UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

Área Académica de Computación y Electrónica

Selección de ítems en ambiente CAT usando intervalos de dificultades

TESIS

Que para obtener el grado de

Maestro en Ciencias Computacionales

PRESENTA

Randolfo Alberto Santos Quiróz

Director de tesis: Joel Suárez Cansino

Diciembre 2014, Pachuca de Soto, Hidalgo, México

Agradecimientos y dedicatorias

Agradecimientos:

A Dios, por todo lo que me ha dado, quizás sin merecerlo,
Al Dr. Joel por su amistad y su ayuda incondicional,
A mis maestros de la maestría, por sus enseñazas y consejos.

Dedicatorias:

A mi esposa Mercedes por su amor, apoyo y comprensión,
A mis hijos Rando, Beto y Brenda, motor e inspiración en mi vida,
A mis padres Randolfo (†) y Gloria, por darme la vida,
A mis hermanos por su cariño y respeto,
A la memoria de Roberto, Esperanza, Raquel y Engracia que en paz descancen.

Resumen

En este trabajo se aborda el problema de la tasa de exposición de items en una evaluación adaptable por computadora y su relación con un banco de ítems. Un ítem es una estructura que consiste de componentes reales o virtuales conteniendo posiblemente elementos textuales, imágenes, audio y/o video con los que se puede construir un contexto sobre el cual se plantea una pregunta, y de elementos o mecanismos diversos de adquisición de información útiles para proporcionar una respuesta a esta pregunta. Cada ítem tiene una dificultad asociada que depende de lo fácil que es responder a la pregunta sobre el contexto definido. Un banco de ítems es un depósito de este tipo de estructuras y, en este trabajo, la estructura de banco de ítems se define en términos de índices estadísticos que surgen de la teoría de la estadística de orden unidimensional , es decir, el índice de vecino más próximo y la variable aleatoria normal de la curva normal, y otro del concepto de compacidad de intervalos de números reales. En este sentido, el problema que considera este trabajo es relacionado con la tasa de exposición de la dificultad de ítems suponiendo que el banco de Ãtems es, de hecho, definido como un conjunto discreto finito de elementos de dificultad. Por lo tanto, se da el énfasis en la dificultad de ítems y se supone que el número de ítems por dificultad en el banco de ítems es ilimitado. Los resultados experimentales se obtienen a través de un entorno de simulación que tiene en cuenta la definición de la estructura de un banco de ítems, la definición de un objeto de prueba y la definición de un contexto de administración de ítems. Por lo tanto, los resultados son principalmente experimentales, más que teóricos, aunque la validación del entorno de simulación se basa en los resultados teóricos de otros autores en el campo.

Índice general

Índice de figuras

Índice de tablas

Índice de Algoritmos

CAPÍTULO

1 Introducción

Diversas plataformas computacionales para la administración del aprendizaje, contienen una componente de evaluación con el objeto de medir el nivel alcanzado por cada uno de los evaluados en relación con tópicos previamente especificados. Por sí misma, esta componente de evaluación define un campo bastante amplio de investigación que, a su vez, incluye diversos aspectos relacionados con la administración de los ítems empleados.

En este capítulo, se presenta una justificación del problema que es objeto de estudio en este documento y que se relaciona con el análisis de la administración de ítems en una evaluación adaptable computarizada, se plantea el problema mismo y una posible metodología para encontrar su respuesta, así como una justificación de ésta, para finalmente plantear el objetivo general y los objetivos específicos de la investigación realizada.

1.1. Planteamiento del problema

La evaluación adaptable por computadora es actualmente una alternativa a otros diferentes métodos de evaluación como, por ejemplo, el lineal, el lineal al vuelo, el testlets y el mastering. En cada uno de estos métodos existen diversos problemas cuyas posibles soluciones han estado bajo profunda revisión por parte de diferentes grupos de investigación [Patelis, 2000].

En la evaluación adaptable por computadora, la cual podría afirmarse es el método a la vanguardia en el desarrollo de sistemas de evaluación basados en computadora, existen diversos problemas que van desde la administración mis-

ma de la base de ítems, hasta la especificación del criterio óptimo de selección de ítems que debe emplearse en tiempo real.

La administración de ítems en tiempo real es una de las componentes esenciales de un evaluador adaptable, y en su diseño y construcción se han propuesto diferentes modelos de estimación inicial de habilidad, de selección de ítems, de control de exposición de los mismos, y diferentes modelos psicométricos, entre otros. Los problemas de selección y control de exposición de ítems están muy relacionados, y ellos tienen una interacción directa con el depósito de ítems, lugar desde donde se extraen éstos.

Dado un contexto de evaluación, el proceso de selección de ítems consiste en aplicar una función que mapea valores de habilidad de un sujeto de prueba hacia valores de dificultad de ítems, con el objeto de presentarle al sujeto los ítems que mejor información proporcionan acerca de su real nivel de habilidad.

En este sentido, si χ define al conjunto de habilidades y Δ es una colección de subconjuntos de dificultades de ítems, entonces $f : \chi \longmapsto \Delta$ define a la función f que, como se introduce más adelante, tiene una regla de correspondencia no muy simple.

Idealmente, el depósito de ítems debería contar con una cantidad suficiente de ellos (con una distribución aceptable de dificultades) para evitar que éstos sean sobreexpuestos en algún instante de tiempo. Sin embargo, esto no es así en aplicaciones reales. Un depósito con cantidad aceptable de ítems acerca de un tema específico debe contener al menos unos 300 ítems de diferente dificultad si se desea reducir el problema de exposición. Aún así, no deja de haber un número finito de ellos (y sus correspondientes dificultades) y siempre deben estar presentes políticas de administración de los ítems (considerando principalmente el costo de construcción de los mismos).

