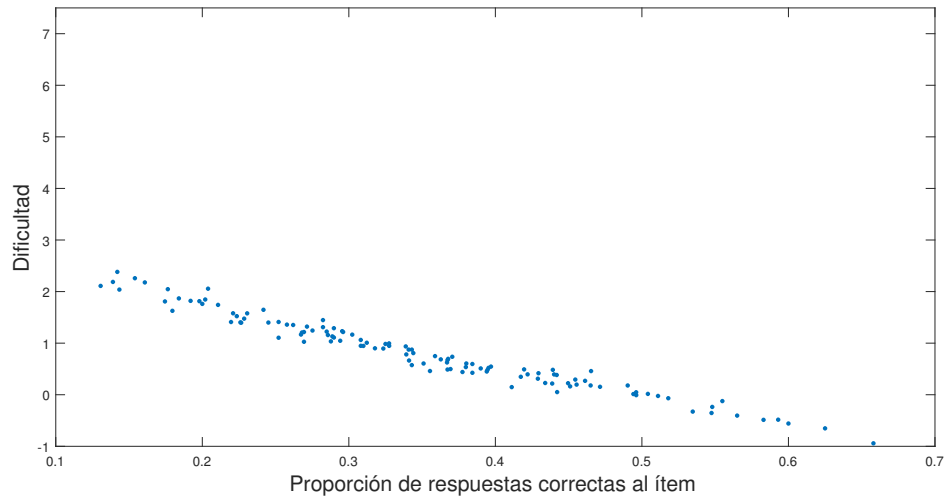


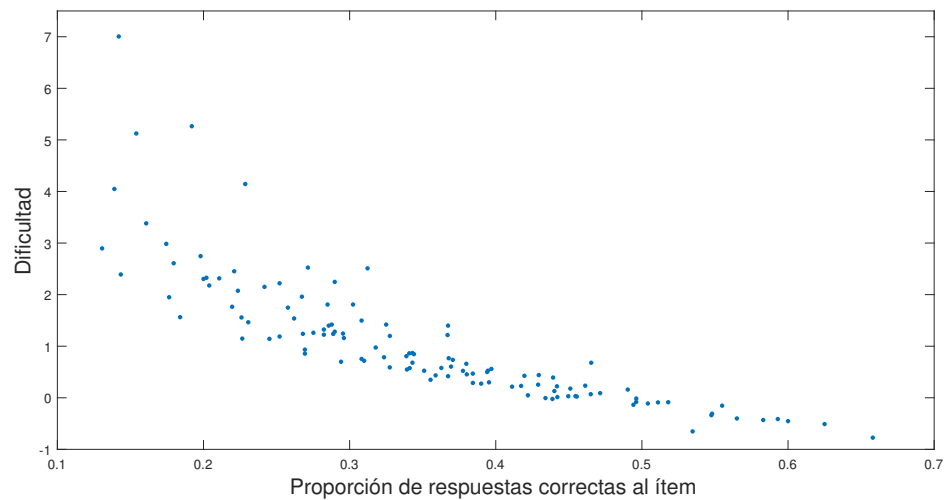
Figura 3.5: Diferencia entre estimaciones 2PL y Rasch para la dificultad.

En el gráfico anterior (figura 3.5) se observa que, en general, la estimación de la dificultad es mayor en el caso de 2PL que en Rasch. Los casos que resultan más llamativos son el del ítem 74, el 22 y el 59, donde la diferencia es 4.62, 3.45 y 2.87, respectivamente. Por otra parte, las preguntas que presentan una diferencia prácticamente nula en la estimación son la número 57 y la número 60.

Finalmente, en la figura 3.6 se muestra la relación entre la proporción de respuestas correctas de cada ítem y su dificultad estimada por cada modelo.



(a) Relación entre proporción de respuestas correctas de los ítems y la dificultad, modelo Rasch.



(b) Relación entre proporción de respuestas correctas de los ítems y la dificultad, modelo 2PL.

Figura 3.6: Relación entre la proporción de respuestas correctas a los ítems y su dificultad.

En los gráficos anteriores (figura 3.6) se observa que la relación entre la dificultad estimada por el modelo Rasch es prácticamente lineal con respecto a la proporción de respuestas correctas a cada ítem. Esto significa que la dificultad Rasch es consistente con la medida clásica de dificultad.

Por otro lado, para el modelo 2PL dos preguntas con igual proporción de respuestas correctas pueden tener diferencias significativas de dificultad. Observando la figura 3.6b, preguntas con una proporción de 0,2 de respuestas correctas tienen dificultad que varía entre 2 y 5. Es difícil de comprender y explicar tanta diferencia, pero podría deberse al efecto de la *pseudo-adivinanza* que no es considerado en ninguno de los dos modelos aplicados.

3.4. Resultados: Análisis de los parámetros de los individuos.

3.4.1. Aplicación del Modelo Rasch

En esta sección se presenta los resultados obtenidos al aplicar el modelo Rasch al segundo conjunto de datos, que contiene mil observaciones comunes a todas las muestras.

La tabla 3.8 presenta un resumen de este análisis. En ella se exponen los dos individuos de mayor habilidad promedio estimada y los dos individuos de menor habilidad promedio estimada, por cada grupo identificado en la tabla 3.1, y la varianza de la dificultad estimada.

Tabla 3.8: Resumen de resultados del análisis de habilidades, modelo Rasch.

Nºrespuestas correctas	ID Alumno	Habilidad	σ^2
0 - 15	151	-1,7568	0,0003
	89	-1,6569	0,0002
	121	-0,7667	0,0002
	69	-0,7667	0,0002
16 - 30	594	-0,7733	0,0002
	622	-0,7733	0,0002
	510	0,3196	0,0005
	573	0,3196	0,0005
31 - 45	765	0,2777	0,0004
	724	0,2777	0,0004
	883	1,2521	0,0011
	774	1,2521	0,0011
46 - 60	886	1,2007	0,0010
	937	1,2353	0,0010
	910	2,2977	0,0019
	943	2,2977	0,0019
61 - 75	990	2,2632	0,0018
	998	2,2632	0,0018
	992	3,5508	0,0025
	995	4,0099	0,0020

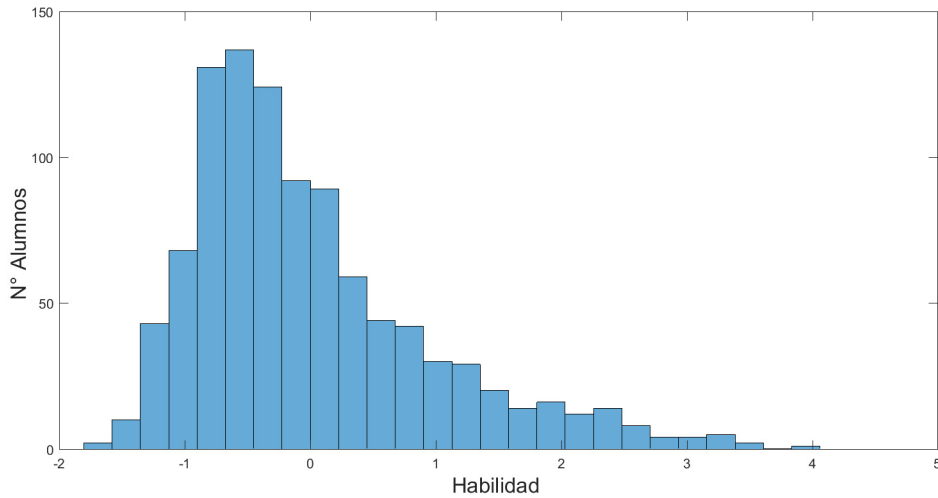
Al igual que en el análisis de las estimaciones de la dificultad de los ítems, es interesante notar que la varianza de la estimación es mayor, al aumentar la habilidad de los examinados. Este fenómeno se presenta visualmente en la figura 3.7b a continuación.

Además, todo parece indicar que la habilidad es una función creciente del número de respuestas correctas obtenidas por los estudiantes, lo cual es consistente con el hecho de que el puntaje obtenido es estadístico suficiente de la habilidad [17].

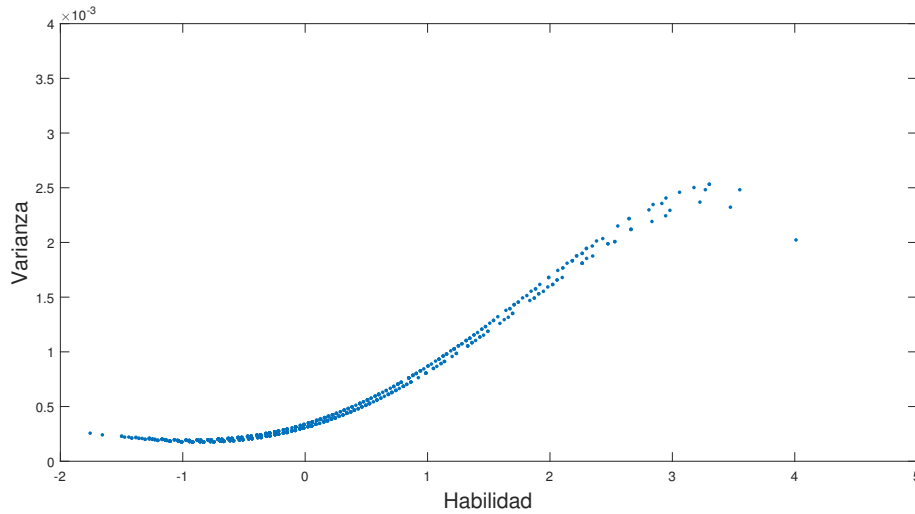
En la figura 3.7a se presenta el histograma de la habilidad promedio estimada con el modelo Rasch. La distribución de ésta resulta ser similar a la de las respuestas correctas

obtenidas por alumno.

