Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería

Área Académica de Computación y Electrónica

Doctorado en Ciencias Computacionales

Efectos de la distribución prior y la entropía en la estimación de habilidades en CAT por Maximum a Posteriori

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Doctor en Ciencias Computacionales

PRESENTA:

Luis Roberto Morales Manilla

DIRECTORES DE TESIS:
Joel Suárez Cansino
Virgilio López Morales

Pachuca de Soto, Hidalgo; Noviembre, 2023

Área Académica de Computación y Electrónica

Mineral de la Reforma Hidalgo, a 31 de octubre de 2023

Número de control ICBI-AACyE/1256/2023
Asunto: Autorización de impresión de tema de tesis

MTRA. OJUKY DEL ROCIÓ ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

El Comité Tutorial de la TESIS del Programa Educativo de Posgrado titulado "**Efectos de la distribución prior y la entropía en la estimación de habilidades en CAT por Maximum a Posteriori**", realizado por el sustentante **M. en C. Luis Roberto Morales Manilla**, con número de cuenta **096606** perteneciente al programa de Doctorado en Ciencias Computacionales, una vez que ha revisado, analizado y evaluado el documento recepcional de acuerdo a lo estipulado en el Artículo 110 del Reglamento de Estudios de Posgrado, tiene a bien extender la presente:

AUTORIZACIÓN DE IMPRESIÓN

Por lo que el sustentante deberá cumplir con los requisitos del Reglamento de Estudios de Posgrado y con lo establecido en el proceso de grado vigente.

UNIVERSIDAD AUTÓNOMA DE HIDALGO

Atentamente
"Amor, Orden y Progreso"

Dr. Joel Suárez Cansino
Director de Tesis

Dr. Virgilio López Morales
Codirector de Tesis

Instituto de Ciencias Básicas e Ingeniería
Área Académica de Computación y Electrónica

Comité Tutorial

Dr. Ismael Dominguez Jiménez	Presidente	UAEH
Dr. Virgilio López Morales	Secretario	UAEH
Dr. Joel Suárez Cansino	Vocal	UAEH
Dra. Martha Idalid Rivera González	Suplente	UAEH

OPC/APL

Ciudad del Conocimiento
Carretera Pachuca-Tulancingo Km. 4.5, Colonia
Carboneras, Mineral de la Reforma, Hidalgo,
México, C.P. 42184
Teléfono: +52 (771) 71 720 00 ext. 2250, 2251
Fax 2109

Quiero dedicar este trabajo a todos los actores que en esta etapa de mi vida han sido sustanciales para poder concluir mis estudios doctorales.

A mis hijos Leonardo y Paulina, porque desde que llegaron a mi vida, iluminaron el camino que debía yo seguir y se volvieron el motor de motivación para superarme día a día. Deseo que este documento sea para ustedes un ejemplo de formación académica y que se motiven a estudiar y alcanzar sus metas sin nunca rendirse.

A mi esposa Sandy, porque has estado a mi lado desde que empecé este camino del doctorado y te ha tocado aguantar mis ausencias y desvelos para poder concluir este trabajo a lo largo de todos los años que me tomó poderlo concluir. Sé que no fue nada fácil para ti y los peques, pero te agradezco tu paciencia y comprensión. Te amo.

A mis dos madres, mi mamá Amelia y mi mamá Lulú, porque siempre estuvieron ahí cuando todo era difícil y me dedicaron solo palabras de aliento para continuar adelante. Porque no me dejaron rendirme y me exhortaron a que concluyera lo que empecé. Las amo.

A mi tío Felipe, porque siempre mantuve en mente tus palabras "debes llegar hasta el final", gracias por esa inspiración que me diste.

A mi asesor el Doc Joel Suárez por demostrarme que por mucho que uno se prepare siempre hace falta ir más allá. Por enseñarme lo valioso que es filosofar y lo importante que es meditar sobre el camino de uno mismo. Pero sobre todo, por querer acompañarme dándome sus enseñanzas hasta el último día que fui su tesista, gracias por su apoyo Doc, sin usted, esto no hubiera sido posible. Fue, es y será siempre un ejemplo de superación para mi.

Finalmente a mi alma mater, cuatro veces heroica, la Universidad Autónoma del Estado de Hidalgo, porque en ella empece mi formación desde la preparatoria y hoy termino en ella mis estudios de doctorado. Ha sido una institución muy bondadosa conmigo y siempre llevaré muy en alto su nombre donde quiera que yo esté.

Reconocimientos

Agradezco a la Universidad Autónoma del Estado de Hidalgo por todas las facilidades para la conclusión de este trabajo de tesis doctoral.

Resumen

En los últimos años, la Teoría de Respuesta al Ítem, o (IRT) por sus siglas en inglés, ha tomado gran auge al ser aplicada en los sistemas informáticos dedicados a hacer evaluaciones o exámenes, estos sistemas son conocidos como Test Adaptables Computarizados (CAT en inglés).

Dentro de esta área de la informática, la forma en la que se realiza la evaluación incluye la selección de un modelo psicométrico. Existen diferentes modelos, algunos de ellos son de tipo logístico, los cuales se encuentran definidos en términos de parámetros, cada parámetro tiene su correspondiente interpretación en los resultados que la evaluación arroja; sin embargo, no existe una razón de peso que indique cuáles de éstos parámetros son los mejores.

Se dice que los modelos paramétricos de estimación de Bayes son mejores dado que aportan mayor información, y dentro de estos los más usados son la Estimación A Posteriori (EAP) y la Estimación Máxima A Posteriori (MAP).

El objetivo de este trabajo se centra en estudiar la técnica MAP, sus principales características y los problemas que presentan a la hora de su aplicación, como son:

(I) Falta de información para garantizar el uso de una distribución *prior* en específico.

(II) Falta de evidencia que relacione la situación real y la simulación.

(III) La elección de la distribución *prior* no es trivial ya que impacta de manera directa sobre la estimación de la habilidad del sujeto que realiza la prueba o test.

Dicho lo anterior, en este trabajo se propone el uso del concepto de entropía como una alternativa para construir distribuciones *prior* de manera teórica, pudiendo estas ser usadas dentro de la técnica MAP para el cálculo de habilidades dentro de sistemas de evaluación adaptables computarizados.

Aunado a lo anterior, se contempla la posibilidad de que dentro del contexto de la evaluación pudieran existir datos que produzcan distribuciones multimodales o asimétricas, es decir, sesgadas. Entonces la aplicación de la técnica MAP conducirá a una estimación de la habilidad, mediante los parámetros de los modelos psicométricos usados por el sistema de evaluación adaptable, mucho más apegada a la realidad y en consecuencia permitirá arrojar datos más confiables en el ámbito de la prueba que se esté realizando.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En este capítulo se presenta la problemática central de este trabajo y la propuesta de solución que se da a la misma. Se plantean los objetivos a perseguir tanto generales como particulares, así como también las metas que se alcanzarán. Por otro lado, se da una semblanza de los capítulos que conforman esta tesis doctoral.

1.1. Presentación

Desde sus inicios en los años 60's hasta la actualidad, la teoría de respuesta al ítem, o (IRT) por sus siglas en inglés, ha experimentado un progreso incesante en diferentes aplicaciones. Una de estas aplicaciones se ha visto reforzada por el acelerado desarrollo tecnológico en el área de la computación, ya que esto ha permitido la construcción de los sistemas informáticos dedicados a hacer evaluaciones o exámenes, los cuales son conocidos como Test Adaptables Computarizados (ó CAT por sus siglas en inglés).

En este tipo de sistemas informáticos se requiere de la selección de un modelo psicométrico para poder realizar la evaluación. La IRT tiene dentro de sus fundamentos la existencia de algunos modelos que son de tipo logístico, los cuales se definen mediante parámetros. Cada parámetro tiene una interpretación que corresponde con los resultados que se obtienen de la evaluación para un contexto dado o área de conocimiento previamente seleccionada.

Los modelos más conocidos son los denominados 1PL, 2PL y 3PL, aunque también se han llegado a proponer los modelos 4PL y 5PL. El número que antecede en cada nombre está asociado con el número de parámetros que definen la estructura del modelo, el cual es de carácter logístico y su principal orientación es reducir el tiempo en los cálculos computacionales durante el proceso de evaluación.

Particularmente interesantes son los modelos 3PL y 4PL, ya que contienen parámetros poco usuales cognitivamente hablando. Por supuesto, la selección del modelo puede sustentarse desde un punto de vista teórico o bien desde un punto de vista experimental. Sin importar qué modelo se escoja, se requiere aplicar un procedimiento apropiado para determinar en forma confiable los parámetros correspondientes.

El problema al cual se enfrenta el administrador de la evaluación consiste en la determinación de los valores de estos parámetros. Para resolverlo, existen al menos dos procedimientos conocidos como Estimación por Máxima Verosimilitud (MLE) y estimación por *Maximum a Posteriori* (MAP). Particularmente, el primero necesita de una función de densidad conjunta, mientras que el último emplea un método de estimación paramétrica basado en inferencia Bayesiana.

El principal inconveniente que se le encuentra a la técnica basada en MLE, es que falla cuando el examinando aporta solamente respuestas correctas o solamente incorrectas a todos los ítems de una prueba. La solución a este problema está dada por el uso de información previa sobre la distribución de la habilidad del examinando, lo que conduce al concepto de distribución *prior*. Típicamente se asume una distribución normal (48).

Estos modelos de estimación paramétrica basados en inferencia Bayesiana son mejores en comparación con los basados en Máxima Verosimilitud y en la literatura sobre el tema se puede encontrar una comparación de ambas técnicas (79). Sin embargo, su empleo requiere también resolver ciertas dificultades, una de las cuales se encuentra asociada a la selección adecuada de una distribución *prior*. Una dificultad adicional es la que se refiere a la posible obtención de distribuciones *a posteriori* que sean multimodales (31, 53).

Como es sabido, la presencia de dos o más modos en una distribución *a posteriori* puede conducir a estimaciones paramétricas que no sean las adecuadas debido a la existencia de un igual número de máximos locales, dificultando así la localización de extremos absolutos y por lo tanto, la determinación de los parámetros adecuados (53).

Hasta donde se sabe, actualmente no se cuenta con criterios que permitan determinar qué estructuras de la distribución *prior* pueden conducir a la obtención de distribuciones *a posteriori* multimodales y con sesgos, sin mencionar los criterios que permitan seleccionar el extremo óptimo de la distribución *a posteriori*.

En este trabajo de investigación se presenta una alternativa para la estimación de parámetros de modelos psicométricos usados en un proceso de evaluación adaptable, aplicando el concepto de entropía para la construcción de distribuciones *prior*, ocupando estas últimas en la técnica *máximum a posteriori*.

1.2. Problemática

En la actualidad no existen criterios que indiquen bajo qué condiciones se obtienen distribuciones *a posteriori* multimodales y/o asimétricas, es decir, sesgadas, especificando previamente las condiciones que llevan a descubrir esta característica. Esto representa una situación que afecta negativamente el proceso de evaluación de un sistema CAT.

Por tal motivo, la propuesta de solución que contempla el uso de modelos psicométricos alternativos y la aplicación de distribuciones *prior*, debe incluir aquellos trabajos en los que se establecen teoremas que provean criterios para detectar condiciones que

indiquen la presencia de múltiples modos en distribuciones *a posteriori*, resaltando el hecho de que el análisis tiene que realizarse usando algunos modelos psicométricos.

Es decir, se debe dar una idea del tipo de distribuciones *prior* que se puedan utilizar en la técnica MAP. Ligado a ello, se tienen que mencionar también aquellos teoremas que detectan la presencia de sesgos positivos o negativos en la distribución *a posteriori* (8, 12, 21, 22, 25, 30, 41, 60, 62, 70, 71, 76).

1.2.1. Planteamiento del problema

La estimación de la habilidad de un sujeto de prueba presenta problemas al inicio del proceso de la evaluación cuando se emplea el método de máxima verosimilitud, específicamente cuando el evaluando responde correcta o bien incorrectamente a todos los ítems de la prueba. Una situación similar ocurre cuando todos los evaluandos responden correcta o incorrectamente a por lo menos alguno de los ítems.

Esto representa un serio problema y dentro de la literatura se puede encontrar que se han propuesto alternativas heurísticas para resolver este problema y también algunas opciones basadas en inferencia Bayesiana(72). De manera particular, la técnica MAP sugiere la introducción del concepto de distribución *prior*.

Sin embargo, aún dentro de este contexto no se tiene suficiente información acerca de la mejor distribución *prior* que debe seleccionarse. Debido a su naturaleza de corte Bayesiana, la técnica MAP requiere un conocimiento previo de la distribución *prior*, la cual contiene información estadística inicial acerca de la habilidad del sujeto examinado.

Típicamente se utiliza una distribución normal (48, 65, 79), pero no hay evidencia de que esto sea necesariamente lo correcto, incluso aparentemente no existe forma alguna que sustente la decisión de optar por un tipo de distribución *prior* en lugar de otro. Este es un punto muy importante ya que la elección inicial de la distribución *prior* afecta de manera directa el cálculo de la estimación de la habilidad y de otros parámetros.

Por supuesto, también debe considerarse en lo anterior la estructura del modelo psicométrico sobre el cual se basa la inferencia Bayesiana. Una selección adecuada de estructura debe proporcionar una interpretación cognitiva de cada uno de sus parámetros, predecir las consecuencias de emplear un modelo psicométrico con las características seleccionadas, así como la relación que guardan estas opciones con la características de multimodalidad y sesgos en la distribución *a posteriori*, la cual se usará finalmente para estimar los parámetros, entre los cuales se encuentra incluida la habilidad del sujeto de prueba y los que definen al modelo psicométrico seleccionado.

En este punto radica la importancia del problema, ya que se debe de saber, qué forma debe tener la distribución *prior* para que se logren estos efectos. Por tal motivo, un concepto sumamente importante es el que se refiere al criterio para seleccionar una buena distribución *prior*. Existen algunos autores que definen algunos de estos criterios y dan ejemplos bastante ilustrativos acerca de la forma en que la selección de la distribución *prior* afecta la distribución *a posteriori* (9, 44). Sin embargo, en esta investigacion se trabaja principalmente con el concepto de entropía, de primera instancia con el concepto propuesto por Shannon (11).

1.2.2. Hipótesis

El concepto de entropía está íntimamente ligado al de distribución *prior* y la distribución *a posteriori*. Una definición general de entropía da un mayor conocimiento acerca de las propiedades que deben de tener las distribuciones *prior* y *a posteriori* mencionadas para obtener estimaciones de habilidad y paramétricas que sean confiables.

En particular, existen definiciones de entropía que incluyen además de la de Shannon, a la de Renyi y algunas generalizaciones.

En el presente trabajo, se propone la utilización de diferentes distribuciones *prior* para la técnica MAP, con la finalidad de establecer la relación entre diferentes conjuntos *prior* y la aparición de varios modos y sesgos en las distribuciones *a posteriori*, y cómo ello afecta el cálculo de los parámetros requeridos. La decisión que permitirá seleccionar la distribución prior más adecuada estará basada en el concepto de entropía aplicado a la información previa que se tenga a la mano sobre la población de examinandos a evaluar.

En otras palabras, aquella distribución que contemple la evidencia con la que se cuenta y que máximize el grado de entropía será una buena elección (33). Debido a que en la literatura no existe información suficiente que permita establecer la relación entre la elección de una distribución *prior* y la aparición de modos o sesgos, resulta conveniente su estudio para su posterior experimentación sobre un sistema CAT real.

1.3. Objetivos

El objetivo general de este trabajo es definir desde un punto de vista teórico las condiciones bajo las cuales las distribuciones *prior* conducen a la existencia de distribuciones *a posteriori* con efectos de multimodalidad y/o sesgo, aplicando para ello criterios entrópicos, usando estos conceptos en la estimación de parámetros de modelos psicométricos generalizados de Rasch.

Los objetivos específicos son:

1. Evaluar el impacto de algunas distribuciones prior en el cálculo de la habilidad del sujeto de prueba, aplicando los principios de la teoría de respuesta al ítem al diseño de un sistema de evaluación adaptable computarizado.

2. Analizar el comportamiento de algunas distribuciones prior en el cálculo de los parámetros del modelo logístico al utilizar los modelos 3PL, 4PL o modelos de más parámetros al aplicar la técnica MAP.

3. Proponer modelos más generales a los que existen en la literatura que puedan considerar la multimodalidad y el sesgo dentro de los parámetros que definen su estructura.

4. Determinar los criterios bajo los cuales se selecciona correctamente la distribución prior aplicando el concepto de entropía.

5. Diseñar y desarrollar una propuesta de arquitectura para un sistema de evaluación adaptable computarizado denominado Ariya, siguiendo las últimas tendencias en desarrollo de sistemas basados en la web, usando como infraestructura una plataforma de cómputo en la nube.

1.4. Motivación

Ya se sabe que la determinación de la distribución *prior* es experimental o a través de consulta a expertos. Sin embargo, poco se ha hecho en relación con el papel que juega el concepto de entropía en la búsqueda de una distribución *prior* adecuada. En este sentido, se requiere del conocimiento previo acerca de la forma en que la distribución *prior* se comporta o, en ausencia de esto, las propiedades que ésta pueda tener (el caso más simple es aquel en el que se sabe que la probabilidad total tiene que ser igual a la unidad, pero puede haber algunas otras propiedades asumidas como conocidas) (10, 33).

Menos aún se ha experimentado con diferentes conceptos de entropía (además de la de Shannon) y el efecto de cada una de ellas en la estimación de los parámetros y las formas de distribución *prior* y, consecuentemente, en las distribuciones *a posteriori*.

El concepto de distribución *prior* juega un papel fundamental en la inferencia Bayesiana, por lo que es importante primeramente determinar experimentalmente cómo obtener dichas distribuciones y también cuáles serían los métodos teóricos para obtener algo similar. Para la construcción de una distribución *prior*, se necesita primero especificar la variable aleatoria, en el caso que se plantea en este trabajo existen varias posibilidades en este sentido.

Sin embargo, en una primera instancia se asume que las distribuciones *prior* deben estar relacionadas con los parámetros del modelo psicométrico seleccionado y la variable o rasgo latente que se desea evaluar asociada al examinando. Desde este punto de vista, en la literatura existen algunos trabajos relacionados con la construcción experimental de las correspondientes distribuciones *prior* (3, 15, 40, 66).

Adicionalmente, y como alternativa al caso de construcción experimental de una distribución *prior*, se opta también por consultar expertos en el dominio del conocimiento que se está evaluando. Con base en una información apropiada, estos expertos son capaces de opinar acerca de la forma o estructura que debe tener la distribución *prior*(3, 40).

Algunos intentos teóricos para determinar la estructura de la distribución *prior* empleando el concepto de entropía (33) se han realizado, aunque no conectados con la idea de evaluación adaptable computarizada. Adicionalmente, no se conoce algún procedimiento que integre los resultados del proceso experimental, además de posiblemente la opinión de expertos, y que con una base teórica, se logren especificar las características o condiciones bajo las cuales se obtengan distribuciones *prior* adecuadas; es decir, que

conduzcan a distribuciones *a posteriori* sin sesgos, sin multimodalidad y a estimaciones que sean confiables.

Son bastante conocidas algunas de las bondades y dificultades a las que conduce la selección de una distribución *prior*, algunas de las cuales ya han sido comentadas en los párrafos anteriores. Desde un punto de vista teórico, existen algunas contribuciones en las que se trata el tema sobre distribuciones *prior* informativas y no informativas, y que son utilizados como ejemplos académicos para demostrar los efectos que tiene la distribución *prior* en la distribución *a posteriori* (74).

1.5. Metodología

La manera en la que este trabajo se realizó tomó como base el método científico. Partiendo de la investigación en el área de los sistemas evaluadores adaptables computarizados, se llegó a la formulación de la hipótesis, la cual surge a través de la curiosidad por aportar mejoras a este tipo de sistemas ocupando metodologías alternativas que permitan enriquecer el desempeño computacional de los algoritmos involucrados en el proceso de evaluación adaptable.

Tras realizar el estudio pertinente del estado del arte sobre los modelos psicométricos involucrados en el proceso de calibración de los ítems y de la evaluación misma, fue posible proponer nuevos modelos de 6 parámetros, que aportan mejoras en el manejo de los posibles escenarios reales que un administrador de la evaluación pudiera encontrar.

Aunado a esto, la experimentación realizada con datos simulados, hizo posible la comprobación de la hipótesis inicial sobre la relación del concepto de entropía con la construcción de distribuciones *prior* la cual arrojó interesantes y novedosos resultados con el apoyo del concepto de índice de divergencia de Kullback–Leibler cuando se utilizan las técnicas *Maximum A Posteriori* ó Expectation A Posteriori. El proceso se puede observar en la figura 1.1.

En suma a lo anterior, no se debe dejar de lado el hecho de que el seguimiento en la aplicación de los conceptos de linearización, entropía, distribución *prior* y distribución *a posteriori*, indicaron las pautas a seguir para en su momento tomar o corregir la línea por la cual habría de ir este trabajo.

Esto es, poder implementar la técnica de *Maximum A Posteriori* con las distribuciones *prior* que se encontraron a través del uso del concepto de entropía junto con los nuevos modelos 6PL que se obtuvieron como una generalización a los ya existentes, mismos que permiten ajustarse a situaciones reales donde existe la posibilidad de obtener distribuciones asimétricas y además multimodales.

1.5.1. Administración de la calidad

El trabajo se dividió en cuatro tópicos fundamentales; a saber, modelo psicométrico, distribución *prior*, distribución *a posteriori* y entropía. Este orden no fue arbitrario, parte considerando los temas básicos y termina con los más complejos. Esto facilitó la

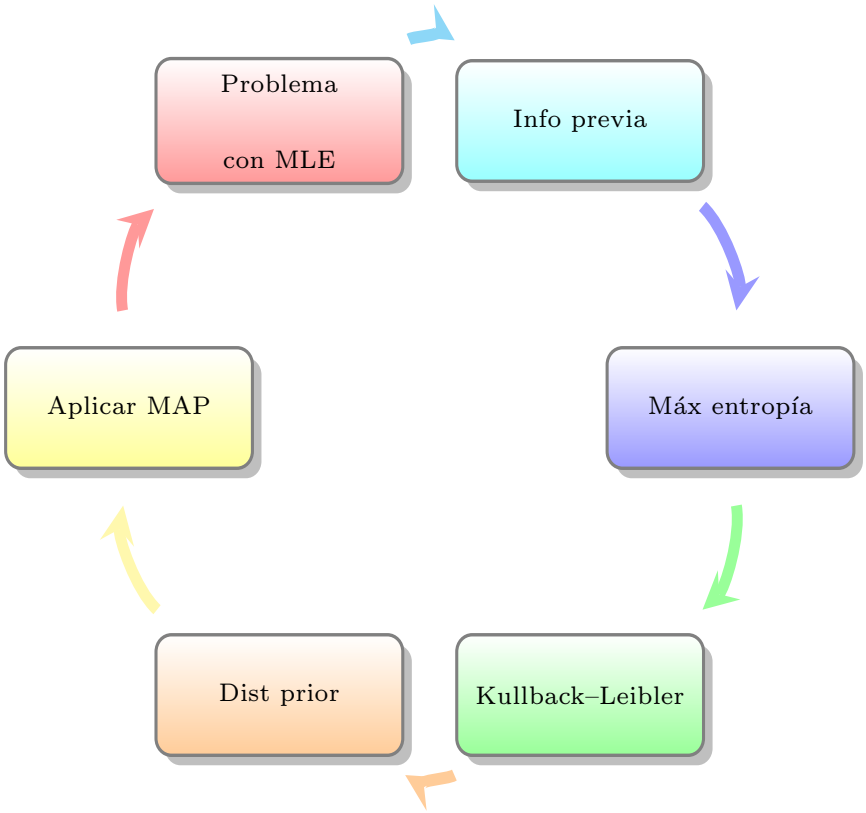

Figura 1.1: Metodología de solución usada en este trabajo de tesis

comprensión y por ende, el desarrollo de la investigación, ya que se fue trabajando de manera tal que al final todos los conceptos estuvieran totalmente relacionados.

Además, cada uno de los tópicos fue analizado en el estado del arte siempre tratando de encontrar su relación con el área de la IRT si es que era posible. La idea era encontrar el soporte teórico para cada uno de los conceptos que se fueron trabajando.

Por otro lado, se consideraron otros aspectos importantes tales como, la pertinencia de la investigación dentro del ámbito de la IRT, la existencia de un soporte rico en materiales bibliográficos que ayude a sustentar la formalidad del trabajo de investigación, la claridad de la redacción de las ideas, su aportación y la validación de las mismas mediante la publicación de al menos un artículo con arbitraje.

1.5.2. Infraestructura

Para la ejecución del proyecto se contó con el equipamiento del Laboratorio de Procesamiento Paralelo Virtual e Inteligente, perteneciente al Cuerpo Académico de Computación Inteligente, en el cual hay tres estaciones de trabajo, dos servidores donde se encuentra alojado el prototipo del sistema de evaluación llamado Ariya de manera local. Se implementó también el prototipo del sistema de evaluación adaptable dentro de un servicio cómputo en la nube para poder utilizar características de alto rendimiento

y crecimiento bajo demanda.

El servicio que se utiliza y donde actualmente quedó implementado el prototipo es la nube ofrecida por la empresa Microsoft llamada Azure. Asimismo, se contó con una conexión de banda ancha de Internet, una computadora personal con características de última generación. Todo el software que se utilizó se encuentra reglamentado de acuerdo al tipo de licencia aplicable a cada uno de ellos.

Para el desarrollo del prototipo del sistema CAT dentro del servicio de cómputo en la nube se contó con la participación de alumnos de la Ingeniería en Sistemas Computacionales de la Universidad Politécnica de Tulancingo, quienes estuvieron liderados por el titular de este proyecto para realizar la implementación del diseño propuesto.

1.6. Contribuciones

El presente trabajo será de gran aportación dentro del área de CAT e IRT ya que a la fecha no se tiene conocimiento de que exista algo similar en el estado del arte. Los modelos propuestos que contienen 6 parámetros son un aporte al área de los sistemas evaluadores adaptables y serán también de utilidad en ambientes de e-learning y sistemas tutoriales inteligentes, por mencionar algunos. Además, las propiedades de estos modelos para poder realizar comparaciones entre examinandos y entre los propios ítems de la prueba son otra aportación que enriquece a este trabajo.

Aunado a esto, está el hecho de que los modelos de 6 parámetros pueden ajustarse mejor a condiciones de evalución donde se presenten distribuciones asimétricas, es decir sesgadas y ademas exista multimodalidad. Este ajuste dará la posibilidad de una interpretación cognitiva de los parámetros del modelo al ámbito del área de conocimiento que se esté evaluando.

Además de lo anterior, también es importante mencionar la propuesta de algunas distribuciones prior, las cuales consideran en su estructura algún tipo de información previa que se puede llegar a tener referente a los examinandos. Cabe destacar la importancia del uso del concepto de entropía para poder construir tales distribuciones garantizando que aporten la mayor información posible. Tales distribuciones son relevantes para poder utilizar la técnica MAP en la estimación de los parámetros de los modelos y así obtener estimaciones más confiables desde el punto de vista cognitivo.

En suma a lo anterior, la experimentación se sometió a la evaluación de expertos del área mediante la redacción de por lo menos un artículo de alto impacto que dio sustento al trabajo realizado y además validó de cierta manera las aportaciones que se describen en esta investigación.

1.7. Estructura de la tesis

El presente documento se encuentra dividido para una mejor lectura y comprensión en las secciones que se describen a continuación.

En el capítulo 2 se concentra el marco teórico que da sustento al trabajo que se realizó, mencionando las técnicas estudiadas durante la realización de esta investigación. Haciendo un énfasis en el método *Maximum A Posteriori* y en los modelos psicométricos utilizados dentro de la Teoría de Respuesta al Ítem que sirvieron como base para obtener modelos más generales. Igualmente es importante mencionar los conceptos de linearización, distribución *prior* y entropía entre otros que se ocuparon a lo largo del trabajo de investigación.

A continuación en el capítulo 3 se detalla el desarrollo del trabajo y las aportaciones teóricas del mismo. La propuesta de nuevos modelos psicométricos más generales que tratan con los efectos de multimodalidad y sesgo de las distribuciones a posteriori. Asimismo, se menciona el planteamiento de la construcción de las distribuciones prior utilizando el concepto de entropía clásica de Shannon y también otros conceptos de entropía que aparecen en la literatura tales como la entropía de Renyi, la entropía residual acumulada, entre otras.

En el capítulo 4 se muestran los resultados que se obtuvieron en esta investigación aplicando las distribuciones *prior* generadas usando entropía y los modelos psicométricos propuestos en la investigación. Se menciona el análisis realizado así como la validación de los resultados obtenidos de manera sintética.

Finalmente, se comentan las conclusiones a las que se llegaron después de haber hecho la presente investigación y se dan algunas propuestas de trabajo futuro sobre el área de estudio que sirvió como base para este trabajo de tesis doctoral.

Capítulo 2

Marco teórico

El presente trabajo de tesis se desarrolla en el marco de las Ciencias de la Computación. Dentro de este campo del saber se encuentra un área que en los últimos tiempos, debido a la popularidad de la digitalización de procesos, ha tomado gran auge. Esta área se denomina "Evaluación Adaptable Computarizada", también conocida en inglés como CAT, y trata de la aplicación de nuevas técnicas informáticas para realizar el proceso de aplicación de un examen.

En otras palabras, lo que comunmente se hace a lápiz y papel, ahora se hace en computadora. Pero la novedad no termina ahí, sino que además este examen digitalizado permite que el evaluador pueda introducir nuevas mécanicas de calificación del examen, incluso algunas de ellas orientadas a la aplicación de técnicas de Inteligencia Artificial que permitan no solo obtener una calificación subjetiva numericamente hablando, sino que aparte se puedan establecer ciertos criterios que proporcionen mayor información sobre la situación de la persona o sujeto que está realizando la prueba o examen.

Aspectos tales como la distracción, la severidad de las preguntas de la prueba, la adivinación de respuestas, por mencionar algunos. Mediante la aplicación de modernas técnicas se pueden considerar variables que sin duda alguna juegan un papel muy importante a la hora de responder un examen.

2.1. La Teoría de Respuesta al Ítem

La Teoría de Respuesta al Ítem, o IRT por sus siglas en inglés, aporta algunas de estas técnicas novedosas a los evaluadores adaptables. Aunque no es la unica técnica que se utiliza para la creación de evaluadores adaptables, ya que existen otras, particularmente interesantes aquellas que son no paramétricas (19). La IRT es de especial interés ya que permite la utilización de técnicas basadas en estadistica inferencial utilizando el Teorema de Bayes.

La IRT habla en especial del uso de la probabilidad condicional como parte esencial del proceso de calificación de un examen. La probabilidad se aplica a la hora de calificar incluso una pregunta o reactivo, la cual adopta el nombre de ítem y está conformado por

dos partes principales, el tallo y las respuestas. Cuando se tiene un conjunto suficientemente grande de ítems en un tópico específico se le denomina pool y debe pasar por un proceso de calibración para conocer los grados de dificultad de cada ítem mediante la construcción de sus curvas características.

Además, se tiene también la aplicación de un modelo psicométrico el cual contiene un número de parámetros los cuales se habrán de estimar para poder calcular la habilidad del examinando dentro del proceso de evaluación. Para tener una idea más clara, la figura 2.1 muestra los componentes principales de manera general del proceso de aplicación de una prueba usando la IRT.

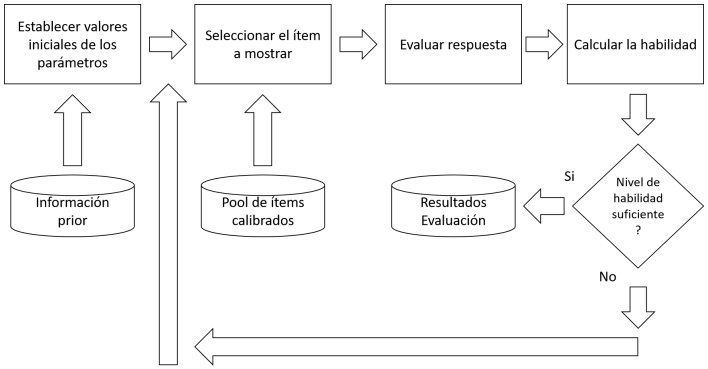

Figura 2.1: Secuencia de pasos de una prueba en la IRT.

Según lo que se plantea en el estado del arte, los elementos matemáticos requeridos para sustentar el presente proyecto que se propone incluyen a la Teoría de la Probabilidad, Inferencia Bayesiana, Estadística Paramétrica y no Paramétrica, Diseño de Experimentos, Teoría de la Información (Entropía), Optimización (Análisis de Regresión Lineal y no Lineal), Teoría de Transformaciones y Teoría de la Medición.

Por otro lado, también se requiere un conocimiento sólido en la parte de desarrollo tecnológico, como son Lenguajes de Programación (como son Java, C++, PHP, ActionScript, Flash), Manejadores de Bases de Datos Relacionales (MySQL), Sistemas Operativos (Linux, Windows), Seguridad (Firewall) y documentos PDF.

Todos estos conocimientos especializados serán requeridos para diseñar y construir una plataforma CAT que permita realizar las pruebas y en consecuencia obtener los resultados que demuestren la certeza de las hipótesis de esta tesis.

2.2. Modelos Psicométricos

Una de las fases iniciales de todo sistema CAT es la que se conoce como Calibración. Los sistemas CAT que tienen como base la IRT ocupan modelos psicométricos para realizar este proceso que consiste en dar respuesta a todos los ítems relacionados con el tema que se quiere evaluar, lo cual permite ajustar los valores de dificultad de cada uno de ellos.

El modelo psicométrico contiene una cantidad de parámetros cuyos valores deben ser encontrados para sintonizar los ítems con el modelo seleccionado. Esta fase es analoga a la fase de entrenamiento de una red neuronal ya que el sistema CAT debe contar con los valores iniciales de los parámetros del modelo antes de poder efectuar la evaluación adaptable.

Los modelos psicométricos de respuesta al ítem son una pieza importante no solamente para la IRT, sino también para la implementación de evaluadores adaptables computarizados. Estos modelos tienen un sustento estadístico y los más conocidos son los denominados 1PL, 2PL y 3PL, aunque se han llegado a proponer los modelos 4PL y 5PL.

El número que aparece como prefijo en el nombre, se refiere a la cantidad de parámetros que definen al modelo, el cual se considera logístico, y está orientado a disminuir los recursos de cómputo (tiempo) durante un proceso de evaluación, aunque sus parámetros tienen una interpretación cognitiva real. En las siguientes secciones se abordará la descripción particular de los modelos piscométricos ya que es de gran relevancia conocer su estructura y los parámetros que los conforman.

Una formulación Bayesiana empleando estos modelos permite determinar los valores de estos parámetros, aunque a costa de introducir requerimientos adicionales, como la construcción apropiada de distribuciones *prior* y también de desarrollar técnicas de selección de modelos, para decidir cuál de ellos es el más adecuado en el ajuste de un conjunto de datos dado (64).

La teoría Bayesiana se puede aplicar en principio empleando directamente el modelo psicométrico seleccionado, aunque existen otras corrientes que aseguran que es mejor hacerlo sobre una representación lineal del mismo. Esta linearización es relativamente fácil de obtener para los modelos 1PL y 2PL, y no queda claro cómo se puede hacer para los modelos 3PL y 4PL y mucho menos para el 5PL.

Si se utiliza directamente el modelo seleccionado para aplicar el teorema de Bayes, con el cual se estimen los parámetros, entonces aparecen de manera natural los conceptos de distribución prior y distribución a posteriori, como se muestra a continuación (10).

2.2.1. Modelo 1PL o modelo de Rasch

El modelo clásico de la IRT se conoce como 1PL o modelo de Rasch en honor al físico danés Goerge Rasch quien lo propuso (57). El término 1 precisa el hecho de que este modelo contiene solo un parámetro el cual está asociado con la dificultad del ítem. El modelo 1PL es el modelo psicométrico más simple y está definido por la ecuación (2.1)

$$p_i(x = 1|\theta, \mu_i) = \frac{1}{1 + e^{-(\theta - \mu_i)}}, \tag{2.1}$$

donde $p_i(x = 1|\theta, \mu)$ denota la probabilidad de respuesta correcta de un examinado al i–ésimo ítem, dado que θ es la habilidad del examinado y μ_i es la dificultad del ítem.