Así, aunque el dominio χ de f puede ser un intervalo real, no sucede lo mismo para la colección Δ que, en una situación concreta, sus elementos serán conjuntos numerables de dificultades. Lo que complica aún más la situación es que la cardinalidad de estos conjuntos es incluso variable cuando se considera que f juega el papel de un mapeo, en forma casi similar a como lo hace un mapeo del intervalo.

Si $\mathscr{B}$ denota al depósito de ítems, entonces la cardinalidad finita de este depósito (el número de dificultades con las que cuenta, ¡no el número de ítems con estas dificultades!) es $|\mathscr{B}|$. Para cada valor μ_i de dificultad en el depósito se tiene un número m_i de ítems con esta dificultad, lo que daría un total

$$\sum_{k=1}^{|\mathscr{B}|} m_k$$

de ítems en el depósito.

Así, metafóricamente cada dificultad etiqueta un recipiente que tiene $m_i(t)$ ítems en un instante de tiempo dado t, y el papel de la función f consiste en seleccionar el conjunto de dificultades (y por supuesto el correspondiente ítem) que mayor información proporciona acerca de la habilidad real del sujeto de prueba, dado que se conoce una estimación de ésta.

La implementación de un sistema de evaluación adaptable computarizado requiere de un conocimiento detallado del comportamiento del depósito de ítems a lo largo de las diferentes intervenciones de los sujetos de prueba. Por ejemplo, es conveniente determinar la frecuencia de uso de cada uno de los ítems en el depósito con el propósito de anticiparse a una sobreexposición o subexposición de éstos, características que, de no cuidarse, tendrían un fuerte impacto en los resultados de las pruebas y en el costo del mantenimiento del depósito.

Aunque diferentes autores han hecho contribuciones en esta dirección, y algunos de sus resultados se están considerando en las etapas actuales de desarrollo de un sistema de evaluación adaptable dentro del Cuerpo Académico de Computación Inteligente, no se ha determinado aún el comportamiento del depósito de ítems cuando la función f, que define el proceso de selección de ítems, está definida en términos de vecindades centradas en el valor del estimado actual de habilidad, denotado por $\hat{\theta}$.

La forma relativamente simple de seleccionar un ítem dentro del depósito consiste en aplicar directamente los criterios de contenido informativo (Fisher, Kullback–Leibler, entre otros), pero existen también otras alternativas que incluyen la posibilidad de seleccionar el ítem una vez que se ha definido un intervalo de valores posibles de dificultad (función de información de Fisher por intervalo [Barrada JR, 2004]).

La investigación que se propone en este trabajo de tesis está orientada a estudiar el comportamiento del depósito de ítems cuando los ítems a ser seleccionados se toman dentro de una vecindad alrededor del estimador actual de habilidad $\hat{\theta}$, tomando en cuenta solamente el modelo logístico de un solo parámetro o modelo de Rasch. El interés por este tema surge de la necesidad de contar con diferentes alternativas de selección de ítems y conocer las posibles consecuencias que ellas tienen en el comportamiento del depósito, tomando en cuenta entornos de aplicación relativamente grandes.

Figura 1.1 ilustra la forma en que interactúan algunos de los actores principales en el problema que se está planteando. El conjunto de habilidades θ's posibles de los sujetos de prueba se toma de un intervalo real, mientras que el depósito $\mathscr{B}$ cuenta con un conjunto finito de dificultades de ítems.

Para simplificar, el análisis considera que cada dificultad en el depósito es simplemente una etiqueta para recipientes de ítems que contienen un número ilimi-

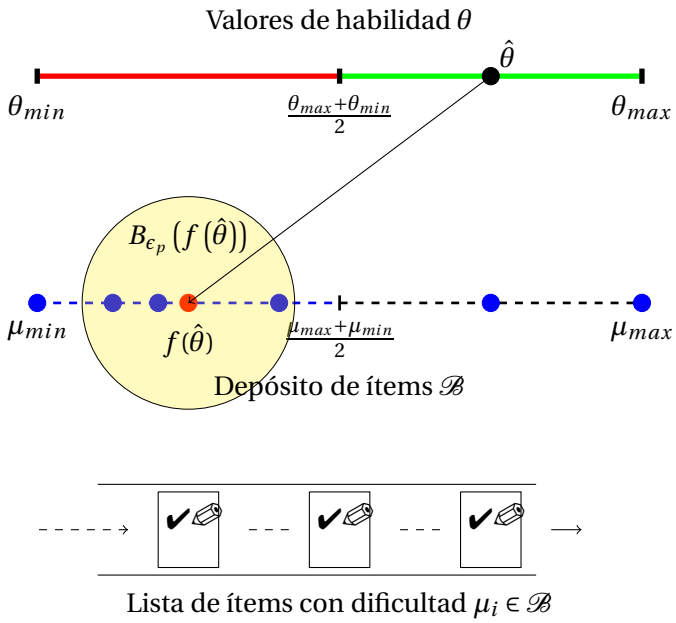


Figura 1.1: Los valores del estimador de habilidad $\hat{\theta}$ se emplean para seleccionar la siguiente dificultad dentro del depósito $\mathscr{B}$ de ítems. Esta selección toma en cuenta los ítems dentro de una vecindad de prueba $B_{\epsilon_p}(f(\hat{\theta}))$. El número de ítems por cada dificultad en $\mathscr{B}$ es ilimitado, no así los valores de dificultad posibles.

tado de ellos. Así, la única limitación que se tiene es la finitud del conjunto de valores de dificultad en el depósito de ítems.

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos valores del estimador de habilidad en dos instantes de tiempo diferentes, entonces puede ocurrir que

$$B_{\epsilon_p}(f(\hat{\theta}_1)) \bigcap B_{\epsilon_p}(f(\hat{\theta}_2)) = \varnothing$$

o bien

$$B_{\epsilon_p}(f(\hat{\theta}_1)) \bigcap B_{\epsilon_p}(f(\hat{\theta}_2)) \neq \varnothing$$

Aparentemente, ocurrirá que mientras mayor cardinalidad tenga la intersección (en caso de que ésta no sea vacía) habrá más posibilidades de sobreexponer las dificultades implicadas, ya que éstas serán accedidas por medio de dos vías; a saber, áquella dada por la vecindad de $f(\hat{\theta}_1)$ y áquella de $f(\hat{\theta}_2)$.