Finalmente, en la figura 3.7b se grafica la habilidad promedio contra la varianza de las estimaciones. Se observa que la varianza crece junto con la habilidad. Es decir, a medida que aumenta la habilidad, mayor es el error de medición; esto se traduce en que hay menor precisión en la estimación de los puntajes más altos. Sin embargo, dado que la varianza es del orden de 10^{-3} (pequeño en comparación con los valores de la habilidad) este error de medición no es significativo.



(a) Histograma de la habilidad promedio estimada según el modelo Rasch.



(b) Gráfico Varianza Muestral vs. Habilidad estimada promedio.

Figura 3.7: Contraste Habilidad y varianza, modelo Rasch.

3.4.2. Aplicación del Modelo 2PL

En esta sección se presentan los resultados obtenidos al aplicar el modelo 2PL al segundo conjunto de datos, al igual que en la sección precedente.

La tabla 3.9 expone el resumen de este análisis. En ella se presentan los dos individuos de mayor habilidad promedio estimada y los dos individuos de menor habilidad promedio estimada, por cada grupo identificado en la tabla 3.1, y la varianza de la dificultad estimada.

Tabla 3.9: Resumen de resultados 2PL

Nºrespuestas correctas	ID Alumno	Habilidad	σ^2
0 - 15	23	-1,5408	0,0006
	89	-1,4868	0,0007
	71	-0,5991	0,0005
	33	-0,5493	0,0004
16 - 30	665	-1,0252	0,0005
	674	-1,0108	0,0005
	272	0,3723	0,0009
	525	0,4279	0,0011
31 - 45	767	0,0360	0,0007
	786	0,0549	0,0006
	858	1,2812	0,0018
	788	1,3155	0,0019
46 - 60	920	1,1003	0,0017
	922	1,1180	0,0017
	941	2,3601	0,0031
	888	2,3984	0,0031
61 - 75	970	2,2752	0,0033
	990	2,3191	0,0030
	975	3,9253	0,0036
	995	4,3483	0,0025

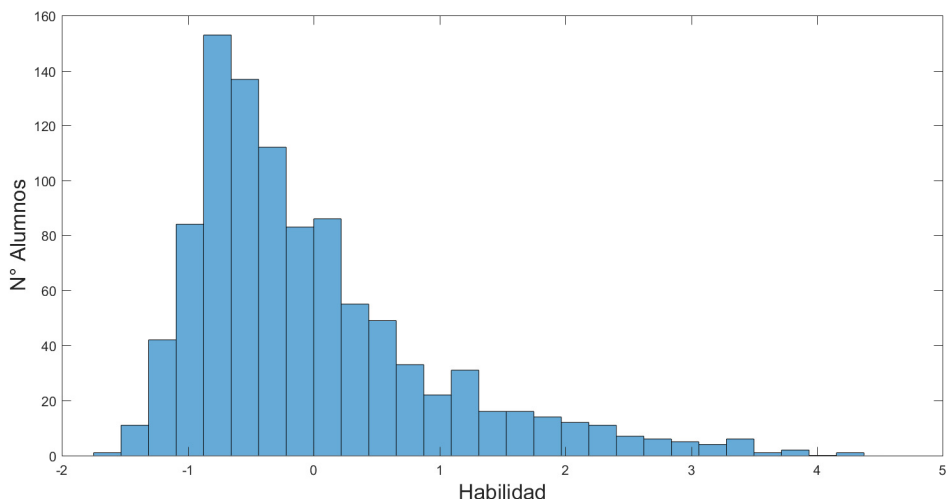
Al igual que para el caso del modelo Rasch, se observa que la varianza es, en general, mayor al aumentar la habilidad. Esta tendencia se observa en la figura 3.8b.

En este caso, si bien la habilidad presenta una tendencia a crecer de acuerdo con el número de respuestas correctas, no se trata de una función creciente. Basta con notar que el alumno 525 tiene una habilidad mayor que el 767, a pesar de tener menor cantidad de respuestas correctas.

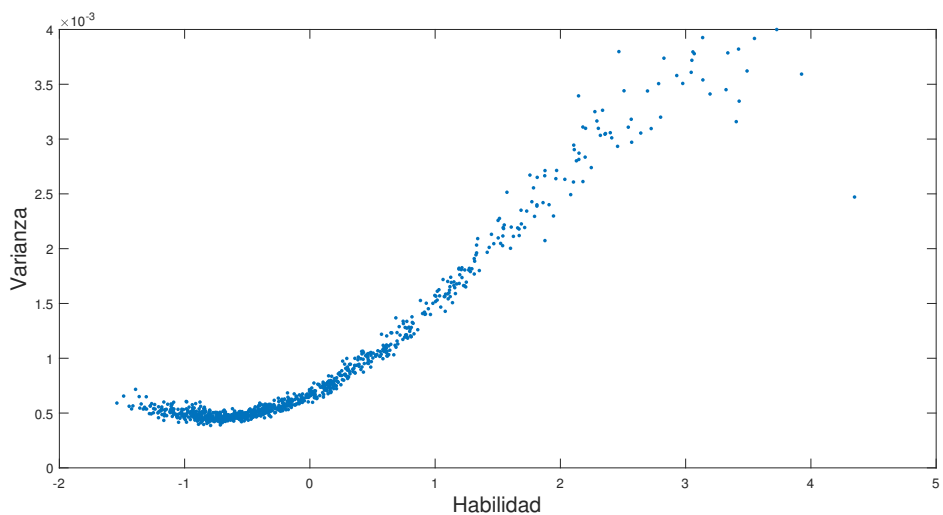
En la figura 3.8a se presenta el histograma de la habilidad promedio estimada con el modelo 2PL. Cabe señalar que, a diferencia de la estimación de las dificultades de los ítems, la distribución de las habilidades estimadas con el modelo 2PL y Rasch son muy similares.

Sin embargo, al comparar las figuras 3.7b y 3.8b se observa que la varianza de las estimaciones de las habilidades con el modelo 2PL es más elevada que con el modelo Rasch. En la

figura 3.8b, se tiene que la varianza aumenta a medida que aumenta la habilidad, pero este aumento es considerablemente mayor que para el caso del modelo Rasch. Esto significa que la precisión de la estimación de habilidades altas es mucho menor que al estimar puntajes bajos. No obstante, como el orden de la varianza es de 10^{-3} , el error no es significativo.



(a) Histograma de la habilidad promedio estimada según el modelo 2PL.



(b) Gráfico Varianza Muestral vs. Habilidad estimada promedio.

Figura 3.8: Contraste Habilidad y varianza, modelo 2PL.

3.4.3. Comparación de resultados

Después de haber revisado los resultados de la estimación de la habilidad de los individuos según cada modelo, cabe preguntarse si existe alguna relación entre ellos. El objetivo de esta sección es estudiar esto.

En la figura 3.9 se presenta la relación entre la habilidad estimada usando el modelo Rasch

y la habilidad estimada según el modelo 2PL.

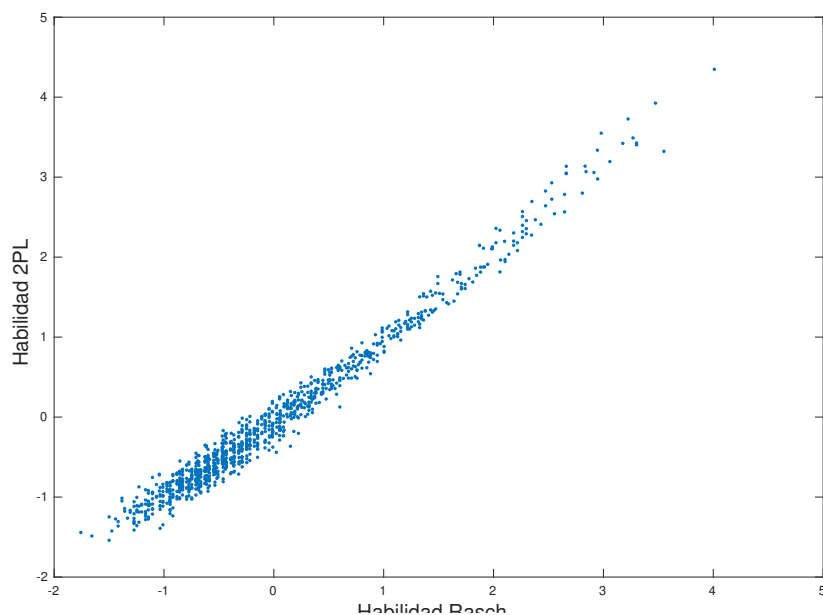


Figura 3.9: Gráfico comparativo habilidad Rasch vs. habilidad 2PL.