Especificamente, en el modelo original de Rasch, la probabilidad de una respuesta correcta está modelada como una función logística de la diferencia entre la habilidad del examinado y la dificultad del ítem. La gráfica resultante de esto es una curva logística o curva con forma de s y un ejemplo es el que se muestra en la figura 2.2.

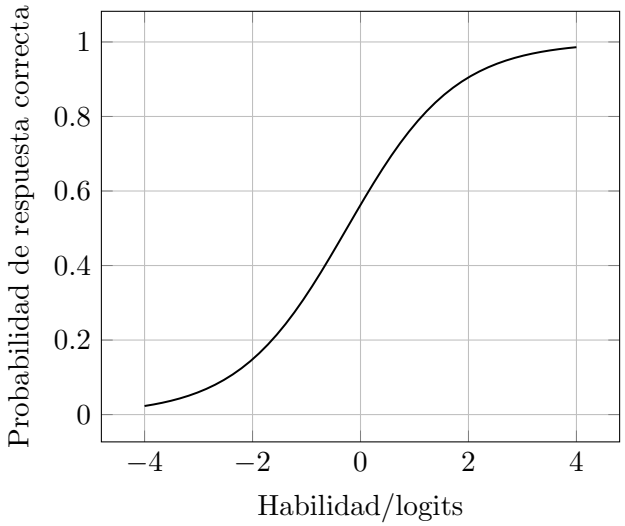

Figura 2.2: Curva logística del modelo de Rasch.

A pesar de ser el más sencillo de los modelos por contener solo un parámetro asociado a la dificultad del ítem, este modelo de Rasch es muy robusto debido a que cumple con varias propiedades siendo una de las más importantes la llamada "comparación invariante"también conocida como "objetividad específica". Esta característica permite relizar una comparación entre el desempeño de dos estudiantes cualquiera o el desempeño entre dos ítems cualquiera dentro de una prueba(32, 67).

En aras de establecer una definición formal de la objetividad específica, se debe recordar lo siguiente. El modelo psicométrico verifica que $p : E \times I \longrightarrow (a, b)$, donde E e I son los conjuntos de examinados y de ítems dentro de la prueba, respectivamente, y $(a, b) \subseteq [0, 1]$.

La objetividad específica del modelo asume la existencia de una función, $\chi : (a, b) \times (a, b) \longrightarrow \mathbb{R}$, y define una función vectorial multivariable $\boldsymbol{p} : (E \times I) \times (E \times I) \longrightarrow (a, b) \times (a, b)$, donde $\boldsymbol{p}((r, u), (s, v)) = (p(r, u), p(s, v))$, de tal manera que la composición de las funciones $c = \chi \circ \boldsymbol{p} : (E \times I) \times (E \times I) \longrightarrow \mathbb{R}$ compara el par (r, u) con el par (s, v) bajo las siguientes condiciones:

1. La comparación de cualesquiera dos objetos $r, s \in E$ es independiente de la elección $u, v \in I, u = v$, y de cualquier otro elemento $t \in E, t \neq r, s$.

2. La comparación de cualesquiera dos objetos $u, v \in I$ es independiente de la elección $r, s \in E, r = s$, y de cualquier otro elemento $w \in I, w \neq u, v$.

La función de comparación c es objetivamente específica dentro del marco de referencia definido por E, I y p. La función c de la primer condición no necesita ser igual a la función c de la segunda condición; *i.e.*, la función χ de la primer condición puede ser muy diferente de la función χ de la segunda condición. Sin embargo, el modelo psicométrico p es siempre el mismo.

2.2.2. Otros modelos con más parámetros

Existen otros modelos que consideran otros parámetros con su correspondiente interpretación cognitiva. Estos modelos pueden ser considerados generalizaciones del modelo original de Rasch, pues cada uno contiene al anterior dentro de su definición. Por ejemplo, la deficnión del modelo 2PL considera dentro de si al modelo 1PL original.

Modelo 2PL

Así como el parámetro del 1PL está asociado a la dificultad del ítem, el modelo 2PL o modelo de Birnbaum generaliza al modelo de Rasch permitiendo a los ítems variar no sólo en términos de la dificultad μ_i sino también en la capacidad de discriminación α entre examinados con diferentes habilidades (49).

La función de distribución acumulativa o CDF de sus siglas en inglés del modelo 2PL está definida por la ecuación 2.2.

$$p(x = 1|\theta, \mu, \alpha) = \frac{1}{1 + e^{-\alpha(\theta-\mu)}} \tag{2.2}$$

Modelo 3PL

En la realidad, cuando se aplica un examen a una población de estudiantes y la prueba presenta ítems dicotómicos de opción multiple, considerando de todas las opciones que solo una es la correcta, es facil suponer que puede existir la probabilidad de que la respuesta para tal ítem pueda darse de manera aleatoria, es decir, tratando de adivinar la respuesta correcta de entre todas las opciones.

Este fenómeno en donde cabe la posibilidad de que el examinado encuentre la respuesta correcta adivinando no está considerado ni por el modelo 1PL ni por el 2PL. Añadiendo otro parámetro al modelo 2PL que represente la contribución de la adivinación o también llamado pseudo-azar a la probabilidad de respuesta correcta, se puede corregir está situación (28).

En 1968, Allan Birnbaum(5) fue el primero en proponer está modificación al modelo logístico de dos parámetros. Él añadió el parámetro c que representa la forma binomial del parámetro de pseudo-azar como se muestra en la ecuación 2.3.

$$p(x = 1|\theta, \mu, \alpha, c) = c_i + (1 - c_i)\frac{1}{1 + e^{-\alpha_i(\theta-\mu_i)}}, \tag{2.3}$$

donde c_i es el parámetro de pseudo-azar para el ítem i o también conocido como la probabilidad de contestar correctamente el ítem i por mera adivinación. Con $0 < c_i < 1$ aunque algunos autores afirman que en la práctica se tiene $0 < c_i < 0.35$ (16).

Modelo 4PL

En 1981, Barton y Lord (6) exploraron de manera empírica la posibilidad de incluir un cuarto parámetro al modelo 3PL existente. La razón de esto está justificada por el hecho de que el modelo 3PL castiga considerablemente a aquellos estudiantes que poseen una alta habilidad pero se equivocan al contestar un ítem con baja dificultad.

En otras palabras, el modelo 3PL ubica las habilidades de los estudiantes menos competentes que adivinan correctamente un ítem difícil, pero asigna de manera efectiva una probabilidad de cero a los estudiantes más competentes al contestar incorrectamente un ítem fácil. Para abordar este supuesto, los autores estimaron el modelo 4PL mediante la ecuación 2.4.

$$p(x = 1|\theta, \mu, \alpha, c) = c_i + (d - c_i)\frac{1}{1 + e^{-\alpha_i(\theta - \mu_i)}}. \tag{2.4}$$

El parámetro d es la asíntota superior y toma valores por debajo de 1, es decir, $d < 1$ y cuando $d = 1$, el modelo se comporta como el 3PL. En la literatura se puede encontrar que este modelo es poco usado, principalmente por la dificultad de implementación que supone el modelo en si; sin embargo, con los recientes avances en cómputo de alto procesamiento se está retomando el interés en el potencial de aplicación del modelo, *cf.* por ejemplo en (54).

2.2.3. Estimación Bayesiana de Parámetros

Todos los modelos psicométricos, de los cuales se ha venido hablando en las secciones anteriores contienen dentro de su estructura un conjunto finito de parámatros, los cuales toman una interpretación cognitiva de acuerdo al contexto dentro del cual se esté aplicando el modelo. Sin embargo, la estimación de los valores de dichos parámetros no es trivial y existen al menos dos enfoques para realizar éste proceso.

El primero enfoque es utilizar un método experimental basado en el análisis de los datos con los que se esté trabajando. El segundo enfoque es usar algún método basado en inferencia Bayesiana. Uno de los métodos más usados es el de Máximum A Posteriori (MAP), el cual toma como base al Teorema de Bayes.

Ésta técnica permite hacer una estimación tomando en cuenta la información previa que se tenga sobre el evento del cual estemos trabajando. Para dar una idea más clara, se tiene la siguiente ecucación 2.5

$$P(\mathcal{A} \cap \mathcal{B}|\mathcal{C}) \quad = \quad P(\mathcal{B}|\mathcal{A} \cap \mathcal{C})P(\mathcal{A}|\mathcal{C}), \tag{2.5}$$

puesto que $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$, al aplicar la regla del producto se tiene 2.6,

$$P(\mathcal{B}|\mathcal{A} \cap \mathcal{C})P(\mathcal{A}|\mathcal{C}) \quad = \quad P(\mathcal{A}|\mathcal{B} \cap \mathcal{C})P(\mathcal{B}|\mathcal{C}), \tag{2.6}$$

lo que conduce al Teorema de Bayes

$$P(\mathcal{A}|\mathcal{B}\cap\mathcal{C}) = \frac{P(\mathcal{B}|\mathcal{A}\cap\mathcal{C})P(\mathcal{A}|\mathcal{C})}{P(\mathcal{B}|\mathcal{C})}. \tag{2.7}$$

La ecuación (2.7) da la definición de la distribución *a posteriori* en términos de la distribución *prior*, la cual se encuentra en el numerador del lado derecho de la ecuación. El trabajo del autor Bretthorts (10) es interesante porque liga el concepto de entropía con la determinación de los parámetros requeridos en la distribución *a posteriori* a través de la especificación de la información dada por la distribución *prior*, dando así una alternativa al método clásico (MLE).

En este sentido, es también interesante investigar acerca de las diferentes definiciones de entropía, más allá de la definición dada por Shannon, y las implicaciones que cada una de ellas tiene sobre la determinación de las distribuciones *prior*. Este tema es tratado en parte por el autor Jaynes (33).

Según la ecuación (2.7), la forma de la distribución *prior* es importante porque afecta a la distribución *a posteriori*. Por ejemplo, podrían existir distribuciones *prior* que produzcan distribuciones *a posteriori* multimodales o incluso sesgadas (77).

Dentro de este trabajo de investigación, el uso de la técnica MAP supone un conjunto de elementos necesarios para su correcta aplicación dentro de un sistema CAT, tales como el concepto de entropía, los modelos psicométricos o el proceso de calibración entre otros. Dichos elementos se pueden observar en la figura 2.3.

2.3. Entropía y Distribución Prior

Existen diferentes ejemplos de distribución *prior* y entre ellos se pueden encontrar a las distribuciones Beta, Gamma, Dirichlet y Normal (74). Sin embargo, no existe dentro de la teoría de probabilidades y dentro de la inferencia Bayesiana, la cual se supedita solamente a la regla producto, la regla suma y el Teorema de Bayes, un método que enuncie cómo asignar probabilidades, a no ser que se haga experimentalmente o, a la manera Laplaciana, la cual sugiere que estas probabilidades se asocian a cierto grado de creencia (10).

Opcionalmente, se pueden emplear también algunos principios de optimización de funciones que de alguna forma dependen de la distribución de probabilidad que se intenta encontrar. Un ejemplo de una función con esta característica es la entropía y el interés consiste en determinar las condiciones bajo las cuales es máxima (1, 2, 10, 11, 13, 14, 18, 20, 23, 24, 26, 29, 33, 34, 35, 36, 38, 39, 42, 43, 45, 50, 51, 52, 59, 73, 75).

El concepto de entropía propuesto por Shannon (2) es el más usado dentro de la literatura, aunque como ya se ha mencionado en otras secciones de este trabajo, existen muchos otros conceptos generalizados de entropía. En este trabajo, los conceptos entrópicos fueron parte sustancial para poder llegar al proceso de creación de distribuciones *prior* tal y como se puede observar en el artículo que se desarrolló y que se encuentra en el anexo A.1.

Figura 2.3: Elementos para realizar una estimación usando MAP dentro de un CAT

El concepto de entropía esta fuertemente ligado al uso de la probabilidad, ya que una de sus funciones es manejar una gran cantidad de eventos que son casuales, pero que de manera conjunta pueden ser predecibles en términos de probabilidad. Esto va de la mano con el hecho de que para construir una distribución prior se debe partir de información que provenga de la muestra de individuos que serán evaluados dentro de un CAT, y tal información será mucho más rica en tanto mayor sea su grado de desorden, es decir, en tanto mayor sea la cantidad de eventos casuales que estén relacionados con la población a evaluar, esto dará una gran aportación para poder plantear una distribución prior y que se ajuste al contexto de la población que se está evaluando.

2.4. Bimodalidad

El concepto de bimodalidad se refiere a la existencia de una distribución de probabilidad continua con dos modos diferentes. Un ejemplo de ello se puede apreciar en la figura 2.4.

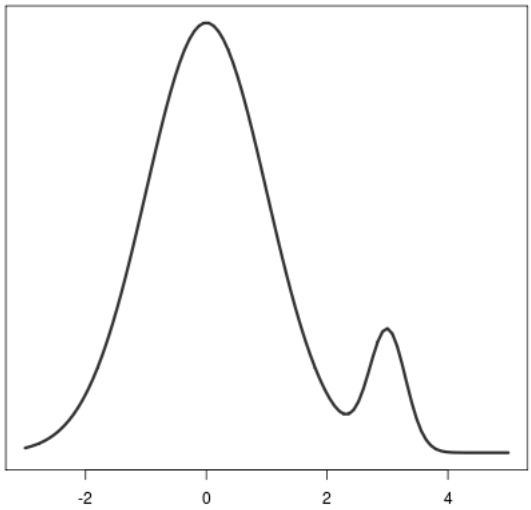

Figura 2.4: Distribución de probabilidad con bimodalidad.

Existen dentro de la literatura, algunos trabajos que proponen pruebas de bimodalidad a los datos. Algunas de estas pruebas son, por ejemplo, la prueba de la figura compleja de Rey-Osterrieth, la prueba de nombres de Boston, la prueba Dip de Hartigans, la prueba de exceso de masa, la prueba de existencia de modos, entre otras (53).

2.5. El proceso de calibración

El procedimiento de calibración de ítems en los test adaptables computarizados tiene una gran importancia, ya que es incuestionable su influencia en la determinación confiable de la estimación de la habilidad de los sujetos examinados en un área dada del conocimiento. El procedimiento de calibración de ítems es una etapa importante en el desarrollo y aplicación de un sistema de evaluación adaptable computarizado y el nivel de importancia puede ser comparado con el proceso de entrenamiento en la construcción de una red neuronal artificial, por ejemplo, donde un excelente entrenamiento es una condición necesaria para un desempeño confiable de la red.

En este sentido, el proceso de calibración de ítems puede ser identificado como el proceso de entrenamiento de un sistema de evaluación adaptable computarizado cuya finalidad es ajustar los valores de los parámetros de los módelos psicométricos usados para ser usados al principio del proceso de evaluación. Este proceso permite generar una clasificación de ítems con base en su complejidad, nivel de discriminación, entre otras posibles características a elegir dependiendo el algoritmo utilizado en este proceso.

Así como el proceso de entrenamiento supervisado en las redes neuronales artificiales, la calibración de ítems requiere que existan datos previos con la intención principal de ajustar esos datos a un modelo psicométrico dado. En el área de las redes neuronales artificiales, el entrenamiento supervisado asume el conocimiento del estimulo de entrada y el correspondiente estimulo de salida, con el próposito principal de comparar la salida teórica de una red neuronal previamente definida en arquitectura con el estimulo de salida deseado, cuando alguno de los estimulos de entrada activa la dinámica de la red propuesta.

La perfecta descripción de los datos de entrada y salida es la mayor ventaja en este caso del entrenamiento supervisado, el cual no se encuentra directamente en el proceso de entrenamiento de un sistema de evaluación adaptable computarizado.

Antes que nada, el proceso de calibración de ítems necesita la realización de un experimento con una muestra representativa de la población de examinados, donde se asume serán evaluados en un área del conocimiento previamente seleccionada, con la idea central de medir un constructo dado o rasgo latente, que para efectos de este trabajo será el relacionado con la habilidad conceptual sobre el manejo de tipos primitivos de datos en el lenguaje de programación Java.

De acuerdo a la Teoría de Respuesta al Ítem, un examinado tiene una cantidad del rasgo latente que está siendo medido y el experimento está orientado a determinar su valor. Además, el experimento también asume la existencia de un conjunto propio de ítems, el cual define el instrumento de medición del constructo dado.

El conjunto de ítems que define al instrumento de medición es útil para obtener las respuestas que vienen de cada uno de los examinados y esas respuestas son los indicadores directamente relacionados con el constructo que se está midiendo. La respuesta depende del tipo de ítem, por ejemplo una prueba construida con ítems dicotómicos debería aceptar configuraciones de respuestas definidas por valores falso o verdadero (0 ó 1). Los ítems politómicos definen pruebas con configuraciones de respuestas mucho

más complejas. En este trabajo solo se consideran ítems dicotómicos.

Para un ítem dado en la prueba, el cluster de respuestas correctas e incorrectas es la base para desarrollar la calibración de los ítems, la cual asume un modelo psicométrico previamente propuesto para definir teóricamente la probabilidad de una respuesta correcta, dada la habilidad del examinado y el conjunto de parámetros que definen al modelo psicométrico.

En este sentido, a lo largo del proceso de calibración se obtiene la curva característica del ítem para cada uno de los ítems de la prueba. En este trabajo de investigación se buscó la manera de tener un modelo psicométrico lo más general posible, por está razón se propusieron dos nuevos modelos de 6 parámetros como se describe en secciones posteriores.

En suma, el conjunto de respuestas provenientes de cada examinado define una colección de configuraciones, la cual se usa para estimar la habilidad del sujeto y la probabilidad de respuesta correcta para cada ítem en la prueba. El conjunto de pares ordenados (habilidad, probabilidad) se usa entonces para ajustar los datos obtenidos al modelo psicométrico propuesto. Los valores de los parámetros para cada ítem en la prueba proveen el resultado final de la calibración. Esos son los tópicos a ser analizados en las secciones siguientes acerca de la propuesta de diseño experimental.

Sin embargo, el proceso de calibración asume una selección propia del modelo psicométrico, la cual se debe ajustar a los datos experimentales obtenidos. En términos generales, se aplican los modelos psicométricos tradicionales por su relativa simplicidad para calcular computacionalmente los valores de los parámetros y por el cómputo rápido de las habilidades durante la fase de evaluaciones reales.

La falta de una completa flexibilidad es uno de los principales problemas de los modelos psicométricos tradicionales. Esta característica hace difícil la aproximación de algunos tipos de datos experimentales o simulados donde el sesgo y/o la multimodalidad pueden existir. De aquí que, al menos el análisis teórico requiera la propuesta de nuevos modelos psicométricos donde se pueda abordar este problema.

Capítulo 3

Desarrollo y Formalización

3.1. Propuesta de nuevos modelos psicométricos

La selección de un modelo psicométrico apropiado al contexto en el cual se esté realizando la evaluación es una parte clave dentro de un sistema CAT. No obstante, la elección puede llegar a complicarse cuando se trata de modelar situaciones de la vida real donde se encuentran presentes los efectos de multimodalidad y sesgo. Esta complicación surge como consecuencia del hecho de que los modelos clasicos basados en Rasch no llegan a poder modelar estos comportamientos. Incluso los modelos de 4 parámetros no contemplan dentro de sus estructuras la posibilidad de manejar datos con tales características. Existe un modelo de 5 parámetros (27) con el cual se pretende modelar el sesgo de los datos. Sin embargo, no contempla la posibilidad de la multimodalidad. Resulta conveniente entonces pensar, que si se añaden más parámetros a un modelo es posible considerar la multimodalidad dentro de la estructura, lo cual conducirá a que la curva ajuste mejor en aquellos conjuntos de datos cuya naturaleza involucre la existencia de la multimodalidad y los sesgos. Claro, se debe tener en cuenta el hecho de que se puede caer en una sobre-parametrización lo cual puede volver muy difícil el cálculo de los parámetros del propio modelo que se proponga.

3.1.1. Modelos Generalizados

Algunos autores han explicado las razones por la cuales proponer formulaciones alternativas más sofisticadas al modelo de Rasch, con la principal intención de incluir el posible sesgo de los datos experimentales (7). En (32) se propone una extensión del concepto de objetividad específica, dando la posibilidad de comparar tres o más elementos en los conjuntos de estudiantes E o de ítems I, e incluso en (67) el concepto de objetividad específica se excluye como un requerimiento necesario, lo cual conduce a la idea de modelos de pseudo-Rasch.

Un módelo general relativamente simple se puede proponer a partir de una ligera

modificación a otra función definida por (27),

$$p(x=1|\theta,\mu,\alpha,a,c,d,g) = d + (a-d)p^g\left(x=1\left|\theta,\mu+\frac{1}{\alpha}\ln c,\alpha\right.\right), \tag{3.1}$$

donde la Función de Distribución Cumulativa (CDF), p, en el lado derecho de (3.1), está dada por el modelo 2PL, el cual se define por la ecuación 3.2

$$p(x=1|\theta,\mu,\alpha) = \frac{1}{1+e^{-\alpha(\theta-\mu)}}. \tag{3.2}$$

3.2. Análisis de los modelos propuestos

En este escenario, multiples alternativas de modelos psicométricos son posibles. Sin embargo, para mantener la fortaleza del modelo de Rasch (56, 58) se deben introducir algunas restricciones. Por ejemplo,

i) el modelo propuesto debe ser parte de un marco de referencia con objetividad específica , aún con el modelo extendido, el cual admite la comparación entre dos, tres o más elementos de E o de I,

ii) el modelo también debe ser muy flexible para admitir sesgo y algún tipo de multimodalidad que pudiera posiblemente mostrar la habilidad y,

iii) finalmente, el modelo propuesto debe verificar que el modelo de Rasch cae como un caso particular.

El modelo dado por (3.1) contiene seis parámetros μ, α, a, c, d y g, y verifica el cumplimiento de las restricciones anteriores. De hecho, la interpretación de los parámetros μ, α, a y d, coincide con la que se da a los parámetros en los ya bien conocidos modelos 1PL, 2PL, 3PL y 4PL (27, 28). En suma a lo anterior, el modelo de Rasch, así como los modelos 2PL, 3PL y 4PL, son casos particulares de este modelo más general 6PL definido por (3.1).

Para ilustrar la comprobación de la restricción iii), se puede obtener de (3.1) el modelo de Rasch si se toma $\alpha = 1, a = 1, c = 1, d = 0, g = 1$ y $\mu \in (-\infty, +\infty)$. Cuando el valor requerido de $\alpha \in (0, +\infty)$ se toma en lugar de $\alpha = 1$, se obtiene el modelo 2PL. También a partir de (3.1) el modelo 3PL se puede obtener cuando $a = 1, c = 1, g = 1$ y $\mu \in (-\infty, +\infty), \alpha \in (0, +\infty), d \in (0, 1)$. Finalmente, cuando $0 < d < a < 1$, entonces se obtiene el modelo 4PL también.

Además. con el modelo dado por (3.1), llamado de aquí en adelante modelo 6PL extendido de Rasch, la interpretación de los nuevos parámetros c y g incluye el concepto de sesgo de los datos experimentales, ya que, el parámetro c implica una corrección al término de la dificultad μ. Sin embargo, se puede probar que el modelo no produce sesgos simétricos (el sesgo a la izquierda no es una imagen espejo del sesgo a la derecha) cómo se puede apreciar en la figura 3.1.

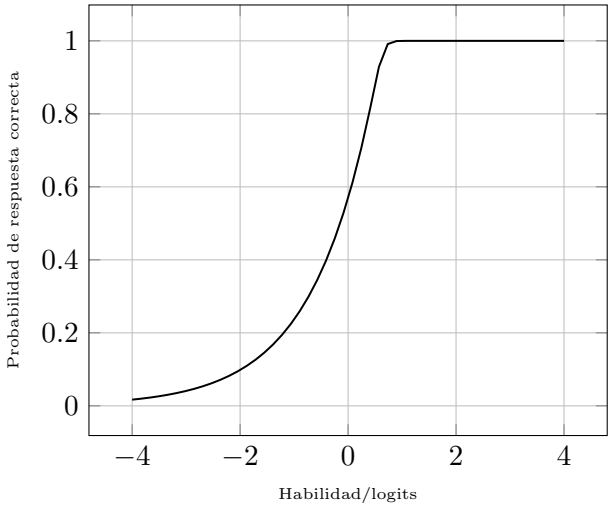

Figura 3.1: Curva característica del ítem usando el modelo 6PL Extendido de Rasch.

3.2.1. Modelo 6PL Extendido

Uno de los primeros avances que se obtuvo dentro de la presente investigación, fue la construcción de un modelo psicométrico de seis parámetros que se pudiera adaptar al concepto orginal del modelo de Rasch, es decir, que cumpliera con las restricciones del propio modelo original y que en consecuencia fuese una generalización del mismo. Si bien es cierto, la idea de tener un mayor número de parámetros con respecto a los modelos ya existentes representa una complejidad aún mayor para su implementación, pero el hecho de que a costa de estos nuevos parámetros se pueda contemplar dentro del modelado la posible existencia de resultados con naturaleza bimodal y sesgada es algo novedoso en el campo de la IRT y de los sistemas CAT.

El cambio de concavidad y el comportamiento simétrico relativo a las asíntotas superior e inferior, respectivamente definido por las ecuación 3.3

$$\lim_{\theta \to \pm\infty} p(x = 1|\theta, \mu, \alpha, a, d, c, g) = \begin{cases} a \\ d \end{cases}, \tag{3.3}$$

son dos puntos importante a considerar en el comportamiento del modelo 6PL extendido de Rasch. En esta parte de la discusión, se analizan las condiciones para aplicar de manera exitosa esta CDF como un modelo psicométrico propio, matemáticamente hablando.

Para hacer esto, en el análisis del comportamiento de la función se involucran la primera y segunda derivadas y se puede probar fácilmente que la CDF (3.1) es una función incremental. Por otro lado, el cambio de concavidad ocurre en el punto simple $\theta = \mu + \frac{1}{\alpha}\ln(cg)$ dentro del dominio de la CDF (3.1) y $d + \frac{a-d}{(1+g^{-1})^g}$ es el valor de la función en este punto. Ahora, la condición de crecimiento rápido de la CDF esta establecida a través de la definición de un parámetro positivo κ tal que,

$$\frac{\kappa}{\alpha g(a-d)} \leq p^{g}\left(x=1\left|\theta,\mu+\frac{1}{\alpha}\ln c,\alpha\right.\right)-p^{g+1}\left(x=1\left|\theta,\mu+\frac{1}{\alpha}\ln c,\alpha\right.\right). \tag{3.4}$$

Entonces, el conjunto de habilidades θ que satisfacen esta desigualdad se vuelven el intervalo donde la CDF (3.1) crece rápidamente. Ya que esta CDF es una función incremental, entonces las raíces de la igualdad en la inecuación (3.4) definen el *infimum* y el *supremum* del intervalo. La desigualdad (3.4) también puede ser vista como la especificación del limite inferior dado por $\frac{\kappa}{\alpha g(a-d)}$ al polinomio $f(x)=x^{g}-x^{g+1}$ en el intervalo real $(0,1)$, con potencia real $g, 0<g$. La derivada de este polinomio esta dada por 3.5

$$\frac{d}{dx}f(x) \quad = \quad (g+1)x^{g-1}\left(\frac{g}{g+1}-x\right), \tag{3.5}$$

la cual es siempre positiva en el intervalo $\left(0,\frac{g}{g+1}\right)$ (función incremental f) y negativa en el intervalo $\left(\frac{g}{g+1},1\right)$ (función decremental f).

Si $0<g<1$, entonces la función f no tiene cambio de concavidad y es siemre concava hacia abajo (ya que su segunda derivada es siempre negativa). Por otro lado, si $1<g$, entonces la función f cambia de ser concava hacia arriba a ser concava hacia abajo en el punto crítico $\theta=\frac{g-1}{g+1}$. El punto $\theta=\frac{g}{g+1}$ en el dominio de f es un punto crítico donde donde la función f tiene un valor *maximum* para valores arbitrarios del parámetro g, $0<g$. Esto significa, que la función definida por la igualdad en (3.4) tiene solamente dos raices reales si la constante $\frac{\kappa}{\alpha g(a-d)}$ satisface las condiciones que se mencionan posteriormente. De hecho, el punto crítico $\frac{g}{g+1}$ está asociado con el *maximum* de $f(x)$ en el intervalo $(0,1)$, y este *maximum* acquiere el valor $\frac{1}{g+1}\left(\frac{g}{g+1}\right)^{g}$. Por lo tanto, la condicion para la existencia de las dos raíces reales está dada por 3.6

$$0<\frac{\kappa}{\alpha(a-d)}<\left(\frac{g}{g+1}\right)^{g+1}. \tag{3.6}$$

El comportamiento asimétrico de la función f asegura que las dos raíces son asimétricas con respecto al punto $\frac{g}{g+1}$. Este comportamiento de f implica la posibilidad de obtener una Función de Densidad de Probabilidad (PDF) sesgada para la CDF (3.1); sin embargo, el comportamiento de este sesgo no es completamente simétrico, en el sentido ya explicado al principio de esta sección.

Para probar la existencia de un sesgo asimétrico para el modelo 6PL extendido de Rasch se considera el valor de esta función en el punto $\theta=\mu+\frac{1}{\alpha}\ln(cg)$, donde aparece el cambio de concavidad. La distancia entre el punto $(\mu+\frac{1}{\alpha}\ln(cg),d)$ en la asíntota inferior y el punto donde aparece el cambio de concavidad esta dada por la expresión $\frac{a-d}{(1+g^{-1})^{g}}$, la cual significa que la distancia es proporcional a $a-d$, donde la proporción está definida por la expresión $\frac{1}{\left(1+\frac{1}{g}\right)^{g}}$.

Notese que el parámentro g tiene un limite inferior igual a cero, pero puede crecer sin limite, así que es interesante saber los limites de esta proporción. Usando la definición

de la base del logaritmo natural y la regla de L'Hopital's, es relativamente sencillo notar que esos limites son 1 y $\frac{1}{e}$, donde $g \to 0$ y $g \to +\infty$, respectivamente.

Entonces, la distancia entre el punto en la asíntota inferior, $(\mu + \frac{1}{\alpha}\ln(cg), d)$, y el punto en la CDF donde ocurre el cambio de concavidad, puede ser tan grande como $a-d$ o tan pequeño como $\frac{1}{e}(a-d)$, lo cual significa que el sesgo a la derecha de la CDF no es necesariamente simétrico al sesgo a la izquierda del mismo modelo de la CDF.

3.2.2. El modelo 6PL Extendido de Rasch y la objetividad específica

El comportamiento de este modelo es muy complejo, y es difícil, si no imposible, manejar de manera directa el concepto de objetividad específica. Sin embargo, la función se puede aproximar mediante una función por partes definida como a continuación se muestra en 3.7 en algunos intervalos de las habilidades

$$p(x=1|\theta,\mu,\alpha,a,c,d,g) \approx \begin{cases} \frac{1}{1+\exp(-\alpha(\theta-\mu-\frac{1}{\alpha}\ln c))}, & \text{si } \mu+\frac{1}{\alpha}\ln c \ll \theta \\ \frac{1}{\exp(-\alpha g(\theta-\mu-\frac{1}{\alpha}\ln c))}, & \text{si } \theta \ll \mu+\frac{1}{\alpha}\ln c \\ \frac{1}{2^g\left(1-\frac{1}{2}g\right)} \cdot \frac{1}{1+\exp\left(-\alpha\left(\theta-\mu-\frac{1}{\alpha}\ln\frac{cg}{2\left(1-\frac{1}{2}g\right)}\right)\right)}, & \text{en otro caso.} \end{cases} \tag{3.7}$$

La definición por partes de la función $p(x=1|\theta,\mu,\alpha,a,c,d,g)$ y la idea de una objetividad específica, permiten comparar tres habilidades escogidas arbitrariamente y dos ítems seleccionados arbitrariamente. Se debe resaltar que la elección de cualesquiera dos expresiones de la definición por partes de la función $p(x=1|\theta,\mu,\alpha,a,c,d,g)$ representa exactamente la misma CDF, así que el concepto de objetividad específica está correctamente aplicado en tal sentido.

Luego entonces, la comparación de las tres habilidades se puede realizar mediante la definición de la función 3.8 con $\chi : (0,1)\times(0,1)\times(0,1) \longrightarrow \mathbb{R}$,

$$\chi(x_1,x_2,x_3) = \frac{\ln\left(\frac{x_1}{1-x_1}\cdot\frac{1-x_2}{x_2}\right)}{\ln\left(\frac{x_1}{1-x_1}\cdot\frac{1-x_3}{x_3}\right)}, \tag{3.8}$$

y a través de asumir que la objetividad específica elimina cualquier factor de escala en cualquier expresión de la definición por partes de la CDF. Así que, por ejemplo, $c((r,u),(s,u),(t,u)) = \frac{\theta_r-\theta_s}{\theta_r-\theta_t}$. De manera similar, dos ítems se pueden comparar a través de la definición de la función 3.9 con $\chi : (0,1)\times(0,1)\times(0,1)\times(0,1) \longrightarrow \mathbb{R}$,

$$\chi(x_1,x_2,x_3,x_4) = \frac{\ln\left(\frac{x_1}{1-x_1}\cdot\frac{1-x_2}{x_2}\right)}{\ln\left(\frac{x_3}{1-x_3}\cdot\frac{1-x_4}{x_4}\right)}, \tag{3.9}$$

y la idea de objetividad específica ya sugerida. Además, dos ítems pueden ser comparados tomando en cuenta la misma expresión, o cualesquiera dos expresiones de la definición por partes de la CDF, $c((r,u),(s,u),(r,v),(r,v)) = \frac{\alpha_u}{\alpha_v}$ ó $c((r,u),(s,u),(r,v),(r,v)) = \frac{\alpha_u g_u}{\alpha_v}$.

Estos resultados implican que un simple ítem se puede comparar consigo mismo a través de definir subdominios, dejando en claro que un ítem puede tener capacidades diferentes de discriminación. Además, dos ítems diferentes (u, v) pueden ser comparados tomando alguno de los siguientes índices $\frac{\alpha_u}{\alpha_v}, \frac{\alpha_u g_u}{\alpha_v}, \frac{\alpha_v}{\alpha_u g_u}, \frac{\alpha_v}{\alpha_u}$ y un solo ítem u con los índices $g_u, \frac{1}{g_u}$.

3.2.3. Modelo 6PL Flexible

En (61) se propone otra CDF con seis parámetros en un contexto diferente, y esta función está definida por,

$$p(x = 1|\theta, \mu, \alpha, \beta, k, a, d) \quad = \quad d + \frac{a - d}{1 + \frac{e^{-\alpha(\theta-\mu)}}{1+e^{k(\theta-\mu)}} + \frac{e^{-\beta(\theta-\mu)}}{1+e^{-k(\theta-\mu)}}}, \tag{3.10}$$

donde la definición de $k = \frac{2\alpha\beta}{|\alpha+\beta|}$ especifíca una restricción a los posibles valores de k. Sin embargo, el modelo que se propone en este trabajo requiere solamente que $0 \leq d < a \leq 1, \mu \in (-\infty, +\infty)$ y no impone restricciones en los posibles valores de k. Es de notar que el modelo satisface los dos comportamientos asintóticos cuando $\theta \to \pm\infty$ y que el modelo de Rasch se puede obtener como un caso particular cuando $a = 1, d = 0, k = 0$ y $\alpha = \beta$. Los posibles valores de α y β son deducidos a partir del análisis del comportamiento asintótico de la función (3.10). Este análisis muestra que $0 < \alpha$ y $0 < \beta$.

El Modelo 6PL Flexible y la Objetividad Específica

La condición de objetividad específica está de alguna forma relacionada intimamente con el concepto de función inversa. Así que, dada la CDF (3.10), uno de los posibles medios para encontrar la transformación adecuada, que conduzca a la propiedad objetividad específica, consiste en encontrar las raices de la ecuación (3.11)

$$1 + \frac{1}{1 + e^{k(\theta-\mu)}} e^{-\alpha(\theta-\mu)} + \frac{1}{1 + e^{-k(\theta-\mu)}} e^{-\beta(\theta-\mu)} = \frac{a - d}{p(x = 1|\theta, \mu, \alpha, \beta, k, a, d) - d}, \tag{3.11}$$

la cual resulta tras algunas manipulaciones sobre la ecuación (3.10).

A primera vista, puede resultar muy difícil encontrar una expresión analítica de las posibles raices de (3.11). Sin embargo, el análisis asintótico arroja alguna luz sobre el comportamiento del lado izquierdo de (3.11) en los limites $k \longrightarrow \pm\infty$ y $\theta \longrightarrow \pm\infty$.