Nótese que debido a que no existe restricción alguna en el número de ítems etiquetados por una dificultad en el depósito, el comportamiento de este último

será dado en términos de la frecuencia de uso de cada una de las dificultades que lo definen. Surgen entonces algunas preguntas cuyas respuestas tendrán influencia significativa en el diseño e instrumentación del método de selección de ítems mismo, por ejemplo ¿qué radios de vecindad o vecindades de prueba son las adecuadas?¿deben ser éstas uniformes o no; es decir, mantenerse con radio constante?¿aleatorias o siguiendo un patrón de comportamiento que depende del grado de avance de la prueba?

1.2. Hipótesis

En este documento, el criterio escogido para estudiar los efectos que tiene la vecindad de prueba en el desempeño del depósito, está basado en la tasa de exposición de cada dificultad y no del ítem, ya que se asume que estos últimos están disponibles en una cantidad ilimitada por cada valor de dificultad en el depósito de ítems.

En una primera aproximación, considerando a la tasa de exposición de la dificultad como una función de la vecindad o vecindades de prueba seleccionadas, podría esperarse que para vecindades pequeñas la tasa de exposición sea grande, y que lo mismo suceda para vecindades mayores. Esto es así porque, en el primer caso, los miembros de cada vecindad tendrán todos una probabilidad muy alta de ser escogidos (cardinalidad de la vecindad disminuye cuando $\epsilon_p \to 0$). En el segundo caso, la probabilidad de seleccionar de nuevo a un mismo reactivo aumenta porque habrá al menos dos vías diferentes de acceder a ellos ($\epsilon_p \to \infty$), ya que los subconjuntos pueden ahora intersectarse.

Lo anterior hace sospechar que existe una transición de fase, especificada por un valor del radio ϵ_p^* de la vecindad de prueba, en donde la tasa de exposición es mínima, lo que Figura 1.2 ilustra esquemáticamente como un comportamiento asimptótico para vecindades cercanas a cero y crecimiento monótono para vecindades grandes, ¿cuál es el valor de ϵ_p^* si este es el caso?¿realmente existe este fenómeno?¿se puede verificar experimentalmente que esto ocurre, o no ocurre, al menos en forma parcial?

1.3. Objetivos

El objetivo general está enfocado al análisis del comportamiento de la tasa de exposición de dificultades en un ambiente simulado. Para ello, se cuenta con el desarrollo de las interfaces de un sistema en el que fácilmente se pueden integrar opciones de simulación en este sentido. Así pues, lo que sigue se sustenta en la existencia preliminar de un ambiente de evaluación.

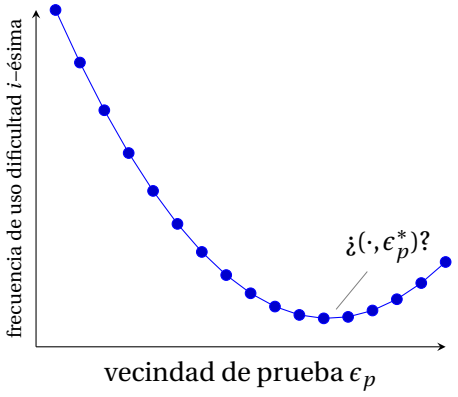


Figura 1.2: Esquema de comportamiento de tasa de exposición de dificultad i–ésima en depósito $\mathscr{B}$ como función de vecindad de prueba ϵ_p. ¿Existe una vecindad de prueba ϵ_p^* para la cual la frecuencia de uso del ítem es mínima?

1.3.1. Objetivo General

Analizar el desempeño del depósito de ítems como una función de la vecindad de prueba para diferentes condiciones de las variables y parámetros que describen y definen, respectivamente, un ambiente de evaluación adaptable computarizada, empleando para ello solamente el modelo psicométrico logístico de un solo parámetro (modelo de Rasch) y la información de Fisher.

1.3.2. Objetivos Específicos

1. Selección del modelo psicométrico.

2. Integrar simulación de medio ambiente de evaluación. En ella se deben simular las habilidades, las dificultades en el depósito de ítems y la respuesta a la pregunta de cada ítem.

3. Construcción simulada de depósito de ítems con al menos ochenta de ellos.

4. Experimentar numéricamente con depósito de ítems para diferentes radios de vecindades de prueba.

5. Describir el comportamiento de la función de selección y la tasa de exposición de dificultades como resultado del punto inmediato anterior.

1.4. Estrategia

Llevar a cabo experimentos reales, en los que los ítems son aplicados a sujetos de prueba concretos, es una tarea que requiere de la existencia de un sistema de evaluación que, además, sea adaptable. En este preciso momento, dentro del CA de Computación Inteligente, se encuentra en etapa de desarrollo un sistema con estas características, lo que hace difícil llevar experimentos de este tipo. Sin embargo, se cuenta con las herramientas y modelos computacionales que permiten llevar a cabo una simulación del proceso de evaluación, aunque en este trabajo se consideran condiciones mínimas, algunas de las cuales ya se han comentado y se profundizan con mayor detalle más adelante en este documento.

Por estas razones, la estrategia a seguir, para resolver el problema que se ha planteado, consiste en construir los algoritmos necesarios que sustenten la simulación de un proceso de evaluación bajo condiciones mínimas necesarias que se especifican más adelante. Esto conlleva la construcción misma del simulador, cuya validación estará sustentada en la verificación de convergencias, y en el cumplimiento de resultados teóricos que han sido demostrados por otros autores.