A diferencia del caso de la estimación de la dificultad de los ítems, donde se apreciaban grandes diferencias entre las estimaciones de los dos modelos estudiados, en el caso de las habilidades los resultados se encuentran en la misma escala y son, como se observa en la figura 3.9, bastante lineales. Por lo tanto, los resultados son coherentes entre sí, en el sentido de que los alumnos con mayor habilidad para un modelo son también, en general, los de mayor habilidad en el otro.

En la figura 3.10 se grafica la diferencia entre la habilidad estimada con el modelo 2PL y la habilidad estimada con el modelo Rasch, para identificar qué alumnos tienen mayores diferencias y encontrar características comunes que justifiquen este comportamiento.

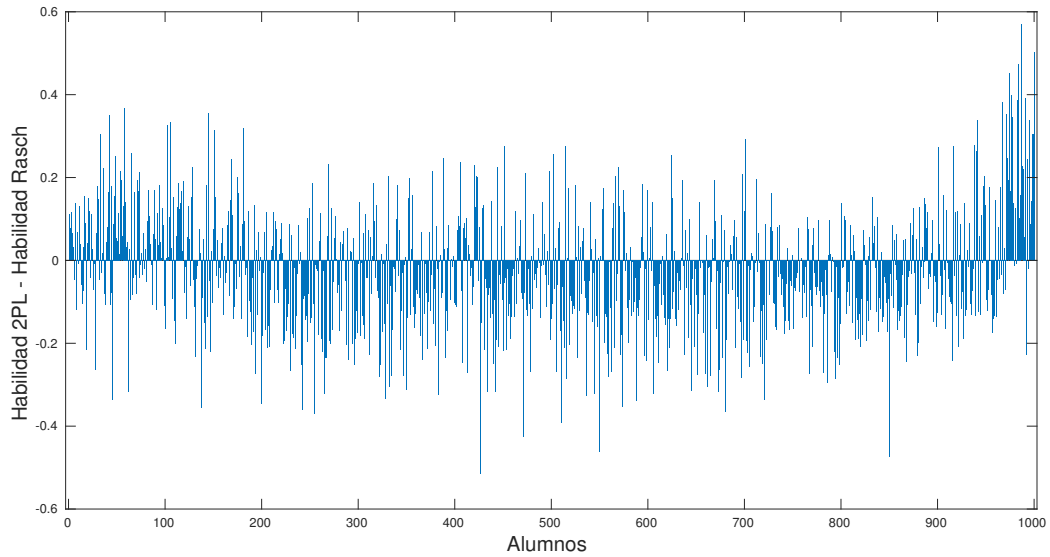


Figura 3.10: Contraste entre estimaciones 2PL y Rasch para la habilidad.

En la figura 3.10, en contraste con el caso de la estimación de las dificultades, no hay una tendencia clara de que uno de los modelos estime habilidades mayores que el otro. Sin embargo, al lado derecho del gráfico, en la zona donde se encuentran los individuos que respondieron correctamente más preguntas, se puede ver que el modelo 2PL tiende a dar una estimación mayor para la habilidad que el modelo Rasch.

Los alumnos que presentan mayor diferencia absoluta son el 1000, el 427 y el 987, donde la habilidad estimada difiere en 0.50, 0.52 y 0.57, respectivamente. Por otro lado, los que presentan diferencia de estimación prácticamente nula son los alumnos 832 y 955.

En las figuras 3.11 y 3.12 se muestra la relación entre el número de respuestas obtenida por cada estudiante y la habilidad estimada según cada modelo, respectivamente.

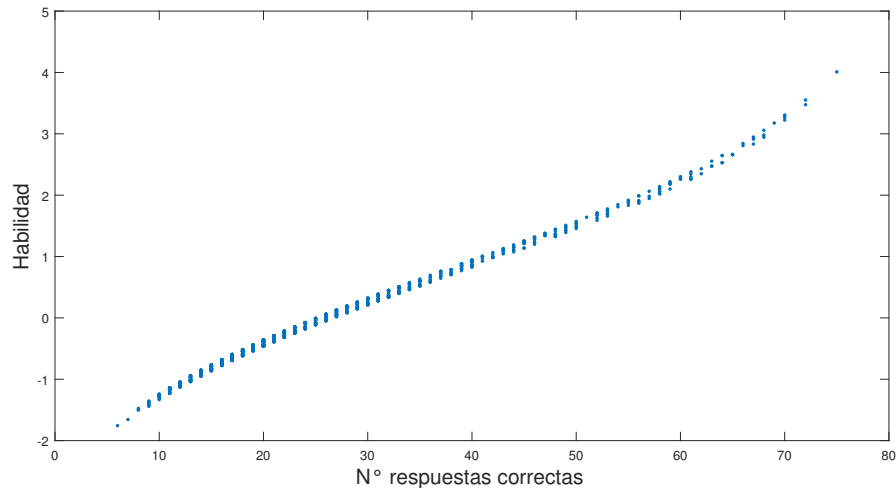


Figura 3.11: Relación entre el número de respuestas correctas de cada alumno y la habilidad estimada, modelo Rasch.

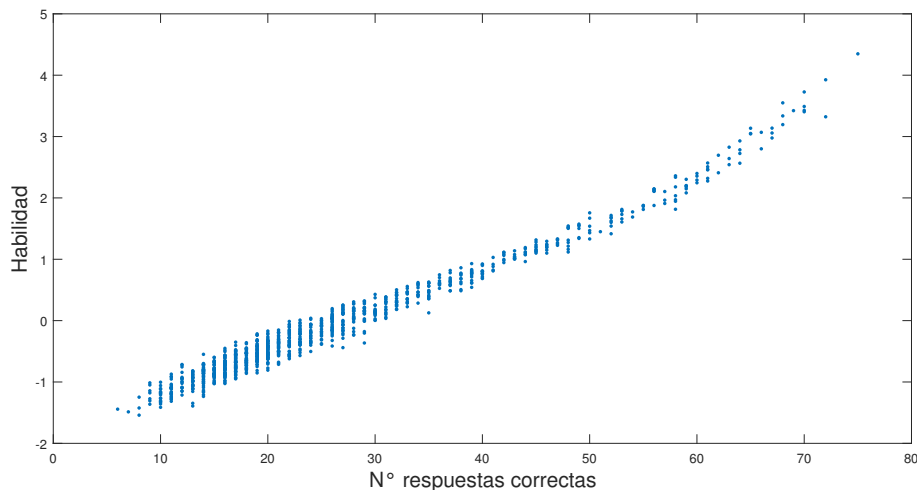


Figura 3.12: Relación entre el número de respuestas correctas de cada alumno y la habilidad estimada, modelo 2PL.

Para el modelo Rasch, se observa que la relación entre el número de respuestas correctas respondidas por un alumno y su habilidad estimada es prácticamente lineal. Es decir, medir uno o el otro es indiferente.

Por otro lado, en el modelo 2PL, personas con el mismo número de respuestas correctas pueden tener gran diferencia en términos de habilidad. Por ejemplo, mirando el nivel de 20 respuestas correctas, se observa que las habilidades varían casi un punto en la escala de habilidad. Asimismo, personas con habilidad estimada igual a 0 pueden tener desde 20 respuestas correctas a 32 respuestas correctas.

Lo anterior significa que, al menos la habilidad estimada con el modelo 2PL, no se re-

comienda como clasificador de estudiantes, ya que comunicacionalmente resulta difícil de explicar a la población que alumnos con igual número de respuestas correctas sean evaluados con distinta habilidad. En este sentido, los resultados de los análisis IRT sólo podrían complementar los análisis de la teoría clásica, pero no deberían reemplazarlos, a menos que se demuestre que las habilidades estimadas con IRT son mejores predictores del rendimiento universitario que los puntajes PSU actuales. Desde una perspectiva pragmática, como el propósito de la PSU es la admisión a las universidades, si esto se demostrara sería legítimo usar la habilidad IRT como indicador de puntaje PSU, pero sería necesario informar a la población previamente cuál sería la ponderación de cada ítem en el cálculo del puntaje.

Finalmente, en las figuras 3.13 y 3.14 se muestra la relación entre el puntaje PSU y la habilidad estimada según cada modelo.

En ellas se puede realizar las mismas observaciones que en las figuras precedentes.