Las condiciones especificadas en la Tabla 3.1 dicen que, para algunos parámetros adecuados α, β, k, existen regiones en el dominio de la función $p(x = 1|\theta, \mu, \alpha, \beta, k, a, d)$ donde la objetividad específica se logra. Así que, por ejemplo, las condiciones $\alpha > 0, \beta > 0$ y $k \gg 1$ definen el comportamiento asintótico $1 + e^{-\alpha(\theta-\mu)}$ del lado izquierdo de (3.11) en un intervalo $(-\infty, \theta^*)$, donde $\theta^* < \mu$.

Tabla 3.1: Comportamiento asintótico del lado izquierdo del modelo 6PL Flexible.

Aunque ir significa 'valor irrestricto', se asume que $\alpha > 0$ y $\beta > 0$ para satisfacer el comportamiento asintótico de la función (3.10).

k	θ	α	β	Comportamiento Asintótico	Comentarios
$+\infty$	$-\infty$	$+$	ir	$1 + e^{-\alpha(\theta-\mu)}$	Estas condiciones implican la existencia de intervalos a la izquierda y derecha de $\theta = \mu$ donde la CDF dada por (3.10) se vuelve incremental y con comportamiento complejo en un vecindario de $\theta = \mu$
	$+\infty$	ir	$+$	$1 + e^{-\beta(\theta-\mu)}$	
$-\infty$	$-\infty$	ir	$+$	$1 + e^{-\beta(\theta-\mu)}$	Comentarios similares a la fila anterior
	$+\infty$	$+$	ir	$1 + e^{-\alpha(\theta-\mu)}$	

Un análisis similar en algún intervalo $(\theta^{**}, +\infty)$, donde $\mu < \theta^{**}$, muestra también la existencia de la objetividad específica. Por simetría, uno debería esperar resultados similares cuando $k \ll -1$, con el comportamiento asintótico específicado por la segunda fila de la Tabla 3.1. Así que, tres regiones en el dominio de la definición de la CDF (3.10) especifican el comportamiento del modelo.

Entonces, el modelo 6PL flexible puede ser aproximado por una función exponencial por partes como sigue en 3.12,

$$p(x = 1|\theta, \mu, \alpha, \beta, k, a, d) \approx \begin{cases} \frac{1}{1+e^{-\alpha(\theta-\mu)}} & \text{if } \theta \in I_\alpha(\mu, k), \\ \frac{1}{1+e^{-\frac{\alpha+\beta}{2}(\theta-\mu)}} & \text{if } \theta \in I_{\frac{\alpha+\beta}{2}}(\mu, k), \\ \frac{1}{1+e^{-\beta(\theta-\mu)}} & \text{if } \theta \in I_\beta(\mu, k), \end{cases} \tag{3.12}$$

donde los intervalos $I_\alpha(\mu, k), I_{\frac{\alpha+\beta}{2}}(k), I_\beta(\mu, k)$ dependen de los parámetros μ y k. Es de notar que, $I_\alpha(\mu, k) \cap I_{\frac{\alpha+\beta}{2}}(\mu, k) = \emptyset, I_\alpha(\mu, k) \cap I_\beta(\mu, k) = \emptyset, I_\beta(\mu, k) \cap I_{\frac{\alpha+\beta}{2}}(\mu, k) = \emptyset$ y también $I_\alpha(\mu, k) \cup I_{\frac{\alpha+\beta}{2}}(\mu, k) \cup I_\beta(\mu, k) = \mathbb{R}$. Esta representación sugiere que el ítem contiene tres parámetros discriminantes, dados por α, β y $\frac{\alpha+\beta}{2}$.

Entonces, la función de comparación es similar a la función del modelo 6PL extendido de Rasch y tres habilidades y dos ítems pueden ser comparados usando las funciones definidas en la ecuación (3.8) y la ecuación (3.9), respectivamente. Sean $\mu_u, \alpha_u, \beta_u, k_u$ los parámetros del ítem u y $\theta_r, \theta_s, \theta_t$, las habilidades arbitrarias de tres examinandos diferentes. Entonces la función de comparación se evalúa como sigue en 3.13,

$$c((r,u),(s,u),(t,u)) \quad = \quad \frac{\theta_r - \theta_s}{\theta_r - \theta_t} \tag{3.13}$$

De manera similar, para poder comparar dos ítems, se consideran solamente los examinandos r y s y los ítems u and v. Entonces,

$$c(p(r,u),p(s,u),p(r,v),p(s,v)) \quad = \quad \frac{\gamma_u}{\gamma_v}, \tag{3.14}$$

3.14 compara dos ítems arbitrarios, sin considerar las características de los examinandos. La comparación se puede obtener de la siguiente manera, $\frac{\alpha_u}{\left(\frac{\alpha_v+\beta_v}{2}\right)}$, $\frac{\beta_u}{\left(\frac{\alpha_v+\beta_v}{2}\right)}$, $\frac{\alpha_u}{\beta_v}$, $\frac{\beta_u}{\alpha_v}$, $\frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\beta_v}$, $\frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\alpha_v}$, $\frac{\alpha_u}{\alpha_v}$, $\frac{\beta_u}{\beta_v}$, $\frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\left(\frac{\alpha_v+\beta_v}{2}\right)}$ o incluso dentro del mismo ítem puede haber comparaciones por regiones, $\frac{\alpha_u}{\left(\frac{\alpha_u+\beta_u}{2}\right)}$, $\frac{\beta_u}{\left(\frac{\alpha_u+\beta_u}{2}\right)}$, $\frac{\alpha_u}{\beta_u}$, $\frac{\beta_u}{\alpha_u}$, $\frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\beta_u}$, $\frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\alpha_u}$.

3.3. Construcción de Distribuciones Prior

El teorema de Bayes como se ha comentado en el Capítulo 2, contempla el uso de una distribución llamada *prior*, con la cual es posible calcular la llamada distribución *a posteriori*. No obstante, la selección de la distribución *prior* no es trivial y se debe garantizar que aporte la mayor cantidad de información sobre los examinandos. Para tal efecto, el concepto de entropía ha sido utilizado en algunos trabajos [referencia] como una alternativa para la construcción de distribuciones prior informativas. Para enriquecer estas propuestas, se puede utilizar el concepto de máxima entropía, lo cual supone la utilización de alguna técnica de optimización para maximizar la información que arrojará la entropía sobre la información previa con la que se cuente respecto a los examinandos en la muestra que se esté evaluando.

Un ejemplo podría ser identificar a buenos programadores dentro de un grupo de alumnos de algún programa de estudios en Ciencias Computacionales. La calificación promedio puede ser importante para saber si un sujeto tiene mayor habilidad en programación. Típicamente, aquellos examinandos con mejores promedios de calificación, tienden a ser buenos programando, pero esto es solo una suposición empírica, se requiere mayor formalidad para sustentar tal aseveración. Es aquí donde el concepto de entropía puede arrojar mayor claridad sobre esta información previa.

Retomando el ejemplo anterior, se podría considerar la proporción de hombres y mujeres en la muestra de examinandos como otro dato relevante además del promedio. Aquí nuevamente el concepto de entropía puede arrojar resultados favorables al respecto que aporten una mayor claridad para la construcción de la distribución prior.

La formulación de esto que se ha dicho en parrafos anteriores, está dada en principio por la aplicación del concepto de entropía de Shannon (33), el cual esta dado por la siguiente ecuación,

$$\mathcal{L} = \max_{p_i} \left\{ -\sum_{i=1}^{n} p_i \ln p_i + \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) \right\}, \tag{3.15}$$

donde $-\sum_{i=1}^{n} p_i \ln p_i$ es la entropía de Shannon y $\sum_{i=1}^{n} p_i - 1$ es la condición de que la suma de las probabilidades es 1.

Con la ecuación anterior lo que se pretende es poder agregar más restricciones que representen cierta información obtenida de manera preliminar de los sujetos de la muestra.

Por ejemplo el género es un factor importante a considerar, ya que las proporciones de hombres y mujeres dentro de una población no es la misma. Y esto se puede ver mucho más marcadamente cuando se habla de una carrera del área de la Ingeniería, como por ejemplo la ciencias computacionales. Si se toma esta información obtenida a partir de una muestra de la población de estudiantes, se podría decir que

$$\mathcal{L} = \max_{p_i} \left\{ -\sum_{i=1}^{2} p_i \ln p_i + \lambda_0 \left(\sum_{i=1}^{2} p_i - 1 \right) + \lambda_1 \left(p_1 - a p_2 \right) \right\}. \tag{3.16}$$

Asumiendo que Mujeres = 1 y Hombres = 2, y donde a es una constante de proporcionalidad y $a = \frac{p_1}{p_2}$.

Sin embargo, la manera de obtener información previa de los examinandos puede llegar a ser muy diversa, por ejemplo, si se aplicara una encuesta, se podría llegar a tener una mezcla entre tipos de datos cualitativos (categóricos) y cuantitativos (numéricos), lo cual hace complejo la inclusión de más restricciones. En tal caso lo que se sugiere hacer es realizar un Análisis de Componentes Principales (PCA) sobre los datos de la encuesta a fin de reducir la dimensionalidad y extraer los datos más representativos. Esto tampoco resulta trivial ya que la mezcla de los datos complica más la aplicación de dicha técnica. No obstante existe una alternativa, la cual implica el estudio de la técnica PCAmix (), la cual según la literatura puede separar los componentes principales de ambos tipos de datos y ayudar a reducir considerablemente la dimensionalidad del problema. Este es un posible camino experimental que se sugiere explorar como trabajo futuro a esta tesis.

Capítulo 4

Experimentación y Resultados

4.1. Resultados

Los hallazgos que se fueron realizando durante la realización de este trabajo de tesis son muy ricos en posibilidades para ser tomados como nuevos puntos de partida en investigaciones posteriores. Además, la posibilidad de contar con una propuesta al menos teórica sobre el cálculo de distribuciones prior, ha permitido realizar algunos cálculos sintéticos sobre el uso de estas distribuciones en conjunto con la técnica MAP o EAP aplicando los modelos psicométricos de Rasch generalizados.

El cálculo de las habilidades de examinados en un sistema CAT requiere, como se ha venido platicando de la selección un módelo psicométrico a ser resuelto a la hora de aplicarse los ítems dentro de la prueba. Tomando los resultados que se han encontrado durante la presente investigación, es pertinente mencionar que los módelos psicométricos propuestos con 6 parámetros arrojan luz sobre la posibilidad de que sirvan de apoyo para trabajar con aquellas circusntancias donde exista una multimodalidad o sesgo en los resultados que arroje la prueba que se esté realizando.

Para poder realizar con mayor efectividad la prueba se debe considerar un buen diseño experimental. Existen algunos trabajos que consideran procedimientos estándar para plantear un buen diseño como se menciona en (17, 37, 80), donde se discute sobre los materiales necesarios y se describe como estos deben ser preparados.

Modelos Psicométricos

Para efectos de este trabajo, si se hablará de utilizar un diseño experimental, se puede considerar la forma general que incluye seis parámetros en la siguiente ecuación

$$P(X = 1|\theta, a, b, c, d, f, g) = d + \frac{a - d}{\left(1 + ce^{b(\theta - f)}\right)^g}, \tag{4.1}$$

donde $f \in (-\infty, +\infty), 0 \leq d < a \leq 1, b < 0, 0 < c, 0 < g$ y a, d representan las asíntotas horizontales superior e inferior respectivamente. Esta ecuación es llamada

modelo psicométrico 6PL e incluye a los otros modelos llamados 1PL, 2PL, 3PL, 4PL y 5PL.

Cada ítem en la prueba tiene una ecuación como ésta y el proceso de calibración se encarga de encontrar los valores adecuados para los parámetros. La interpretación de cada parámetro de la ecuación es particularmente importante. Por ejemplo, el parámetro f representa la dificultad del ítem, el cual mide cuanta dificultad encuentra el examinando para contestar correctamente a tal ítem, cuando se compara con su habilidad.

El parámetro b es llamado el discriminante y mide que tanto el ítem discrimina entre los examinados. Valores altos de este parámetro discriminan mejor que los valores bajos. Por ejemplo, valores bajos de este parámetro hacen dificil discriminar entre dos examinandos con habilidades notablemente diferentes, ya que las probabilidades de responder correctamente al ítem son casi iguales. Ítems con un alto valor discriminante son más valiosos que los que tienen un valor bajo.

Por otro lado, otras alternativas de funciones de distribución acumulada parecen ser mucho más complicadas (4), mientras que otras están entre los dos puntos de vista como por ejemplo la que se menciona en (61).

La segunda referencia en el párrafo anterior propone una función de distribución acumulativa en la forma,

$$\begin{aligned} f(x) &= d + \frac{a-d}{1 + \frac{e^{p_3(x-p_4)}}{1+e^{k(x-p_4)}} + \frac{e^{k(x-p_4)}e^{p_5(x-p_4)}}{1+e^{k(x-p_4)}}}, \\ &= d + \frac{(a-d)\left(1+e^{k(x-p_4)}\right)}{1 + e^{k(x-p_4)} + e^{p_3(x-p_4)} + e^{k(x-p_4)}e^{p_5(x-p_4)}}, \end{aligned} \tag{4.2}$$

donde los autores del artículo sugieren que $k = \frac{2p_3p_5}{|p_3+p_5|}$ y ésta propuesta de diseño experimental requiere que $0 \leq d < a \leq 1, p_3 < 0, p_4 \in (-\infty, +\infty)$, y $p_5 < 0$. Se ha de notar también que este modelo satisface los dos principales requerimientos de esta propuesta de diseño experimental; a saber

$$\lim_{x \longrightarrow -\infty} f(x) = d \tag{4.3}$$

y

$$\lim_{x \longrightarrow +\infty} f(x) = a \tag{4.4}$$

Además, este diseño experimental no introduce ninguna restricción dentro del parámetro k, y este parámetro k irrestricto provee más flexibilidad al modelo, haciendo posible obtener comportamientos complejos de la función de distribución cumulativa y la función de densidad de probabilidad, donde esta complejidad está definida por la presencia de sesgos y multimodalidad.

Algoritmos y heurísticas

La lista de materiales de un buen diseño experimental incluye los algoritmos o heurísticas para procesar los datos crudos. Existen al menos dos caminos para procesar

los datos. Un camino para procesar los datos no requiere más consideraciones y calcula directamente las variables importantes considerando los datos experimentales (las respuestas a cada ítem provenientes de cada examinando) (69).

El segundo camino incluye la especificación de un modelo psicométrico y la idea de máxima verosimilitud. Es importante decir que el patrón de respuestas por examinado no tiene ningún efecto en ambos casos de procesamiento de datos, sin embargo, sería muy interesante incluir está característica en un trabajo futuro.

Como un ejemplo de datos crudos, la tabla 4.1 muestra los resultados después de aplicar, de manera simulada, una prueba o test definido por veinte ítems dicotómicos a un conjunto de quince examinados. Las habilidades de los estudiantes fueron seleccionadas de manera aleatoria dentro del intervalo $[-3, 3]$ de manera uniforme, donde las unidades de las habilidades están dadas en logits. Las respuesta del examinando fue producida a través de la aplicación del modelo logístico de dos parámetros y la selección aleatoria de un número real en el intervalo $[0, 1]$ de manera uniforme.

La simulación requiere veinte valores del par de parámetros llamados dificultad del ítem y discriminante del examinando; en otras palabras, un valor para este par de parámetros por cada ítem en la prueba. Por supuesto, en una situación real se asume que estos parámetros no son conocidos y el proceso de calibración debe estimar su valor.

Dentro de la simulación, la dificultad del ítem esta definida mediante una selección aleatoria de un número real en el intervalo $[-3, +3]$ de manera uniforme, donde los logits son las unidades de la dificultad del ítem. De manera similar, el discriminante del examinando esta definido mediante una selección aleatoria de un número real en el intervalo $(0, 5)$ de manera uniforme y este parámetro es adimensional.

No obstante, ¿cómo puede uno obtener una distribución de $0's$ y $1's$ para lograr un tipo dado de distribución? La respuesta es relativamente simple, y esta dada por el Algoritmo 1

¿Cómo obtener un histograma simulado que represente la posible frecuencia de las habilidades de un grupo de examinados respondiendo a un ítem? Esta es otra pregunta interesante. Otra vez, el ítem en particular está representado por un conjunto conocido de parámetros $\mu, \alpha, a, d, \beta, c, g$ que definen al correspondiente modelo psicómetrico. La respuesta a esta pregunta está dada por el algoritmo 2.

La discusión anterior permite trabajar con datos simulados, haciendo posible asumir las circunstancias necesarias de una prueba en las que el beneficio de contar con un buen diseño experimental apoya de manera contundente la propuesta de usar modelos psicométricos de seis parámetros para el cálculo de las habilidades de los examinados.

Algoritmo 1 `Algoritmo` para obtener respuestas con una distribución acumulativa dada.

`Entrada:` Una CDF dada $P(X = 1|\theta, \boldsymbol{p})$.

`Salida:` Conjunto de respuestas correctas e incorrectas siguiendo la distribución dada $P(X = 1|\theta, \boldsymbol{p})$.

1: **procedure** DISTRIBUCIÓN($P(X = 1|\theta, \boldsymbol{p})$)
2: $\boldsymbol{\theta} \leftarrow \{\theta_1, \theta_2, \ldots, \theta_m\}$ %define uniformemente el conjunto de habilidades
3: $\boldsymbol{p} \leftarrow \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n\}$ %define uniformemente el conjunto de vectores de los parámetros
4: **for** $i = 1$ to $Nitems$ **do**
5: **for** $j = 1$ to $Nexaminandos$ **do**
6: $r \leftarrow rand$ %selecciona uniformemente un número en el intervalo $(0, 1)$
7: $p_{ij} = P(X = 1|\theta_j, \boldsymbol{p}_i)$ %calcula la probabilidad de respuesta correcta
8: **if** $p_{ij} < r$ **then**
9: $resp \leftarrow 0$
10: **else**
11: $resp \leftarrow 1$ %correcto si la probabilidad de éxito es mayor que r

Tabla 4.1: Un ejemplo simulado de respuestas de los examinandos a ítems dicotómicos en una prueba.

Las respuestas a los ítems definen los datos crudos, mientras que las habilidades y dificultades necesitan ser estimadas en una situación real.

Habilidad	Dificultad del ítem (logits)																		
(logits)	0.66	2.35	−0.48	−0.60	0.32	0.58	0.04	0.52	−2.20	1.57	0.91	−0.89	−1.12	−1.35	−2.10	−0.99	2.00	−1.60	−2.58	−0.03
1.15	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
2.10	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1
0.07	1	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	0	1	1	1
2.13	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
−0.21	1	0	1	1	1	0	0	0	1	0	0	1	1	1	1	1	0	1	1	0
−1.41	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	1	0
0.78	0	0	0	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1
0.38	0	0	1	1	0	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1
0.25	0	0	0	1	0	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1
−0.64	0	0	1	1	1	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0
−0.48	1	0	1	1	0	0	0	0	1	0	0	1	1	1	1	0	0	1	1	0
−2.18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
−1.25	1	0	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	1	1	0
−2.96	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2.60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
	Respuesta a un ítem dado, donde 0 significa error y 1 siginifica acierto																			

Algoritmo 2 `Algoritmo` para obtener habilidades de los examinandos con alguna frecuencia a un ítem dado.

`Entrada:` Dada una CDF $P(X = 1|\theta, \boldsymbol{p})$.

`Salida:` El conjunto de habilidades siguiendo la distribución $P(X = 1|\theta, \boldsymbol{p})$ dada.

1: **procedure** DENSIDAD($P(X = 1|\theta, \boldsymbol{p})$)
2: $\boldsymbol{P} \leftarrow \{P_1, P_2, \ldots, P_m\}$ %define uniformemente el conjunto de probabilidades en $(0, 1)$
3: $\boldsymbol{p} \leftarrow \{p_1, p_2, \ldots, p_n\}$ %define uniformemente el vector de parámetros
4: **for** i = 1 to Nprobs **do**
5: $P_i = P(X = 1|\theta, \boldsymbol{p})$ %cálcula las raices de esta ecuación, variable θ
6: $abilities \leftarrow roots$ %almacena las habilidades
7: **return** el conjunto de habilidades

Por otro lado, es importante mencionar que dentro de los procesos de la prueba CAT se requiere realizar la calibración de los ítems. En este sentido, los factores que participan a lo largo del proceso de calibración son múltiples y variados. Tales factores incluyen las respuestas del examinado, la estructura de los ítems (complejidad, la cual incluye a la dificultad), el número de ítems, el número de examinados, el área de conocimiento a evaluar, el nivel de conocimiento de los estudiantes, el examinador, etc.

Además, las variables independientes están definidas por las respuestas a cada ítem, el contenido de los ítems, el número de ítems, el número de examinados, el tiempo de la prueba, etc. Los parámetros de los ítems y la probabilidad de respuesta correcta al ítem define las variables dependientes.

Sin embargo, estas variables dependientes son útiles para definir variables más estructuradas con las cuales, al mismo tiempo, se vuelven las nuevas variables dependientes. Por ejemplo, la suma total de errores cuadrados, donde el error está dado por la diferencia entre los valores teóricos y los valores experimentales de la probabilidad P.

4.2. Cálculo de la distribución *prior* usando el concepto de entropía

Cuando se quiere explicar alguna relación entre rasgos latentes de los individuos, por ejemplo, características o atributos no observables y sus manifestaciones (resultados observados, respuestas o desempeño), entonces la IRT se convierte en una valiosa herramienta formal.

IRT como se ha mencionado anteriormente dentro de este documento, es una familia de modelos para analizar y predecir el comportamiento de las variables involucradas, y sus aplicaciones cubren diferentes escenarios de evaluación, exposición del ítem, el control y la calibración de ítems y la generación automática de ítems son solo algunos ejemplos de estos escenarios y las variables involucradas.

Desde un punto de vista teórico, existen varios modelos conocidos para analizar estos temas y herramientas potencialmente útiles para proponer soluciones novedosas (46). Estos modelos consideran un conjunto de ítems para definir un instrumento de medición. Un conjunto de parámetros especifíca las características del ítem que dependen de la aplicación en particular.

Los resultados de la aplicación del instrumento de medición proporcionan información sobre el rasgo latente del examinado. La IRT supone que los valores del constructo latente (por ejemplo, estrés, conocimiento, valores de actitudes) y los valores de los parámetros de algunos ítems son organizados en un continuo inobservable como variables aleatorias. Así, la IRT ayuda a establecer la posición o valor del rasgo latente del examinado en ese continuo considerando las características de los ítems y la calidad de las respuestas a ellos (60, 78).

Incluir ítems con diferentes formatos de presentación y respuesta, examinados y examinadores, son sólo una parte del escenario de evaluación. En particular, los sistemas CAT son una área donde la aplicación de la IRT es de gran utilidad para automatizar las evaluaciones de desempeño. En general, la IRT y su aplicación en los sistemas CAT suponen la existencia de un conjunto de ítems que por construcción tiene cardinalidad finita.

El proceso CAT supone que, a través de la experimentación, el pool contiene ítems calibrados (grupo de ítems calibrados o CIP en inglés), es decir, un experimento previo proporciona información sobre los valores de los parámetros de los ítems que definen su características correspondientes.

El proceso de calibración de ítems implica un análisis estadístico de las respuestas que surgen de un conjunto de sujetos de prueba y el ajuste de los datos experimentales a un modelos de función de distribución cumulativa sigmoidea (CDF) asignado a cada ítem. Este procedimiento define los parámetros correspondientes al ítem y, por lo tanto, la curva característica del ítem (ICC), la cual depende de la habilidad como rasgo latente.

Los parámetros de una ICC tienen una interpretación en términos de la dificultad del ítem y su capacidad de discriminación, entre otros, e influyen en la determinación del valor del rasgo latente que se está analizando.

Hay algunos modelos de ICC bien conocidos, como el 1PL, 2PL y 3PL, aunque se pueden encontrar 4PL y modelos 5PL también. Naturalmente, el modelo 1PL (modelo logístico de un parámetro) y 2PL (modelo logístico de dos parámetros) son los modelos más simples ya que sus parámetros tienen una interpretación directa y sus relaciones con el proceso de búsqueda del rasgo latente son claras. Tales modelos ya fueron mencionados en las ecuaciones 2.1 y 2.2 en el Capítulo 2 de este trabajo.

Geométricamente, la curva característica de un ítem con dificultad μ_1 difiere de otro

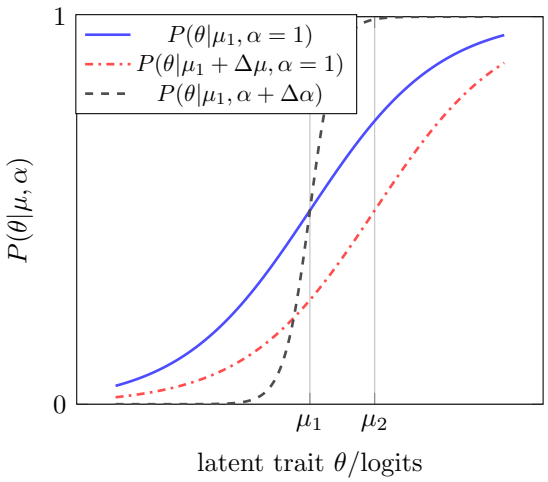

Figura 4.1: Efecto del valor de dificultad μ y el valor discriminante α de una CCI en el caso de los modelos de rasgos latentes 1PL y 2PL.

asociado con un elemento con dificultad μ_2 mediante un simple desplazamiento hacia la izquierda o hacia la derecha en el dominio de la ICC, que viene dado por los valores de los rasgos latentes, dependiendo de si $\mu_1 < \mu_2$ o $\mu_2 < \mu_1$, respectivamente, como se muestra en la Figura 4.1.

Por otro lado, el modelo de rasgo latente 2PL tiene una regla de correspondencia dada por la Ecuación (2.2), donde el parámetro α representa la capacidad discriminatoria del ítem; es decir, qué tan bien diferencia entre los examinados que tienen un rasgo latente mayor que la dificultad μ y aquellos que tienen una habilidad menor que μ.

En este caso, las gráficas de dos ítems difieren no sólo por el desplazamiento producido por el parámetro de dificultad μ, sino también por la tasa creciente de la función, que es proporcional al parámetro α (ver Figura 4.1).

Después de que el sistema CAT plantea el primer ítem al examinado y obtiene la respuesta, el siguiente paso es seleccionar un segundo ítem dentro del CIP con ciertas características dependiendo de si la respuesta es correcta o incorrecta. Después de responder el segundo ítem, el sistema selecciona el tercer ítem en función de las respuestas dadas a los dos primeros ítems, cuyas respuestas ya se tienen registradas.

Las configuraciones posibles están en el siguiente conjunto: f(correcto, correcto), (correcto, incorrecto), (incorrecto, correcto), (incorrecto, incorrecto), y así sucesivamente hasta finalizar el proceso de evaluación de un examinado.

Cuando un examinado ha respondido n ítems, el número de configuraciones de n ensayos de tipo Bernoulli que son elementos del conjunto derivado es 2^n. En cada caso, una trayectoria *sui generis* conduce al valor estimado del rasgo latente *theta* asociado con ese examinado en específico. Naturalmente, y como se ha dicho a lo largo de este trabajo, el tipo de ítems que se han manejado son dicotómicos.

Instituciones de educación superior

La Universidad Autónoma del Estado de Hidalgo y la Universidad Politécnica de Tulancingo son las dos instituciones de educación superior que de manera conjunta comenzaron con el desarrollo de un prototipo de sistema evaluador adaptable para poder ser usado en investigaciones futuras.

Alumnos de la Ingeniería en Sistemas Computacionales de la Universidad Politécnica de Tulancingo, participaron conformando diferentes equipos de trabajo y en múltiples fases del desarrollo de este prototipo, el cual fue construido con una orientación a la web, siguiendo los estándares de arquitectura y diseño de sistemas actuales en la industria del desarrollo de software.

Las dos instituciones están ubicadas en el Estado de Hidalgo, México. Específicamente la Universidad Autónoma del Estado de Hidalgo está localizada en la ciudad de Pachuca de Soto, mientras que la Universidad Politécnica de Tulancingo está situada en la ciudad de Tulancingo de Bravo.

La Universidad Autónoma del Estado de Hidalgo es más vieja que la Universidad Politécnica de Tulancingo y la carrera de Ciencias Computacionales ha sido ofrecida por casi veinte años. Por otro lado, la Universidad Politécnica de Tulancingo ha ofertado la carrera profesional de Sistemas Computacionales por los últimos quince años.

Los datos anteriores sirven como contexto para ubicar la alianza que se realizó entre investigadores de ambas instituciones con la finalidad de construir un prototipo actualizado de sistema CAT que siguiera en su diseño y construcción las tendencias más actuales de desarrollo de software aplicables a este tipo de sistema. Cabe señalar que el desarrollo del software estuvo a cargo del autor de este trabajo y forma parte de los resultados logrados a lo largo del desarrollo de esta tesis.

En párrafos posteriores se abordará un poco más a fondo el prototipo desarrollado de sistema CAT.

4.3. Prueba de Distribuciones Prior

La posibilidad de encontrar información que presente en su naturaleza una forma de distribución bimodal, puede arrojar luz en el camino de la búsqueda de distribuciones *prior* que se puedan aplicar en la estimación de habilidades utilizando la técnica de *Maximum Aposteriori*.

En la búsqueda realizada en el marco de este trabajo de tesis para encontrar diferente tipos de distribuciones *prior*, fue necesario investigar sobre aquellos tipos de pruebas cuyos resultados presentaran algún comportamiento en la distribución de los datos de sus resultados que no sean de tipo normal.

Por ejemplo, la presencia de bimodalidad está reportada en la literatura en los resultados de la llamada prueba de la figura compleja de Rey-Osterrieth (55). Por otro lado, se han encontrado resultados al menos bimodales en la pruebas conocidas como Boston Naming y Mouse Tracking entre otras (47).

Durante la realización de este trabajo, se realizaron pruebas a una población de estudiantes del área de informática de dos instituciones de educación superior. Se tomó una muestra de alumnos de la carrera de Licenciatura en Ciencias Computacionales perteneciente al Instituto de Ciencias Básicas e Ingenierías de la Universidad Autónoma del Estado de Hidalgo. Asimismo, se tomó una muestra de alumnos de la carrera de Ingeniería en Sistemas Computacionales perteneciente a la División de Ingenerías de la Universidad Politécnica de Tulancingo.

Las evaluaciones realizadas a ambas poblaciones consistieron en la aplicación de la conocida prueba de la figura compleja de Rey-Osterrieth, la cual, de acuerdo a lo reportado en la literatura, se puede encontrar un comportamiento bimodal en el análisis de sus resultados.

La aplicación de esta prueba fue una manera de poder obtener datos reales con bimodalidad para garantizar que una prueba podia contener este tipo de naturaleza en su distribución de resultados y que esto sirviera como punto de partida para la creación de distribuciones prior de manera experimental.

No obstante, aunque los resultados no fueron del todo favorables, en el sentido de que fue muy pequeña la población que realizó la prueba, estos resultados sirvieron para trabajar en la comprobación de la existencia de bimodalidad, el cual fue un trabajo de investigación que sirvió como tesis de titulación de maestría de un alumno de la Universidad Politécnica de Tulancingo, titulado "Detección de multimodalidad aplicando un estimador de densidad de npucleo en resultados de la prueba de Rey-Osterrieth".

En tal investigación se propone el uso de estimadores de densidad de núcleo para poder garantizar la naturaleza bimodal de una distribución de datos, entre estos estimadores destaca la ecuación de Silverman (68). Cabe mencionar que el asesor de dicho trabajo de tesis fue el autor de este proyecto de investigación doctoral.

4.3.1. La prueba de Rey-Osterrieth

Entre los resultados que se obtuvieron al trabajar con el estimador de densidad de núcleo para los datos obtenidos de la aplicación de la prueba de Rey-Osterrieth a los alumnos de las instituciones de educación superior mencionadas párrafos arriba destaca la comprobación de la existencia de bimodalidad en los datos obtenidos de la prueba.

Lo anterior sirve como punto de partida para poder formular una propuesta de trabajo futuro, en la que se pueda tomar esta prueba como parte de un experimento posible dentro de un sistema CAT, aplicando los modelos psicométricos de seis parámetros. En el caso de está investigación realizar tal experimentación quedaba fuera del alcance propuesto para este proyecto.

No obstante, el prototipo de sistema CAT desarrollado se preparó de tal manera que pudiera permitir la aplicación de este tipo de prueba, al menos en el sentido de mostrar vía computadora la figura, y poder capturar los dibujos realizados a través de un dispositivo táctil o de dibujo digital.

En la figura 4.3 se pueden apreciar los resultados obtenidos al realizar una regresión lineal utilizando los datos de las pruebas de la figura compleja provenientes de

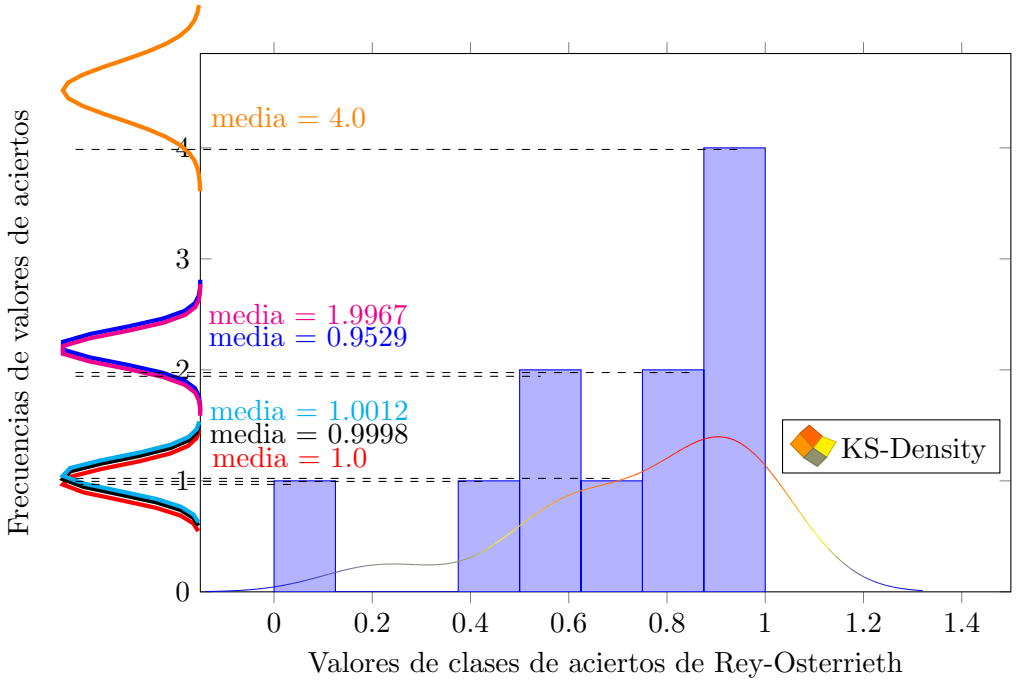

Figura 4.2: Comparativa de resultados de la experimentación aplicando el test de la figura compleja de Rey-Osterrieth

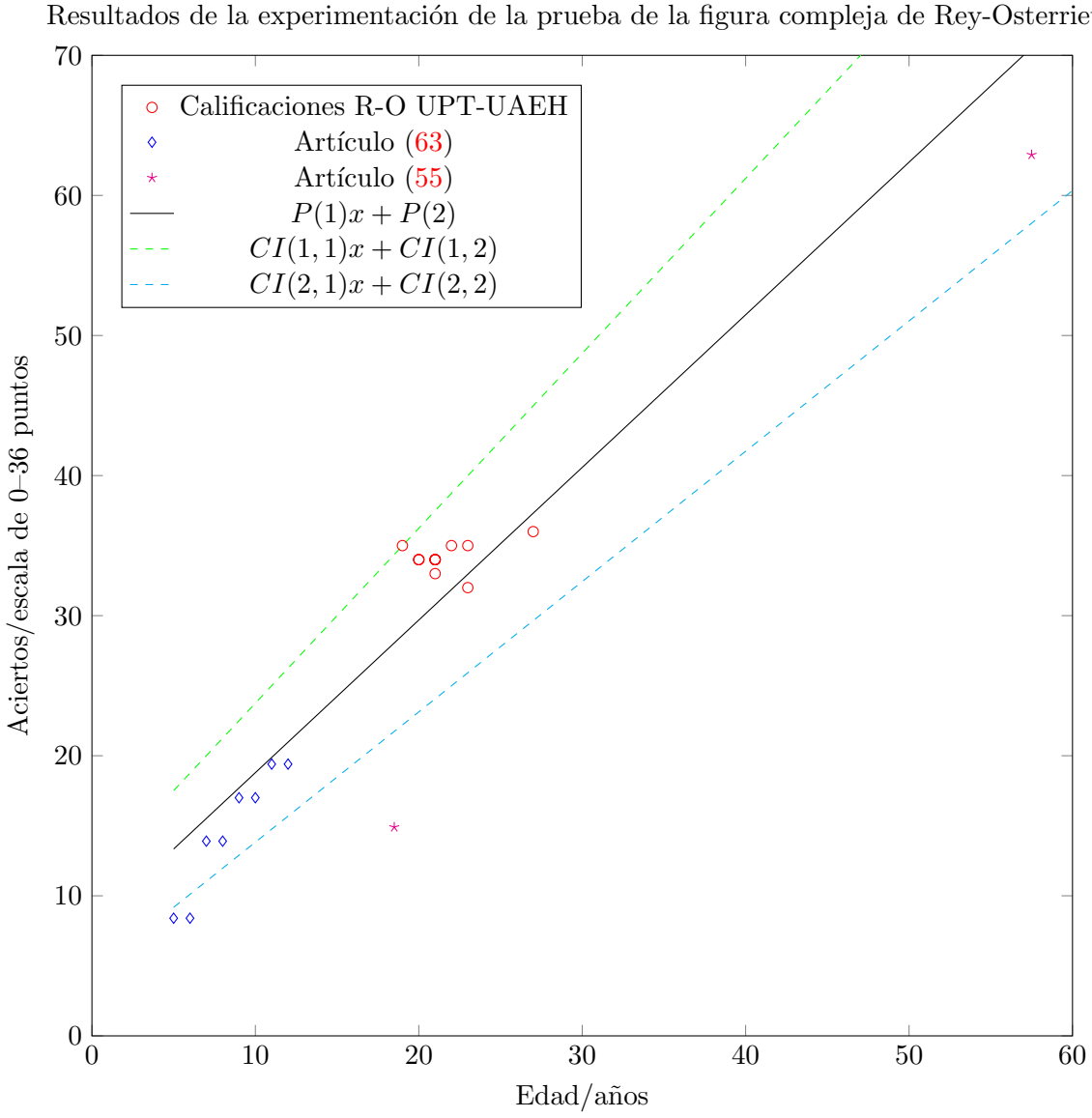

Figura 4.3: Correlación de resultados de la experimentación aplicando el test de la figura compleja de Rey-Osterrieth

la exerimentación que se llevó a cabo con estudiantes de la carrera de Ingeniería en Sistemas Computacionales de la Universidad Politécnica de Tulancingo y de la carrera de Licenciatura en Ciencias Computacionales de la Universidad Autónoma del Estado de Hidalgo.