Asimismo, la estrategia contempla la realización de diversos experimentos con el simulador, de tal forma que con ellos se pueda determinar el grado de convergencia hacia el valor de habilidad en función de diversas características del depósito de ítems, el impacto que la vecindad de selección tiene en el desempeño de éste último y las dependencias posibles que tienen la precisión en la estimación de habilidad con algunas pruebas estadísticas basadas en estadísticas de orden.

CAPÍTULO

2 Estado del arte

Sin duda, la estructura del depósito de ítems tiene un fuerte impacto en la precisión con que se determine la habilidad de un sujeto de prueba. Por ejemplo, un depósito con muy pocos ítems en él, principalmente con valores de dificultad separados en forma apreciable uno del otro, seguramente producirá valores de habilidad cuya precisión puede no ser aceptable, sobre todo para un sujeto de prueba que cuente con una habilidad real ubicada entre dos valores de dificultad en el depósito.

Así, la estructura del depósito de ítems se encuentra definida por el número de ítems que hay en él, la distribución de dificultades asociadas a estos ítems o bien el número de ítems por dificultad, el intervalo de valores posibles para estas dificultades, el tipo de ítems, etc.

El método de selección es afectado también por la estructura del depósito. Es evidente que, sin importar qué método de selección se esté empleando, un depósito con una distribución de dificultades bastante raquítica hará que éstas sean sobreexpuestas en un tiempo bastante corto, sobre todo si la frecuencia de uso del depósito es alta.

Al momento actual, se han hecho ya múltiples estudios acerca de las relaciones que guardan la estructura del depósito de ítems con algunos de los métodos de selección, la tasa de exposición de los ítems (subexposición y sobreexposición) y la precisión con que se obtiene la habilidad de un sujeto de prueba.

A continuación, se lleva a cabo una revisión crítica-propositiva de algunos enfoques relacionados con la forma en que se seleccionan los ítems, su relación con la precisión del estimado de habilidad, la tasa de exposión de los ítems y la estructura del depósito de ítems. Asimismo, se incluyen aspectos relacionados con

la simulación de un proceso de evaluación y cómo el concepto de estadística de orden puede ser de utilidad para especificar la bondad de una estructura de depósito de ítems.

2.1. Selección de ítems por intervalo

Usando máxima información de Fisher y el modelo psicométrico logístico de un solo parámetro o modelo de Rasch, la forma más simple de selección de ítem sugiere seleccionar aquel ítem que proporcione mayor información acerca de la habilidad real del sujeto de prueba. Para algunos autores este es un criterio que basa la selección del ítem en la evaluación de la información en un solo valor del estimador $\hat{\theta}$ [Olea and Ponsoda, 2003].

Sin embargo, si se considera al depósito $\mathscr{B}$ de ítems como vecindad de cada uno de los valores de dificultad que ahí se encuentran, este criterio de selección también se puede tomar como un caso extremo de un criterio de selección basado en intervalo.

Diversos autores han enfocado su atención al estudio de reglas de selección en las que el intervalo dentro del cual se puede escoger el siguiente ítem en un instante del proceso de evaluación adaptable, es de radio más reducido. Estas propuestas son realmente críticas al método de selección basado solamente en la información de Fisher.

Sin embargo, uno de los inconvenientes de estas propuestas, incluida la basada solamente en la información de Fisher, es que se obtienen altas y bajas tasas de exposición de los ítems en el depósito, o tienden a afectar la exactitud de los estimados de habilidad, o bien conducen hacia ambas cosas [Cheng and Liou, 2002].

En el análisis del desempeño

Ojo: La precisión con que se determina el valor de la habilidad está en función del tamaño del depósito de ítems [Olea and Ponsoda, 2003]. En todo caso, lo que habría que experimentar es el efecto que tiene sobre la precisión del valor de habilidad el hecho de que se seleccione información de Fisher por intervalo por un lado y la selección uniforme por el otro. Ver por ejemplo [Glas, 2005]. Otro aspecto que resulta interesante investigar es el efecto que tiene sobre la precisión del valor del estimado de habilidad la separación entre cada valor de dificultad en el depósito de ítems. A todo lo anterior habría que agregar por supuesto el grado de convergencia que el método sigue hasta conseguir el valor de habilidad (consultar artículos ya encontrados).

2.2. Simulación de proceso de evaluación

Existen algunas versiones de simuladores de evaluación adaptable computarizada, en las que se simula completamente el proceso de evaluación, creando sujetos de prueba e ítems ficticios, o bien haciendo combinaciones de datos reales con datos ficticios. Las funcionalidades que proporcionan estos simuladores son muy variadas y cambian de un simulador a otro [Choi, 2008, Raîche and Blais, 2005].

Las razones por las que se proponen estos simuladores van desde áquellas que consideran la simulación de una evaluación adaptable computarizada como una etapa previa al desarrollo real de un sistema de evaluación adaptable, hasta áquellas que consideran que los simuladores actuales no proporcionan código abierto con el fin de insertar modelos propios del investigador [Raîche and Blais, 2005].

Además de considerar que estas razones son válidas y suficientes para avocarse al desarrollo de un simulador de evaluación adaptable computarizada, este trabajo hace hincapié en el hecho de que es importante tomar en cuenta dentro del simulador aspectos que consideren la estructura del depósito de ítems, tema que raramente se encuentra incluido en los simuladores existentes. En particular, estadísticos de prueba que permitan determinar ciertos aspectos en la distribución de dificultades, aspectos que tienen que ver con el grado de dispersión o acumulación de éstas.

Por estas razones, dentro del simulador propuesto en este trabajo se incluyen estadísticos de prueba asociados con el concepto de densidad de puntos en un intervalo y con el concepto de índice de vecino más cercano. Sobre todo este último concepto, se asocia con la idea de reconocer en qué momento se tiene una distribución de dificultades que es dispersa o que forma cúmulos. Sin alguna duda, estos aspectos son sumamente importantes en la etapa de administración del depósito de ítems. Por supuesto, el análisis, diseño y construcción del simulador que se propone en este trabajo sigue un estándar mínimo en lo que a contenido se refiere, tema que es analizado más adelante en otros capítulos.