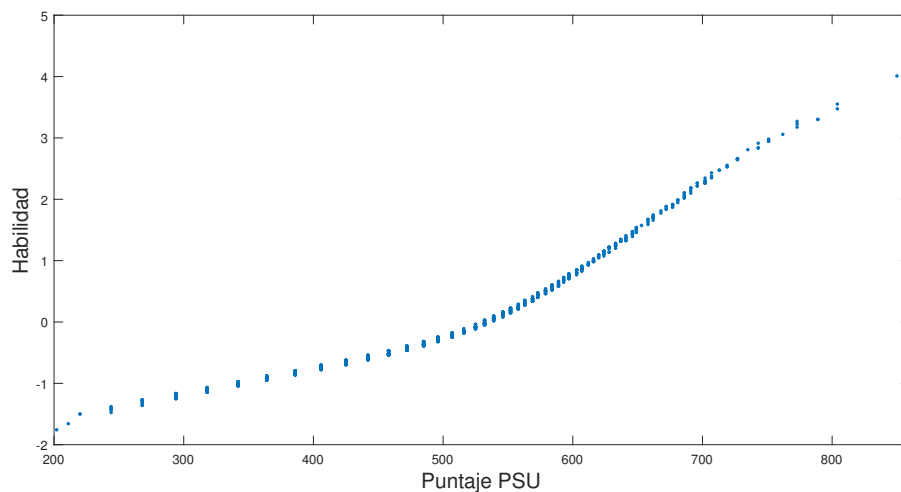


Figura 3.13: Relación entre el puntaje PSU obtenido por cada alumno y la habilidad estimada, modelo Rasch.

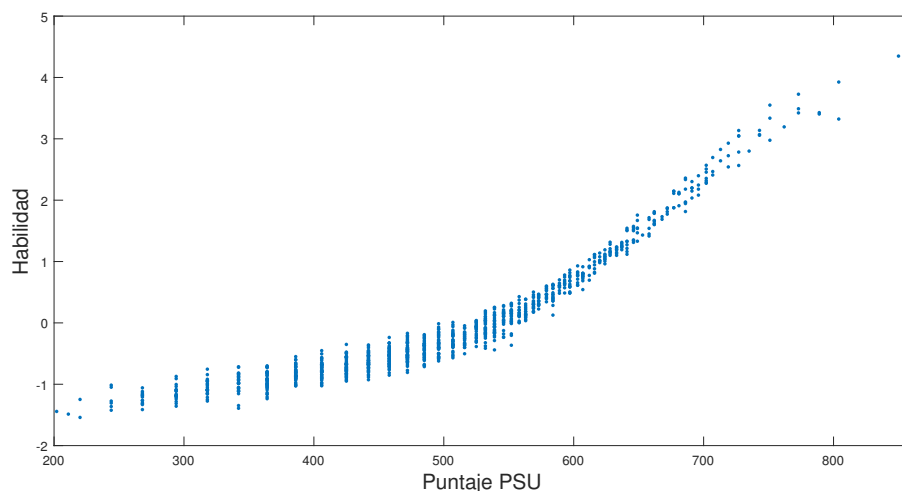


Figura 3.14: Relación entre el puntaje PSU obtenido por cada alumno y la habilidad estimada, modelo 2PL.

3.5. Exploración de Unidimensionalidad

De los resultados anteriores, se tiene que los modelos IRT no entregan información muy diferente de la que se desprende al aplicar la teoría clásica. La habilidad medida con el modelo Rasch es prácticamente equivalente a ver la cantidad de respuestas correctas de un individuo, mientras que la estimación de la dificultad representa esencialmente lo mismo que el inverso de la proporción de respuestas correctas al ítem (en ambos casos hay variaciones, pero son poco significativas). Por otro lado, dada la alta varianza que presentan los resultados del modelo 2PL, no parece confiable utilizarlo para el análisis de la PSU.

A raíz de esto y dado que uno de los objetivos de este trabajo es evaluar si se cumple el supuesto de unidimensionalidad en la PSU, lo cual permitiría el empleo de métodos IRT en la asignación de puntajes de este instrumento, esta sección está dedicada a estudiar este aspecto.

3.5.1. Coeficiente Alfa de Cronbach

En primer lugar, se calcula el coeficiente alfa de Cronbach, puesto que por muchos años fue el índice más utilizado para determinar unidimensionalidad de un test.

En este caso, la PSU de matemática obtiene un coeficiente

$$\alpha = 0,9079 \quad (3.1)$$

Este resultado está muy cerca el máximo valor del coeficiente (que es 1), lo cual es deseable en cualquier test. Sin embargo, esto sólo significa que la PSU es una prueba de alta *confiabi-*

lidad, es decir, sus preguntas tienen alta correlación entre sí, pero no implica necesariamente que sea unidimensional. De hecho, en pruebas de selección y admisión universitarias como la PSU, un índice de 0.90 es el mínimo de confiabilidad deseada, mientras que 0.95 es el estándar que se considera apropiado para este tipo de instrumentos [10].

3.5.2. Análisis de Componentes Principales

Adicionalmente, se realiza un análisis de componentes principales para la PSU de matemáticas para estudiar la existencia de una dimensión que entregue considerablemente más información sobre los datos que las demás, es decir, si existe una dimensión preponderante.

En primer lugar, se realiza un ACP para la matriz completa de respuestas que considera las cuatro formas de la prueba juntas y luego se realiza un ACP a cada forma de la prueba por separado. Los resultados se resumen en la tabla 3.10, en ella se observa los porcentajes de varianza explicados por cada una de las primeras cinco componentes principales.

Tabla 3.10: Resumen Análisis de Componentes Principales.

	Porcentaje de Varianza explicado				
	CP1	CP2	CP3	CP4	CP5
Matriz Completa	13,6021	11,9648	10,6284	2,1223	1,2547
Forma 1	18,0727	3,3443	2,1320	1,7578	1,4393
Forma 2	18,3862	3,3466	2,2724	1,8983	1,4104
Forma 3	17,6546	3,3814	2,2145	1,7827	1,4644
Forma 4	18,1942	3,3262	2,1973	1,8229	1,4236

Estos resultados presentan gráficamente en las figuras 3.15 y 3.16. En la primera se presentan las diez componentes principales de la prueba, considerando todas las formas juntas, y la varianza correspondiente a cada una de las componentes principales. En la segunda se presentan las diez componentes principales de cada forma de la prueba por separado y la varianza explicada por cada componente.

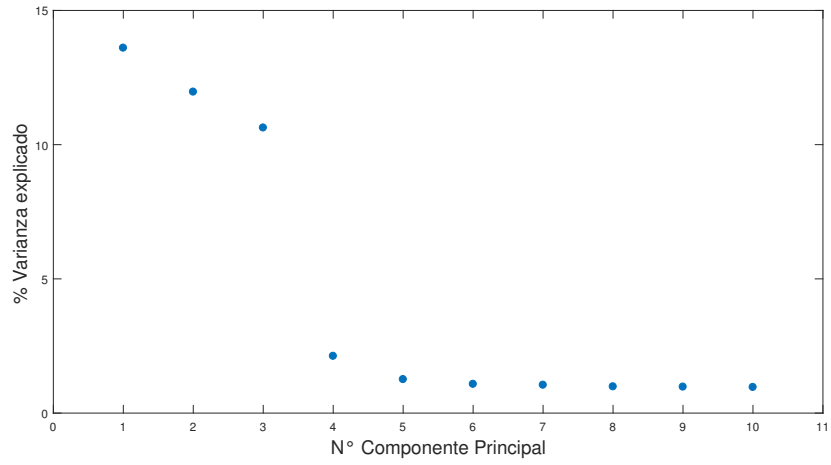
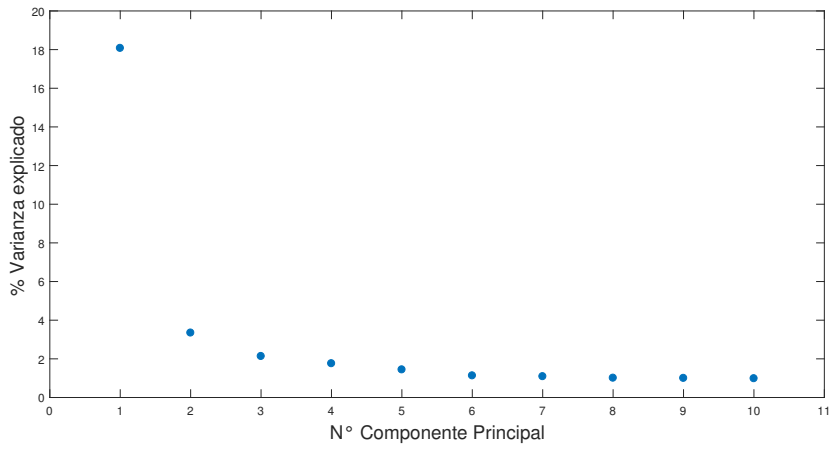
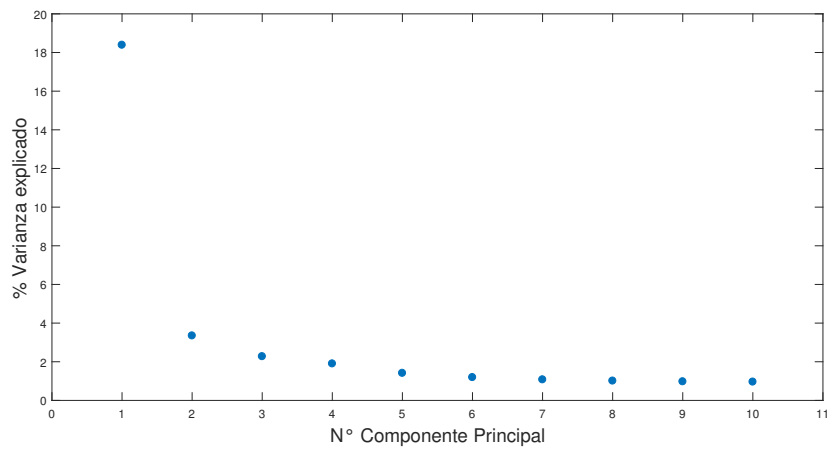