En conjunto con los datos que aparecen en el estado del arte en los árticulos (63) y (55) los cuales mencionan las medias de los resultados en los puntajes de la prueba de la figura compleja de Rey-Osterrieth en la fase de dibujar de memoria la figura.

La experimentación se realizó primeramente tomando un grupo de alumnos de octavo semestre de la carrera de Licenciatura en Sistemas Computacionales de la Universidad Autónoma del Estado de Hidalgo. El experimento consistió en presentarles la figura compleja de Rey-Osterrieth proyectandola mediante la computadora y un cañon proyector.

Se les dio a los alumnos la instrucción de copiar la figura directamente de lo que estaban viendo, esto constituye la primer fase de la prueba. Posteriormente se retiró la proyección y se dio un tiempo de descanso libre a los alumnos de 20 minutos, pasado este tiempo se les pidió que volvieran a dibujar la figura que vieron pero está vez ocupando solo su memoria.

Esta misma prueba se volvió a llevar a cabo en una segunda institución de educación superior, la Universidad Politécnica de Tulancingo, donde también se cuenta con una carrera afin a la computación. Se tomaron tres grupos de la Ingeniería en Sistemas Computacionales. Con cada grupo se comenzó presentandoles la figura compleja de Rey-Osterrieth y permitiendoles que realizaran la fase de copia. Posteriormente se procedio a la fase de dibujo de la figura pero desde la memoria.

Los resultados que se obtuvieron se unieron con los obtenidos de la primer institución y se procedió a realizar la calificación de cada uno de ellos. Para esto se utilizó la escala de 36 puntos según lo indica (63).

Según lo reportado en la literatura la prueba de la Figura Compleja de Rey-Osterrieth arroja resultados cuya naturaleza es bimodal. Lo que se pretende al realizar esta experimentación es tener un sustento real de que al aplicar una prueba a un grupo de alumnos se pueden obtener resultados cuyas distribuciones no sean necesariamente normales, sino que pueden aparecer más de un modo y estos a su vez pueden presentar un sesgo positivo o negativo.

Según algunos investigadores este tipo de distribuciones no se puede presentar en resultados de exámenes basados en IRT; sin embargo, este experimento demuestra que si puede haber contextos donde se encuentren este tipo de distribuciones multimodales y sesgadas y por lo tanto la selección de un modelo psicométrico tendría que considerar esta posibilidad.

Esta selección de dicho modelo podría incluir a aquellos que sean de más de tres parámetros siendo opciones a considerar los modelos psicométricos de cuatro, cinco o incluso seis parámetros como los que se proponen dentro del marco de esta tesis.

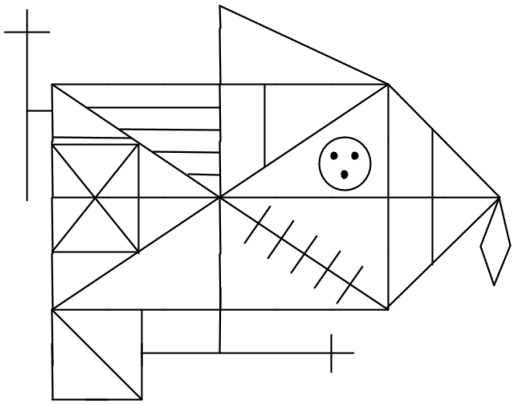

Figura 4.4: La figura compleja de Rey-Osterrieth

4.4. Sistema de Evaluación Adaptable Ariya 3.0

Dentro de los alcances de este trabajo de tesis, se encuentra la propuesta e implementación de una arquitectura de sistema CAT que sirva como punto de partida para realizar investigaciones posteriores sobre los hallazgos encontrados en este trabajo de tesis.

El resultado propuesto es una plataforma de evaluación adaptable computarizada, la cual pudiera ser utilizada para la aplicación de exámenes en línea, aplicando algún modelo de teoría de respuesta al ítem y con algoritmos de calibración de ítems previamente cargados.

Asimismo, la plataforma debería ser capaz de permitir la selección de diferentes modelos psicométricos para realizar el proceso de evaluación, todo esto como trabajo futuro, ya que el objetivo era definir la arquitectura y tener la idea de la infraestructura y su desempeño al estar alojada en un servidor creado dentro de la plataforma de un proveedor de servicios de cómputo en la nube.

La plataforma propuesta se denomina Ariya 3.0, y tiene por objetivo la gestión de instituciones de educación superior, carreras, asignaturas, exámenes, temas, subtemas y constructos para la creación de ítems que permitan hacer evaluación adaptable y determinar la habilidad de los estudiantes en algún tópico o área del conocimiento en particular.

Las siguientes imágenes muestran algunas interfaces de la plataforma Ariya 3.0, desarrollada durante la ejecución de este trabajo de investigación, como una colaboración entre la Universidad Politécnica de Tulancingo y la Universidad Autónoma del Estado de Hidalgo. Cabe hacer la aclaración, que el prototipo es meramente para validar la propuesta de la arquitectura y su correspondiente infraestructura, queda fuera del alcance de este trabajo el realizar la carga de los algoritmos de evaluación adaptable, quedando como un posible trabajo futuro de investigación.

Figura 4.5: Login de la Plataforma Ariya 3.0 para Evaluación Adaptable

En primera instancia y por manejo de seguridad, el sistema cuenta con un inicio de sesión que válida al usuario como registrado y con un rol específico para tener acceso a los recursos de la plataforma como se muestra en la figura 4.5.

Para poder dar de alta a los usuarios, la plataforma debe registrar a través del administrador del sistema a un administrador por institución educativa, quien será el responsable de dar de alta a su vez a profesores y estudiantes de su institución. Para realizar tal registro, se deben llenar algunos datos de identificación de la institución, con la finalidad de poder crear su propio espacio de trabajo dentro del sistema. Esto se muestra en la figura 4.6.

Como se menciona en párrafos arriba, el proyecto de plataforma Ariya 3.0 fue desarrollado por alumnos de la Universidad Politécnica de Tulancingo como un prototipo de cascaron no funcional para poder ser montado dentro del proveedor de servicios de cómputo en la nube. En particular se eligió la plataforma Microsoft Azure como proveedor por ser en la que más experiencia se tiene. Algunos de los alumnos que fueron parte del desarrollo se pueden observar en la figura 4.7.

El diagrama de la arquitectura propuesta en este trabajo esta dado por la siguiente imagen 4.8, en el cual se pueden observar todos los elementos que están involucrados en la infraestructura que debe ser montada dentro de los servidores de nube, a fin de poder garantizar el correcto desempeño de la plataforma a la hora de trabajar con grandes cantidades de alumnos, pertenecientes a diversas instituciones de educación o capacitación.

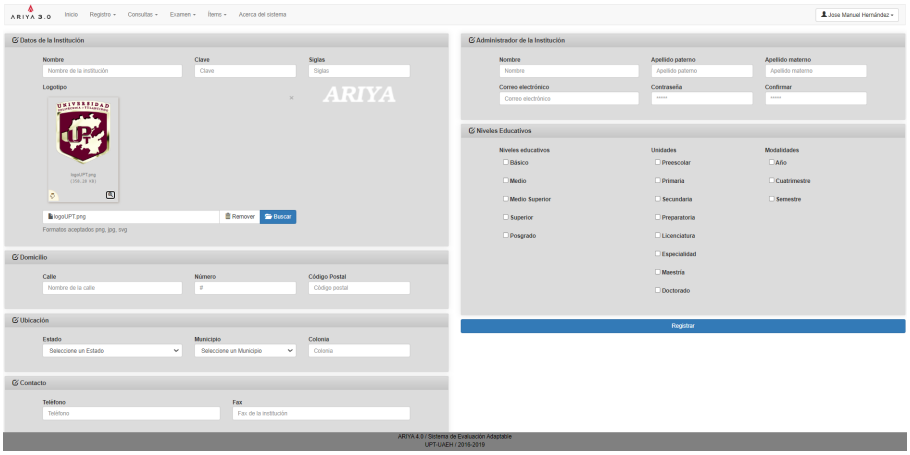

Figura 4.6: Registro de insttuciones en la Plataforma Ariya 3.0 para Evaluación Adaptable

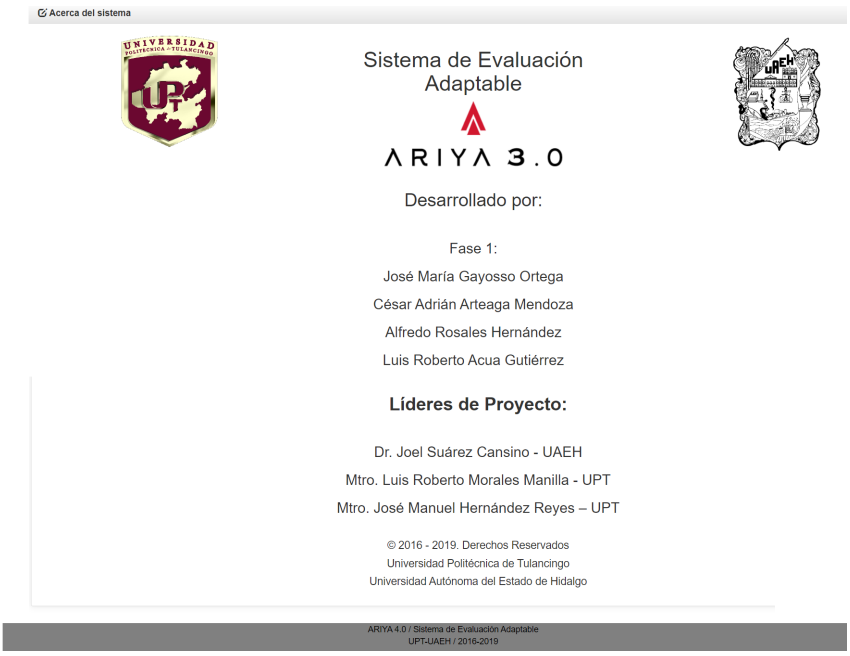

Figura 4.7: Desarrolladores a lo largo del tiempo de la Plataforma Ariya 3.0 para Evaluación Adaptable

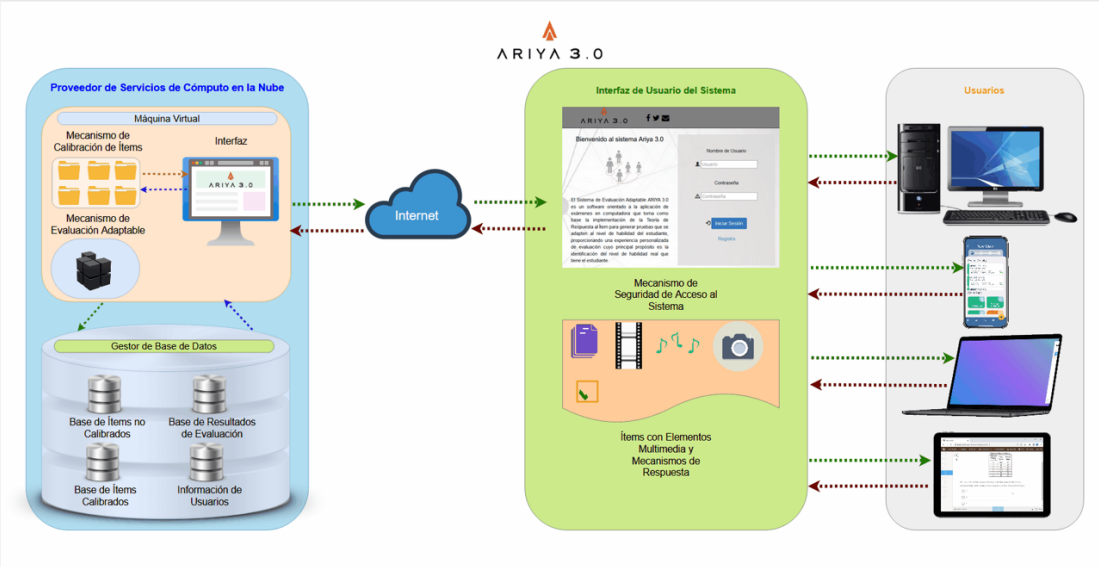

Figura 4.8: Arquitectura de implementación dentro de un proveedor de servicios de nube para un sistema CAT.

4.4.1. Alternativas de posibles trabajos futuros

Como parte de los trabajos que se fueron realizando durante este proyecto de investigación, se fueron abriendo caminos diversos en los cuales se presentaron ideas interesantes para continuar la investigación; sin embargo, de este cumulo de ideas que se originaron, algunas tuvieron que ser dejadas para trabajo futuro dados los tiempos previstos para la culminación de esta tesis.

No obstante, y para no perder estás ideas que resultan muy interesantes para ser exploradas, se creó un mapa mental que permita almacenarlas para poder trabajarlas posteriormente y no perder todos los avances que se lograron en estos tópicos hallados durante esta investigación.

Estos trabajos futuros se abordarán a más detalle en el siguiente Capítulo.

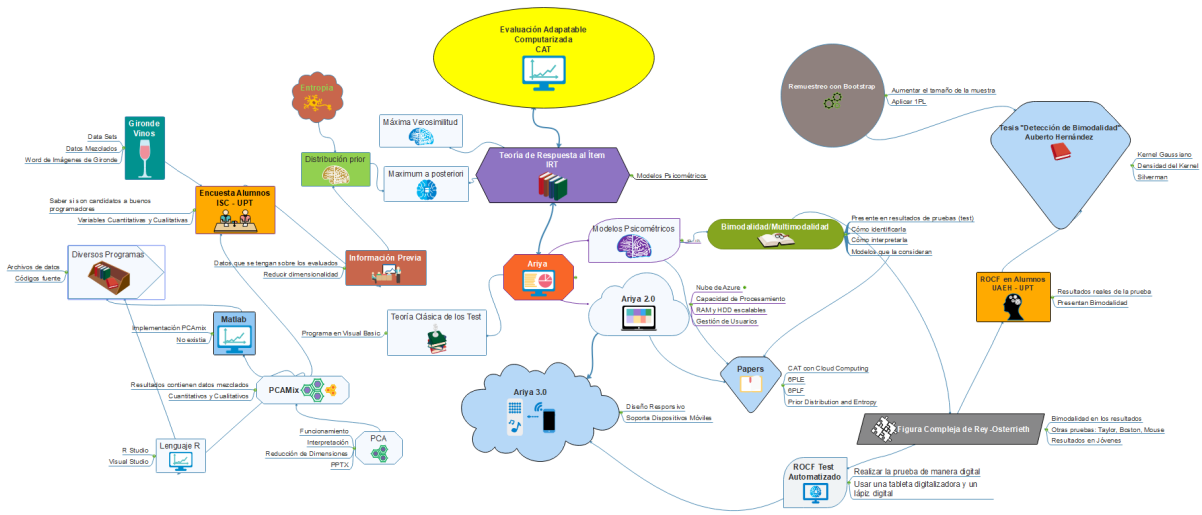

Figura 4.9: Mapa mental que concentra los conceptos explorados a lo largo de la realización de este trabajo de investigación de tesis doctoral.

Capítulo 5

Conclusiones y Trabajo Futuro

En el contexto actual de la computación aplicada a la evaluación, existen los sistemas de evaluación adaptable computarizada, los cuales tienen como principal objetivo romper con el paradigma de que el alumno debe ajustarse al examen. La idea detrás de estos sistemas es que el examen se adapte al alumno propiciando con esto que el resultado de una evaluación sea más confiable y se apegue al nivel de habilidad o conocimiento que posee realmente un sujeto que está siendo examinado.

El desarrollo de estos sistemas no es trivial ya que como se menciona a lo largo de este documento, se requiere hacer uso de modelos psicométricos cuya complejidad radica en el número de parámetros que posean. Matemáticamente hablando, entre mayor sea el número de parámetros que maneje el modelo, mayor será su complejidad. En términos de cómputo, eso implicará mayor tiempo de procesamiento para resolver el modelo, el cuál se realiza con cada ítem que el estudiante responda para poder así determinar cuál será el nuevo ítem a responder.

Actualmente, la capacidad de los equipos de cómputo ha crecido favorablemente haciendo posible realizar muchos más cálculos complejos en menor tiempo, lo que beneficia directamente el uso de modelos de más de un parámetro en un sistema CAT. Con esto en mente, se puede decir que, es factible usar modelos psicométricos de varios parámetros que se ajusten a aquellos escenarios de evaluación cuya naturaleza presente datos multimodales o sesgados, pues la capacidad de cómputo ya no será un impedimento, sino tan solo un elemento más a considerar.

En este trabajo de investigación, se realizaron algunas sugerencias interesantes desde el punto de vista de la innovación en el ámbito de los sistemas evaluadores adaptables, entre ellas destacan la propuesta de nuevos módelos psicométricos cuyos parámetros están orientados a trabajar con la multimodalidad. Asimismo, la iniciativa de usar el concepto de entropía para el cálculo de distribuciones *prior* que sirvan como semilla de inicio para el cálculo de los parámetros logísticos de módelos psicométricos en evaluaciones donde se utilice la estimación por *Maximum A Posteriori*.

Y no se debe olvidar que el propósito de estos hallazgos es poder ser usados dentro de un sistema CAT, para lo cual se sugiere dentro de este documento, la arquitectura basada en cómputo de nube para su implementación, de tal manera que se puedan

aprovechar las capacidades de escalamiento en cuanto a procesamiento, almacenamiento y memoria, ofrecidas por los proveedores de servicios de nube.

Lo anterior con la finalidad de poder enfrentar escenarios donde miles de usuarios estén resolviendo exámenes dentro de la plataforma de manera simultanea, garantizando que cada examen se ajustará eficientemente al nivel de habilidad con el que cuente cada estudiante.

Dicho lo anterior, se puede mencionar que, el presente trabajo de tesis resultó muy interesante en su realización ya que abre diversos caminos para continuar con las investigaciones de manera experimental sobre el uso de distribuciones *prior* usando la técnica MAP en un sistema CAT implementado con la arquitectura propuesta en este documento.

Como se puede observar, las aportaciones que se han encontrado favorecen bastante los objetivos iniciales, entre los que se encontraban ahondar sobre el manejo de modelos psicométricos que puedan lidiar con la multimodalidad y el sesgo en pruebas que resulten tener esta naturaleza en la obtención de los datos de las habilidades de los sujetos de evaluación. Lo anterior dentro de las una pruebas realizadas en un sistema CAT, y para tal propósito se proponen dos nuevos modelos psicométricos ricos en características que pueden ser exploradas con mayor profundidad de manera experimental.

La definición de dos nuevos modelos generalizados de Rasch que contienen 6 parámetros que se adaptan a la multimodalidad y sesgos de los datos, cumpliendo con la propiedad de objetividad especifica propuesta por Rasch, es uno de los principales aportes de este trabajo de investigación. Esto se complementa con la posibilidad que ofrecen estos módelos de poder hacer comparaciones a mayor profundida, como por ejemplo, comparar un ítem consigo mismo a través de su capacidad discriminante en subdominios.

La posibilidad de comparar examinados o ítems a través de la objetividad especifica aún cuando no pertenezcan a la misma región asintótica también es otro aporte que se debe resaltar entre otros puntos que fueron reportados mediante la publicación de un artículo de congreso internacional, mismo que se anexa en este documento A.1.

Por otro lado, es muy importante resaltar como un gran hallazgo, el aporte del cálculo de distribuciones prior usando el criterio de entropía, esto es muy relevante, ya que abre la puerta a posibles exploraciones sobre el uso de otras definiciones entrópicas que puedan crear distribuciones *prior* que al ser usadas en MAP dentro de un CAT arrojen resultados que pudieran adaptarse de mejor manera a la multimodalidad y el sesgo.

Esto dará un mayor grado de confiabilidad a los exámenes en donde sea implementado, pues los resultados estarán mucho más apegados a la naturaleza de la prueba misma y a los niveles reales de habilidad de los estudiantes evaluados, pues la información previa usada en la construcción de tales distribuciones *prior* favorece en gran medida este ajuste en los resultados.

En este apartado cabe hacer mención, que se logró demostrar a través de la teoría de maximización de la entropía, un conjunto dado de restricciones, y aplicando experimentación númerica, que es posible realizar el cálculo de una distribución *apriori* que

pueda ser usada en un sistema CAT mediante inferencia Bayesiana.

También se demostró que con el concepto de entropía aunado a la apropiada selección de restricciones se pueden resumir los datos experimentales a través de la especificación de funciones índice que estén relacionadas con los hábitos de estudio, los niveles de comprensión, la deserción del curso escolar y la reprobación de los cursos.

Esta aportación de las funciones índice es particularmente interesante, pues dichas funciones sugieren formas de cómo construir cuestionarios que aporten información del propósito del índice, por ejemplo, la deserción o permanencia de los estudiantes dentro de un curso o asignatura determinada.

Lo dicho en los párrafos anteriores, se reportó y validó mediante la publicación de un artículo de revista internacional con arbitraje y factor de impacto, mismo que se encuentra anexo a este trabajo. Esto da certeza, validez y originalidad al trabajo realizado en el sentido del cálculo de las distribuciones *prior* dentro de este documento de tesis A.1.

No menos importante es también, la propuesta de arquitectura de un sistema CAT tomando como infraestructura un servicio de cómputo en la nube para potencializar la capacidad de cómputo que se puede ofrecer al usar modelos psicométricos de cuatro o más parámetros con la finalidad de arrojar resultados mucho más confiables en las evaluaciones realizadas dentro del sistema CAT.

En este caso, la arquitectura o marco de trabajo propuesto lleva por nombre Ariya 3.0 y fue el resultado de analizar versiones anteriores del sistema dentro de un ambiente de cómputo en la nube, para poder optimizar la arquitectura y poder proponer así una infraestructura que permita su escalamiento de acuerdo al exponencial crecimiento de uso que puede llegar a tener un sistema de esta naturaleza. Esto se logró reportar también gracias a la publicación de un artículo de revista nacional, mismo que se anexa en este documento.

Finalmente y como trabajo futuro se propone realizar experimentación para probar el funcionamiento de los índices de comparación propuestos por el modelo 6PL Flexible, ya que sería novedoso encontrar índices que se adapten mucho mejor a las condiciones de problemas reales de evaluación. Dada la capacidad de estos índices para poder realizar comparaciones entre ítems y entre examinandos, resulta interesante realizar más investigación al respecto de su utilidad en un caso experimental.

Asimismo, queda abierta la posibilidad de continuar la investigación sobre la línea del análisis de componentes principales con datos mezclados usando el método PCAmix, ya que esto podría arrojar luz en el cálculo de manera experimental de distribuciones prior aplicando el concepto de entropía, para ser usadas en el sistema CAT de acuerdo a lo propuesto en este trabajo.

Esto último teniendo la posibilidad de ser integrado como nuevas funcionalidades al marco de trabajo propuesto mediante el sistema de evaluación adaptable Ariya 3.0. Donde se podría crear un test usando la figura compleja de Rey-Osterrieth para obtener resultados experimentales que presenten multimodalidad y poder aplicar los nuevos modelos psicométricos en conjunto con las distribuciones prior propuestas bajo el concepto de entropía.

Bibliografía

[1] Albers, C. J. (2000). How to assign probabilities if you must, addition on article. Technical report iwi–2000–5–04, University of Groningen. 17

[2] Albers, C. J. and Schaafsma, W. (2001). How to assign probabilities if you must. *Statistica Neerlandica*, 55(3):346–357. 17

[3] Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., and Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis*, 7:503–532. 5

[4] Asgharzadeh, A., Esmaeili, L., Nadarajah, S., and Shih, S. H. (2013). A Generalized Skew Logistic Distribution. *REVSTAT – Statistical Journal*, 11(3):317–338. 34

[5] Baker, F. B. (2001). *The basics of item response theory*. Education Resources Information Center ERIC. 15

[6] Barton, M. A. and Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1). 16

[7] Bazán, J. L., Branco, M. D., Bolfarine, H., et al. (2006). A skew item response model. *Bayesian analysis*, 1(4):861–892. 23

[8] Boeck, P. D., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., and Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, 39(12):1–28. 3

[9] Botje, M. (2006). Introduction to Bayesian inference. Lecture notes, NIKHEF National Instituut voor subatomaire fysica, Amsterdam, the Netherlands. 3

[10] Bretthorst, G. L. (1990). An introduction to parameter estimation using Bayesian probability theory. In Fougère, P. F., editor, *Maximum Entropy and Bayesian Methods*, pages 53–79. Kluwer Academic. 5, 13, 17

[11] Bromiley, P. A., Thacker, N. A., and Bouhova-Thacker, E. (2010). Shannon entropy, renyi entropy, and information. Technical report no. 2004–004, Imaging Science and Biomedical Engineering, School of Cancer and Imaging Science. 3, 17

[12] Campbell, G. (2012). Item response models for dichotomous and polytomous data in the context of generalized linear models with applications. Master science thesis, The University of Texas at Tyler, Department of Mathematics, College of Arts and Sciences, Tyler, Texas, USA. 3

[13] Conrad, K. (2005). Probability distributions and maximum entropy. Expository Paper on Analysis at the Mathematics Department of the University of Connecticut, USA. 17

[14] Coolen, F. P. A. and Newby, M. J. (1994). Bayesian reliability analysis with imprecise prior probabilities. *Reliability Engineering and System Safety*, 43:75–85. 17

[15] Dayanik, A., Lewis, D. D., Madigan, D., Menkov, V., and Genkin, A. (2006). Constructing informative prior distributions from domain knowledge in text classification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 493–500. ACM. 5

[16] Desa, Z. N. D. M. and Latif, A. (2007). Probability theory and application of item response theory. 16

[17] Dicks, H. M. (2006). Introduction to Experimental Design. Lecture Notes on Biometry 222, Department of Statistics and Biometry, University of Natal, Pietermaritzburg, South Africa. 36 pages. 33

[18] Dodier, R. (2009). On representing prior information as an asymmetric prior distribution over weights. CiteSeer Computer and Information Science Publications Collection. 17

[19] Duanli Yan, Charles Lewis, M. S. (2002). Adaptive testing without irt in the presence of multimodality. Technical report, Educational Testing Service - Statistics and Research Division - Princeton, NJ. 11

[20] Everitt, B. (2008). *Chance Rules, An Informal Guide to Probability, Risk and Statistics*. Springer. 17

[21] Feddag, M. L. and Mesbah, M. (2006). Approximate estimation in generalized linear mixed models with applications to the rasch model. *Elsevier – Computer and Mathematics with Applications*, 51:269–278. 3

[22] Fox, J. (2002). Linear mixed models. In *An R and S-Plus Companion to Applied Regression*, chapter Appendix to An R and S-Plus Companion to Applied Regression. SAGE Publications. 3

[23] Garrido, A. (2011). Classifying entropy measures. *Symmetry*, 3:487–502. 17

[24] Gelman, A. (2002). Prior distribution. In El-Shaarawi, A. H. and Piegorsch, W. W., editors, *Encyclopedia of Environmetrics*, volume 3, pages 1634–1637. John Wiley and Sons. 17

[25] Gelman, A. (2004). Parameterization and Bayesian modeling, review article. *Journal of the American Statistical Association*, 99(466):537–545. 3

[26] Goldwater, S. (2011–2012). Bayesian modelling. Notes for Lecture 12, Computational Cognitive Science, School of Informatics, University of Edinburgh. 17

[27] Gottschalk, P. G. and Dunn, J. R. (2005). The five-parameter logistic: a characterization and comparison with the four-parameter logistic. *Analytical biochemistry*, 343(1):54–65. 23, 24

[28] Hambleton, R. K. (1982). Item response theory: The three-parameter logistic model. Cse report no. 220, Center for the Study of Evaluation, University of California. 15, 24

[29] Harwell, M. R. and Baker, F. B. (1991). The use of prior distributions in marginalized bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, 15(4):375–389. 17

[30] Hedeker, D. (2002). Generalized linear mixed models. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley and Sons. 3

[31] Ho, A. D. and Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, 75(3):365–388. 2

[32] Irtel, H. (1995). An extension of the concept of specific objectivity. *Psychometrika*, 60(1):115–118. 14, 23

[33] Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions On Systems Science and Cybernetics*, 4(3):227–241. 4, 5, 17, 30

[34] Jaynes, E. T. (1988). The relation of bayesian and maximum entropy methods. In Erickson, G. J. and Smith, C. R., editors, *Maximum Entropy and Bayesian Methods in Science and Engineering*, volume 1, pages 25–29. Kluwer Academic Publishers. 17

[35] Jaynes, E. T. (2011). *Probability Theory: The Logic of Science, Chapter 11: Discrete Prior Probabilities The Entropy Principle*. Cambridge University Press, 8th edition. 17

[36] Jordan, M. I. (2010). Jeffreys priors and reference priors. Bayesian Modeling and Inference, Lecture Note 7, Department of Statistics, University of California, USA. 17

[37] Kallet, R. H. (2004). How to write the methods section of a research paper. *Respiratory Care*, 49(10):1229–1232. Daedalus Enterprises. 33

[38] Kass, R. E. and Wasserman, L. (1994). Formal rules for selecting prior distributions: A review and annotated bibliography. Technical report, Journal of the American Statistical Association. 17

[39] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules, review paper. *Journal of the American Statistical Association*, 91(435):1343–1370. 17

[40] Keller, L. A. (2000). Ability estimation procedures in computerized adaptive testing. Technical report, American Institute of Certified Public Accountants. 5

[41] Kim, S. H. (*circa* 2009). Irt modeling using the generalized linear mixed model: A longitudinal study with ordinal scale items. Technical report, Teachers College, Columbia University. 3

[42] Kumar, R., Kumar, S., and Kumar, A. (2010). A new measure of probabilistic entropy and its properties. *Applied Mathematical Sciences*, 4(28):1387–1394. 17

[43] Kvam, P. H. and Vidakovic, B. (2007). Bayesian statistics. In Kvam, P. H. and Vidakovic, B., editors, *Nonparametric Statistics with Applications to Science and Engineering*. John Wiley and Sons. 17

[44] LaValle, S. M. (2006). *Planning Algorithms*. Cambridge University Press. 3

[45] Lee, J. J. and Shin, W. S. (1990). Prior distributions using the entropy principles. *The Korean Journal of Applied Statistics*, 3(2):91–105. 17

[46] Li, H., Zhang, N., and Chen, Z. (2012). A simple but effective maximal frequent itemset mining algorithm over streams. *J. Softw.*, 7(1):25–32. 39

[47] Linda E. Nicholas, Robert H. Brookshire, D. L. M. J. G. S. and Porrazzo, S. A. (1989). Revised administration and scoring procedures for the boston naming test and norms for non-brain-damaged adults. *Aphasiology*, 3(6):569–580. 41

[48] Lord, F. (1986). Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23:157–162. 2, 3

[49] Lord, F. M., Novick, M. R., and Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *In Statistical Theories of Mental Test Scores*, pages 395–479, Addison–Wesley. 15

[50] Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book, A Practical Introduction to Bayesian Analysis, Chapter 5: Prior Distributions*. CRC Press/Chapman and Hall. 17

[51] Martin, A. D. (*circa* 2005). Bayesian analysis. In *The Oxford Handbook of Political Methodology*. 17

[52] Meeden, G. (2007). Fuzzy set representation of a prior distribution. IMS Lecture Notes – Monograph Series, Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh. 17

[53] Mitrushina, M. (2005). *Handbook of Normative Data for Neuropsychological Assessment*. Oxford University Press, USA. 2, 19

[54] Osgood, D. W., McMorris, B. J., and Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance i: Item response theory scaling. *Journal of Quantitative Criminology*, 18(3):267–296. 16

[55] Philip S. Fastenau, N. L. D. and Hufford, B. J. (1999). Adult norms for the rey-osterrieth complex figure test and for supplemental recognition and matching trials from the extended complex figure test. *The Clinical Neuropsychologist*, 13(1):30–47. PMID: 10937646. 41, 44, 45

[56] Rasch, G. (1960a). On objectivity and models for measuring. In *Stene, J. (ed.). Lecture notes*. Stene, J. (ed.). 24

[57] Rasch, G. (1960b). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Danish Institute for Educational Research, Denmark. 13

[58] Rasch, G. (1966). An individualistic approach to item analysis. In *Reading in mathematical social science*, pages 89–107. Cambridge: MIT Press. 24

[59] Rényi, A. (1961). On measures of entropy and information. In of Calif. Press, U., editor, *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pages 547–561. 17

[60] Revelle, W. (2013). The "New Psychometrics"– item response theory. In *An Introduction to Psychometric Theory with Applications in R*, chapter 8, pages 241–264. 3, 39

[61] Ricketts, J. H. and Head, G. A. (1999). A five parameter logistic equation for investigating asymmetry of curvature in baroreflex studies. *American Journal of Physiology*, 277(R441–R454):441–454. 28, 34

[62] Rijmen, F., Tuerlinckx, F., Boeck, P. D., and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8(2):185–205. 3

[63] Rosselli, M. and Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and cognition*, 52(3):326–333. 44, 45

[64] Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, (72):217–232. 13

[65] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17. 3

[66] Sanghyun S. Jeon, S. Y. W. S. I. F. (2011). Deriving prior distributions for bayesian models used to achieve adaptive e-learning. *Knowledge Management & E-Learning: An International Journal (KM&EL)*, 3:251–270. 5

[67] Scheiblechner, H. H. (2009). Rasch and pseudo-rasch models: suitableness for practical test applications. *Psychological Test and Assessment Modeling*, 51(2):181. 14, 23

[68] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press. 42

[69] Sinharay, S. (2003). Bayesian Item Fit Analysis for Dichotomous Item Response Theory Models. Research and Development RR–03–34, ETS, Princeton, NJ, USA. 52 pages. 35

[70] Stephens, M. (1999). Dealing with multimodal posteriors and non–identifiability in mixture models. *Submitted to Journal of the Royal Statistical Society, series B*. 3

[71] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 62(4):795–809. 3

[72] Swaminathan, H. and Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51(4):589–601. 3

[73] Syversveen, A. R. (1998). Noninformative bayesian priors, interpretation and problems with construction and applications. CiteSeer Computer and Information Science Publications Collection. 17

[74] Tchourbanov, A. (2002). Prior distributions. Technical report, Department of Biology New Mexico State University Road Runner Gnomics Laboratories. 6, 17

[75] Tektas, D. and Günay, S. (2008). A bayesian approach to parameter estimation in binary logit and probit models. *Hacettepe Journal of Mathematics and Statistics*, 37(2):167–176. 17

[76] Van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, 23(1):21–29. 3

[77] Van der Linden, W. J. and Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In Van der Linden, W. J. and Glas, C. A. W., editors, *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, chapter 1. Springer. 17

[78] Veldkamp, B. P. and Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação*, 21(78):57–82. 39

[79] Wang, T.; Vispoel, W. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(2):109–135. 2, 3

[80] Wu, C. and Chen, J. (2006). Sampling and Experimental Design. Lecture Notes on Stat 322/332/362, Department of Statistics and Actuarial Science, University of Waterloo. 73 pages. 33

Apéndice A

Código/Manuales/Publicaciones

A.1. Apéndice

Artículos generados dentro de este trabajo de investigación.

entropy

MDPI

Article

Prior Distribution and Entropy in Computer Adaptive Testing Ability Estimation through MAP or EAP

Joel Suárez-Cansino [1], Virgilio López-Morales [1,*], Luis Roberto Morales-Manilla [2], Adrián Alberto-Rodríguez [1] and Julio César Ramos-Fernández [3]

[1] Basic Sciences and Engineering Institute, Systems and Information Technologies Research Center, Intelligent Computing Research Group, Autonomous University of Hidalgo State, Col. Carboneras, Mineral de la Reforma 42184, Hidalgo, Mexico
[2] Department of Software Development, Advanced Computing and Innovation Research Group, Polytechnic University of Tulancingo, Tulancingo 43629, Hidalgo, Mexico
[3] Department of Mechatronics, Smart Technologies Applied to Social Development Research Group, Polytechnic University of Pachuca, Zempoala 43830, Hidalgo, Mexico
* Correspondence: virgilio@uaeh.edu.mx

Abstract: To derive a latent trait (for instance *ability*) in a computer adaptive testing (CAT) framework, the obtained results from a model must have a direct relationship to the examinees' response to a set of items presented. The set of items is previously calibrated to decide which item to present to the examinee in the next evaluation question. Some useful models are more naturally based on conditional probability in order to involve previously obtained hits/misses. In this paper, we integrate an experimental part, obtaining the information related to the examinee's academic performance, with a theoretical contribution of maximum entropy. Some academic performance index functions are built to support the experimental part and then explain under what conditions one can use constrained prior distributions. Additionally, we highlight that heuristic prior distributions might not properly work in all likely cases, and when to use personalized prior distributions instead. Finally, the inclusion of the performance index functions, arising from current experimental studies and historical records, are integrated into a theoretical part based on entropy maximization and its relationship with a CAT process.