2.3. Estadística de orden y estructura de depósito

Existen diferentes aplicaciones de la teoría de estadística de orden en muchas ramas de la ciencia, principalmente biología, geología, etc. [Clark and Evans, 1954]. Aquí se emplea también el concepto de vecino más próximo, pero de una manera diferente a como se hace en el caso del control de vecindades de dificultades para un proceso de selección de ítems, tema ya introducido en Sección 2.1.

La estructura del depósito de ítems generalmente se da en términos del número de ítems que contiene, el formato o tipo de éstos y sus dificultades correspon-

dientes. Se asume que el número de ítems proporciona implícitamente la distribución de dificultades. La propuesta que se hace en este trabajo radica en poner más atención a la distribución de dificultades en lugar del número de ítems, suponiendo que éstos se encuentran disponibles ilimitadamente.

Aunque sin ninguna duda el número de ítems resulta importante para contar con un depósito que se desempeñe en forma aceptable, la especificación de este valor no garantiza que se cuente con un buen número de dificultades y con una buena distribución de ellas, condiciones indispensables para lograr precisiones óptimas en los estimados de habilidad, y contar con una política efectiva de control de exposición de los ítems. Por ello, resulta más importante hablar de número de ítems por dificultad, término que aparece de manera natural cuando se considera la distribución de dificultades como un problema de agrupamiento en el campo de minería de datos.

El análisis de la distribución de dificultades se ve entonces como un problema de clusterización unidimensional que puede ser descrito por medio de índices bastante conocidos dentro de la mimería de datos, sobre todo cuando se habla de un método conocido como el de vecino más cercano. Específicamente, el índice de vecino más cercano. Con este índice se puede describir cuantitativamente el grado de dispersión de los puntos en un intervalo de valores previamente especificado.

En lo que respecta a la separación existente entre las dificultades en el depósito de ítems, algunos autores introducen el concepto de propiedades de segundo orden o locales como complemento de propiedades de primer orden o globales que representan patrones importantes de distribución (media, varianza, modalidad, etc.) mientras que propiedades locales describen patrones vecinales de la distribución completa [Levine, 2004].

Por ejemplo, la distancia entre puntos vecinos más cercanos se emplea para definir algunos índices en este sentido, siendo este índice un caso particular de áquel que considera los próximos k vecinos, $k \geq 1$.

Adicionalmente, se define un índice de no aleatoriedad para diferentes valores de escala denominado estadístico K de Ripley. Este índice proporciona un estadístico de vecino más cercano de mayor escala, dando una prueba de aleatoriedad para cada distancia desde la más pequeña hasta el límite mayor especificado. Se le conoce algunas veces como medida reducida del segundo momento, significando que se diseña para medir tendencias de segundo orden (es decir, agrupamiento local en oposición a un patrón general sobre la región). Sin embargo, también se encuentra sujeto a efectos de primer orden por lo que no es estrictamente una medida de segundo orden [Levine, 2004].

La teoría que sustenta todo lo anterior para el caso unidimensional se conoce como estadística de orden y a continuación se proporcionan algunas definiciones importantes que son de mucha utilidad más adelante,

Definición 2.3.1. (Distancia entre vecinos más cercanos)
Considérese una distribución unidimensional espacialmente aleatoria de n puntos $x_1, x_2, \ldots, x_n$, ordenados en forma creciente (sin importar en este momento el tipo de la distribución). Sea d_{ij} la distancia entre el punto i y su punto vecino más cercano j, entonces la distancia al vecino más cercano para el punto i se define como sigue

$$d_{NN}(i) = min\left\{d_{ij} \mid j = i \pm 1\right\}, \forall i = 1,2,\ldots,n$$

y el promedio de estas distancias como

$$\hat{d} = \frac{1}{n}\sum_{i=1}^{n} d_{NN}(i) \tag{2.1}$$

□

La utilidad de este promedio de distancias, $\hat{d}$, se hace evidente cuando se compara con su promedio $\langle\hat{d}\rangle$ respecto a un intervalo (a,b) previamente especificado, y que es dado por la ecuación siguiente si se asume que la distribución de los puntos es uniforme (ver Apéndice A),

$$\langle\hat{d}\rangle = \frac{n+2}{2n(n+1)}(b-a), \forall n \geq 2 \tag{2.2}$$

Ecuaciones (2.1) y (2.2) sirven para definir el denominado índice de vecinos más cercanos, η, como sigue,

$$\eta = \frac{\hat{d}}{\langle\hat{d}\rangle} \tag{2.3}$$

y que, según el intervalo de valores en que se ubique η, es un indicador del grado de agrupamiento, dispersión o aleatoriedad que tienen los puntos dentro del intervalo (a,b).

Es interesante investigar cómo este índice podría ser de utilidad para determinar la bondad de la estructura de un depósito de ítems, tema que, hasta donde se tiene conocimiento, no ha sido tratado por trabajos en esta dirección. Por ejemplo, los efectos que tendría una distribución de dificultades sumamente dispersa en el valor del estimado de habilidad y en la tasa de exposición de los ítems, aún cuando en una estructura tal se cuente con un número aceptable de ítems por dificultad. Esquema que deja claro que existe una diferencia entre número de ítems y distribución de valores de dificultad.

Este enfoque evidencia que el problema del control de la tasa de exposición abarca solamente el control de presentación de ítems por dificultad, y que el problema de la precisión en el valor del estimador de habilidad está íntimamente relacionado con la distribución de dificultades en el depósito.