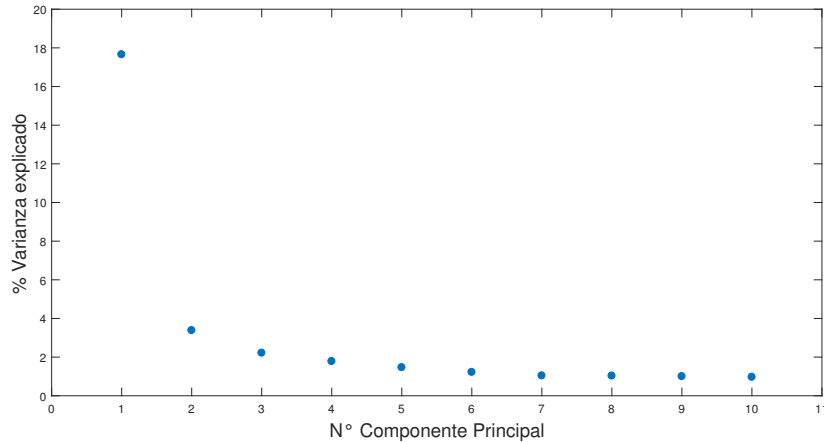
Figura 3.15: ACP aplicado a la matriz de respuestas considerando todas las formas.



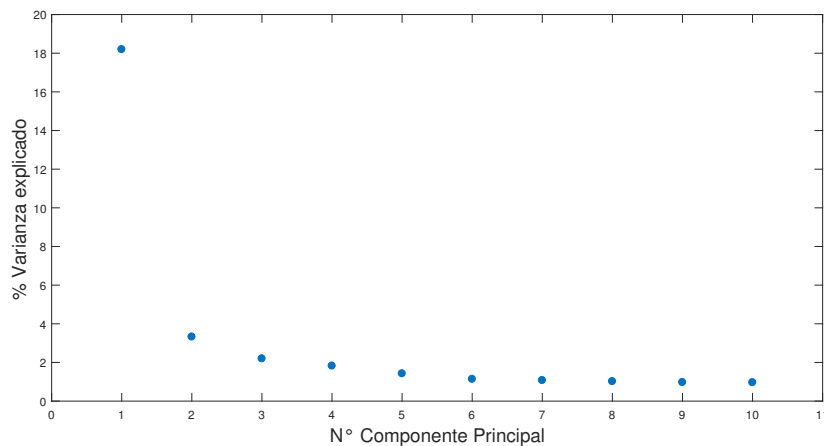
(a) ACP aplicado a la forma 1.



(b) ACP aplicado a la forma 2.



(c) ACP aplicado a la forma 3.



(d) ACP aplicado a la forma 4.

Figura 3.16: Resultados ACP sobre cada forma por separado.

De la tabla 3.10 y la figura 3.15, cuando se estudia la matriz completa de respuestas con todas las formas juntas, se tiene que existen tres componentes principales que proporcionan, en conjunto, el 36,19 % de la varianza de los datos. Sin embargo, no es posible identificar una sola dimensión que sea preponderante, pues no hay una diferencia significativa entre estas tres componentes.

Por otro lado, los resultados observados en la tabla 3.10 y en la figura 3.16 para cada una de las formas indican que, al considerar cada forma por separado, sí hay una componente preponderante que aporta significativamente más a la varianza de los datos que las demás componentes. En cada forma la primera componente principal representa un porcentaje del orden del 18 % de la varianza, mientras que la segunda componente representa un porcentaje del orden del 3 %, una diferencia importante. Esto significa que es posible considerar que cada forma de la prueba es aproximadamente unidimensional.

Cabe mencionar que los resultados observados del ACP cambian radicalmente al considerar

toda la matriz de respuestas, con todas las formas juntas, que al analizar cada una de las formas por separado. Lamentablemente, al considerar la matriz con todas las formas se pierde mucha información, pues no todas las preguntas han sido respondidas por todos los alumnos.

3.5.3. Análisis Paralelo Modificado

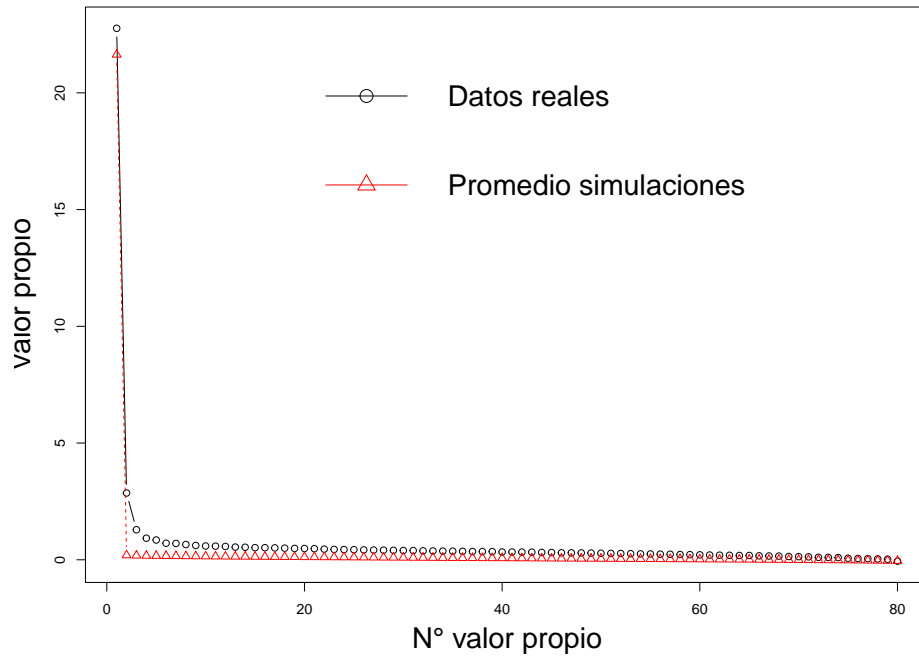
Como se expuso en el capítulo 1 (ver sección 1.5), los estadísticos que suelen ser utilizados para determinar la unidimensionalidad de un test no poseen una base teórica que los sustente.

Por esto, se realiza un estudio empírico que permite determinar si la prueba es *suficientemente unidimensional* para poder aplicar los modelos IRT. Se utiliza el método de análisis paralelo modificado (*modified parallel analysis*) que fue descrito en la sección 1.5.1.

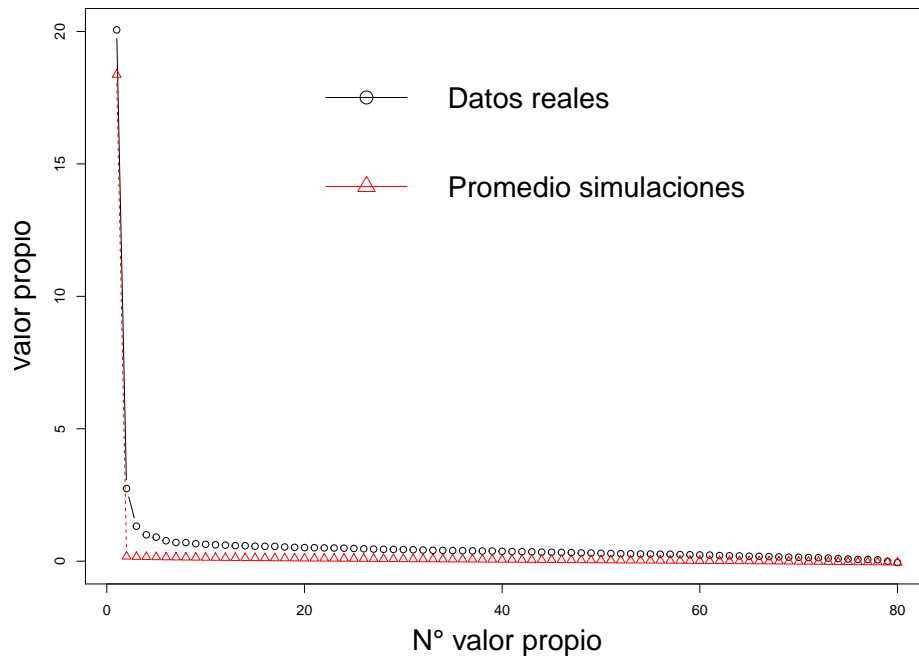
Para aplicar satisfactoriamente este método, es necesario que la matriz de datos, que en este caso corresponde a la matriz de respuestas de los alumnos a la PSU, esté completa. No obstante, el hecho de que la PSU tenga cuatro formas distintas y no todas las preguntas estén presentes en todas las formas implica que no todos los alumnos deben responder los mismos ítems (de hecho, sólo responden los 75 correspondientes a su forma, del total de 120 preguntas que existen en las cuatro formas) y no es posible tener una matriz de datos completa. Por este motivo, se decide aplicar el análisis paralelo modificado a cada forma por separado.