Keywords: entropy; CAT; Kullback–Leibler divergence; likelihood function; maximum a posteriori; expectation a posteriori; Bayesian inference; performance index function; item response theory; item characteristic curve

Citation: Suárez-Cansino, J.; López-Morales, V.; Morales-Manilla, L.R.; Alberto-Rodríguez, A.; Ramos-Fernández, J.C. Prior Distribution and Entropy in CAT Ability Estimation through MAP or EAP. *Entropy* **2023**, *25*, 50. https://doi.org/10.3390/e25010050

Academic Editor: Antonio M. Scarfone

Received: 20 September 2022
Revised: 29 November 2022
Accepted: 14 December 2022
Published: 27 December 2022

1. Introduction

When one wants to explain some relationship between latent traits of individuals, for instance, unobservable characteristics or attributes and their manifestations (observed outcomes, responses, or performance), then item response theory (IRT) becomes a valuable formal tool. IRT is a family of models to analyze and predict the behavior of the involved variables, and their applications cover different assessment scenarios. Item exposure control, item calibration, and automatic item generation are only some examples of these scenarios and the involved variables. From a theoretical point of view, there are several well-known models to analyze these topics and potentially helpful tools to propose novelty solutions [1–5]. These models consider a set of items to define a measurement instrument. One set of parameters specifies the item's characteristics that depend on the particular application. The results of the application of the measurement instrument provide information about the examinee's latent trait. IRT assumes that the latent construct values (e.g., stress, knowledge, attitudes values) and some items' parameter values are organized in an unobservable continuum as random variables. Thus, IRT helps to establish

the position or value of the examinee's latent trait on that continuum by considering the items' characteristics and the quality of responses to them [6,7].

Items with different presentation and answer formats, examinees, and examiners are just a part of the assessment scenario. Particularly, computer adaptive testing (CAT) is an area where the application of IRT is highly useful to automatize performance assessments.

In general, IRT and its application to CAT assume the existence of a pool of items that by construction has finite cardinality. The CAT process supposes that, through experimentation, the pool contains calibrated items (calibrated items pool or CIP), i.e., a previous experiment provides information about the values of the items' parameters that define its corresponding characteristics.

The item calibration process entails a statistical analysis of the responses arising from a set of test subjects, and the fitting of the experimental data to a sigmoidal cumulative distribution function (CDF) model assigned to every item. This procedure defines the corresponding item's parameters and, therefore, the item characteristic curve (ICC) which depends on the latent trait as *ability*. The parameters of an ICC have an interpretation in terms of the item's difficulty and the item's discrimination capability, among others, and they influence the determination of the value of the latent trait under analysis. There are some well-known ICC models, such as 1PL, 2PL, and 3PL, although one can find 4PL and 5PL models, too. Naturally, 1PL (one-parameter logistic model) and 2PL (two-parameter logistic model) are the simplest models since their parameters have a direct interpretation and their relationships with the process of searching for the latent trait are clear.

Equation (1) defines the general structure for the 2PL model and gives the conditional probability of correctly answering an item with known difficulty and discriminant:

$$P(\theta|\mu,\alpha) = \frac{1}{1+e^{-\alpha(\theta-\mu)}}, \tag{1}$$

where the parameters μ and α represent, respectively, the item's difficulty and the item's discriminant, and θ refers to the examinee's ability.

An examinee participating in a CAT has to pass through three steps.

1. Assignment of an initial estimate of the examinee's ability or the item's difficulty, since the system needs to know the characteristics of the first item in the evaluation process.
2. The system saves the examinee's response, decides if the examinee gave a correct answer or not, and builds the response pattern for this specific testing process.
3. The system considers the response pattern and the selected latent trait model to build a likelihood function, intending to decide what proper item (calibrated) must come next. There are several methods to do this, and here we apply the definition of the prior distribution. After deciding what item comes next, the CAT procedure poses this item to the examinee, and the testing returns to the second step again. Our main contribution aims to solve some problems in this step.

A reliable estimate of the next question to be presented, for instance, through the concept of the maximum likelihood function, requires at least two responses to the presented items in the evaluation process. One of the items needs to have a right answer, and the other an incorrect one. Only in this case, the likelihood function will have an extreme point in the set of values of abilities and, therefore, a maximum value at this point.

Note that in the event that all the answers obtained were correct (or incorrect), the likelihood function is just a sigmoid (the ICC of the items) that does not have extreme points in the domain of the examinee's ability. Because of this reason, it becomes impossible to compute the next estimation of the examinee's ability.

To overcome the difficulty of estimating the next examinee's ability by just using the likelihood function, some authors have proposed different options:

1. The use of two fictitious items with high and low probabilities to ensure that the examinee answers alternatively correctly and incorrectly to the items.

2. The use of heuristic formulas to estimate the examinee's latent trait until a maximum likelihood makes it possible to estimate the value.
3. To define prior distributions until one can apply a likelihood function to estimate the examinee's ability. This proposal relates directly to the statement of the research problem in this paper.

The first option has the inconvenience that the estimated latent trait value after applying the first non-fictitious item reaches very extreme values, and the second non-fictitious item of the CAT process provides more information for that extreme value of ability. Thus, the second non-fictitious item becomes less informative for the final ability value and it does not contribute considerably to the test precision [8,9].

The second option has the inconvenience that in some circumstances the CAT process does not converge, although when the increment (or decrement) of the latent trait value is variable this phenomenon does not occur [8,9].

Finally, the third option has several inconveniences:

(i) A general use of prior information in educational assessment appears to be inhibited solely by the assumption that including a priori information on test scores in performance assessment may be unfair to students [10].
(ii) Assuming that regular evaluation practices include information provided by the examinee (regarding past experiences) or collected from multiple sources in the assessment procedure without specifying the type of the sources [10].

Additionally, there is not much information about the potential risks when prior information is not perfectly accurate. Overconfidence in inaccurate prior information may in fact increase test length and/or lead to severely biased final latent trait estimates. In this event, then the system could, for example, select an incorrect starting point or introduce bias in the trait estimation process, and provide items that do not match the participant's trait level [10]. On the other hand, the level trait does not depend solely on the examinee's performance but on the values of mean and variance that one assigns to the trait's prior distribution in the population [8].

From a theoretical point of view, depending on the established a priori distribution, one can obtain a multimodal posterior so that the Bayesian MAP estimation might refer to a local maximum [8].

Finally, in some cases, the Bayesian procedures provide estimation with a specific regression toward the mean of the prior distribution of the latent trait. This phenomenon can favor examinees with low levels and affect examinees with high ability [8].

There are several advantages and drawbacks of introducing information before starting an adaptable evaluation process. The usual way of building prior distributions lends itself to subjectivities, even though the benefits in the administration of the evaluation are undoubted [11,12]. However, the subjectivity inherent in the prior distribution can be minimized as long as reasonable evidence supports the distribution proposal [11].

In this work, we address the role that the entropy can play in the reduction of this subjectivity in the construction of the prior distribution by using a set of proposed constraints related to entropy. Any distribution must satisfy these constraints that consider, for instance, its first and second moments and the academic framework such as, for example, school dropout and failure, among others.

1.1. Problem Statement

The use of Bayesian statistical inference in the CAT process is delicate and has to justify the application of essential components such as the prior distribution [13]. Different theoretical and experimental techniques exist to determine the prior distribution to initialize the CAT process. Some authors suggest that physical, mathematical, engineering, expert opinion models, historical data under similar circumstances, or other reasonable information can support the prior proposal [13]. Thus, we formally introduce the models related to academic performance, for example, the failure rate, the dropout rate, the study

habits index, and the subject comprehension index, among others, to further specify the structure of the prior distribution by using the concept of entropy.

1.1.1. Preliminaries

The estimation of the ability of a test subject presents problems at the beginning of the evaluation process when using the maximum likelihood method and when the examinee responds correctly or incorrectly to all the test items. Several proposals solving this problem have been published and there are some options based on Bayesian inference [14]. In particular, the MAP or EAP techniques use the concept of the prior distribution, with the drawback that the definition of the structure of this distribution can lead to subjectivities.

1.1.2. Originality

Within the given context, there is not enough information about the best prior distribution to be selected. Due to the Bayesian nature, MAP or EAP techniques require previous knowledge of the prior distribution, which contains initial statistical information about the ability of the examined subject.

One typically uses a normal distribution [15–17], but there is no evidence that this is necessarily correct since there is no reliable way to support the decision to opt for one prior type of distribution over another. The initial choice of the a priori distribution is paramount since it directly affects the calculation of the skill estimate and other parameters.

Furthermore, the structure of the psychometric model supporting the Bayesian inference process must be considered. An adequate structure selection provides an appropriate interpretation of each item's characteristics, predicts the consequences of using a psychometric model with the selected characteristics, and ensures the relationship between these options and the multimodality and bias characteristics in the a posteriori distribution that finally helps to estimate the corresponding latent trait [18–21].

1.1.3. Impact

In order to solve the former problems, one must then propose the form of the prior distribution through formal criteria to select good prior distributions. Some authors define some non-formal criteria and give quite illustrative examples of how the selection of a priori distribution affects the posterior distribution [22,23]. However, this research paper works mainly with the concept of entropy and, in a first instance, with the definition proposed by Shannon [24].

1.2. Article Structure

The paper is organized as follows: Section 2 focuses on a short hypothesis or conjecture statement and the paper's objectives. Section 3 contains a brief discussion about some works on the importance of the prior distribution and the most common assumptions that the researchers make on its structure. This part also discusses the role that entropy could play in determining the *a priori* distribution and the previous work in this regard, but not within the framework of a CAT.

Section 4 briefly describes the differences between the 1PL and 2PL latent trait models and explains the meanings of the difficulty and discriminant parameters. Through these models and definitions, the concepts of maximization a posteriori, or MAP, and expectation a posteriori, or EAP, and their relationships with the prior distribution are introduced.

In addition, one recalls Shannon's concept of entropy and states the ansatz (assumptions about the form of an unknown function, made to facilitate the solution of a problem) that give rise to the proposed method for estimating the prior distribution. Section 5 illustrates our numerical experimentation results, and provides and discusses the findings about the structures of the a priori distributions obtained through the proposed method. Finally, Section 6 synthesizes the results from numerical experiments and remarks some comments about the future work within the topic of the paper.

2. Hypothesis or Conjecture Statement

The specification of the prior distribution is a problem that does not have a straightforward solution in a CAT process. Part of this is due to the lack of formal procedures to get an analytical form of the distribution since there is no standard procedure on how the required information, to start the CAT process, can be integrated into a methodology to get an approximation to the model defining acceptable prior distributions.

2.1. Hypothesis Statement

If there is no formal procedure to determine prior distributions to initialize the CAT process, and Shannon's entropy plays the role of the objective function depending on the *a priori* information distribution, which is subject to constraints of normality, mean values, and variance of the ability, in addition to the satisfaction of academic performance constraints considering failure, study habits, subject comprehension, and dropout rates of the course of interest, among others, then the formal finding of a prior distribution to initialize a CAT process is possible.

2.2. Objectives

Our general objective is to build informative prior distribution functions by considering the maximization of Shannon's entropy as a cost function that depends on the distribution of *a priori* information, subject to normality, mean, variance, and academic performance constraints, to obtain formal prior distribution expressions. The specific objectives are the following:

1. To propose an ansatz about the school performance of the examinees, considering that they must be random functions depending on the random variable defined by the latent trait θ and some specific parameters, through the analysis of qualitative results obtained by various authors and, with these results, subsequently introduce distribution constraints based on the proposed assumptions.
2. To build an objective function to maximize entropy by considering the definition of entropy and the proposed ansatz in objective 1, and to obtain a methodology building and applying prior distributions in the CAT process.
3. To obtain experimental results numerically by simulating the behavior of the CAT process, to later make comparisons of the advantages and disadvantages of different scenarios that use prior distribution estimations.

3. State of the Art

The determination of the a priori distribution is experimental or through consultation with experts. However, regarding the role that entropy can play in searching for an adequate prior distribution, one can find a few research works on the topic. In this sense, to know how the *a priori* distribution behaves, one needs prior knowledge of the properties that it may have (the normality of the distribution is the simplest example of this knowledge, but there may be some other properties that are possible to know beforehand) [25,26].

The concept of prior distribution plays a fundamental role in Bayesian inference, so experimental determination of how to obtain these distributions and what theoretical methods should be to get something similar are paramount. To build a prior distribution, it is first necessary to specify representative random variables. In this sense, there are several possibilities that this paper introduces.

In the first instance, one assumes that the prior distributions must be related to the parameters of the selected psychometric model and the examinee's latent trait variable to evaluate as proposed in [27–30] through the experimental construction of the corresponding prior distributions.

Additionally, one can consult experts in the knowledge domain to evaluate in order to obtain an opinion about the form or structure that the a priori distribution should have [28,30].

Despite not being connected to CAT systems, one can find in the literature some theoretical attempts to determine the structure of the prior distribution using the concept of entropy [26]. In addition to possibly getting the expert's opinion, no known procedure integrates the results of the experimental process which, with a theoretical basis, can specify the characteristics or conditions under which one can obtain adequate prior distributions; that is, leading to unbiased posterior distributions, without multimodality, and to reliable latent trait estimates [18–21].

From a theoretical point of view, some contributions have dealt with the topic of informative and non-informative prior distributions, and they apply these definitions as academic examples to show the effects that the *a priori* distribution has over the *a posteriori* distribution [31].

In practice, heuristic distribution applications are analyzed when they are not supported by experimental data. In fact, some authors state that the practical consequences of using a prior distribution can depend on data. A heuristic distribution, such as the uniform or the normal with zero mean and unitary variance, can lead to nonsense inferences even when it has a large sample size. Currently, the study of prior distributions becomes relevant to analyze problems inside the frontiers of applied statistics [32,33].

In this sense, our paper integrates the experimental part of obtaining information related to the examinee's academic performance into the theory of maximum entropy. The structure of the academic performance index functions supports this experimental part, which, as an additional result, explains under what conditions one can use heuristics priors. Additionally, the paper remarks that the heuristic prior distributions could not properly work in all the cases and that one must consider personalized prior distributions instead. Finally, the inclusion of the performance index functions, arising from current experimental studies and historical records, are integrated into a theoretical part based on the entropy maximization and its relationship with a CAT process.

4. Modeling Initialization of the Evaluation Process

Geometrically, the characteristic curve of an item with difficulty μ_1 differs from another associated with an item with difficulty μ_2 by a simple shift to the left or right on the domain of the ICC, which is given by the latent trait values, depending on whether $\mu_1 < \mu_2$ or $\mu_2 < \mu_1$, respectively, as shown in Figure 1. On the other hand, the 2PL latent trait model has a correspondence rule given by Equation (1), where the parameter α represents the discriminatory capacity of the item; that is, how well it differentiates between examinees who have a latent trait greater than difficulty μ and those who have an ability less than μ. In this case, the graphs of two items differ not only by the displacement produced by the difficulty parameter μ but also by the function's increasing rate, which is proportional to the parameter α (see Figure 1).

After the CAT system poses the first item to the examinee and obtains the answer, then the next step selects a second item within the CIP with characteristics depending on whether the answer is correct or incorrect. After answering the second item, the system selects the third item depending on the answers given to the first two items, whose response's configurations are in the following set:

$$\{(correct, correct), (correct, incorrect), (incorrect, correct), (incorrect, incorrect)\},$$

and so on until the evaluation process of an examinee ends.

When an examinee has answered n items, the number of configurations of n Bernoulli-like trials that are elements in the derived set is 2^n. In each case, a *sui generis* trajectory leads to the estimated value of the latent trait θ associated with the specific examinee. Naturally, in this case, the items are dichotomous.

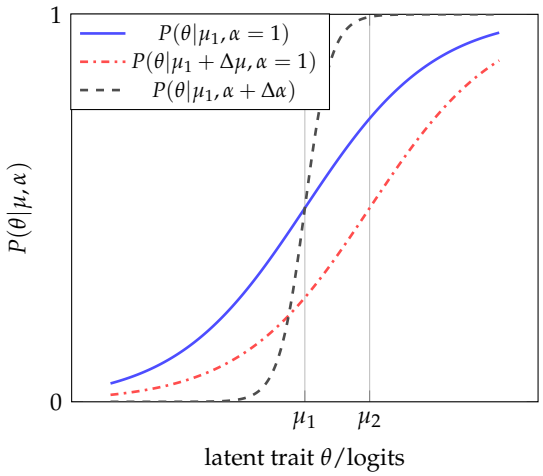

Figure 1. Effect of the difficulty value μ and the discriminant value α of an ICC in the case of the 1PL and 2PL latent trait models.

One of the main characteristics of a CAT is that the test should have the smallest possible number of items and still estimate the value of the specific ability, i.e., the selection of the n items in a particular sequence is not arbitrary. Given the response sequence for the first $n-1$ items, it is possible to estimate the $(n-1)$-th value of the latent trait θ, which has the symbolic representation θ_{n-1}.

By knowing this estimate of the latent trait at iteration $(n-1)$-th, the CIP provides the next most informative item [8,34,35]. Fortunately, the Fisher's information index gives a criterion to select the most informative one (see Equation (2)),

$$I(\theta) = \frac{\left(\frac{d}{d\theta}P(\theta|\vec{p})\right)^2}{P(\theta|\vec{p})Q(\theta|\vec{p})}, \tag{2}$$

where $\vec{p}$ is the vector of parameters defining the structure of the latent trait model correspondence rule and $Q(\theta|\vec{p}) = 1 - P(\theta|\vec{p})$. For the 1PL and 2PL models, $I(\theta)$ is given by Equations (3) and (4), respectively.

$$I(\theta) = P(\theta|\mu)Q(\theta|\mu), \tag{3}$$
$$I(\theta) = \alpha P(\theta|\mu, \alpha)Q(\theta|\mu, \alpha). \tag{4}$$

Under the condition of independence and identical distribution of the items in the CIP, it is possible to build a likelihood function with the first $(n-1)$ ICCs that the CAT system has applied up to the current answered items. In the best case, this likelihood function will have extreme points in the domain given by the latent feature values, implying that the likelihood function has at least one maximum [20,36].

However, the worst-case scenario is that all the first $(n-1)$ items have a correct answer or all have a wrong answer. If one of these situations occurs, then building a likelihood function with a maximum, at least, is impossible. How does one determine the estimate of the latent trait value, in this case, to continue with the adaptive testing process?

There are several solution proposals to this problem, but a natural one [37] involves statistical information before the start of the evaluation process by using a Bayesian procedure. The idea is to use a prior distribution with which it is possible to use Bayesian argumentation to obtain estimates of the latent trait. Algorithm 1 provides a simple outline of this process.

Algorithm 1 Outline of the computerized adaptive assessment process.

1: **procedure** EVALUATIONPROCESS
2: $item_1 \leftarrow$ select the first item with parameters $\vec{p}_1$
3: *top*: *response* $\leftarrow$ reply to item $i - th, 1 \leq i$
4: $\theta \leftarrow$ estimate latent trait based on the pattern of first responses to the i items
5: **if** pattern is all correct or incorrect **then**
6: **return** use *prior distribution* and Bayes
7: **else**
8: **return** use Maximum Likelihood estimation
9: $item_{i+1} \leftarrow$ select item with a higher Fisher information in the Items Pool
10: **goto** *top*

Note that the selection of the first item in step 2 of Algorithm 1 can proceed in at least one of two possible manners, namely

1. To calculate an estimate of the latent trait θ before starting the evaluation process and, with this estimate, to determine the item with the maximum Fisher information within the CIP [37].
2. To compute an estimate of the parameters of the first item (difficulty, discriminant, guessing, etc.) following some of the methods in [37].

Step 5 is central to the Algorithm 1 since Bayes' Theorem requires a *prior* distribution.

Bayes' theorem involves the use of a prior distribution to calculate the so-called posterior distribution. However, selecting an *a priori* distribution is not trivial, and one must ensure that this distribution provides the highest amount of information about each of the examinees.

The following steps are essential to the understanding of our methodology:

1. To know the relationships among the *a posteriori* probabilities, the prior probability and the likelihood function.
2. Find the *a priori* probability and its closest dependence on an academic framework.
3. The analysis of discrete and continuous cases (the latter being of greater interest).

Regarding the first step, and given in Equation (5),

$$p(\theta|\vec{p}) = \frac{L(\vec{p}|\theta)}{p(\vec{p})} \cdot p(\theta), \tag{5}$$

we note that the likelihood function $L(\vec{p}|\theta)$ is the product of the item characteristic curves that arise throughout a specific individual evaluation pattern result. In this case, $p(\theta)$ directly gives the prior distribution.

Thus, the prior distribution is a function of the latent ability or trait θ. Finally, the transition from the discrete case to the continuous one is provided by:

$$\mathcal{S}(p) = -\int_{-\infty}^{+\infty} p(x) \log p(x) dx, \tag{6}$$

which may be subject to constraints of the form

$$\int_{-\infty}^{+\infty} f(x)p(x)dx = \langle f(X) \rangle, \tag{7}$$

where $\langle f(X) \rangle$ is the expectation of the random variable defined as $f(X)$, where X is a random variable whose values x's define the population of interest. Equation (7) provides the general form of the constraints.

At this stage of the CAT process, the prior distribution $p(\theta)$ and the likelihood function $L(\vec{p}|\theta)$ are available to compute the posterior distribution $p(\theta|\vec{u})$ through Equation (5).

Once one calculates the posterior distribution, then estimates the next latent trait value, there are two possibilities:

- Determine the extreme point in the domain of the posterior distribution and compute the maximum value of the distribution at this extreme value.
- Determine the mean value of the latent trait population along the whole domain of the posterior distribution.

In order to sketch how the informative prior distribution can be related to the academic framework of the examinees, we propose several ansatzes.

Ansatz for Different Indices of Student Performance as a Function of the Ability θ

Some works use the concept of entropy [38] as an alternative for the construction of informative prior distributions. In this paper we introduce the maximum entropy through the application of optimization techniques to maximize the information that the entropy will yield concerning the specific examinee.

In addition to the distribution normality constraint, the latent trait mean and variance specifications, we analyze the contribution of special examinees' academic performance constraints to properly determine the population distribution through entropy maximization. In this sense, we apply the concept of an index (a random variable), depending on the ability θ.

By defining entropy as a cost function, entropy maximization considers that this function is subject to a list of constraints other than the constraints based on normality and first and second moments. The additional elements of the list of restrictions include the dropout rate, the failure rate, and the habits of study rates from one or more courses belonging to an examinee's record. Additionally, one can consider the index of understanding of topics that an examinee has in a historical academic record.

To relate study habits rate and its relationship to an ability function, several authors [39–43] have identified some factors between good habits and excellent academic achievement:

1. Attend classes regularly.
2. Take notes while teaching.
3. Concentrate on studying.
4. Study with a view to gaining meaning, not storing facts.
5. Prepare a schedule.
6. Follow the schedule.
7. Have appropriate rest periods.
8. Facing problems considering the home environment and planning.
9. Facing the challenges posed by the school environment.
10. Keep a daily update of the work done.

The statistical results in [39] confirm that the study habits index is indeed an increasing function of ability, as illustrated in Figure 2a. By applying methodologies such as those indicated by the authors in [44,45], one can adequately prepare a questionnaire including questions related to the preceding list.

On the other hand, a lack of academic and social skills leads to the student being unable to process the information transmitted by the instructor [46]. Then, we can infer that the understanding of topics is related to the student's ability, as Figure 2b illustrates.

By means of Figure 2a, we state that the study habits index behaves sigmoidally depending on the examinee's ability with the following correspondence rule

$$f(\theta) = B_h + \frac{1 - B_h}{1 + b_h e^{-a_h \theta}}, 0 < a_h. \tag{8}$$

Meanwhile, Figure 2b states that the rate of topic comprehension by students also has a sigmoidal behavior as follows

$$g(\theta) = B_c + \frac{1 - B_c}{1 + b_c e^{-a_c \theta}}, 0 < a_c. \tag{9}$$

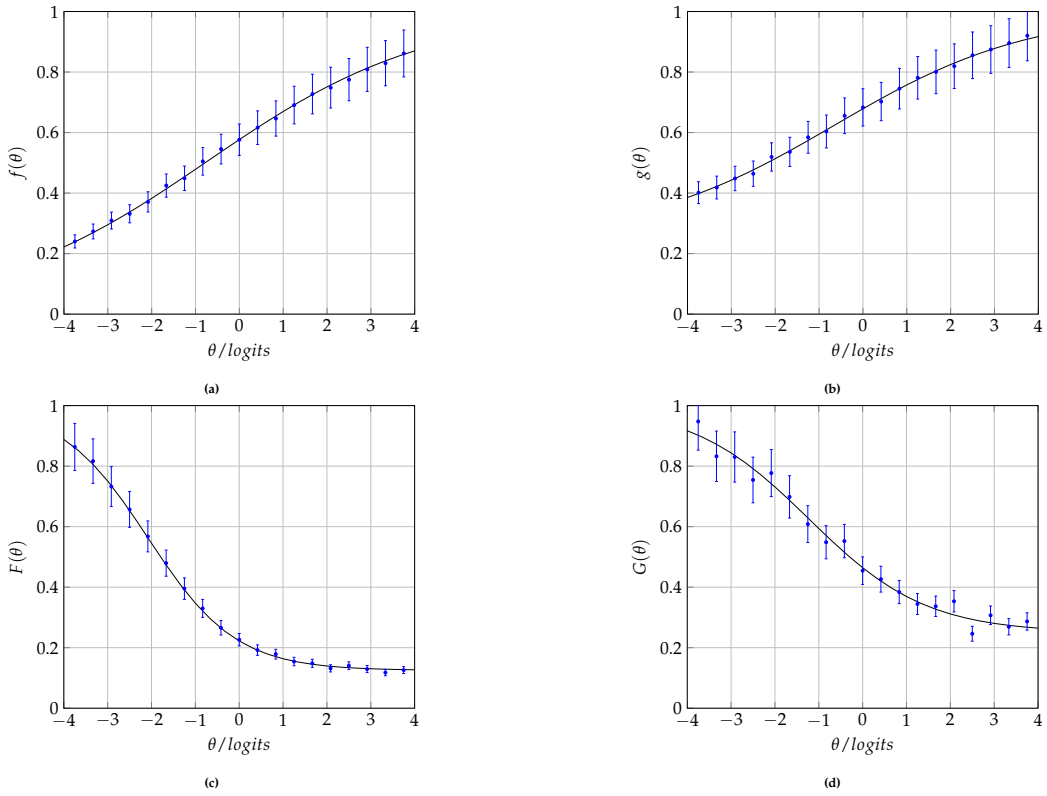

Figure 2. Assumed behavior for some of the indices that can be included as constraints for the determination of the prior distribution by means of entropy. (**a**) Study habit index f as a function of ability θ. (**b**) Subject comprehension index g as a function of ability θ. (**c**) Course dropout rate F as a function of ability θ. (**d**) Course failure rate G as a function of ability θ.

In order to be rational, we consider the good study habits rate in conjunction with the students' failure rate as a function of the ability θ as follows:

1. For a randomly selected group of students, determine their abilities $\theta_1, \theta_2, \ldots, \theta_n$.
2. For each of the selected examinees, as indicated in the former point 1, investigate the total number of failed subjects throughout their academic history.
3. With the assigned ability, the quotient of the total number of failed subjects and the total number of subjects taken or studied (considering even repetitions or recursing) define the failure rate for a specific examinee.

The third step is reinforced by the results published in [47], where they claim that low levels of ability tend to cause dropout from a course, if not from the school itself. The failure rate has an identical behavior and, for all these reasons, Figure 2c and Figure 2d, respectively, postulate that the dropout ($F(\theta)$) and failure ($G(\theta)$) from a course decrease exponentially with the ability of the student. The following correspondence rules illustrate these functions:

$$F(\theta) = 1 - \frac{1 - B_d}{1 + b_d e^{-a_d \theta}}, 0 < a_d, \tag{10}$$

$$G(\theta) = 1 - \frac{1 - B_r}{1 + b_r e^{-a_r \theta}}, 0 < a_r. \tag{11}$$

In all cases, note that the ability θ is a random variable and that the functions f, g, F, and G are, therefore, random variables. In summary, the graphs in Figure 2 are the results of ansatzes here proposed to illustrate the behaviors of the random variables f, g, F, and G.

Taking into account the postulated index functions and Equations (6) and (7), the Lagrangian $\mathcal{L}$ to optimize, is given by Equation (12).

$$\begin{aligned}\mathcal{L}(p(\cdot)) = &-\int_{-\infty}^{+\infty} p(\theta)\log p(\theta)d\theta + \lambda_0\left(\int_{-\infty}^{+\infty} p(\theta)d\theta - 1\right) + \lambda_1\left(\int_{-\infty}^{+\infty} \theta p(\theta)d\theta - \hat{\theta}_p\right)\\ &+ \lambda_2\left(\int_{-\infty}^{+\infty} (\theta-\hat{\theta}_p)^2 p(\theta)d\theta - \hat{\sigma}_p^2\right) + \lambda_3\left(\int_{-\infty}^{+\infty}\left(B_h + \frac{1-B_h}{1+b_h e^{-a_h\theta}}\right)p(\theta)d\theta - C_h\right)\\ &+ \lambda_4\left(\int_{-\infty}^{+\infty}\left(B_c + \frac{1-B_c}{1+b_c e^{-a_c\theta}}\right)p(\theta)d\theta - C_c\right) + \lambda_5\left(\int_{-\infty}^{+\infty}\left(1 - \frac{1-B_d}{1+b_d e^{-a_d\theta}}\right)p(\theta)d\theta - C_d\right)\\ &+ \lambda_6\left(\int_{-\infty}^{+\infty}\left(1 - \frac{1-B_r}{1+b_r e^{-a_r\theta}}\right)p(\theta)d\theta - C_r\right),\end{aligned} \tag{12}$$

where Table A1 in Appendix A provides the meaning of the symbols appearing in the equations.

Without loss of generality and for reasons of simplicity, one only considers the constraint referring to the course failure rate $G(\theta)$ so that the maximization of entropy solves the set of non-linear equations defined by Equations (13)–(16) through:

$$\int_{-\infty}^{+\infty} e^{-1+\lambda_0+\lambda_1\theta+\lambda_2(\theta-\hat{\theta}_p)^2+\lambda_6\left(1-\frac{1-B_r}{1+b_r e^{-a_r\theta}}\right)}d\theta - 1 = 0, \tag{13}$$

$$\int_{-\infty}^{+\infty} \theta e^{-1+\lambda_0+\lambda_1\theta+\lambda_2(\theta-\hat{\theta}_p)^2+\lambda_6\left(1-\frac{1-B_r}{1+b_r e^{-a_r\theta}}\right)}d\theta - \hat{\theta}_p = 0, \tag{14}$$

$$\int_{-\infty}^{+\infty} (\theta-\hat{\theta}_p)^2 e^{-1+\lambda_0+\lambda_1\theta+\lambda_2(\theta-\hat{\theta}_p)^2+\lambda_6\left(1-\frac{1-B_r}{1+b_r e^{-a_r\theta}}\right)}d\theta - \hat{\sigma}_p^2 = 0, \tag{15}$$

$$\int_{-\infty}^{+\infty}\left(1 - \frac{1-B_r}{1+b_r e^{-a_r\theta}}\right) e^{-1+\lambda_0+\lambda_1\theta+\lambda_2(\theta-\hat{\theta}_p)^2+\lambda_7\left(1-\frac{1-B_r}{1+b_r e^{-a_r\theta}}\right)}d\theta - C_r = 0. \tag{16}$$

Note that Equation (6), both for the discrete (with summatory symbol instead of integral) and continuous cases can be considered as a measure of the misinformation (un-informativeness) that the prior distribution $p(\theta)$ provides about how the latent trait θ distributes [38,48]. This result is also supported by a research paper [48] where they state that the entropy maximization due to the constrained non-uniform prior distribution being equivalent to minimizing the distance between this distribution and an unconstrained uniform a priori distribution with no other constraint than a normalization process.

5. Results

Algorithm 2 illustrates our general procedure to select the a priori distribution. Line 1 of the algorithm assigns an initial estimation of the latent trait average $\hat{\theta}_p$ and variance average $\hat{\sigma}_p^2$. Additionally, the performance index function structures are defined as closely related to the experimental procedures (academic) already mentioned. Finally, lines 5 to 8 find the conditions to properly select an a priori distribution satisfying the normalization, ability's average value, variance's average value, and average values of expected performance index functions.

There are assumptions or inconveniences that appear when one applies a MAP or EAP technique at the initialization of the CAT process, considering that one does not know something about the prior distribution when the CAT system provides the first item. However, fortunately, there are several ways to solve this problem [8,35,49]. In this work, one proposes an initial ability equal to the value $\hat{\theta}_p$ given to the constraint in Equation (14). Thus we can apply Algorithm 3 to simulate the CAT process.