En el caso general, el índice de vecino más cercano converge a cero conforme aumenta el número de puntos en el intervalo (a,b); es decir,

$$\begin{aligned}\lim_{n\to+\infty}\eta &= \frac{2}{b-a}\lim_{n\to+\infty}\frac{(1+\frac{1}{n})}{(1+\frac{2}{n})}\sum_{i=1}^{n}d_{NN}(i)\\ &= 0\end{aligned}$$

por lo que el límite inferior de η es cero. El caso extremo superior del valor de η ocurre cuando se tienen solamente dos dificultades $\mu_1 \leq \mu_2$ en el intervalo (a,b), lo que significa que $n=2$, el índice de vecinos es simplemente

$$\eta = 3\frac{\mu_2-\mu_1}{b-a}$$

y ello indica que el valor máximo del índice de vecino más cercano es igual a tres, siendo esto cierto cuando μ_1 está muy cerca de a y μ_2 cerca de b.

Obsérvese que valores de η cercanos a cero indican agrupación de los puntos en el intervalo (a,b), sin que ello signifique necesariamente una distribución uniforme a lo largo del intervalo. Por otro lado, valores de η cercanos a tres indican generalmente una fuerte dispersión de los puntos dentro del intervalo (a,b). La correspondencia que puede haber con la forma en que se distribuyen las dificultades en el depósito de ítems es clara, y resulta interesante profundizar en el tópico.

2.4. Densidad de puntos

Por supuesto, el índice de vecino más cercano η solamente proporciona información acerca del grado de dispersión de los puntos dentro del intervalo (a,b), pero no dice nada acerca de la densidad de éstos en comparación con la densidad del mismo intervalo; es decir, cómo tienden a compactarse a lo largo de él.

Por ejemplo, el índice de vecinos más cercanos puede indicar un alto grado de dispersión o agrupamiento, sin dejar claro si esto ocurre a lo largo de todo el intervalo (a,b) o solamente sobre una porción de él.

En forma muy similar a como se define el índice de vecino más cercano, se define el índice de densidad de un conjunto discreto de puntos. En este sentido es necesaria una ligera modificación de la definición de distancia que se ha dado anteriormente a través de Ecuación (2.1),

Definición 2.4.1. (Distancia promedio entre puntos)
Considérese una distribución unidimensional espacialmente aleatoria de n puntos $x_1, x_2, \ldots, x_n$, ordenados en forma creciente (sin importar en este momento el tipo de la distribución). Sea D_{ij} la distancia entre el punto i y el punto j, definida como sigue,

$$D_{ij} = \left| x_i - x_j \right|, \forall\, i, j \in \{1, 2, \ldots, n\}$$

entonces el promedio de las distancias asociadas al punto i es dado por la ecuación siguiente,

$$\hat{D}_i = \frac{1}{n} \sum_{j=1}^{n} D_{ij}, \forall\, i \in \{1, 2, \ldots, n\}$$

y el promedio de estas distancias promedio como,

$$\begin{aligned} \hat{D} &= \frac{1}{n} \sum_{i=1}^{n} \hat{D}_i \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_i - x_j \right| \end{aligned} \tag{2.4}$$

□

A diferencia del concepto de índice de vecino más cercano, se trata ahora de verificar qué tan compacto o denso se encuentra el conjunto de puntos con respecto a la densidad del intervalo (a, b) que los contiene, el cual es un conjunto compacto. Es bien sabido que el valor esperado del promedio de las distancias para cada punto en este intervalo es dado por la expresión siguiente, si se asume que la distribución de los puntos es uniforme (ver Apéndice A),

$$\langle \hat{D} \rangle = \frac{1}{3}(b - a)$$

por lo que una definición natural del índice de densidad es dada por la ecuación

$$\delta = \frac{\hat{D}}{\langle \hat{D} \rangle} \tag{2.5}$$

Una descripción más completa de la estructura del depósito de ítems requiere acompañar el índice del vecino más cercano con el índice de densidad. Al igual que el índice de vecino más cercano, los valores posibles del índice de densidad δ se encuentran en un intervalo cuyos extremos inferior y superior son, respectivamente, cero y la unidad. El primer caso ocurre cuando se tienen solamente dos puntos agrupados dentro del intervalo, mientras que el segundo cuando se tiene un suficiente número de puntos distribuidos uniformemente por todo el intervalo (agrupados).

CAPÍTULO

3 Simulador y diseño experimental

Para encontrar una respuesta al problema planteado en este trabajo, se requiere construir un simulador de evaluación adaptable computarizada, aunque sin entrar en los detalles más profundos de un sistema con estas características. Al final de cuentas, un simulador es un modelo que se propone para explicar de alguna forma el comportamiento del sistema real.

En este sentido, no solamente se debe proponer la forma en que se integran cada una de las componentes del simulador, sino también proponer diseños experimentales que permitan verificar y validar el funcionamiento del mismo. Estos son los tópicos que se analizan en este capítulo.

3.1. Estructura del simulador

La realización de un experimento real de evaluación adaptable computarizada requiere de una planeación cuidadosa, y el costo de los recursos de tiempo, espacios, humanos, tecnológicos y materiales puede resultar sumamente alto. Esta situación hace bastante complicado que la validación de nuevas ideas y modelos, asociados al proceso de evaluación adaptable, sean probados en una forma rápida, eficiente y confiable.

El diseño de la interfaz del simulador hace posible experimentar un proceso de evaluación adaptable computarizada, aunque en esta primera etapa contiene solamente los servicios mínimos requeridos. En particular, el diseño de la interfaz permite simular la respuesta del sujeto de prueba usando solamente el modelo psicométrico de un solo parámetro o modelo de Rasch, aunque deja abierta la

posibilidad de integrar otras diferentes, como por ejemplo los modelos de 2, 3 y 4 parámetros.