A continuación, en la figura 3.17 se presentan los resultados de la aplicación de este método. En ella se contrastan los valores propios de la matriz de correlaciones tetracóricas para los datos reales y para datos simulados. El estadístico que se usa para aceptar o refutar la unidimensionalidad de los datos es el segundo valor propio.

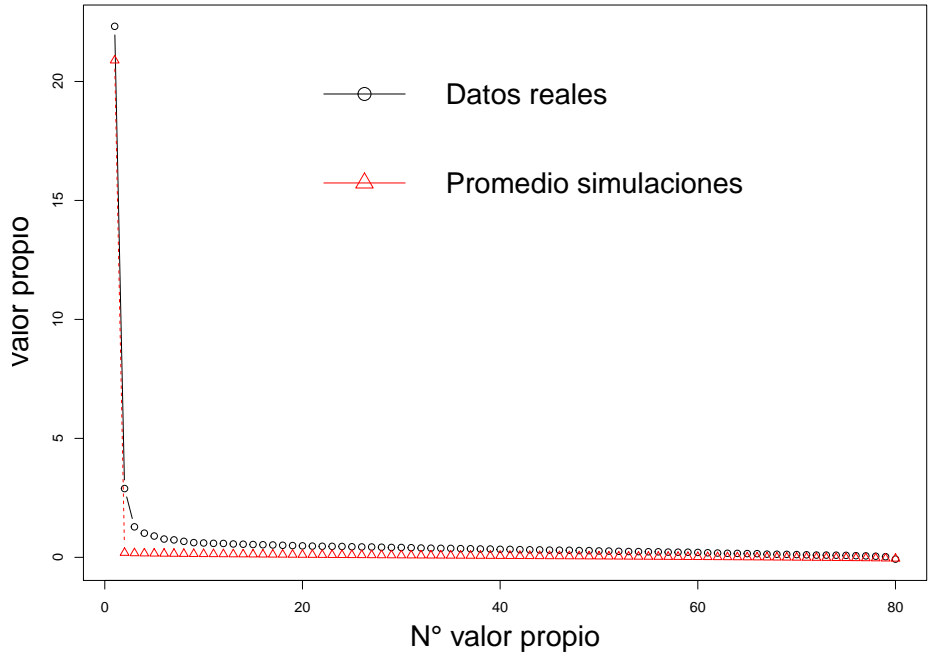
En cada uno de los cuatro gráficos presentados en la figura 3.17 se observa que el segundo valor propio de los datos reales es considerablemente mayor que el segundo valor propio de los datos simulados bajo la hipótesis de unidimensionalidad. Esto significa, según el criterio de MPA, que ninguna de las cuatro formas la PSU de matemática es suficientemente unidimensional para poder aplicar los modelos IRT.



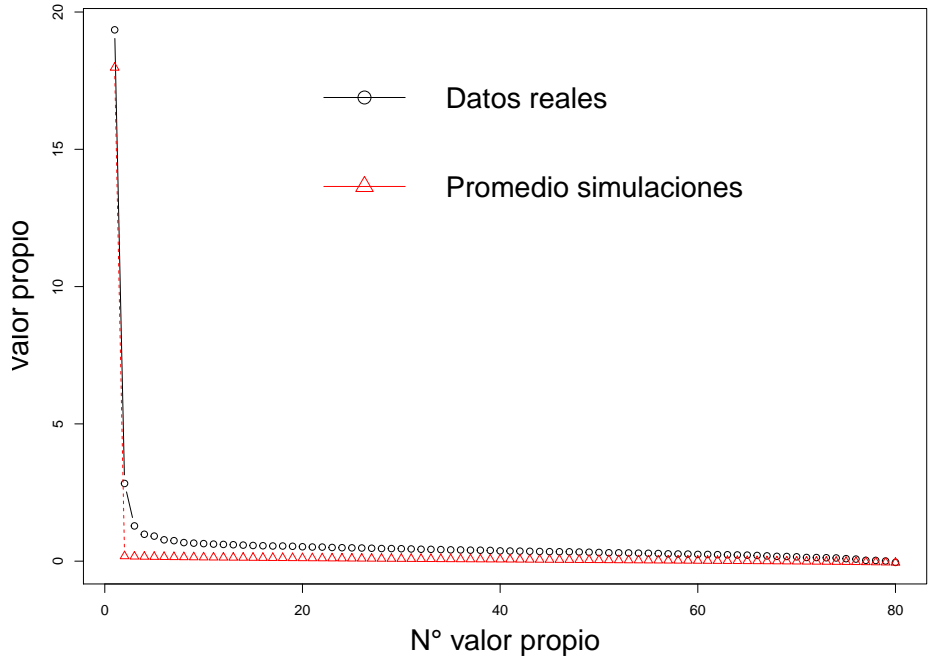
(a) MPA aplicado a la forma 1.



(b) MPA aplicado a la forma 2.



(c) MPA aplicado a la forma 3.



(d) MPA aplicado a la forma 4.

Figura 3.17: Resultado de MPA, aplicado a cada forma de la prueba por separado.

3.6. Aplicación de Modelos MIRT

Con los resultados expuestos en las secciones precedentes y dado que, según el MPA realizado, los datos no son suficientemente unidimensionales para aplicar los modelos IRT, resulta natural preguntarse si la aplicación de modelos IRT multidimensionales podría enmendar esta situación y aportar mayor información al análisis de los resultados de la PSU.

Para comenzar a responder esta interrogante, se ha decidido considerar un enfoque distinto al anterior. En este caso, se desea contrastar los resultados obtenidos de un análisis MIRT con los que se obtienen con IRT unidimensional sobre grupos específicos de estudiantes.

Con este propósito, se realizará el análisis de los datos usando el modelo 2PL y, en paralelo, la extensión multidimensional del modelo 2PL, con cuatro dimensiones. Se ha escogido el modelo 2PL por ser un modelo simple (más simple que 3PL, lo que significa menor tiempo de cómputo) y que, en su extensión multidimensional, posee una gran ventaja sobre la extensión del modelo Rasch, pues no hay que dar valores a priori sobre los valores de las componentes del vector \mathbf{a} , sino que estos son parámetros a estimar dentro del modelo (ver sección 2.1.1). Por otro lado, se ha decidido utilizar la extensión multidimensional considerando cuatro dimensiones de acuerdo con la existencia de cuatro ejes temáticos dentro de la prueba, sin embargo, ésta fue una decisión arbitraria a falta de un criterio sólido para determinar la cantidad de dimensiones a considerar y podría no ser la dimensión óptima.

En los párrafos que siguen se presentan los grupos de estudiantes analizados y los resultados obtenidos.

Históricamente se ha observado que el rendimiento general del grupo femenino es menor que el del grupo masculino en la PSU de matemática. Se trata de un fenómeno que se advierte en pruebas educacionales que se aplican en diversas partes del mundo, lo que justifica el estudio desagregado por sexo. De la población total de alumnos que rindieron la PSU de matemática el año 2015, el 53.1% son mujeres y 46.9%, varones. Considerando esto, se genera una muestra aleatoria formada por 531 mujeres y 469 hombres.

Los resultados del análisis de esta muestra a través de distintas técnicas se resumen a continuación en la tabla 3.11.

Tabla 3.11: Comparación de resultados por género, usando IRT y MIRT

	Teoría Clásica	IRT		MIRT			
	N°respuestas correctas	θ_{Rasch}	$\theta_{2\text{PL}}$	θ_1	θ_2	θ_3	θ_4
Mujeres	26.0301	-0.0192	-0.0554	0.1795	0.0068	-0.0800	-0.0027
Hombres	27.1748	0.0596	0.0212	0.0931	0.0202	0.0337	-0.0200

Si bien la teoría clásica y los modelos IRT unidimensionales indican que, en promedio, las mujeres rinden menos que los hombres en la PSU, el análisis a través del modelo MIRT muestra que existen dimensiones donde las mujeres tienen mayor habilidad que los hombres. Lo que falta es comprender cuál es la interpretación de estas dimensiones, ¿en qué aspectos de la habilidad el género femenino supera al género masculino?.

De manera similar a las desigualdades observadas entre géneros, es sabido que existen diferencias concretas en el rendimiento de estudiantes de distintos tipos de establecimientos educacionales. En general, los resultados de alumnos provenientes de colegios particulares superan a los resultados de colegios subvencionados, mientras que los colegios municipales suelen tener resultados por debajo de los dos anteriores. Por este motivo, se incluye en esta sección el análisis de una muestra de estudiantes considerando el tipo de establecimiento del que provienen.

Del total de alumnos que rindieron la prueba, el 10.6 % provenía de colegio particular, el 54.7 % cursó sus estudios en colegio subvencionado, mientras que el 34.7 % lo hizo en colegio municipal. De esta manera, se genera una muestra aleatoria conformada por 106 estudiantes de colegio particular, 547 de colegio subvencionado y 347 de colegio municipal.