Algorithm 2 Diagram of the prior distribution search process.

procedure SEARCHPRIOR
  $\theta_p^* \leftarrow \hat{\theta}_p$ ▷ assign average of given skill in restriction (14)
  $\sigma_p^{2*} \leftarrow \hat{\sigma}_p^2$ ▷ assigns expected variance in constraint (15)
  $\vec{p} \leftarrow$ defines index parameters used in constraints
top:
  $s \leftarrow$ solve system of nonlinear equations defined by constraints
  if unsatisfied constraints **then**
    goto *top*
  return $\vec{l}$ ▷ returns Lagrange multipliers

Algorithm 3 Scheme of the CAT process using Bayesian estimation with prior distribution.

procedure ADAPTABLEEVALUATIONPROCESSWITHPRIOR
  $responses \leftarrow []$
  $\theta^* \leftarrow \hat{\theta}_p$ ▷ allocates average of given skill in constraint (14)
  $p(\theta) \leftarrow$ determine prior distribution maximizing entropy with constraints
top:
  $i \leftarrow \max_{Pool}\{I(\theta^*)\}$
  $r \leftarrow$ reply to item $i, 1 \leq i$
  $responses \leftarrow concat(responses, r)$ ▷ update response history
  $L(\vec{p}|\theta) \leftarrow$ determines the likelihood function as a product of ICCs
  if *responses* are all correct or incorrect **then** ▷ uses Bayesian inference
    $p(\theta|\vec{p}) \leftarrow L(\vec{p}|\theta)p(\theta)$
    $p(\theta|\vec{p}) \leftarrow kp(\theta|\vec{p})$ ▷ normalizes *posterior* distribution
    $\theta^* \leftarrow \int_{-\infty}^{+\infty} \theta p(\theta|\vec{p})d\theta$ ▷ compute average skill with new distribution
    $p(\theta) \leftarrow p(\theta|\vec{p})$
  else
    $\theta^* \leftarrow$ use Maximum Likelihood estimation as usual
  goto *top*.
  return θ^*

After running a sequence of simulations under the directions of Algorithm 3, one obtains as examples the corresponding CAT processes that Figures 3–5 show. Table 1 gives some numerical results, whereas the third experiment shows the complete running. The following list synthesizes the obtained results of the corresponding simulation process.

1. After several iterations, the CAT system always tends to the maximum Fisher's information index, regardless of the intermediate value of the estimated ability θ. Thus, the final selected item has a difficulty μ (see Figures 3–5).
2. When the study habits index function discriminates well and plays the role of one constraint in entropy maximization, one can expect a bimodal a priori distribution as acceptable (see Figure 3)
3. A possible behavior in the initialization of the CAT process when the discriminating power of the study habits index function is not high or low can be found in Figure 4. Note that the a priori distribution shows some non-null skewness.
4. A failure rate with a lower discrimination index provides an initial prior distribution with almost null skewness. So, in practice, when one takes a normal or Gaussian prior distribution $\mathcal{N}(\theta; \mu, \sigma)$ with a high variance σ^2 [50] or a uniform distribution $\mathcal{U}(a, b), a \ll b$, one also assumes that the examinees' failure records are the same.

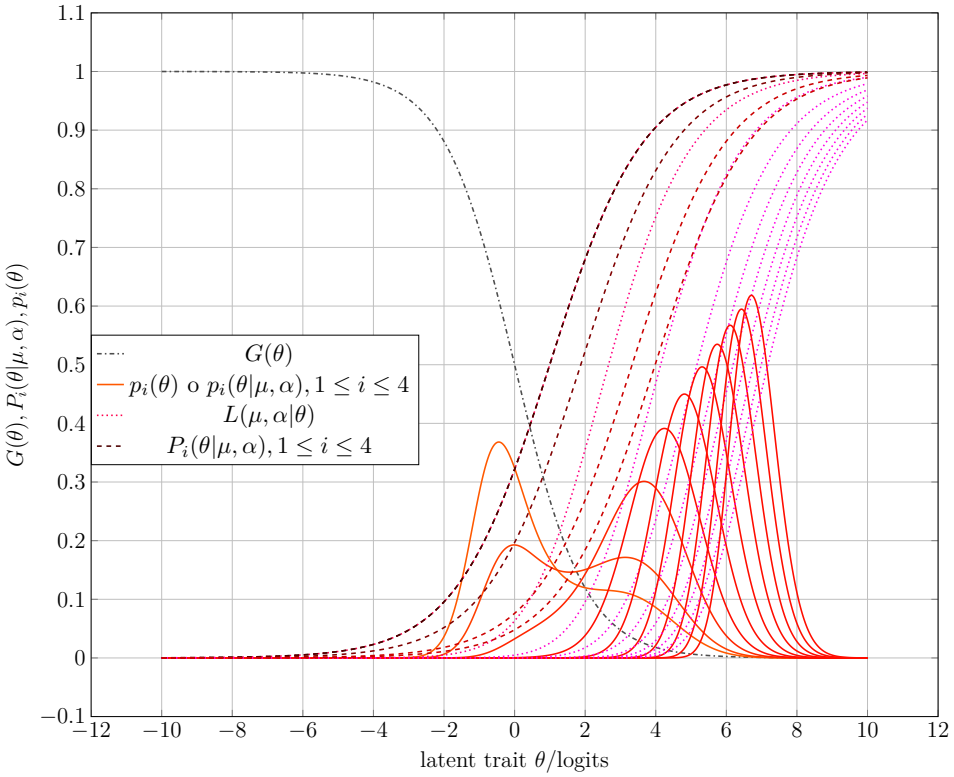

Figure 3. Starting of the CAT for the first experiment. Table 1 shows the expected and computed parameters from entropy maximization and the simulation process.

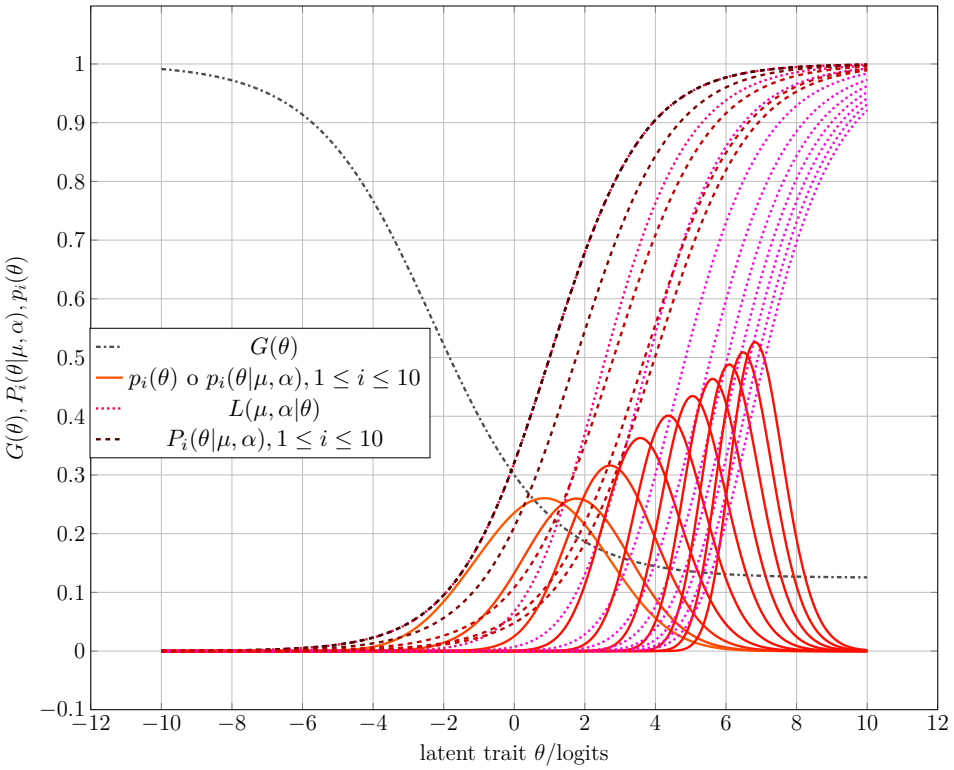

Figure 4. Starting of the CAT for the second experiment. Table 1 shows the expected and computed parameters from entropy maximization and the simulation process.

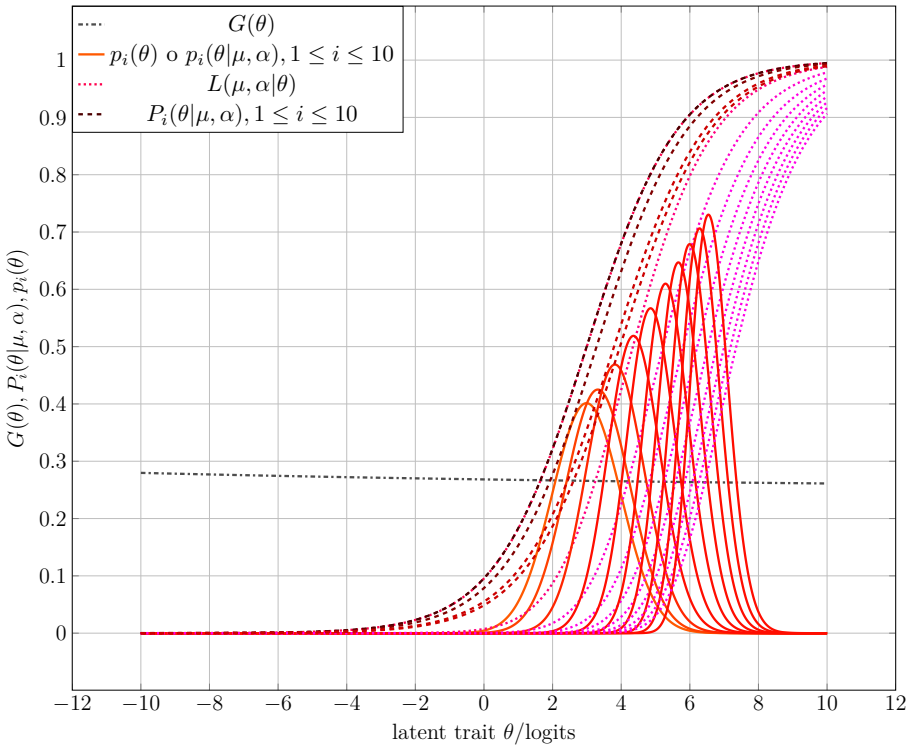

Figure 5. Starting of the CAT for the third experiment. Table 1 shows the expected and computed parameters from entropy maximization and the simulation process.

Table 1. Results of numerical experimentation. The parameters values of the Start Parameters column assume that they come from estimations and/or experiments. The parameters values of the prior properties column assume that they come from the computation of the corresponding integral expression containing the computed prior distribution (first terms in left-hand sides of Equations (13)–(16)). The numerical experimentation of a running for a CAT in Experiment 3 assumes the existence of a pool of 1000 calibrated items.

Experiment	Start Parameters						*prior* properties					Lagrange's multipliers			
1	θ_p^*	σ_p^{*2}	a_r	b_r	B_r	C_r^*	normality	θ	σ^2	C_r	skewness	λ_1	λ_2	λ_3	λ_4
	1	4	1	1	0	0.6	1.18	0.98	3.99	0.47	3.66	−6.60	2.12	−0.40	13.74
Experiment	Start Parameters						Properties *prior*					Lagrange's multipliers			
2	θ_p^*	σ_p^{*2}	a_r	b_r	B_r	C_r^*	normality	θ	σ^2	C_r	skewness	λ_1	λ_2	λ_3	λ_4
	1	4	0.6	0.25	0.125	0.6	1.20	0.96	3.99	0.34	−2.00	−1.34	0.16	−0.19	3.62
Experiment	Start Parameters						Properties *prior*					Lagrange's multipliers			
3	θ_p^*	σ_p^{*2}	a_r	b_r	B_r	C_r^*	normality	θ	σ^2	C_r	skewness	λ_1	λ_2	λ_3	λ_4
	3	1	0.05	0.025	0.25	0.3	1.00	3.00	1.00	0.27	−0.04	−3.49	0.00	−0.5	13.46
	1000 difficulties in items Pool														
	iteration		θ_p^*				μ_p^*				normality	map		eap	
	1		3				2.9978				1.00411027	2.99		3.00159777	
	2		3.3222				3.3207				1.00	3.32		3.32218745	
	3		3.8357				3.8430				1.00	3.82		3.83572950	
	4		4.3893				3.9851				1.00	4.36		4.38926317	
	5		4.8986				3.9851				1.00	4.85		4.89863112	
	6		5.3412				3.9851				1.00	5.29		5.34123495	
	7		5.7224				3.9851				1.00	5.67		5.72237081	

There are several fine details to work out when one uses a priori distributions [51]; however, in this paper, we provide a unified approach to derive prior distributions with a

less subjective selection of the distribution when the initialization of the CAT process uses Bayesian estimation [52,53].

There is a large number of research papers published about the advantages and drawbacks on the use of a priori distributions topic, but the techniques used there are based on heuristics to build the Bayesian inference procedure within the initialization of the CAT process in some special cases [51,54].

In order to compare likely differences between the results of heuristic techniques and our methodology, a useful tool to be used is the Kullback–Leibler (KL) divergence index. This index measures the divergence of the expected amount of extra information required to obtain population samples that follow the prior distribution $p(\theta)$ when using population samples that follow a distribution $q(\theta)$ [55].

The KL divergence measure is defined by Equation (17).

$$D_{KL}(p(\theta)||q(\theta) = \int_{-\infty}^{+\infty} p(\theta) \ln \frac{p(\theta)}{q(\theta)} d\theta. \tag{17}$$

In this sense, the information is more ordered when one applies the prior distribution obtained with our method than with the popular unconstrained heuristic distributions. One should expect this result since the introduction of constraints orders the information under analysis. In this manner, $p(\theta)$ represents a "realistic" data distribution or a precisely calculated theoretical distribution and the typical distribution $q(\theta)$ represents a description or approximation of $p(\theta)$ (see [55]).

Through Table 1, a correspondence rule for the a priori distribution $p(\theta)$ can be defined such as Table 2 illustrates. Therefore, the two measures given by Equations (18) and (19) considering the normal and uniform distributions, respectively, can be calculated.

$$D_{KL}(p(\theta)||\mathcal{N}(\theta;\mu,\sigma)) = \int_{-\infty}^{+\infty} p(\theta) \ln \frac{p(\theta)}{\mathcal{N}(\theta;\mu,\sigma)} d\theta, \tag{18}$$

$$D_{KL}(p(\theta)||\mathcal{U}(\theta;a,b)) = \int_{-\infty}^{+\infty} p(\theta) \ln \frac{p(\theta)}{\mathcal{U}(\theta;a,b)} d\theta. \tag{19}$$

Table 2. A priori distributions distances with respect to heuristics ones for every experiment in Table 1.

Experiment	$p(\theta)$	$q(\theta)$	KL Distance
1	$e^{-1+\lambda_1+\lambda_2\theta+\lambda_3\left(\theta-\theta_p^*\right)^2+\lambda_4\left(1-\frac{1-Br}{1+bre^{-a_r\theta}}\right)}$	$\mathcal{N}(\theta;1,2)$	3.5402
		$\mathcal{U}(\theta;1-2.225,1+2.225)$	0.73339
2	$e^{-1+\lambda_1+\lambda_2\theta+\lambda_3\left(\theta-\theta_p^*\right)^2+\lambda_4\left(1-\frac{1-Br}{1+bre^{-a_r\theta}}\right)}$	$\mathcal{N}(\theta;1,2)$	2.4119
		$\mathcal{U}(\theta;1-2.385,1+2.385)$	0.077474
3	$e^{-1+\lambda_1+\lambda_2\theta+\lambda_3\left(\theta-\theta_p^*\right)^2+\lambda_4\left(1-\frac{1-Br}{1+bre^{-a_r\theta}}\right)}$	$\mathcal{N}(\theta;3,1)$	0.069902
		$\mathcal{U}(\theta;3-1.7375,3+1.7375)$	0.084054

From Table 2, note that the distance given by the KL divergence in the first experiment when comparing the prior distribution with the Gaussian distribution $\mathcal{N}(\theta;\mu,\sigma)$ suggests that to analyze the population with this last distribution, one should expect an amount of 3.5402 extra information to include the data population related to the first distribution. Figure 6a–c compare the three distributions for every experiment in Table 2, and show their respective KL measures.

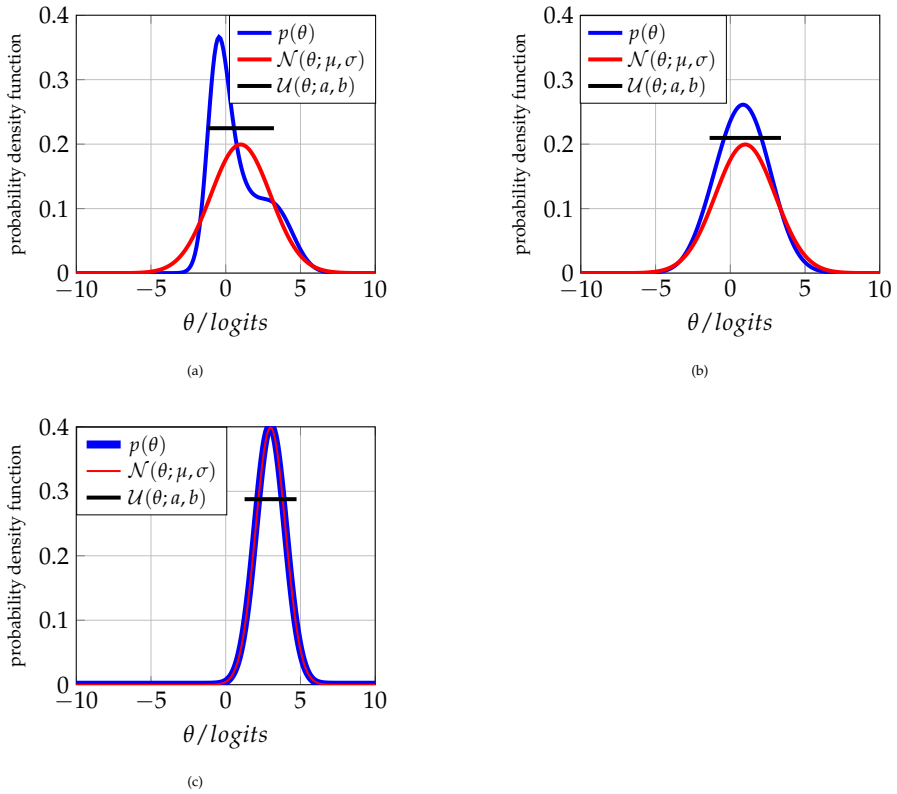

Figure 6. KL distance between each of the initial prior distributions $p(\theta)$ from each experiment and the classic distributions $\mathcal{N}(\theta;\mu,\sigma)$ and $\mathcal{U}(\theta;a,b)$. (**a**) First experiment: $p(\theta)$ vs. $\mathcal{N}(\theta;1,2)$ and $\mathcal{U}(\theta;a,b)$. (**b**) Second experiment: $p(\theta)$ vs. $\mathcal{N}(\theta;1,2)$ and $\mathcal{U}(\theta;a,b)$. (**c**) Third experiment: $p(\theta)$ vs. $\mathcal{N}(\theta;3,1)$ and $\mathcal{U}(\theta;a,b)$.

In the first experiment, when one compares the distribution $\mathcal{U}(\theta;1-2.225,1+2.225)$, the a priori distribution $p(\theta)$ also has a KL divergence equal to 0.73339. On the other hand, for the second experiment, KL divergences become equal to 2.4119 and 0.077474 when one, respectively, approximates through $\mathcal{N}(\theta;1,2)$ and $\mathcal{U}(\theta,1-2.385,1+2.385)$. Finally, for the third experiment, the KL divergences equal to 0.069902 and 0.084054 when one, respectively, approximates through $\mathcal{N}(\theta;3,1)$ and $\mathcal{U}(\theta;3-1.7375,3+1.7375)$.

Intervals (a,b) for every uniform distribution are calculated by looking for the lower distance between the corresponding $p(\theta)$ and $\mathcal{U}(\theta;a,b)$ distributions. Note that the third experiment results agree with the heuristic suggestion of using the normal, or uniform distributions, as good approximations to the prior distribution. So, the alternative is acceptable when the course failure index function does not discriminate well.

6. Conclusions

In this paper, we demonstrate that through the theory of entropy maximization, a given set of constraints, and under numerical experimentation, the computation of an a priori distribution to initialize a CAT process by using Bayesian inference can be carried out. Furthermore, the examinee's performance index functions define the constraints, and they complement the usual distribution constraints (normality, first and second moment, etc.).

We also show that through the entropy theory, the selection of appropriate constraints summarizes experimental data through the specification of index functions related to study habits, comprehension levels, course dropout, and lecture failure.

A given set of constraints can produce a set of acceptable or unacceptable a priori distributions, so one needs to look for a stop criterion in searching for the optimal set of parameters that defines the distributions through entropy maximization. To define the stop criterion, we verify how the estimated set of distribution parameters and those that define the constraints are close enough to the expected values used in the constraints

definition. Thus, the most appropriate distribution is chosen and, under the assumption of responding correctly to the first items in the testing process, we can verify its latent trait prediction capability.

Index functions playing the role of constraints with acceptable discrimination properties produce a priori distributions with bimodality, as one can expect, so that the obtained distribution estimates reasonable latent trait values along the simulation of the CAT process.

In summary, entropy maximization can be used inside the frame of a CAT to derive more generalized a priori distributions through constraint specifications related to index functions. This method can provide a unified approach to derive a priori distributions for initializing the CAT process through a Bayesian inference procedure.

Author Contributions: Conceptualization, J.S.-C.; Methodology, J.S.-C. and V.L.-M.; Software, J.S.-C. and L.R.M.-M.; Validation, J.S.-C.; Formal analysis, J.S.-C. and V.L.-M.; Investigation, V.L.-M., L.R.M.-M., A.A.-R. and J.C.R.-F.; Resources, A.A.-R.; Data curation, J.C.R.-F.; Writing—original draft, V.L.-M.; Writing—review & editing, V.L.-M.; Visualization, A.A.-R. and J.C.R.-F.; Supervision, V.L.-M. All the authors have equally contributed and worked in this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like, through these lines, to really thank the anonymous reviewers and the journal's staff for their valuable support to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1PL	One-Parameter Logistic
2PL	Two-Parameter Logistic
CAT	Computer Adaptive Testing
CIP	Calibrated Item Pool
EAP	Expectation a posteriori
IRT	Item Response Theory
ICC	Item Characteristic Curve
KL	Kullback–Leibler
MAP	*maximum a posteriori*
w.r.to	with respect to

Appendix A. Table of Symbols

Table A1. Summary of all the symbols used in the paper, their location and meaning.

Symbol	Meaning	Pages Where the Symbol Appears
e	Base of the natural logarithm	2
θ	Examinee's ability or latent trait to be estimated in the CAT process	2–6, 10, 12–16
θ_i	Examinee's ability at i-th iteration	6
μ	Item's difficulty playing the role of a parameter in the definition of the ICC that the CDF P defines	2, 5, 6, 15, 16
$\mu_i, i = 1, 2, \ldots$	Difficulty of item i	5
α	Item's discriminant playing the role of a parameter in the definition of the ICC that the PDF P defines	5, 6, 13, 14

Table A1. *Cont.*

Symbol	Meaning	Pages Where the Symbol Appears
P	Conditional cumulative distribution function for computing the probability that the examinee with ability θ gives a correct answer to an item, given that the item's difficulty is μ and (possibly) the discriminant is α	2, 5, 6
Q	It is equal to $1 - P$; in other words, it represents the probability that the examinee with ability θ gives an incorrect answer to an item, given that the item's difficulty is μ and (possibly) the discriminant is α	6
$\Delta\mu$	Increment of item's difficulty	6
$\Delta\alpha$	Increment of item's discriminant	5
$n = 1, 2, \ldots$	Number of items at iteration n, or n-th iteration, in the CAT process	6
I	Fisher's information, as a function of the examinee's ability θ, is useful to compute the amount of information that a given item provides about the θ estimate by knowing the current value of θ	6
$\vec{p}$	Vector of parameters (item's difficulty, item's discriminant, etc.) that defines the model P	6, 12
p	If p depends on θ, then it represents a prior distribution; otherwise, if p depends on θ given the vector of parameters $\vec{p}$, then it represents a posteriori distribution	7, 16
L	Likelihood function depends on the vector of parameters $\vec{p}$ given the examinee's ability θ, $L(\vec{p}_i, i = 1, 2, \ldots, n\|\theta) \propto \prod_{i=1}^{n} P_i^k(\theta\|\vec{p}_i) Q_i^{n-k}(\theta\|\vec{p}_i)$, where k represents the number of correct answers that an examinee gives to n items	7
f, g, F, G	Performance indexes are functions of a random variable X	7
$\mathcal{S}$	Entropy	7
$\mathcal{L}$	Lagrangian	8, 11
λ	Lagrange multiplier	8, 11
$B_i, b_i, a_i,$ where $i = h, c, d, r$	Parameters defining performance index function: study habit, subject comprehension, course dropout rate, and course failure rate, respectively	10
$\hat{\theta}_p$	Initial estimation of the latent trait θ, just before starting the CAT process, the subindex p comes from the word prior	11, 12
$\hat{\sigma}_p^2$	Initial estimation of the variance σ^2 of the distribution of the latent trait θ, the subindex p comes from the word *prior*	11, 12
s	It contains the solution to entropy maximization under the given constraints	12
$\vec{l}$	Vector of Lagrange multipliers after satisfying entropy maximization under given constraints	12

Table A1. *Cont.*

Symbol	Meaning	Pages Where the Symbol Appears
θ^*	Initial estimation of the latent trait θ, just before starting the CAT process, or expectation a posteriori of the latent trait along the CAT process, or the extreme point of the ability values θ, along the CAT process, where the a posteriori distribution has a *maximum*	12, 13
$\mathcal{U}$	Uniform distribution	13, 15, 16
$\mathcal{N}$	Normal distribution	13, 15, 16
a, b with $a < b$	lower and upper bounds, respectively, that define the uniform distribution	15, 16
σ^2, σ	Variance and standard deviation estimation, respectively, for a priori distribution	13, 15
C_r	Estimation of the course failure rate average by means of the estimated a priori distribution	11, 13
C_r^*	Course failure rate estimation coming from the experimental results for obtaining the course failure rate index function	11, 13

References

1. Haifeng, L.; Ning, Z.; Zhixin, C. A Simple but Effective Maximal Frequent Itemset Mining Algorithm over Streams. *J. Softw.* **2012**, *7*, 25–32.
2. Li, M.; Han, M.; Chen, Z.; Wu, H.; Zhang, X. FCHM–Stream: Fast Closed High Utility Itemsets Mining over Data Streams. Research Article, Research Square. 2022, 19p. Available online: https://assets.researchsquare.com/files/rs-1736816/v1_covered.pdf?c=1655222833 (accessed on 14 December 2022).
3. Liu, J.; Ye, Z.; Yang, X.; Wang, X.; Shen, L.; Jiang, X. Efficient strategies for incremental mining of frequent closed itemsets over data streams. *Expert Syst. Appl.* **2022**, *191*, 116220. [CrossRef]
4. Caruccio, L.; Cirillo, S.; Deufemia, V.; Polese, G. Efficient Discovery of Functional Dependencies from Incremental Databases. In Proceedings of the 23rd International Conference on Information Integration and Web Intelligence, IIWAS2021, Linz, Austria, 29 November–1 December 2021; pp. 400–409.
5. Hu, K.; Qiu, L.; Zhang, S.; Wang, Z.; Fang, N. An incremental rare association rule mining approach with a life cycle tree structure considering time–sensitive data. *Appl. Intell.* **2022**. [CrossRef]
6. Revelle, W. Chapter 8 The "New Psychometrics"– Item Response Theory. In *An Introduction to Psychometric Theory with Applications in R*; 2013; pp. 241–264. Available online: http://personality-project.org/courses/405.syllabus.html (accessed on 14 December 2022).
7. DeMars, C. *Item Response Theory*; Oxford University Press: Cary, NC, USA, 2010.
8. Olea, J.; Ponsoda, V. *Tests Adaptativos Informatizados*; Universidad Nacional de Educación a Distancia (UNED) Ediciones: Madrid, Spain, 2003.
9. Revuelta, J.; Ponsoda, V. Una Solución a la estimación inicial en los Tests Adaptativos Informatizados. *Rev. Electrónica De Metodol. Apl.* **1997**, *2*, 1–6.
10. Frans, N.; Braeken, J.; Veldkamp, B.P.; Paap, M.C.S. Empirical Priors in Polytomous Computerized Adaptive Tests: Risks and Rewards in Clinical Settings. *Appl. Psychol. Meas.* **2022**, *47*, 48–63. [CrossRef]
11. O'Hagan, A.; Luce, B.R. *A Primer on BAYESIAN STATISTICS in Health Economics and Outcomes Research*; Bayesian Initiative in Health Economics & Outcomes Research, Center for Bayesian Statistics in Health Economics, MEDTAP International: Bethesda, MD, USA, 2003.
12. Veldkamp, B.P.; Matteucci, M. Bayesian Computerized Adaptive Testing. *Ensaio Avaliação e Políticas Públicas em Educação* **2013**, *21*, 57–82. [CrossRef]
13. Liu, X.; Lu, D. A MAP method with nonparametric priors for estimating P–S–N curves. In Proceedings of the Fifth International Symposium on Life–Cycle of Engineering Systems: Emphasis on Sustainable Civil Infrastructure Conference, Delft, The Netherlands, 16–19 October 2016; pp. 2120–20124.
14. Swaminathan, H.; Gifford, J.A. Bayesian estimation in the three-parameter logistic model. *Psychometrika* **1986**, *51*, 589–601. [CrossRef]
15. Wang, T.; Vispoel, W.P. Properties of ability estimation methods in computerized adaptive testing. *J. Educ. Meas.* **1998**, *35*, 109–135. [CrossRef]

16. Lord, F.M. Maximum likelihood and Bayesian parameter estimation in item response theory. *J. Educ. Meas.* **1986**, *23*, 157–162. [CrossRef]
17. Samejima, F. Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychom. Monogr.* **1969**, *17*, 100. Available online: http://www.psychometrika.org/journal/online/MN17.pdf (accessed on 14 December 2022). [CrossRef]
18. Mitrushina, M.; Boone, K.B.; Razani, J.; D'Elia, L.F. Statistical and Psychometric Issues. In *Handbook of Normative Data for Neuropsychological Assessment*, 2nd ed.; Mitrushina, M., Boone, K.B., Razani, J., D'Elia, L.F., Eds.; Oxford University Press: New York, NY, USA, 2005; pp. 33–56.
19. Ho, A.D.; Yu, C.C. Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educ. Psychol. Meas.* **2015**, *75*, 365–388. [CrossRef] [PubMed]
20. Mitrushina, M. *Handbook of Normative Data for Neuropsychological Assessment*; Oxford University Press: New York, NY, USA, 2005; pp. 38–39.
21. Stephens, M. Dealing with multimodal posteriors and non–identifiability in mixture models. *J. R. Stat. Soc. Ser. B* **1999**, *62*, 795–809. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c0d3e6bf574ec653f5b81d39149413c7a3aa949a (accessed on 14 December 2022). [CrossRef]
22. Botje, M. *Introduction to Bayesian Inference. Lecture Notes at NIKHEF National Instituut Voor Subatomaire Fysica*; Publisher National Instituut voor Subatomaire Fysica: Amsterdam, The Netherlands, 2006.
23. LaValle, S.M. *Planning Algorithms*; Cambridge University Press: Cambridge, UK, 2006.
24. Bromiley, P.A.; Thacker, N.A.; Bouhova-Thacker, E. *Shannon Entropy, Renyi Entropy, and Information*; Technical Report No. 2004-004; Imaging Science and Biomedical Engineering, School of Cancer and Imaging Science: London, UK, 2010.
25. Bretthorst, G.L. An Introduction to Parameter Estimation Using Bayesian Probability Theory. In *Maximum Entropy and Bayesian Methods*; Kluwer Academic, P.F.F., Ed.; Springer: Dordrecht, The Netherlands, 1990; pp. 53–79.
26. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241. [CrossRef]
27. Jeon, S.S.; Su, S.Y.W. Deriving Prior Distributions for Bayesian Models Used to Achieve Adaptive E–Learning. *Knowl. Manag. E-Learn. Int. J. (KM EL)* **2011**, *3*, 251–270.
28. Albert, I.; Donnet, S.; Guihenneuc-Jouyaux, C.; Low-Choy, S.; Mengersen, K.; Rousseau, J. Combining Expert Opinions in Prior Elicitation. *Bayesian Anal.* **2012**, *7*, 503–532. [CrossRef]
29. Dayanik, A.; Lewis, D.D.; Madigan, D.; Menkov, V.; Genkin, A. Constructing informative prior distributions from domain knowledge in text classification. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 493–500.
30. Keller, L.A. *Ability Estimation Procedures in Computerized Adaptive Testing*; Technical Report; American Institute of Certified Public Accountants: Durham, NC, USA, 2000.
31. Tchourbanov, A. *Prior Distributions*; Technical Report; Department of Biology New Mexico State University Road Runner Gnomics Laboratories. 2002. Available online: https://datajobs.com/data-science-repo/Conjugate-Priors-[Alexandre-Tchourbanov].pdf (accessed on 14 December 2022).
32. Gelman, A.; Simpson, D.; Betancourt, M. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy* **2017**, *19*, 555. [CrossRef]
33. Matteucci M.; Veldkamp, B.P. Bayesian Estimation of Item Response Theory Models with Power Priors. In Proceedings of the Statistical Conference, Advances in Latent Variables, Methods, Models and Applications, SIS 2013, Brescia University, Owensboro, KY, USA, 19–21 July 2013. (Special Session)
34. Muñiz, J. Introducción a la Teoría de Respuesta a los Ítems. Ediciones Pirámide, Colección Psicología, Sección Psicometría , Madrid. 1997. Available online: https://www.semanticscholar.org/paper/Introducci%C3%B3n-a-la-teor%C3%ADa-de-respuesta-a-los-%C3%ADtems-Fern%C3%A1ndez/4bd320747e2df2f34ba71e61199ff49e93df007e (accessed on 14 December 2022).
35. van der Linden, W.J.; Pashley, P.J. Chapter 1 Item Selection and Ability Estimation in Adaptive Testing. In *Computerized Adaptive Testing: Theory and Practice*; van der Linden, W.J., Glas, C.A.W., Eds.; Springer: Dordrecht, The Netherlands, 2000; pp. 1–25.
36. Yao, L. Item Selection Methods for Computer Adaptive Testing With Passages. *Front. Psychol.* **2019**, *10*, 240. [CrossRef]
37. Olea, J.; Ponsoda, V. Capítulo 4 Algoritmos Adaptativos. In *Tests Adaptativos Informatizados*; Universidad Nacional de Educación a Distancia (UNED) Ediciones: Madrid, Spain, 2003; pp. 47–66.
38. Consonni, G.; Fouskakis, D.; Liseo, B.; Ntzoufras, I. Prior Distributions for Objective Bayesian Analysis. *Bayesian Anal.* **2018**, *13*, 627–679. [CrossRef]
39. Siahi, E.A.; Maiyo, J.K. Study of the relationship between study habits and academic achievement of students: A case of Spicer Higher Secondary School, India. *Int. J. Educ. Adm. Policy Stud.* **2015**, *7*, 134–141.
40. Ebele, U.F.; Olofu, P.A. Study habit and its impact on secondary school students' academic performance in biology in the Federal Capital Territory, Abuja. *Educ. Res. Rev.* **2017**, *12*, 583–588. [CrossRef]
41. Andrich, D. A Structure of Index and Causal Variables. *Trans. Rasch Meas. SIG Am. Educ. Res. Assoc.* **2014**, *28*, 1475–1477.
42. Andrich, D.; Marais, I. Chapter 4 Reliability and Validity in Classical Test Theory. In *A Course in Rasch Measurement Theory, Springer Texts in Education, Measuring in the Educational, Social and Health Sciences*; Berlin/Heidelberg, Germany, 2019; pp. 41–53. [CrossRef]
43. Stenner, A.J.; Stone, M.H.; Burdick, D.S. Indexing vs. Measuring. *Rasch Meas. Trans.* **2009**, *22*, 1176–1177. Available online: https://www.rasch.org/rmt/rmt224b.htm (accessed on 14 December 2022).