Adicionalmente, el diseño permite construir diferentes depósitos de ítems con distribución uniforme de dificultades, aunque desigualmente espaciadas. Por el momento, se asume que por cada dificultad existe una cantidad ilimitada de ítems, funcionalidad que debe ser modificada en un futuro próximo. Por supuesto, el usuario también puede decidir acerca del número de dificultades en el depósito, y si éste se guarda o no en un archivo.

El método de selección de ítems que ofrece el diseño es el que emplea el concepto de máxima información de Fisher, aunque deja abierta la posibilidad de integrar otros más, como por ejemplo el de Kullback-Leibler. Por supuesto, el usuario tiene la posibilidad de seleccionar entre todas estas opciones. En conexión con esto, el usuario puede también seleccionar uno de diferentes contextos de selección de ítems (por ejemplo incluir vecindades de dificultades, tal como se entiende este concepto en este trabajo, o vecindades basadas en criterios de máxima información de Fisher).

Asimismo, la técnica que emplea el diseño para determinar el valor del estimado de habilidad es la que se basa en el concepto de máxima verosimilitud, la cual se obtiene a través de la construcción dinámica (en tiempo real) de la función de verosimilitud, tomando en cuenta el historial de respuestas del sujeto de prueba correspondiente, las dificultades de los ítems empleados y el modelo psicométrico de Rasch. Sin embargo, también en este caso, el diseño deja abierta la posibilidad de integrar otras técnicas diferentes, como por ejemplo MAP (Maximum a Posteriori).

El diseño también permite que el usuario seleccione el número de pruebas a realizar, el valor de habilidad inicial de cada sujeto de prueba, diferentes radios de vecindad dentro del contexto de selección de ítems, el valor de habilidad del sujeto de prueba que el sistema debe estimar, además de integrar como único criterio de terminación de la prueba el grado de precisión con que se obtiene el estimado de habilidad, según el valor arrojado por la función de verosimilitud y no en comparación con los valores de dificultad en el depósito de ítems.

A cada depósito de ítems definido por el usuario, el diseño muestra los correspondientes valores de índice de vecino más cercano, índice de densidad e índice de variación estándar de la curva normal. Indicando en cada caso la bondad de la estructura del depósito de ítems. Como un complemento visual, el diseño también muestra gráficamente el comportamiento del depósito, según se van empleando cada una de las dificultades en el mismo, lo que ciertamente indica cuántos ítems deben estar asociados mínimamente a cada una de éstas.

Finalmente, el diseño permite experimentar con el impacto que tienen los índices de vecino más cercano, densidad y variación estándar en la precisión del estimado de habilidad, la convergencia a este estimado y la tasa de exposición de cada dificultad.

3.2. Análisis

Las componentes de la interfaz están definidas por los diferentes actores que intervienen en el proceso de evaluación. Estos actores, junto con sus elementos integrantes, son los siguientes,

1. Depósito de ítems. Esta componente la definen el rango de dificultades e incluso un subrango dentro de éste, el número de dificultades que contiene y, para una futura extensión del simulador, el número de ítems por dificultad. Se asume también que la distribución de estas dificultades es uniforme, pero nuevamente se deja la posibilidad de que sea de otra forma, por ejemplo Gaussiana.

2. Sujeto de prueba. Esta componente la definen el número de sujetos de prueba, el modelo psicométrico de respuestas con la posibilidad de integrar otros, la habilidad real de cada uno de los sujetos y también la habilidad inicial de cada uno de ellos, planteando la posibilidad de que ésta varíe de un sujeto a otro o que permanezca fija para todos. En cada caso, se cuenta con diferentes modelos de estimación de habilidad inicial como una futura extensión de las funcionalidades del simulador.

   El número de sujetos de prueba especifica cuántos sujetos van a realizar una prueba bajo las mismas condiciones de estructura de depósito y contexto de evaluación (mismas vecindades para selección de ítems). Se da pie a la posibilidad de que cada sujeto tenga un estimado de habilidad inicial completamente distinto a la de los otros sujetos. Además de que sus habilidades reales, a determinar por el proceso de simulación, pueden ser completamente distintas entre sí.

3. Administrador de ítems. Esta componente la definen el método de selección de ítems, enriquecido con varias posibilidades para futuras extensiones del simulador. También incluye un especificador de forma de terminación de una prueba particular, con varias posibilidades para futuras extensiones. Dentro de este actor, se encuentran también elementos que permiten definir el contexto de la evaluación, sobre todo la definición de vecindades para selección de ítems.

4. Presentador de resultados. Esta componente la definen graficadores de resultados, una para el índice de vecino más cercano, otra para el índice de densidad y una más para el índice de desviación de la curva normal, orientadas a describir la relación que tienen con la precisión en el estimado de habilidad, la convergencia a este estimado y la tasa de exposición de las dificultades en el depósito. Como complemento a esto, también se integra un graficador del comportamiento del depósito de ítems que se esté usando.

5. Certificador de entradas. Esta componente está enfocada a validar las diferentes entradas al sistema, manejando ventanas con mensajes que orienten acerca de los errores cometidos y la forma de corregirlos.

La idea subyacente en el diseño consiste en simular considerando bloques de sujetos de prueba y bloques de condiciones de administración para una misma estructura de depósito de ítems. Esto puede resultar redundante, pero se prefiere contar con un mayor número de resultados. Se espera llevar a cabo una mejora del simulador en este sentido en futuros trabajos.

El diseño de la interfaz adopta una filosofía modular o funcional, en el sentido de que en ella se muestran cuatro grandes bloques; a saber, uno dedicado a la definición del depósito de ítems, un segundo dedicado a la definición de los sujetos de prueba, un tercero dedicado a la definición de las condiciones de administración de las pruebas y, finalmente, un cuarto dedicado a la presentación de resultados. Figura 3.1 ejemplifica el funcionamiento de la interfaz gráfica.

Una parte importante del bloque de presentación de resultados es la que se refiere a la exposición de cada uno de las dificultades en el depósito, a lo largo de todas las pruebas de que consta un experimento. Esta parte está orientada a dar soporte visual acerca del comportamiento de cada una de las dificultades (su tasa de exposición) por cada experimento.