Los resultados se muestran en la tabla 3.12 a continuación.

Tabla 3.12: Comparación de resultados por tipo de establecimiento educacional, usando IRT y MIRT

	Teoría Clásica	IRT		MIRT			
	N°respuestas correctas	θ_{Rasch}	$\theta_{2\text{PL}}$	θ_1	θ_2	θ_3	θ_4
Particular	39.3302	0.8942	0.9273	-1.0157	-0.8010	0.8790	0.4472
Subvencionado	26.2303	-0.0216	-0.0516	0.1640	0.0307	-0.0845	-0.0599
Municipal	23.5216	-0.2121	-0.2528	0.4443	0.1573	-0.2302	-0.1659

Al igual que en el caso de la comparación por géneros, los resultados de la teoría clásica y de IRT indican que los alumnos provenientes de colegios particulares tienen un mejor rendimiento que los de colegios subvencionados, quienes a su vez superan a los de colegios municipales. Por su parte, el modelo MIRT revela que existen dimensiones donde este orden se invierte y los colegios municipales tienen más desarrolladas ciertas dimensiones de la habilidad que los otros dos tipos de instituciones.

Los resultados de esta sección muestran la potencia que podría tener el análisis de la PSU a través de modelos MIRT, ya que permitiría tener una mayor comprensión sobre cómo abordan las preguntas ciertos grupos de estudiantes o incluso la creación de preguntas que potencien ciertas habilidades. Sin embargo, se hace necesario un estudio de la relación entre cada dimensión de habilidad y algún rasgo observable de los estudiantes que proporcione una interpretación concreta de estos parámetros, por ejemplo, si se trata efectivamente de los cuatro ejes temáticos de la prueba o si se trata de otro tipo de habilidades como comprensión de lectura, concentración u otro.

3.7. Comparación de Modelos

En la sección precedente (3.6) se ha observado que el análisis de los datos a través de modelos multidimensionales (en este caso, cuatro dimensiones) permite abordar aspectos de los resultados de la PSU que escapan del alcance de la teoría clásica y de los modelos IRT

unidimensionales. Por este motivo se requiere una mayor comprensión de los parámetros que dan forma a los modelos MIRT.

Sin embargo, antes de comenzar a investigar la interpretación de cada dimensión de la habilidad y su relación con rasgos observables, es pertinente realizar un test que indique cuál modelo, unidimensional o multidimensional, se ajusta mejor a los datos. El método utilizado en este trabajo está inspirado en el test de razón de verosimilitud, realizando simulaciones de Montecarlo. Considerando como hipótesis nula, H_0 , que los datos se ajustan mejor al modelo unidimensional, la idea es encontrar una aproximación de la distribución del estadístico diferencia de log-verosimilitud bajo H_0 , que es equivalente a la razón de verosimilitud. Una vez encontrada dicha distribución, se puede obtener un intervalo de confianza y testear si la diferencia de log-verosimilitud de los datos reales se encuentra dentro del intervalo y así decidir si rechazar o no la hipótesis nula.

Formalmente, sea ℓ_0 la log-verosimilitud obtenida al aplicar el modelo unidimensional a los datos y sea ℓ_1 la log-verosimilitud obtenida al aplicar el modelo multidimensional a los datos, entonces la hipótesis nula y la hipótesis alternativa se expresan, respectivamente, como:

$$H_0 : \ell_1 - \ell_0 = 0 \quad \text{y} \quad H_1 : \ell_1 - \ell_0 > 0.$$

El método se detalla a continuación:

- (i) Se realizan simulaciones de Montecarlo bajo la hipótesis nula. Para esto, se consideran los 120 ítems de la PSU con los parámetros estimados según el modelo 2PL (ver sección 3.3.2) y se considera una muestra de estudiantes hipotéticos, cuyas habilidades se encuentran equidistribuidas en el intervalo $[-4,4]$. Para este estudio, se han simulado 200 estudiantes.
- (ii) Con esta información, se calcula la probabilidad de cada estudiante de responder correctamente al ítem i , según su habilidad θ , de acuerdo a la ecuación (1.6). Conociendo esa probabilidad, se puede simular la respuesta de un estudiante a cierto ítem como una variable Bernoulli, obteniendo una matriz de respuestas simulada.
- (iii) Este proceso se repite según el número de simulaciones deseadas. Para este caso en particular, se realizaron 1000 simulaciones. Cabe destacar que las simulaciones están bajo H_0 , pues sólo se ha considerado una dimensión de habilidad y la probabilidad de respuesta correcta a un ítem ha sido calculada según el modelo 2PL unidimensional.
- (iv) Las simulaciones son sometidas a un análisis IRT unidimensional con el modelo 2PL y, paralelamente, a un análisis MIRT con la extensión a cuatro dimensiones del modelo 2PL. En ambos casos, se registra el valor de la log-verosimilitud obtenida.
- (v) Se calcula la diferencia entre la log-verosimilitud del modelo multidimensional y la log-verosimilitud del modelo unidimensional para cada muestra (que es equivalente a calcular el logaritmo de la razón de verosimilitud). El estadístico observado tiene media empírica $\bar{x} = 252,0566$ y varianza empírica $s^2 = 1013,3904$.

- (vi) Se considera cada una de las mil realizaciones del experimento como una variable aleatoria X_i , con $i \in \{1, \dots, 1000\}$. Donde cada X_i es suma de 200 variables aleatorias independientes (pues las respuestas de los 200 estudiantes fueron simuladas independientemente) que se pueden considerar normales.

Es decir:

$$X_i = \sum_{j=1}^{200} Y_{ij}, \quad \text{donde } Y_{i,j} \sim \mathcal{N}(\mu, \sigma^2) \quad \forall i \in \{1, \dots, 1000\}. \quad (3.2)$$

- (vii) Gracias a los puntos (v) y (vi) se puede obtener una estimación para μ y para σ^2 , pues gracias a la independencia de las variables $\{Y_{ij}\}_{j=1}^{200}$, se tiene que:

$$\mathbb{E}(X_i) = 200\mu, \quad \text{Var}(X_i) = 200\sigma^2, \quad \forall i \in \{1, \dots, 1000\}. \quad (3.3)$$

Con esto, se obtiene:

$$\mu \approx \frac{\bar{x}}{200} = 1,2603, \quad \sigma^2 \approx \frac{s^2}{200} = 5,0670. \quad (3.4)$$

- (viii) Considerando ahora la población completa, es decir, $N = 252745$ estudiantes. La diferencia de log-verosimilitud se representa mediante la variable aleatoria X_p , que distribuye como $\mathcal{N}(\mu_{X_p}, \sigma_{X_p}^2)$.

Bajo la hipótesis nula, el valor de $\ell_1 - \ell_0$ debería ser muy pequeño, por lo que se considera $\mu_{X_p} = 0$.

Para aproximar la varianza $\sigma_{X_p}^2$ se utiliza las simulaciones de Montecarlo. Similarmente al paso (vi), se puede considerar X_p como suma de variables aleatorias independientes Y_{pj} .

$$X_p = \sum_{j=1}^N Y_{pj}, \quad \text{donde } Y_{pj} \sim \mathcal{N}(0, \sigma^2). \quad (3.5)$$

De esta manera, $\sigma_{X_p}^2 = N\sigma^2$. En consecuencia, $X_p \sim \mathcal{N}(0, N\sigma^2)$ y se obtiene que el intervalo de confianza al 95 % para el valor de $\ell_1 - \ell_0$ es:

$$I = (-\infty ; z_{0,05}\sigma\sqrt{N}] = (-\infty ; 1861, 41] \quad (3.6)$$

- (ix) Finalmente, se calcula la diferencia de log-verosimilitudes entre el modelo multidimensional y el unidimensional, $\ell_1 - \ell_0$, aplicados a los datos reales de los $N = 252745$ estudiantes. Y se verifica si el estadístico se encuentra dentro del intervalo de confianza para rechazar o no la hipótesis nula.

En este caso, se obtienen los siguientes resultados:

$$\ell_0 = -1,0374 \times 10^7 \quad (3.7)$$

$$\ell_1 = -1,0339 \times 10^7 \quad (3.8)$$

$$\ell_1 - \ell_0 = 35188,69 \quad (3.9)$$

Por lo tanto, $\ell_1 - \ell_0 \notin I$.

Además, el p-valor calculado es $p \approx 0$, un valor extremadamente pequeño. Por lo tanto, la hipótesis nula es rechazada fuertemente.

Conclusión

El desarrollo e implementación de pruebas estandarizadas en el ámbito educacional es un tema relevante a nivel social, ya que es el medio de evaluación y medición del aprendizaje y del nivel de conocimiento de los escolares. En el caso de la PSU, se trata del sistema para realizar un ranking de los estudiantes del país, según su nivel de conocimiento, al finalizar su enseñanza media. El propósito de este proceso es la admisión de los alumnos en las universidades.

Históricamente, el análisis de este instrumento se ha llevado cabo a través de la teoría clásica. Sin embargo, son conocidas las limitaciones de esta teoría, por lo que existe un creciente interés de parte de los especialistas en implementar nuevas técnicas que complementen sus resultados [21]. En particular, progresivamente se ha ido incorporando la teoría de respuesta al ítem (IRT) como método de análisis de la PSU [10].