44. Ramakrishnan, S.; Robbins, T.W.; Zmigrod, L. Research Article—The Habitual Tendencies Questionnaire: A tool for psychometric individual differences research. *Personal. Ment. Health* **2022**, *16*, 30–46. [CrossRef] [PubMed]
45. Abed, B.K. Study Habits Used by Students at the University of Technology. *J. Educ. Coll.* **2016**, *1*, 537–558. Available online: https://www.iasj.net/iasj/article/113575 (accessed on 14 December 2022).
46. Eleby, C. The Impact of a Student's Lack of Social Skills on their Academic Skills in High School. Master's Thesis, Marygrove College, Detroit, MI, USA, April 2009.
47. Cruz-Sosa, E.M.; Gática-Barrientos, L.; García-Castro, P.E.; Hernández-García, J. Academic Performance, School Desertion And Emotional Paradigm In University Students. *Contemp. Issues Educ. Res.* **2010**, *3*, 25–36. [CrossRef]
48. Kapur, J.N. Chapter 1 Maximum–Entropy Probability Distributions: Principles, Formalism and Techniques. In *Maximum–entropy Models in Science and Engineering*; Jagat Narain Kapur (Revised Edition); Wiley: Hoboken, NJ, USA, 1993; pp. 1–29.
49. Meijer, R.R.; Nering, M.L. Computerized Adaptive Testing: Overview and Introduction. *Appl. Psychol. Meas.* **1999**, *23*, 187–194. [CrossRef]
50. Raîche, G.; Blais, J.G.; Magis, D. Adaptive estimators of trait level in adaptive testing: Some proposals. In Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing, 8 June 2007; Weiss, D.J., Ed. Available online: https://www.researchgate.net/publication/228435165_Adaptive_estimators_of_trait_level_in_adaptive_testing_Some_proposals (accessed on 14 December 2022).
51. Cengiz, M.A.; Öztük, Z. A Bayesian Approach for Item Response Theory in Assessing the Progress Test in Medical Students. *Int. J. Res. Med. Health Sci.* **2013**, *3*, 15–19.
52. Van der Linden, W.J. Empirical Initialization of the Trait Estimator in Adaptive Testing. *Appl. Psychol. Meas.* **1999**, *23*, 21–29. [CrossRef]
53. Chen, L.; Singh, V.P. Entropy–based derivation of generalized distributions for hydrometeorological frequency analysis. *J. Hydrol.* **2018**, *557*, 699–712. [CrossRef]
54. Gelman, A. Objections to Bayesian statistics. *Bayesian Anal.* **2008**, *3*, 445–450. [CrossRef]
55. Han, J. Chapter 2 Know Your Data (Additional Material) Kullback–Leibler Divergence. Lecture Notes (3rd ed.) CS412 Fall 2008 Introduction to Data Warehousing and Data Mining at the Department of Computer Science, University of Illinois, August 2017. Available online: http://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf (accessed on 14 December 2022).

On the Extended Specific Objectivity of Some Pseudo–Rasch Models Applied in Testing Scenarios

Joel Suárez–Cansino[1(✉)], Luis R. Morales–Manilla[2], and Virgilio López–Morales[1]

[1] Information and Systems Technologies Research Center, Autonomous University of the State of Hidalgo, Mineral de la Reforma, Hidalgo, Mexico
jsuarez@uaeh.edu.mx

[2] Applied Software Development Research Group, Engineering Division, Polytechnic University of Tulancingo, Tulancingo, Hidalgo, Mexico

Abstract. The design of a Computer Adaptive Testing (CAT) system assumes the existence of an item pool containing properly calibrated items. The calibration is based on an Item Characteristic Curve (ICC). In this paper two mathematical ICC models, and how these models properly fit into the concept of extended Rasch specific objectivity, are under analysis. The results make clear that the comparison between two items depends on subdomains of the complete domain of the corresponding ICC's. The introduced models are also useful to describe the characteristics of skewness and bimodality in the population, where classical models commonly fail.

Keywords: Computer Adaptive Testing · Specific objectivity · Item Characteristic Curve · Skewness · Bimodality

1 Introduction

Computer Adaptive Testing (CAT) is an example of a Computer Based Test (CBT) and is one of the main trending topics in the area of knowledge testing [1] and, more recently, in e–learning or in Intelligent Tutoring Systems scenarios [2]. The Item Response Theory (IRT) defines the theoretical basis of a CAT implementation [3], which assumes the existence of a repository of items that is used during the testing process of a particular examinee. Every item in the repository must be calibrated at the initial steps of the implementation of the CAT system, based on the specification of a psychometric model. The calibration process is achieved to determine the parameters values defined by a psychometric model previously defined and, in general, the parameters values change from one item to another. In fact, the calibration stage can be compared with the learning stage of an artificial intelligent system, and this is one of the main reasons to make a reliable calibration process.

O. Pichardo Lagunas et al. (Eds.): MICAI 2015, Part II, LNAI 9414, pp. 232–248, 2015.
DOI: 10.1007/978-3-319-27101-9_17

The psychometric model used in the process depends on the specific scenario where the CAT system is applied. Usually, its structure is defined by a sigmoidal or logistic function, which depends on the examinee's ability and contains one or more parameters. Hence, an item designer could find models of one, two, three or four parameters, depending on the chosen model called 1PL, 2PL, 3PL and 4PL model, respectively.

For 1PL model, also known as the Rasch's model, ideal experimental conditions are assumed, while for the remaining models some additional item properties are highlighten which are useful to describe the item's capability to clearly distinguish among the examinees' abilities, the degree of item guessing and the degree of item inattention.

However, when considering some simulated or real experimental results, these models can have a lower performance when skewed and/or multimodal behaviors are included in the statistical characteristics of the population. The knowledge of the distribution behavior becomes a very important question when student achievement is involved. A distribution with bimodality in this kind of context can be useful to predict failures of the lower sub–population in future testing processes, and to make more reliable item calibrations. There are some interesting examples in the literature where dealing with these failures is a very important question for solving several academic problems [4–6].

The choose of a suitable psychometric model structure and its parameters is a *sine qua non* condition in order to provide reliable information concerned with the examinee's ability, item's and test's difficulty, among others.

Furthermore, the values of the item's parameters are related to this kind of information, and once that the parameters values are obtained, they are very useful in e–learning scenarios, Intelligent Tutoring Systems or Computer Adaptive Testing Systems, where an immediate and reliable diagnostic is required for giving a support to the teaching–learning or testing process [7].

The finding of meaningful interpretations of the parameters in the model can be also very useful to make the best decisions in this sense. The constraint models (for instance, 2PL/MML framework), nonparametric function estimates and others flexible models, such as Ramsay–curves and splines, do not provide a direct way of doing this. On the other hand, the cited examples are real contexts where constraint models like 2PL/MML are not enough to describe them.

As a matter of fact, it is well known that traditional psychometric models provide an interpretation of every item's parameter in terms of the item's difficulty, item's discriminant, degree of item guessing, and so on [8]. In this paper, the properties of two generalized sigmoidal psychometric models are analyzed and shown to be more flexible than the previous ones, meaning that they could be applied in more complex testing scenarios, where some aspects of skewness and bimodality can be included.

The analysis of our models is made on the basis of the mathematical behavior, which is described not only by the latent trait variable or examinee's ability, but by the parameters, as well. The concept of specific objectivity is also used as a mean of validation of our models.

1.1 Standard Psychometric Models and Specific Objectivity

The 1PL model is the simplest psychometric model and is defined by the Equation (1)

$$p_i(x = 1|\theta, \mu_i) = \frac{1}{1 + e^{-(\theta - \mu_i)}} \tag{1}$$

where $p_i(x = 1|\theta, \mu)$ denotes the probability of a correct response of the examinee to the i–th item, given that θ is the examinee's ability and μ_i is the item's difficulty. An important characteristic of this model is that the specific objectivity is verified, since it allows a comparison of both the performance of any two examinees and of any two items in one test [9,10].

In order to state a formal definition of specific objectivity, let us recall the following. The psychometric model verifies $p : E \times I \longrightarrow (a, b)$, where E and I are the sets of examinees and items in the test, respectively, and $(a, b) \subseteq [0, 1]$. The specific objectivity of the model assumes the existence of a function, $\chi : (a, b) \times (a, b) \longrightarrow \mathbb{R}$, and defines a multivariable vector function $\boldsymbol{p} : (E \times I) \times (E \times I) \longrightarrow (a, b) \times (a, b)$, where $\boldsymbol{p}((r, u), (s, v)) = (p(r, u), p(s, v))$, such that the composition of functions $c = \chi \circ \boldsymbol{p} : (E \times I) \times (E \times I) \longrightarrow \mathbb{R}$ compares the pair (r, u) with the pair (s, v) under one of the following conditions [9,10],

1. The comparison of any two objects $r, s \in E$ is independent of the choice $u, v \in I, u = v$, and of any other element $t \in E, t \neq r, s$.
2. The comparison of any two objects $u, v \in I$ is independent of the choice $r, s \in E, r = s$, and of any other element $w \in I, w \neq u, v$.

Comparing function c is specifically objective within the frame of reference defined by E, I and p. The function c of the first condition does not need to be equal to the function c of the second condition; *i.e.*, the function χ of the first condition can be quite different of the function χ of the second condition. However, the psychometric model p is always the same.

1.2 Generalized Models

Some authors have explained the reasons for proposing more sophisticated alternatives formulations to the Rasch model, with the main intention of including the possible skewness of the experimental data [11]. In [9] it is proposed an extension of the specific objectivity concept, giving the possibility of comparing three or more elements in the sets E or I, and even in [10] the specific objectivity concept is excluded as a necessary requirement, which leads to the idea of pseudo–Rasch models.

Example 1. A relatively simple general model can be proposed, which is a slight modification of another function by [12],

$$p(x = 1|\theta, \mu, \alpha, a, c, d, g) = d + (a - d)p^g \left(x = 1 \left| \theta, \mu + \frac{1}{\alpha} \ln c, \alpha \right. \right) \tag{2}$$

where the Cumulative Distribution Function (CDF), p, on the right hand side of (2), is given by the 2PL model, as defined by (3),

$$p(x = 1|\theta, \mu, \alpha) = \frac{1}{1 + e^{-\alpha(\theta-\mu)}}. \tag{3}$$

2 Analysis of the Proposed Models

In this scenario, several alternatives of psychometric models are possible. However, in order to keep the strength of Rasch's model [13,14] some constraints are introduced. For instance,

(i) the proposed model must be part of a frame of reference with specific objectivity, even with the extended one, which admits the comparison among two, three or more elements of E or I,
(ii) the model also must be quite flexible to admit skewness and some multimodality that the ability could possibly show and,
(iii) finally, the proposed model must verify the Rasch's model as a particular case.

The model given by (2) contains six parameters μ, α, a, c, d and g, and it verifies the constraints mentioned before. As a matter of fact, the interpretation of parameters μ, α, a and d, coincides with that given to the parameters in the very well known 1PL, 2PL, 3PL and 4PL models [12,15]. In addition to that, the Rasch's model, along with the 2PL, 3PL and 4PL models, are particular cases of the more general 6PL model defined by (2).

Furthermore, with the model given by (2), hereafter called extended 6PL Rasch's model, the interpretation of the new parameters c and g includes the concept of skewness of the experimental data, since the parameter c implies a correction term to the difficulty μ. However, it can be proved that the model does not produce symmetrical skews (the left skew is not a mirror image of the right skew).

2.1 Behavior of the Extended 6PL Rasch's Model

The change of concavity and the symmetrical behavior relative to the upper and lower asymptotes, respectively defined by the equations (4),

$$\lim_{\theta \to \pm\infty} p(x = 1|\theta, \mu, \alpha, a, d, c, g) = \begin{cases} a \\ d \end{cases} \tag{4}$$

are two important points to be considered in the behavior of the extended 6PL Rasch's model. At this stage of the discussion, let us analyze conditions to successfully apply this CDF as a proper psychometric model, mathematically speaking.

In order to do so, in the analysis of the function behavior first and second derivatives are involved and it can be easily proved that the CDF (2) is an increasing function. On the other hand, the change of concavity occurs at the single point

$\theta = \mu + \frac{1}{\alpha}\ln(cg)$ in the domain of the CDF (2) and $d + \frac{a-d}{(1+g^{-1})^g}$ is the value of the function at this point. Now, the condition of rapid growing of the CDF is established through the definition of a positive parameter κ such that,

$$\frac{\kappa}{\alpha g(a-d)} \leq p^g\left(x=1\left|\theta,\mu+\frac{1}{\alpha}\ln c,\alpha\right.\right) - p^{g+1}\left(x=1\left|\theta,\mu+\frac{1}{\alpha}\ln c,\alpha\right.\right) \tag{5}$$

Then, the set of abilities θ satisfying this inequality becomes the interval where the CDF (2) grows rapidly. Since this CDF is an increasing function, then the roots of the equality in inequation (5) define the *infimum* and *supremum* of the interval. Inequality (5) can also be seen as the specification of the lower bound given by $\frac{\kappa}{\alpha g(a-d)}$ to the polynomial $f(x) = x^g - x^{g+1}$ in the real unit interval $(0,1)$, with real power $g, 0 < g$. The derivative of this polynomial is given by (6)

$$\frac{d}{dx}f(x) = (g+1)x^{g-1}\left(\frac{g}{g+1} - x\right), \tag{6}$$

which is always positive in the interval $\left(0, \frac{g}{g+1}\right)$ (increasing function f) and negative in the interval $\left(\frac{g}{g+1}, 1\right)$ (decreasing function f).

If $0 < g < 1$, then the function f does not have a change of concavity and is always concave downward (since its second derivative is always negative). On the other hand, if $1 < g$, then the function f changes from being concave upward to be concave downward at the critical point $\theta = \frac{g-1}{g+1}$. The point $\theta = \frac{g}{g+1}$ in the domain of f is a critical point where the function f has a *maximum* value for arbitrary values of the parameter g, $0 < g$. This means that the function defined by the equality in (5) has only two real roots if the constant $\frac{\kappa}{\alpha g(a-d)}$ satisfies the conditions (7),

$$0 < \frac{\kappa}{\alpha(a-d)} < \left(\frac{g}{g+1}\right)^{g+1} \tag{7}$$

The asymmetrical behavior of the function f ensures that the two roots are asymmetrical with respect to the point $\frac{g}{g+1}$. This behavior of f implies the possibility of obtaining a skewed Probability Density Function (PDF) for the CDF (2); however, the behavior of this skewness is not completely symmetrical, in the sense already explained at the beginning of this section.

The prove of the existence of an asymmetrical skewness for the extended 6PL Rasch's model considers the value of this function at the point $\theta = \mu + \frac{1}{\alpha}\ln(cg)$, where the change of concavity appears. The distance between the point $(\mu + \frac{1}{\alpha}\ln(cg), d)$ on the lower asymptote and the point where the change of concavity appears is given by the expression $\frac{a-d}{(1+g^{-1})^g}$, which means that the distance is proportional to $a-d$, where the proportion is defined by the expression $\frac{1}{\left(1+\frac{1}{g}\right)^g}$.

Note that the parameter g has a lower bound equal to zero, but it can grow without limit, so that it is interesting to know the bounds of this proportion. By using the definition of the basis of the natural logarithm and the L'Hopital's

rule, it is relatively easy to notice that these limits are 1 and $\frac{1}{e}$, when $g \to 0$ and $g \to +\infty$, respectively.

Thus, the distance between the point on the lower asymptote, $(\mu+\frac{1}{\alpha}\ln(cg), d)$, and the point on the CDF where the change of concavity occurs, can be as long as $a-d$ or as short as $\frac{1}{e}(a-d)$, which means that the right skewness of the CDF is not necessarily symmetrical to the left skewness of the same model of the CDF.

2.2 The Extended 6PL Rasch's Model and Specific Objectivity

The behavior of this model is quite complex, and it is difficult, if not impossible, to handle in a direct way the concept of specific objectivity. However, the function can be approximated by the piecewise function (8), which is defined in some interval of abilities,

$$p(x=1|\theta,\mu,\alpha,a,c,d,g) \approx \begin{cases} \frac{1}{1+\exp\left(-\alpha\left(\theta-\mu-\frac{1}{\alpha}\ln c\right)\right)}, & \text{if } \mu+\frac{1}{\alpha}\ln c \ll \theta \\ \frac{1}{\exp\left(-\alpha g\left(\theta-\mu-\frac{1}{\alpha}\ln c\right)\right)}, & \text{if } \theta \ll \mu+\frac{1}{\alpha}\ln c \\ \frac{1}{2^g\left(1-\frac{1}{2}g\right)} \cdot \frac{1}{1+\exp\left(-\alpha\left(\theta-\mu-\frac{1}{\alpha}\ln\frac{cg}{2\left(1-\frac{1}{2}g\right)}\right)\right)}, & \text{otherwise} \end{cases} \tag{8}$$

The piecewise definition of the function $p(x=1|\theta,\mu,\alpha,a,c,d,g)$ and the idea of an extended specific objectivity permit to compare three arbitrarily chosen abilities and two arbitrarily chosen items. It should be noticed that the choice of any two expressions of the piecewise definition of the function $p(x=1|\theta,\mu,\alpha,a,c,d,g)$ represents to exactly the same CDF, so that the concept of specific objectivity is properly applied in this sense.

Hence, the comparison of three abilities can be made by means of the following definition of the function $\chi : (0,1)\times(0,1)\times(0,1) \longrightarrow \mathbb{R}$,

$$\chi(x_1,x_2,x_3) = \frac{\ln\left(\frac{x_1}{1-x_1}\cdot\frac{1-x_2}{x_2}\right)}{\ln\left(\frac{x_1}{1-x_1}\cdot\frac{1-x_3}{x_3}\right)} \tag{9}$$

and through the assumption that the specific objectivity gets rid of any scale factor in any expression of the piecewise definition of the CDF. So, for example, $c((r,u),(s,u),(t,u)) = \frac{\theta_r-\theta_s}{\theta_r-\theta_t}$. Similarly, two items can be compared through the following definition of the function $\chi : (0,1)\times(0,1)\times(0,1)\times(0,1) \longrightarrow \mathbb{R}$,

$$\chi(x_1,x_2,x_3,x_4) = \frac{\ln\left(\frac{x_1}{1-x_1}\cdot\frac{1-x_2}{x_2}\right)}{\ln\left(\frac{x_3}{1-x_3}\cdot\frac{1-x_4}{x_4}\right)} \tag{10}$$

and the idea of specific objectivity already suggested. Therefore, two items can be compared considering the same expression, or any two expressions, of the piecewise definition of the CDF, as follows, $c((r,u),(s,u),(r,v),(r,v)) = \frac{\alpha_u}{\alpha_v}$ or $c((r,u),(s,u),(r,v),(r,v)) = \frac{\alpha_u g_u}{\alpha_v}$.

These results imply that one single item can be compared with itself through subdomain definitions, making clear that one item can have different discriminant capabilities. Therefore, two different items (u, v) can be compared taking some of the following indexes $\frac{\alpha_u}{\alpha_v}$, $\frac{\alpha_u g_u}{\alpha_v}$, $\frac{\alpha_v}{\alpha_u g_u}$, $\frac{\alpha_v}{\alpha_u}$ and one single item u with the indexes g_u, $\frac{1}{g_u}$.

2.3 An Improved and More Flexible 6PL Model

The authors in reference [16] propose another CDF with six parameters in a different context, and this function is defined as follows,

$$p(x = 1|\theta, \mu, \alpha, \beta, k, a, d) = d + \frac{a - d}{1 + \frac{e^{-\alpha(\theta-\mu)}}{1+e^{k(\theta-\mu)}} + \frac{e^{-\beta(\theta-\mu)}}{1+e^{-k(\theta-\mu)}}} \tag{11}$$

where the definition $k = \frac{2\alpha\beta}{|\alpha+\beta|}$ specifies a constraint on the possible values of k. However, the model discussed in this work only requires that $0 \leq d < a \leq 1, \mu \in (-\infty, +\infty)$ and does not impose constraints on the possible values of k. Notice also that the model satisfies the two asymptotic behaviors when $\theta \to \pm\infty$ and that the Rasch's model can be obtained as a particular case when $a = 1, d = 0, k = 0$ and $\alpha = \beta$. The possible values of α and β are deduced from an analysis of the asymptotic behavior of the function (11). This analysis shows that $0 < \alpha$ and $0 < \beta$.

2.4 The Flexible 6PL Model and Specific Objectivity

The condition of specific objectivity is in some way intimately related to the concept of inverse function. So that, given the CDF (11), one possible means to find the proper transformation, leading to the property of specific objectivity, consists in finding the roots of the equation (12)

$$1 + \frac{1}{1 + e^{k(\theta-\mu)}} e^{-\alpha(\theta-\mu)} + \frac{1}{1 + e^{-k(\theta-\mu)}} e^{-\beta(\theta-\mu)} = \frac{a - d}{p(x = 1|\theta, \mu, \alpha, \beta, k, a, d) - d} \tag{12}$$

which comes after some manipulation over the Equation (11).

At first sight, it might be quite difficult to find an analytical expression of the possible roots of (12). However, the asymptotic analysis shed some light on the behavior of the left side of (12) in the limits $k \longrightarrow \pm\infty$ and $\theta \longrightarrow \pm\infty$.

Conditions specified by Table 1 say that, for some proper parameters α, β, k, there are regions in the domain of the function $p(x = 1|\theta, \mu, \alpha, \beta, k, a, d)$ where the specific objectivity is achieved. So, for example, the conditions $\alpha > 0, \beta > 0$ and $k \gg 1$ define the asymptotic behavior $1 + e^{-\alpha(\theta-\mu)}$ of the left side of (12) in an interval $(-\infty, \theta^*)$, where $\theta^* < \mu$.

A similar analysis on some interval $(\theta^{**}, +\infty)$, where $\mu < \theta^{**}$, shows the existence of specific objectivity, as well. By symmetry, one should expect similar results when $k \ll -1$, with the asymptotic behavior specified by the second row of the Table 1. So, three regions in the domain of definition of the CDF (11) specify the approximated behavior of the model.

Table 1. The asymptotic behavior of the left side of the Equation (12). Although ur means 'unrestricted value', it is assumed that $\alpha > 0$ and $\beta > 0$ to satisfy the asymptotic behavior of the function (11).

k	θ	α	β	Asymptotic behavior	Comments
$+\infty$	$-\infty$	$+$	ur	$1 + e^{-\alpha(\theta-\mu)}$	These conditions imply the existence of intervals to the left and right of $\theta = \mu$ where the CDF given by (11) becomes increasing and with complex behavior in a neighborhood of $\theta = \mu$
	$+\infty$	ur	$+$	$1 + e^{-\beta(\theta-\mu)}$	
$-\infty$	$-\infty$	ur	$+$	$1 + e^{-\beta(\theta-\mu)}$	Similar comments to previous row
	$+\infty$	$+$	ur	$1 + e^{-\alpha(\theta-\mu)}$	

Thus, the flexible 6PL model can be approximated by the piecewise exponential function (13),

$$p(x = 1|\theta, \mu, \alpha, \beta, k, a, d) \approx \begin{cases} \frac{1}{1+e^{-\alpha(\theta-\mu)}} & \text{if } \theta \in I_\alpha(\mu, k), \\ \frac{1}{1+e^{-\frac{\alpha+\beta}{2}(\theta-\mu)}} & \text{if } \theta \in I_{\frac{\alpha+\beta}{2}}(\mu, k), \\ \frac{1}{1+e^{-\beta(\theta-\mu)}} & \text{if } \theta \in I_\beta(\mu, k), \end{cases} \tag{13}$$

where the intervals $I_\alpha(\mu, k), I_{\frac{\alpha+\beta}{2}}(k), I_\beta(\mu, k)$ depend on the parameters μ and k. Note that $I_\alpha(\mu, k) \cap I_{\frac{\alpha+\beta}{2}}(\mu, k) = \emptyset, I_\alpha(\mu, k) \cap I_\beta(\mu, k) = \emptyset, I_\beta(\mu, k) \cap I_{\frac{\alpha+\beta}{2}}(\mu, k) = \emptyset$ and also $I_\alpha(\mu, k) \cup I_{\frac{\alpha+\beta}{2}}(\mu, k) \cup I_\beta(\mu, k) = \mathbb{R}$. This representation suggests that the item contains three discriminant parameters, as given by α, β and $\frac{\alpha+\beta}{2}$.

Thus, the comparison function is similar to the function of the extended Rasch 6PL model and three abilities and two items can be compared by using the functions defined in the Equation (9) and the Equation (10), respectively. Let $\mu_u, \alpha_u, \beta_u, k_u$ be the parameters of the item u and $\theta_r, \theta_s, \theta_t$, which are the abilities of three arbitrary and different examinees. Then the comparison function is evaluated as follows,

$$c((r, u), (s, u), (t, u)) = \frac{\theta_r - \theta_s}{\theta_r - \theta_t} \tag{14}$$

Similarly, in order to compare two items, let us consider only the examinees r and s and the items u and v. Then,

$$c(p(r, u), p(s, u), p(r, v), p(s, v)) = \frac{\gamma_u}{\gamma_v}, \tag{15}$$

compares two arbitrary items, without considering the examinee's characteristics. The comparisons can be obtained as follows, $\frac{\alpha_u}{\left(\frac{\alpha_v+\beta_v}{2}\right)}, \frac{\beta_u}{\left(\frac{\alpha_v+\beta_v}{2}\right)}, \frac{\alpha_u}{\beta_v}, \frac{\beta_u}{\alpha_v}, \frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\beta_v}, \frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\alpha_v}, \frac{\alpha_u}{\alpha_v}, \frac{\beta_u}{\beta_v}, \frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\left(\frac{\alpha_v+\beta_v}{2}\right)}$ or even within the same item it could be there comparisons by regions, $\frac{\alpha_u}{\left(\frac{\alpha_u+\beta_u}{2}\right)}, \frac{\beta_u}{\left(\frac{\alpha_u+\beta_u}{2}\right)}, \frac{\alpha_u}{\beta_u}, \frac{\beta_u}{\alpha_u}, \frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\beta_u}, \frac{\left(\frac{\alpha_u+\beta_u}{2}\right)}{\alpha_u}$.

3 Simulation Results

In order to get a glimpse of the possible conclusions coming from the theoretical analysis of both models, in the following, a discussion of some results obtained by a numerical simulation, is given.

The simulation can be implemented in two ways that assume the definition of the samples of examinees and items. The examinees' abilities and items' parameters represent the examinees and items, respectively. Given an examinee and an item, the probability of correct response is computed through the *a priori* definition of the psychometric model, as well.

However, these definitions are given with simulation purposes, since in a real calibration process they are unknown and need to be determined. Both approaches have some advantages and drawbacks and proceed as follows.

3.1 Complete Simulation Process

The experimental setting of the simulation considers the number of examinees, M, and items, N, as two variables running into proper sample sizes. The simulation defines a test with these examinees and items, whose abilities and parameters, respectively, are unknown. However, for simulation purposes, their values are randomly generated by a normal or uniform distribution, within *a priori* bounded real intervals or just considering some well–known finite sets of real numbers, *cf.* [17–19].

After defining the mechanism to generate the unknown variables and parameters, the definition of another mechanism to generate the items' responses is performed. This kind of simulation can involve the generation of examinees' responses based on the application of the CDF (11) along with uniformly and randomly generated values in the unit interval $(0, 1)$.

Given an examinee and an item, which are respectively represented by θ and a set of the parameter's values $(\mu, \alpha, \beta, k, a, d)$, the probability of successful response is computed through the model (11). So, the result is compared against r, which is a number randomly selected with uniform distribution from the unit interval $(0, 1)$. If $r \leq p$, then the response is assumed correct, otherwise the response is incorrect.

Responses are then processed to obtain the experimental values of the examinee's abilities, the item's difficulties and the probabilities of successful responses. However, this procedure does not necessarily ensure that the experimental probabilities of successful response are properly fitted into the generating CDF (11). Nevertheless, the acquisition of simulated experimental raw data is one of the main advantages of this procedure. These data can then be fitted into a proper psychometric model to obtain the estimated items' parameters.

3.2 Partial Simulation Process

Unlike the complete simulation process, the partial simulation does not require to generate items' responses. The probabilities of correct responses are not experimentally computed, but they are directly given by the CDF (11), and slightly

modified through a normal or uniformly distributed random noise. In other words, the procedure assumes that a set of responses are previously given and that a calibration process has been made to get the experimental probabilities of a successful response.

For a given simulated ability θ, the corresponding noised probability $p(\theta)$ of a correct response is a random variable with normal distribution $\mathcal{N}(p(\theta), \sigma_c)$ in some subinterval of the unit interval $[0, 1]$, or a random variable with skewed normal distribution $\mathcal{N}(0, \sigma_l)$ in a neighborhood of 0, or a random variable with skewed normal distribution $\mathcal{N}(1, \sigma_r)$ in a neighborhood of 1. The Fig. 1 shows this kind of behavior in the experimental probability with error.

The lack of a specific set of experimental responses, to explain where these probabilities are coming from, is one of the main inconvenience of this method. However, there are some possibilities to analyze the fit of the data to others different psychometric models, by making some comparisons against the generating function $p(x = 1|\theta, \mu, \alpha, \beta, k, a, d)$.

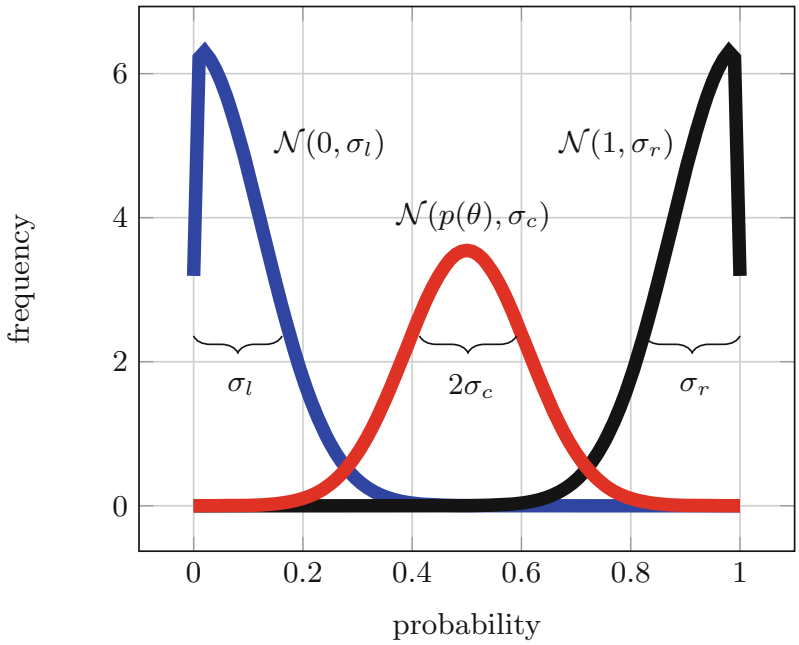

Fig. 1. Experimental probabilities are produced with three different PDF's. The probabilities closer to the asymptotes of the ICC are produced with skew normal distributions.

3.3 Presentation of Experimental Results

Based on some previous results in the literature on the topic [17,19], the examinees' abilities are assumed to have a normal distribution $\mathcal{N}(0, 1)$ and a sample size of 80 examinees has been chosen. Nevertheless, in a real situation, skewness and multimodality appear. Then, the function $p(x = 1|\theta, \mu, \alpha, \beta, k, a, d)$ should be adopted during the simulation. Hereafter, the comparison between the 2PL

model, the extended 6PL Rasch model and the flexible 6PL model considers the data produced by a partial simulation process.

Within ideal circumstances, the experimental data should fit properly the 2PL, the 6PL extended Rasch and the 6PL flexible models. The ideal situation assumes the lack of skewness and multimodality, and this situation is precisely considered by any of the different versions of the original Rasch model (1PL, 2PL, 3PL and 4PL models). Furthermore, in Subsection *Behavior of the extended 6PL Rasch's model* and Subsection *An improved and more flexible 6PL model* it is shown that some kind of skewness and multimodality in experimental data coming from a population, can be properly represented into the extended 6PL Rasch and the flexible 6PL models. In Fig. 2 this behavior is shown, where simulated data were generated with the CDF

$$p(x=1|\theta) = \frac{1}{1+\left(1-\frac{1}{1+e^{-10\theta}}\right)e^{-\theta}+\frac{e^{-5\theta}}{1+e^{-10\theta}}} \tag{16}$$

The PDF associated with this CDF contains some degree of skewness and a bimodality as Fig. 3 illustrates (curve with label 'original').

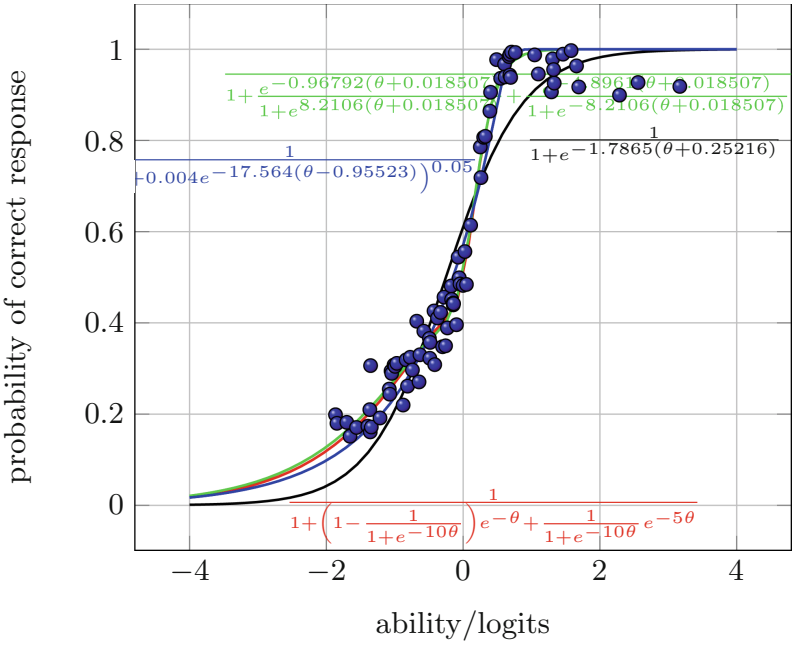

Fig. 2. Experimental simulated probabilities are generated with the CDF (16) and random noise given by PDF in Fig. 1.

Notice that in Fig. 3 the 'flexible' 6PL model with four fixed parameters ($\beta = 5, k = 10, a = 1, d = 0$), and two parameters (μ, α) determined by curve fitting, achieves a better approximation than 2PL and extended 6PL Rasch's models.

On the other hand, the 'extended' 6PL Rasch's model with four fixed parameters ($a = 1, d = 0, c = 0.004, g = 0.05$), and two parameters (μ, α) determined

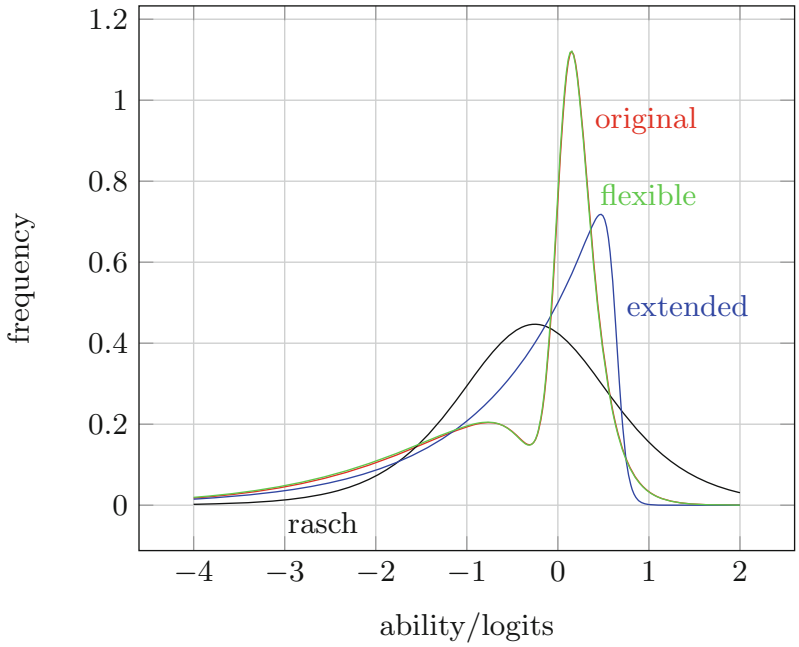

Fig. 3. The PDF's of the corresponding CDF's shown by the Fig. 2. The PDF coming from the fitted 6PL flexible CDF exactly fits the PDF coming from the CDF (16).

by curve fitting, performes better than 2PL Rasch's model, but does not improve over the 'extended' 6PL with four fixed parameters.