En este trabajo de tesis se ha realizado un estudio detallado de los modelos IRT existentes, así como un análisis de los resultados de esta teoría al ser aplicada sobre los datos de la PSU de matemática, admisión 2016, con el objetivo de estudiar si se cumplen los supuestos para la aplicación de estos modelos.

Las principales conclusiones de esta investigación se enumeran a continuación:

1. Se realizó un estudio de la robustez de los modelos Rasch y 2PL, aplicándolos sobre distintas muestras de la población para estudiar el comportamiento de las estimaciones de los parámetros de los modelos. Se observó que el modelo Rasch resulta ser considerablemente más robusto que 2PL pues presenta una variabilidad mucho menor. En este sentido, 2PL no es un modelo confiable y no se recomienda su utilización para el análisis de la PSU. En cambio, Rasch -aplicado a la PSU- presenta la propiedad de invarianza que distingue a los modelos IRT de la teoría clásica.
2. Los estadísticos que representan rasgos como la habilidad de una persona y la dificultad de un ítem tienen el mismo comportamiento en la teoría clásica y en el modelo Rasch. Por lo tanto, en este sentido Rasch no aporta mayor información que la teoría clásica. El modelo 2PL, por su parte, presentó una variabilidad muy alta en las estimaciones, por lo que sus resultados no serían confiables para un proceso de admisión universitaria. Además, como se observó en la sección 3.4.3, la habilidad estimada a través de los modelos IRT no debería ser usada como clasificador de los estudiantes en la PSU, ya que resulta difícil de explicar a la población que alumnos con igual número de respuestas correctas sean evaluados con distinta habilidad. En este sentido, los resultados de los análisis IRT sólo podrían complementar los análisis de la teoría clásica.

3. Los datos de la PSU fueron sometidos a un test de unidimensionalidad, el análisis paralelo modificado, cuyos resultados indican que el conjunto de preguntas que forman esta prueba no es suficientemente unidimensional. En otras palabras, la PSU es un instrumento complejo que no cumple con los supuestos necesarios para poder aplicar los modelos IRT convencionales y los resultados de su aplicación no serían concluyentes.

Cabe destacar que existen alternativas para sortear la violación del supuesto de unidimensionalidad, entre ellas:

- (i) Eliminar ítems hasta obtener un subconjunto unidimensional, como en el método propuesto por Budescu [3].
 - (ii) Usar MIRT, identificando correctamente el número de dimensiones a considerar.
4. Se realizaron pruebas comparativas entre el modelo 2PL unidimensional y la extensión de este modelo a cuatro dimensiones. El test realizado en la sección 3.7 indica que los datos de la PSU se adaptan mejor al modelo multidimensional que al unidimensional. Además, la aplicación de modelos multidimensionales proporcionaría mayor información sobre las habilidades de los estudiantes. Este último punto resulta interesante para entender el comportamiento de grupos específicos de personas. A modo de ejemplo, en la sección 3.6 se presenta una comparación de alumnos según el tipo de institución de procedencia. Cuando se compara sus resultados, la teoría clásica e IRT unidimensional muestran que, en promedio, los alumnos de colegios municipales tienen un rendimiento menor que los alumnos que provienen de colegios particulares; mientras que un análisis MIRT con cuatro dimensiones refleja que existen dimensiones de la habilidad donde los alumnos de colegios municipales superan, en promedio, a los de colegios particulares.

Limitaciones de este trabajo

Hay varios aspectos de este trabajo que podrían extenderse o analizarse desde otra perspectiva. Algunos de ellos se listan a continuación.

- (i) Este trabajo se ha limitado al estudio de la aplicación de los modelos Rasch y 2PL a la PSU de matemática. Sería conveniente extender este estudio al modelo 3PL, considerando que a partir del año 2014, a diferencia de las versiones anteriores de la prueba, las preguntas respondidas incorrectamente no descuentan puntos. Esta medida implicó una disminución significativa de la omisión y, en consecuencia, la *adivinanza* debe haber aumentado. Es importante medir los efectos del tercer parámetro en el análisis IRT de la PSU.
- (ii) En la sección 3.3 se estudiaron las estimaciones de los parámetros de los ítems (dificultad y discriminación) según el eje temático de la pregunta. Sin embargo, este estudio podría extenderse al comportamiento de los parámetros de los ítems según el nivel (1°, 2°, 3° o 4° medio) al que corresponde la pregunta.
- (iii) Para los análisis IRT realizados se consideró siempre la matriz completa de respuestas,

con las 120 preguntas. En particular, en las secciones 3.3.1 y 3.5.2 se evidencian los efectos que genera la falta de información que tiene la matriz completa, ya que no todos los alumnos responden las mismas preguntas. Por este motivo, sería conveniente hacer los mismos análisis considerando cada forma por separado.

- (iv) Respecto a la comparación entre el modelo 2PL unidimensional y su extensión a cuatro dimensiones expuesta en 3.7 hay al menos dos aspectos que se podrían modificar. En el punto (i) se expuso que los datos simulados se generan a partir de 200 estudiantes ficticios, cuyas habilidades están equidistribuidas en $[-4, 4]$. En primer lugar, 200 estudiantes parece ser muy poco, dado que se consideran los 120 ítems. Sería ideal repetir el experimento con una mayor cantidad de estudiantes, de manera que sea óptima para el número de ítems. Este trabajo se limitó a 200 debido a los tiempos de cómputo del modelo de cuatro dimensiones.

Por otro lado, las 200 habilidades se consideraron equidistribuidas, sin embargo, es razonable repetir el experimento con habilidades que siguen una distribución normal.

Trabajo futuro

Existen principalmente tres líneas de continuación de este trabajo, las dos primeras son aspectos de los modelos MIRT que es necesario estudiar para su correcta implementación y utilización futura, y la última guarda relación con la elección entre habilidad IRT y puntaje PSU actual.

- (i) Por una parte, se requiere el desarrollo de alguna técnica para determinar de manera efectiva el número de dimensiones a considerar cuando se lleva a cabo el análisis con los modelos multidimensionales. En este caso, se tomó una decisión arbitraria que coincidiera con el número de ejes temáticos evaluados en la prueba, sin embargo, posiblemente éste no es el número óptimo para realizar el análisis.
- (ii) Una vez indentificado el número óptimo de dimensiones a considerar, es necesario encontrar la relación de cada dimensión con algún rasgo observable en los individuos que permita dar una interpretación concreta de los parámetros del modelo.
- (iii) La elección final de qué indicador usar como *ranking* de los alumnos en la PSU, ya sea la habilidad Rasch, la habilidad 2PL, la habilidad 3PL o el puntaje PSU actual, debería basarse netamente en cuál es el mejor predictor del rendimiento universitario de los alumnos. Lamentablemente, para este trabajo no se disponía de los datos necesarios para hacer un estudio de este aspecto, pero sería muy interesante efectuarlo.

Bibliografía

- [1] Frank B. Baker. *The basics of item response theory*. ERIC, 2001.
- [2] Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.
- [3] David V. Budescu et al. A revised modified parallel analysis (rmpa) for the construction of unidimensional item pools. 1993.
- [4] Gregory Camilli. Origin of the scaling constant $d = 1.7$ in item response theory: Correction. 1995.
- [5] R. Philip Chalmers et al. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012.
- [6] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.
- [7] Rafael Jaime De Ayala. *The theory and practice of item response theory*. Guilford Publications, 2013.
- [8] Fritz Drasgow and Robin I. Lissak. Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied psychology*, 68(3):363, 1983.
- [9] Fritz Drasgow and Charles K. Parsons. Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7(2):189–199, 1983.
- [10] Pearson Education. Final report: Evaluation of the chile psu. Technical report, 2013.
- [11] Ronald K. Hambleton and Hariharan Swaminathan. *Item response theory: Principles and applications*. Springer Science & Business Media, 1985.
- [12] Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. *Fundamentals of Item Response Theory*. Sage Publications, 1991.
- [13] John Hattie. Methodology review: assessing unidimensionality of tests and items. *Applied psychological measurement*, 9(2):139–164, 1985.

- [14] John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.
- [15] Matthew S. Johnson et al. Marginal maximum likelihood estimation of item response models in r. *Journal of Statistical Software*, 20(10):1–24, 2007.
- [16] FM Lord. Applications of item response theory to practical testing problems lawrence erlbaum associates hillsdale. *NJ Google Scholar*, 1980.
- [17] Frederic M. Lord, Melvin R. Novick, and Allan Birnbaum. *Statistical theories of mental test scores*. Addison-Wesley, 1968.
- [18] Stanley A Mulaik. *Linear causal modeling with structural equations*. CRC Press, 2009.
- [19] Mark Reckase. *Multidimensional item response theory*, volume 150. Springer, 2009.
- [20] Dimitris Rizopoulos. ltm: An r package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5):1–25, 2006.
- [21] Educational Testing Services. Evaluación externa de las pruebas de selección universitaria (psu). Technical report, 2005.