Finally, in Fig. 3 it is also shown that the simulated data fit better to the PDF from the flexible 6PL model, which contains skewness and bimodality. On the other hand, the extended 6PL Rasch's model contains the skewed behavior of the experimental data. The 2PL Rasch's model cannot predict these kind of behaviors.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two importante statistical methods to test the performance of fit [20–22]. The first method includes a penalization for overparameterization and, along with BIC, respectively produce the results 151.7, 151.18 and 151.29, when the 2PL Rasch's model, the flexible model and the extended 6PL Rasch's model consider only the two parameters (μ, α), while the rest of the parameters are fixed.

However, the 2PL Rasch's model is outperforming when, in the parameter estimation, the fitting of the whole set of parameters is considered. The 2PL Rasch's model produce similar results by using both AIC and BIC. Since AIC includes a penalization for overparameterization, it should be expected that the 2PL Rasch's model is more efficient. This result changes through the application of the following three steps, where the inattention and guessing parameters are set to 1 and 0, respectively

1. Fit the models with the presence of the corresponding set of fitting parameters; in other words, the fitting process must consider as unknown two parameters for the 2PL Rasch model and four parameters for the extended 6PL Rasch's model, and the flexible 6PL model.

2. Keep the unknown parameters μ, α and fix the rest of the unknown parameters to their values already found in the first step.
3. Fit again the models considering only the parameters μ, α as unknown. In this situation the penalization for overparameterization becomes identical in the three models, and the statistical tests AIC and BIC consider only the mean square error as a criterion for goodness of fit.

3.4 Computation of Item's Discriminant

The results of the simulation can also be useful to illustrate the computation of the item's discriminant in subdomains of abilities. For example, the Fig. 2 shows an item characteristic curve based on the flexible 6PL model, where $\alpha = 0.96792, \beta = 4.8961$, so that the item distinguishes to a greater degree for higher abilities and to a lower extent for smaller abilities; *i.e.*, $\frac{\beta}{\alpha} = 5.0584$, $\frac{\beta}{\gamma} = 1.6699$, $\frac{\gamma}{\alpha} = 3.0292$, where $\gamma = \frac{\alpha+\beta}{2}$.

3.5 Abilities and Parameters Estimation

The proposed models can actually be estimated and the number of observations needed to make the estimation can be acceptably good, since the maximum likelihood method leads to a system of decoupled nonlinear equations; namely, M equations for abilities $\boldsymbol{\theta}$ need to be solved, where some seeds are required for the $6N$ parameters, and $6N$ equations involving the parameters $\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{a}$ and $\boldsymbol{d}$, need to be solved assuming the abilities' values already found in the first step,

Particularly, the estimated ability of the i–th examinee can be computed by finding the roots of the equation i–th in the given system of decoupled nonlinear equations. Of course, the standard assumptions given in the literature must be also applied to get the required results (for instance, every examinee provide a correct response and one incorrect response, at least, to one pair of items in the set of items) [3,23,24].

The decoupled aspect of the nonlinear system of equations concerning the parameters of the items, leads to similar comments to those given at the end of the previous paragraph, although a system of six coupled nonlinear equations per item needs to be solved. In this case, it is very useful to know that the parameters α_i, β_i, a_i and d_i for item i–th need to satisfy the constraints $0 < \alpha_i, 0 < \beta_i$ and $0 \leq d_i < a_i \leq 1$.

Although the topic on root finding is currently under research by the authors of this paper, it is possible to get some information about this problem, based on the comments already made in the previous paragraphs, and the realization of relatively simple simulated examples. One example considers the case where 45 is the number of examinees and items. Assumming that the items' parameters are known, a single iteration estimates the abilities with an rms error equal to 0.03. The second example considers the case where 20 is the number of examinees and 15 is the number of items. The example applies two iterations to estimate the

examinees' abilities, assuming that at the first iteration the items' parameters are known. The estimate of the abilities at the first iteration produces an rms error equal to 0.01 in the values, while the estimate of the parameters in the same iteration produces an rms error equal to 0.09 in the values. Finally, the second iteration produces an estimate of abilities with an rms error equal to 0.07 in the values. The correlation coefficient between the estimated abilities in the first and the second iterations acquires the value 0.93, which is acceptably fine. Figure 4 shows the results after the second iteration in the abilities values for the case of 20 examinees and 15 items.

The extended 6PL Rasch's model has an identifibiality problem for its ability and difficulty parameter θ and $\mu + \frac{1}{\alpha}\ln c$, respectively, since the CDF remains the same when they are substituted by the corresponding expressions $\theta + \delta$ and $\mu + \frac{1}{\alpha}\ln c + \delta, 0 < \delta$ [25]. This is a location identifiability, but there is a scale identifiability as well, since the CDF also remains the same through the scaling $\frac{\alpha}{\kappa}, \theta\kappa, \left(\mu + \frac{1}{\alpha}\ln c\right)\kappa$, where $0 < \kappa$. A popular practice to solve the problem of identifiability defines the mean and standard deviation of the parameters θ's equal to zero and one across the test takers in the sample, respectively [25]. This procedure is usually applied in the cases of the 2PL and 3PL models, where a similar expression as a function of the ability θ and the parameters α and μ produce the identifiability problem, as well. However, there are some other possibilities of useful restrictions giving equally good estimates [25]. On the other hand, the approximation of the flexible 6PL model suggests also that the same set of restrictions should yield acceptable results.

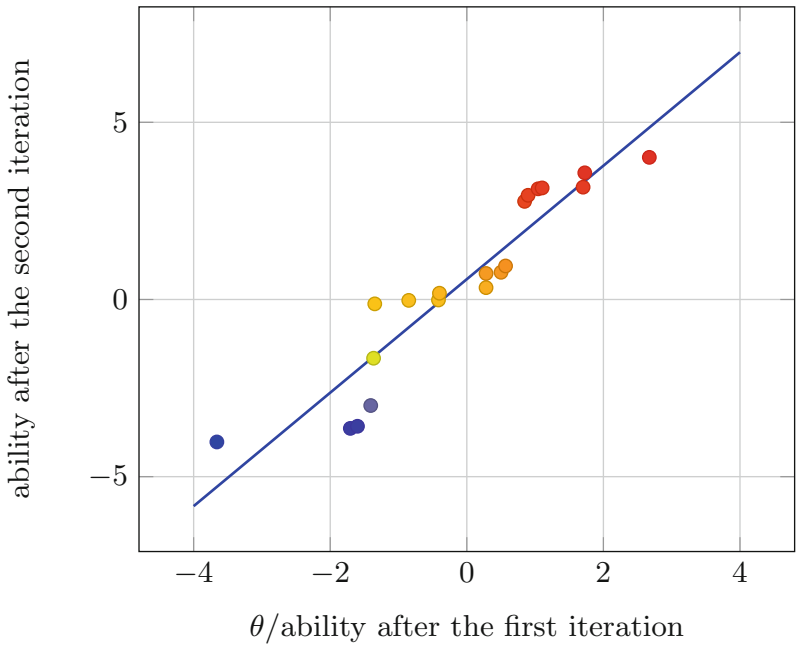

Fig. 4. Correlation between the abilities' estimates in first and second iterations in the case of 20 examinees and 15 items. The correlation coefficient is equal to 0.93 and the linear fit is given by the function $1.6\theta + 0.57$, where θ is the ability after the first iteration.

4 Conclusion

In addition to the possibility of the item's difficulty, the item's discriminant, the item guessing and the item inattention, some complex testing scenarios need to be considered in contexts where multimodality and skewness can be found. Thus, traditional psychometric models are not useful any more to cope with this problem. Particularly, two well–known psychometric models have been analyzed here. Every model is defined by six parameters, where four parameters are associated with the difficulty, discriminant, guessing and inattention, and the others could be associated with distribution skewness and bimodality from population under analysis. A direct extension of the usual psychometric 1PL, 2PL, 3PL and 4PL models is introduced, and the new models so defined use parameters to control the characteristics of bimodality and skewness in the distribution functions.

Naturally, the models performance depend on the accomplishment of a specific objectivity. In this sense, the concept of specific objectivity is also modified giving models that satisfy the specified requirements, at least in some intervals of ability in the worst case, as represented by the extended 6PL Rasch's model. Furthermore, as a drawback, the degree of skewness is constrained under the extended 6PL Rasch's model, which is avoided with the flexible 6PL model.

The identification of two and three values of the item's discriminant for a single item is an interesting result arising from the extended 6PL Rasch's model and the flexible 6PL model, respectively. Another result is that the value of the item's discriminant is found to depend on the interval where the ability belongs to. The redefinition of a specific objectivity suggests that two arbitrary items can be compared in four and nine different ways, respectively, for the 6PL Rasch's and flexible models, which depend on the selected values of the items' discriminants. Particularly, both models give the possibility of comparing one single item with itself, through its discriminant capabilities by subdomains.

The possibility of comparing examinees or items through the constraint of specific objectivity, even though they do not belong to the same asymptotic region, is one important point to remark. The partial specific objectivity in some models, or even the absence of this property, has suggested the introduction of pseudo–Rasch models [10].

The authors of the reference [26] introduce less restrictive ICC's to IRT models through the definition (17),

$$p(x=1|\theta,\mu,\alpha) = d + (a-d)F(m) \tag{17}$$

where $m = \alpha(\theta - \mu)$ and $F(x) = \frac{1}{(1+e^{-x})^{\lambda}}, 0 < \lambda$. On the other hand, the proposed extended 6PL Rasch's model is given by the function (18),

$$p(x=1|\theta,\mu,\alpha,a,c,d,g) = d + (a-d)p^{g}\left(x=1\left|\theta,\mu+\frac{1}{\alpha}\ln c,\alpha\right.\right), 0 < g, c \tag{18}$$

where $p(x=1|\theta,\mu,\alpha)$ is defined through the equation (3).

The value of the item's difficulty is the main difference between the two models. The authors of reference [26] assume that μ is the item's difficulty, while

the extended 6PL Rasch's model assumes that the difficulty is given by the expression $\mu + \frac{1}{\alpha}\ln c$. The presence of the parameter c implies the existence of right or left skewness in the PDF. So that, in this work, just one function for the CDF contains the left and right skewed behavior reported by Bolfarine et. al., who uses two CDF's to describe it. On the other hand, the parameter g plays exactly the same role in both models, which means that the two lead to similar conclusions about the comparison of items or examinees in subdomains.

References

1. Guzmán, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 12–21. Springer, Heidelberg (2004)
2. Kozierkiewicz-Hetmańska, A., Nguyen, N.T.: A computer adaptive testing method for intelligent tutoring systems. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part I. LNCS, vol. 6276, pp. 281–289. Springer, Heidelberg (2010)
3. van der Linden, W.J., Hambleton, R.K.: Handbook of Modern Item Response Theory. Springer, Heidelberg (1997)
4. Yadin, A.: Using unique assignments for reducing the bimodal grade distribution. Stand. Art. Sect. J. ACM Inroads **4**(1), 38–42 (2013)
5. Bonisegni, M.: Bimodal Distributions of Student Achievement. Consulted in 20 January 2015. https://expbook.wordpress.com/page/21/, 14 September (2008)
6. Robins, A.: Learning edge momentum: a new account of outcomes in CS1. Comput. Sci. Educ. **20**(1), 37–71 (2010)
7. Lazarinis, F., Green, S., Pearson, E.: Creating personalized assessments based on learner knowledge and objectives in a hypermedia Web testing application. Comput. Educ. **55**, 1732–1743 (2010)
8. Magis, D.: A note on the item information function of the four-parameter logistic model. Appl. Psychol. Meas. **37**(4), 304–315 (2013)
9. Irtel, H.: An extension of the concept of specific objectivity. Psychometrika **60**(1), 115–118 (1995)
10. Scheiblechner, H.H.: Rasch and pseudo-Rasch models: suitableness for practical test applications. Psychol. Sci. Q. **51**(2), 181–194 (2009)
11. Bazán, J.L., Branco, M.D., Bolfarine, H.: A skew item response model. Bayesian Anal. **1**(4), 861–892 (2006)
12. Gottschalk, P.G., Dunn, J.R.: The five-parameter logistic: a characterization and comparison with the four-parameter logistic. Anal. Biochem. **343**, 54–65 (2005)
13. Rasch, G.: An individualistic approach to item analysis. In: Lazarsfeld, P.F., Henry, N.W. (eds.) Reading in Mathematical Social Science, pp. 89–107. MIT Press, Cambridge (1996)
14. Rasch, G.: On Objectivity and Models for Measuring. In: Stene, J. (ed.) Lecture notes (1960)
15. Hambleton, R.K.: Item Response Theory: The Three-Parameter Logistic Model. In: Graduate School of Education, University of California, Los Angeles. Center for the Study of Evaluation Report No. 220 (1982)
16. Ricketts, J.H., Head, G.A.: A five-parameter logistic equation for investigating asymmetry of curvature in baroreflex studies. Am. J. Physiol. **277**, 441–454 (1999)

17. Engelhard, G. Jr.: A Simulation Study of Computerized Adaptive Testing with a Misspecified Measurement Model. In: Proceedings of the Section on Survey Research Methods 1986. American Statistical Association, pp. 631–636, 18–21 August 1986
18. van Rijn, P.W., Eggen, T.J.H.M., Hemker, B.T., Sanders, P.F.: A Selection Procedure for Polytomous Items in Computerized Adaptive Testing. In: Measurement and Research Department Reports 2000–5, 2000. Central Institute for Test Development - Cito, September 2000
19. Thorpe, G.L., Favia, A.: Data analysis using item response theory methodology: an introduction to selected programs and applications. Psychology Faculty Scholarship, Paper 20, The University of Maine, DigitalCommons@UMaine, July 2012
20. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control AC **19**(6), 716–723 (1974)
21. Maydeu-Olivares, A., García-Forero, C.: Goodness-of-Fit Testing. In: International Encyclopedia of Education 2010, vol. 7, pp. 190–196. Elsevier (2010)
22. Kang, T., Cohen, A.S.: IRT model selection methods for dichotomous items. Appl. Psychol. Meas. **4**(31), 331–358 (2007)
23. Olea, J., Ponsoda, V.: Tests Adaptativos Informatizados. Universidad Nacional de Educación a Distancia, Madrid (2003)
24. Muñiz, J.: Introducción a la Teoría de Respuesta a los Ítems. Psicología Pirámide, Madrid (1997)
25. van der Linden, W.J.: Linking response - time parameters onto a common scale. J. Educ. Meas. **47**(1), 92–114 (2010)
26. Bolfarine, H., Bazan, J.L.: Bayesian estimation of the logistic positive exponent IRT model. J. Educ. Behav. Stat. **35**(6), 693–713 (2010)

Computer Adaptive Testing and Cloud Computing

Joel Suárez–Cansino[1], Luis R. Morales–Manilla[2], and Virgilio López–Morales[1]

[1] Autonomous University of Hidalgo State,
Institute of Exact Sciences and Engineering,
Research Center of Information Technologies and Systems,
Mexico

[2] Polytechnic University of Tulancingo, Engineering Division,
Mexico

{joel.suarez,virgilio.lopez}@uaeh.edu.mx
luisr.morales@upt.edu.mx
http://www.uaeh.edu.mx

Abstract. Computer Adaptive Testing (CAT) is an example of a Computer Based Test (CBT) and is one of the main trending topics in the area of knowledge testing and, more recently, in e–learning or in Intelligent Tutoring Systems scenarios. The Item Response Theory (IRT) defines the theoretical basis of a CAT implementation, which assumes the existence of a repository of properly calibrated items that is used during the testing process of a particular examinee. The calibration and adaptation are based on an Item Characteristic Curve (ICC) related to an specific model, being Rasch's models the most widely used. CAT systems require high computational cost to implement the calibration and evaluation processes and the amount of concurrent users at a time could be large enough. Thus, the platform must support high concurrency and availability to perform a desired level of functionality. Technological tendencies in computing offer each time better platforms to develop and manage big collections of data for its processing and relevant information extraction. This paper presents a perspective of using new technologies in CAT as an alternative of implementation. Particularly, the use of a cloud computing platform as current alternative for online CAT systems using the capabilities of multicore processing and big amount of RAM that offers the cloud, to resolve the proper mathematical equations related to psychometric models and the operations described in their algorithms in a real evaluation scheme.

Keywords: Computer adaptive testing, item response theory, cloud computing, big data, multicore processing

1 Introduction

Computer Adaptive Testing (CAT) is an example of an informatic system oriented to knowledge, skills and behavior explorations, among others questions,

pp. 103–113; rec. 2015-04-09; acc. 2015-07-22 103 *Research in Computing Science 100 (2015)*

related to a person in an specific working area. Recently, this kind of systems has become very well known in online e–learning scenarios, as a result of the several benefits that these platforms provide in comparison with those given by Classical Testing Systems (CTS) [8]. CAT systems assume the existence of a pool of perfectly calibrated items, which are used during the testing of an specific person. The calibration process allows to determine the value of the parameters associated to certain psychometric features of the examinees and the items.

The essential idea in CAT systems consists in presenting to a given examinee one item after another, depending on the responses given to previous items along the testing process. If the response is incorrect, then the next item has a lower difficulty than the current one; on the other hand, if the response is correct, then the next item has a higher difficulty. This characteristic makes the entire process adaptable, which means that the item presentation adapts to the examinees' knowledge, in opposition to what classical testing does.

The implementation of this kind of environment is not trivial, since mathematical models and sophisticated algorithms are used for items' calibration and the adapting testing process itself. Particularly, the calibration process requires to find the solutions of a system of non–linear equations, while the adaptive testing computation needs at least to search for the solution of one non–linear function in *quasi* real time. In addition to this, the users can access the system in a concurrent way, which obviously impacts the system's and the hosting platform's performances.

Currently, there do exist informatics environments where a CAT system can be hosted with the aforementioned features, and they provide services over the Internet. In this sense, technological advances have changed the paradigm of implementing the solution in–house ('on–premises') to implement the solution on the Cloud ('off–premises').

Cloud computing is every day a more often used concept in computer systems. The Cloud–based service providers offer each time more and better management options, as well as benefits related to the accessible cost which depends of the user needs. So that nowadays it is possible that every individual has her/his own personal Cloud, which makes a proper environment for the hosting, creation and maintenance of applications for a wide variety of topics such as, for example, the automatized testing systems [8]. Cloud computing is an information service that offers software, platforms and infrastructure to an organization. Cloud computing technology incorporates different types of private, community, public and hybrid Clouds [8].

In the technology industry, Cloud computing is exponentially changing the implementation of information technology services. This is due to the fact that Cloud computing is a new information technology platform, that will positively change the nature of information management systems in the organizations.

The Cloud, which is another name for Cloud computing, is sometimes referred to as utility computing, since it uses interconnects networked devices to share information resources [12]. The online software and virtual maintenance of Internet infrastructure are among the benefits of Cloud computing for organizations,

and they can be synchronized from any geographic location [1].

Cloud computing uses the power of large computing devices that work on a common software format making parallel networks possible [10]. The large processing power of Cloud computing makes multiple systems on the Internet work by the interaction with virtual physical resources that conform the service architecture. In [1] Cloud computing is defined as Internet–based applications that can provide different information systems services including networking, filing and storage.

The innovation of using Cloud computing as a platform for the implementation of an information system inside an organization is due to the simplicity of configuring and programming the features that the providers offer [12], since by combining different hardware and software modalities in a virtual environment can increase the efficiency [4]. The authors in reference [1] highlight that cloud computing is the on–demand and expandable technology service offered over the Internet from data centers.

The next sections in this paper are organized in the following manner. In the second section the problem of the implementation of a CAT system in on–premises enviroments is explained. The third section deals with the solution proposal that involves the use of Cloud computing. The fourth section gives an introduction to CAT systems and its relationship with Cloud computing is explained. The fifth section shows and explains some screen test of the implementation of a prototype of CAT in Microsoft Azure Cloud platform. Finally, the sixth section provides some conclusions.

2 The Problem

Computer Adaptive Testing (CAT) is a technique that assumes the construction of items related to a previously given knowledge topic, which are then used to evaluate the abilities of a person in the aforementioned topic. This technique allows to finish the testing process in several ways and one of them consists in verifying that the most recent abilitie's values do not change within a given precision.

CAT systems based on IRT require a great computing capacity, since they are platforms that need high processing, availability and concurrence capacity. These needs arise from the fact that a proper CAT system requires the calibration of huge amount of items, which are related to a psychometric model containing a finite number of parameters, whose values depend on the corresponding item. In real scenarios, there does exist the possibility of handling at least one hundred items in a calibration process at a given instant of time, which means that one hundred parameters are required in the simplest case of the 1PL model or Rasch's psychometric model. In addition to this, and as a consequence of mobile devices development, CAT system must allow that the users access the testing service from any smart device, like cellphones or tablets, producing in this way the possibility of a huge demand on the part of the users and, as a consequence, a high concurrence of simultaneous access to the platform.

Even worse, a CAT system considers three main actors or user types; namely, the administrators, the evaluators and the examinees. Furthermore, since the system can offer the service to different educational institutions, which can consider very diverse topics of any knowledge area, several students and evaluators can access the system over the Internet demanding a very robust hardware and software infrastructure for the correct function of the platform, and as a support for the future growing on the demand of the users.

Therefore, CAT systems pose the problem of satisfying the need of great power of processing while granting a high availability, a flexible growing storage, an acceptable bandwidth size to support the high concurrence and a management easiness in the whole resources for the good platform's performance. These needs are not easily obtained under the standard use of personal servers landed on–premises and some services offered by the Internet providers.

3 Use of the Cloud Computing as an Alternative Solution

As it has been written above, the implementation of a CAT system nowadays presents several challenges, mainly those related to the requirements of hardware and software introduced at the end of the previous section. For this to be successfully solved, this paper presents an alternative of solution through the use of a Cloud Computing infrastructure currently offered by Microsoft, which is called Azure. The reason of using this particular cloud computing provider is centered mainly on the author's knowledge about Microsoft's technologies; however, future work implies to make a deeper analysis about the features that other providers offer, in order to determine which one could fit the best for a CAT system implementation.

4 Computer Adaptive Testing Systems and their Relationship with Cloud Computing

CAT systems are platforms for testing some given knowledge areas commonly used in e–learning or Intelligent Tutoring systems (ITR) [5]. A CAT system has a very solid mathematical support tipically defined by the Item Response Theory (IRT), which contains the basis for the implementation of a calibration mechanism for the items that will conform the tests [9]. In other words, the system must initially contain a repository of calibrated items, which are fitted through a calibration algorithm previously selected. The algorithms for calibration require the specification of a psychometric model, which is defined in terms of parameters. In consequence, the calibration process searches for the proper values of the parameters related to every item, based on the selected psychometric model. The selection of the psychometric model depends on the scenario where the test is applied, and the system must give reliable information about the abilities of the examinee, the dificulty of the items and the test in general.

When the values for the parameters are obtained, then these values can be useful in e–learning environments, Intelligent Tutoring Systems and CAT systems, which are examples of scenarios where a reliable and inmediate diagnostic is required to give support to the teaching–learning process.[7].

However, the implementation of a CAT system is not trivial since, as it has been aforementioned in this work, the system must completely cover the specific requirements to grant the correct operation of a platform of this kind. Nowadays, the digital revolution has transformed even the manner of teaching in the classroom and the testing of students, in such a way that a present–day testing mechanism is needed to bring both, the institutions and the students, to carry out these testing process by means of their digital devices. Internet is the more immediate alternative by making use of Cloud Computing service providers. In this way, a CAT system can be configured inside a platform that uses virtual physical resources, which interact in the infrastructure already hosted in a data center having the necessary support to host big amounts of informations systems, giving facilities for the access, configuration and management on the part of the users through a big bandwidth.

4.1 The Cloud and the Virtual Machines

Nowadays the cloud providers offer, as part of their services, the possibility of creating virtual machines with several combinations of resources that include memory sizes, hard disk storage capacity and the number of processor cores to be used. These benefits are important points to consider, when the creation of an online platform allowing the test of abilities in some knowledge topic is desired.

The considerations must include the high disponibility of the platform, which is granted by the right configuration of the virtual machine, the installation of an operating system in server version, the adequate assignment of the storage capacity, given that the system will have an extense number of items inside a repository and these items might contain hypermedia; in other words the items can include, in addition to plain text, embedded images, audio files, video or PDF documents. Then the necessary storage for all those items demands to the platform a big space in virtual hard disk.

On the other hand, it should carry out the right installation and configuration of the database manager, since it must create a database of non–calibrated items, a database of calibrated items, a database of testing results and a database with the information of the users and their access keys to the system. This represents an exponential growing in the databases due to the amount of users per educational institution having access to the platform.

Furthermore, there is the need of processing the CAT algorithms of the tests currently taken, applying the selected psychometric model and calculating the best item item choice along the test process. This leads to think in the concept of a real time system such as in the concurrency of the users, since in a same instant of time, the system could be testing a big amount of persons belonging to different educational institutions located in several geographic points [2], in different topics or knowledge areas, accessing from several types of digital and

mobile devices. By being a platform online that is processing the answers for all the items in an intelligent manner, it requires a huge processing power, which tipically is expressed by the number of cores of the processor that can be assigned in the virtual machine.

If all these things will be implemented in only one physical server, like used to be so, then it will lead to the fast saturation of the capabilities and, in consequence, to the impossibility to bring a testing service of high availability with no possible scaling to cover the needs. Moreover, the considerations of publishing a system of this kind over the Internet must include the security topic, since the security of the already stored information should be granted [11][13], and the privacy of the registered users data should be kept, as well. Some of the concerns and effects occurring when the Cloud Computing is used as a host of a system with educational orientation can be find in some references [1].

4.2 Microsoft Azure as a Cloud Computing Platform

Azure is the cloud portal of Microsoft and offers very good options of services with a relative low cost in accordance to the needs that somebody may have. In other words, Azure allows to the users the creation of virtual machines, web sites, mobile services, among others, in a wide gamma of options that the management portal has. If advanced services to create Internet based systems are required, then Azure is one of the main platforms for Cloud Computing that can be used nowadays.

The present work precisely suggests how well the implementation of a prototype of adaptive testing system, in the cloud of Azure, allows the use of features of scalability that guarantee the correct operation of the platform, the high concurrency and enough space for data storage, which give an advice that this kind of implementations are an excellent alternative to get the computing capabilities that a CAT system really needs.

4.3 Architecture of CAT System in the Cloud

Some research results show an architecture of adaptive testing mounted over a web access platform [6]. However, the work uses the traditional structure consisting of a server in–house which, through a public IP over TCP/IP protocol, makes the system accessible from any device. The present work proposes a similar architecture, but inside the space offered by one Cloud service provider, which allows inmediate scalability, growing flexibility in the virtual physical resources on–demand, in addition to better bandwidths that can satisfy the high traffic and the user's concurrency.

The Fig. 1 shows the architecture that has been designed for the implementation of the Ariya Framework in the Cloud Computing service of Microsoft Azure.

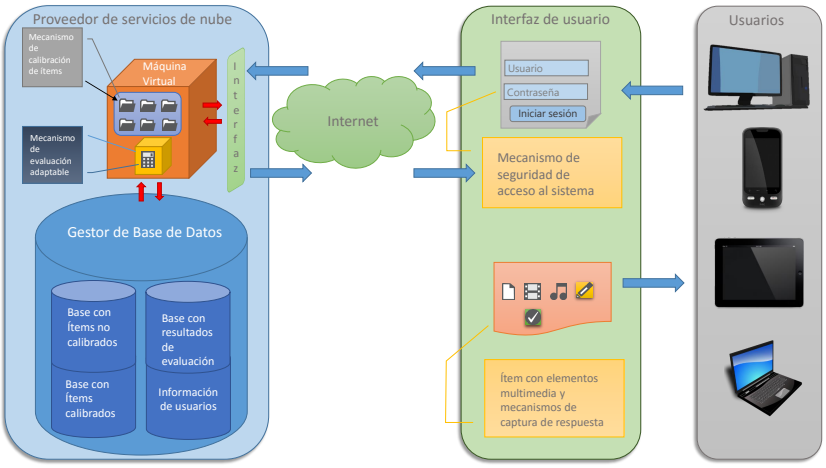

Fig. 1. Architecture diagram of the Framework Ariya in Azure

5 Implementation of the Framework Ariya in the Azure's Cloud

The implementation of the Framework Ariya in the cloud of Azure uses a Bizspark program consumer account to access the Azure management portal. Once inside the portal, the configuration of a virtual machine type A4 with 8 processor cores and 14GB of RAM memory is carried out . There is a fee for using the features of the cloud, but thinking in terms of the capabilities offered, like high processing and big RAM amount, the cost is worth to be payed. On the other hand, if one thinks for a while that the service will be given to different educational institutions to make their corresponding tests over this platform, it makes sense to think that the cost of the initial operation will be covered by the payed cuotes of rent for the Framework Ariya, as well.

With no doubt, the making of a business model to offer the services of the platform Ariya is necessary; however, the discussion about this component is left as a future work since the correct configuration of the prototype, and the making of the corresponding tests to garantee the right operation, should be made beforehand. The Fig. 2 shows the screencapture where the registration of a new cloud service for the Framework Ariya, inside the portal of Azure, appears.

5.1 Configuration of the Virtual Environment in Azure for the Framework Ariya

The Fig. 3 shows the screen capture of the virtual machine configured with the features previously described, relative to the Azure's management portal. The Fig. 4 shows the register of a virtual hard disk asociated to the platform Ariya with a Linux operating system. A version of Ubuntu server as the host for the

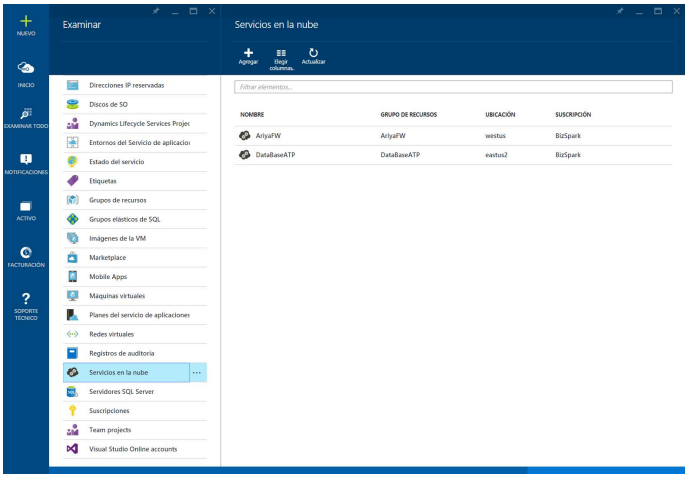

Fig. 2. Ariya cloud service in the portal of Azure

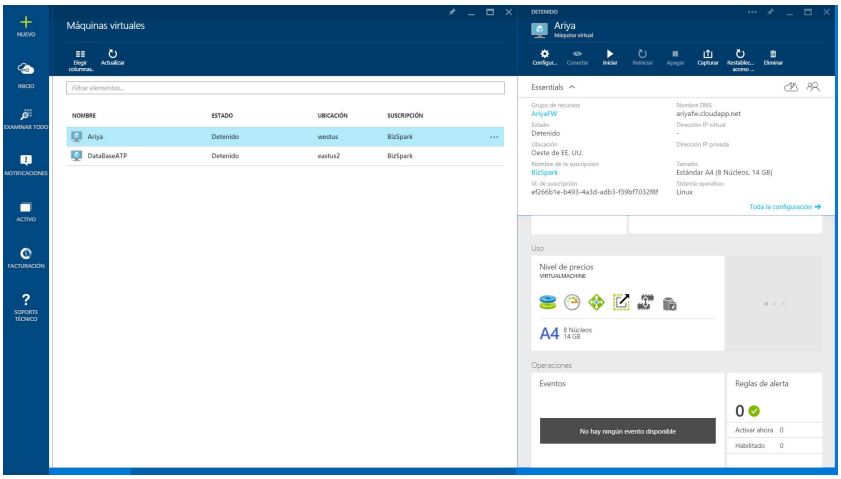

Fig. 3. Configuration of the virtual machine for the Framework Ariya in Azure

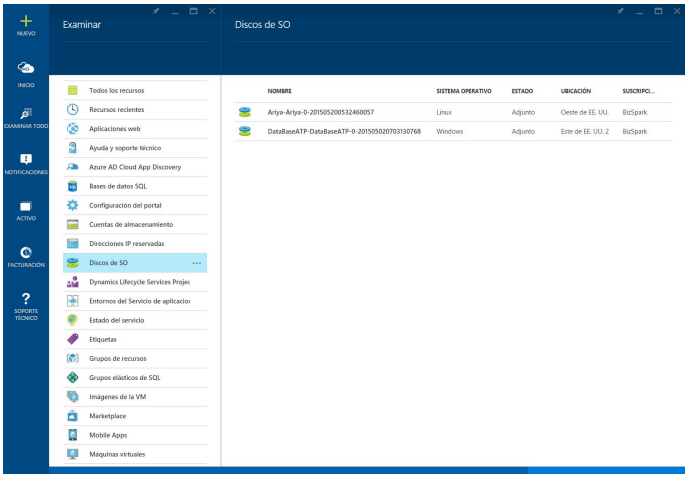

Fig. 4. Disk of the operating system assigned to Ariya's virtual machine

Fig. 5. Start user interface of the Framework Ariya

CAT system has been configured in this implementation.

The start screen of the user interface for the Ariya system, ready to be accessed by a user through her/his user name and password, can be appreciated in the Fig. 5.

6 Conclusions

This work presents a proposal of implementation for a CAT system using the benefits that the Cloud Computing service providers offer nowadays. In particular, the use of the Microsoft Azure cloud service case is shown. The implementation of the Framework Ariya offers a pilot test of how this kind of adaptive testing systems could be worked in order to bring to the users the functionalities when this kind of platforms is used. The easiness for accessing the system from any digital or mobile smart device that the users can already own is among these functionalities, since this characteristic does not require the installation of any additional software, it is enough to have an Internet connection and to access from the web browser of the device.

The implementation of the CAT system in the Cloud allows the use of the great processing power of the multicore processors, from which the virtual machines can be configured, in addition to the sharing of storage space in the database and in the hard disk for the calibrated items, the richness of multimedia contents and the results of the tests. The adaptive testing mechanism will use the resources of processing power in order to calculate, almost in real time, the results of the adaptive testing and the items' calibration.

Furthermore, this paper discusses the implementation of a prototype of the Framework Ariya which intents to have all the elements of a fully operational CAT system, so that the adaptive evaluation tests using different psychometric models can be made in a future work, integrating inclusive more complex and resource demanding mathematical techniques [3].

References

1. Arpaci, I., Kilicer, K., Bardaki, S.: Effects of security and privacy concerns on educational use of cloud services. Computers in Human Behavior 45, 93–98 (2015)
2. Atabekova, A., Gorbatenko, R., Chilingaryan, K.: Student's attitude to cloud-based learning in university diverse environment: a case of rusia. Academic Journals 10(1), 1–9 (2014)
3. Cheng, S.C., Huang, Y.M., Chen, J.N., Lin, Y.T.: Automatic leveling system for e-learning examination pool using entropy-based decision tree. Lecture Notes in Computer Science pp. 273–278 (2005)
4. Grossman, R.: The case for cloud computing. IT Professional 11(2), 23–27 (2009)
5. Guzmn, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. Lecture Notes in Computer Science 3220, 12 (2004)
6. Huang, Y., Lin, Y., Cheng, S.: An adaptive testing system for supporting versatile educational assessment. Computers and Education 52(1), 53–67 (2009)

7. Lazarinis, F., Green, S., Pearson, E.: Creating personalized assessments based on learner knowledge and objectives in a hypermedia web testing application. Computers and Education 55, 1732–1743 (2010)
8. Lin, Y.T., Wen, M.L., Jou, M., Wu, D.W.: A cloud-based learning environment for developing student reflection abilities. Computers in Human Behavior 32(3), 244–252 (2014)
9. van der Linden, W., Hambleton, R.: Handbook of Modern Item Response Theory (1997)
10. Park, E., Kim, K.J.: An ntegrated adoption model of mobile cloud services: Exploration of key determinants and extension of technology acceptance model. Telematics and Informatics 31(3), 376–385 (2014)
11. Stantchev, V., Colomo-Palacios, R., Soto-Acosta, P., Misra, S.: Learning management systems and cloud file hosting services: A study on student's acceptance. Computers in Human Behavior 31, 612–619 (2014)
12. Sultan, N.: Making use of cloud computing for healthcare provision: Opportunities and challenges. International Journal of Information Management 34, 177–184 (2014)
13. Wang, C., Wang, Q., Ren, K., Cao, N., Lou., W.: Toward secure and dependable storage services in cloud computing. IEEE Transactions on Services Computing 5 (2), 220–232 (2012)