Asimismo, en la interfaz se muestran ayudas visuales que permiten verificar gráficamente la validación del funcionamiento del simulador, como un complemento a las validaciones basadas en resultados teóricos que se introducen más adelante .

Los bloques de definición de estructura de depósito, definición de sujetos de prueba y de presentación de resultados, se encuentran respaldados por sus correspondientes algoritmos, cuyas descripciones se dan en Sección 3.3.

3.3. Algoritmos

El algoritmo que sustenta mayormente el funcionamiento de la interfaz gráfica del simulador, emplea los datos introducidos en los bloques asociados a la definición

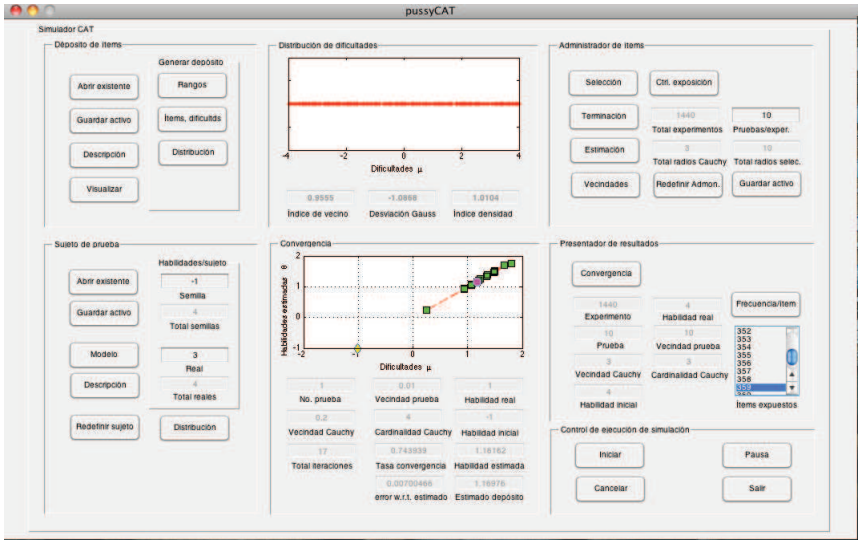


Figura 3.1: Interfaz gráfica de simulador de evaluación adaptable computarizada. Se muestran los cuatro bloques principales orientados al control del depósito de ítems, la definición de los sujetos de prueba, el control de la administración de ítems y el control de la presentación de resultados de simulación.

de la estructura del depósito de ítems, la definición de los sujetos de prueba y de la administración de estas pruebas.

Dentro de este algoritmo principal, se encuentran incluidos módulos destinados a llevar un control de selección de ítems, dar respuesta simulada a las preguntas de estos ítems y determinar el estimado de habilidad. A continuación, cada uno de estos módulos es también analizado a mayor profundidad.

3.3.1. Construcción del depósito de ítems

Esencialmente, el depósito de ítems se define de tal forma que sobre él se puede llevar a cabo un análisis de estadística de orden, con el fin de descubrir niveles de agrupamiento de las dificultades que lo definen. En este sentido, se parte de la definición de un intervalo abierto (a,b), conteniendo dificultades de ítems. Este intervalo se toma como la referencia con respecto a la cual se describirá el grado de agrupamiento, dispersión o aleatoriedad de las dificultades en el depósito.

Esto significa que dentro del intervalo abierto (a,b) se define un subintervalo (c,d) que permite tener control de la forma en que se distribuyen las dificultades en el depósito de ítems. Así, si se desea que estas dificultades se distribuyan sobre

todo el intervalo (a,b), entonces debe ser cierto que $(c,d) = (a,b)$. Si se desea que estas dificultades se distribuyan solamente en una porción compacta de (a,b), entonces $(c,d) \subset (a,b)$.

La forma en que se distribuyen las dificultades dentro del intervalo (c,d) es totalmente uniforme, cuidando que las distancias entre una dificultad arbitraria y sus vecinas más cercanas sean preferentemente heterogeneas. En este sentido, las dificultades $\mu_1, \mu_2, \ldots, \mu_n$ son todas elementos del intervalo (c,d) y se seleccionan en forma aleatoria con distribución uniforme, donde la distribución se define como sigue

$$\mathcal{U}_{(c,d)}(\mu) = \frac{1}{d-c}\chi_{(c,d)}(\mu)$$

siendo χ la función característica

$$\chi_{(c,d)}(\mu) = \begin{cases} 1 & \text{si } c \leq \mu \leq d \\ 0 & \text{en caso contrario} \end{cases}$$

Figura 3.2 muestra una distribución de dificultades dentro del subintervalo $(-3,-1) \subset (-4,+4)$. Las dificultades muestran un alto agrupamiento (índice de vecino más cercano ≈ 0.3307) en relación con el intervalo $(-4,+4)$, además de que con una confianza del 95 % la estructura que muestran las dificultades no es debido al azar (índice de variación estándar ≈ -2.4656). Adicionalmente, la densidad del conjunto de dificultades, otra vez en relación con el intervalo $(-4,+4)$, es bastante baja (índice de densidad ≈ 0.18126).

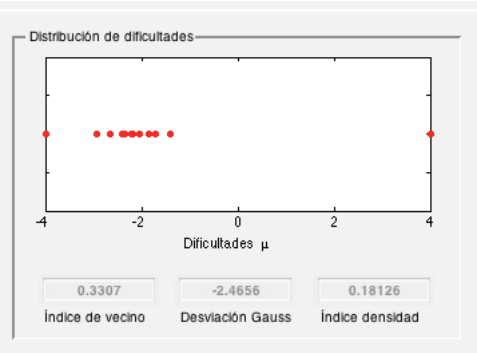


Figura 3.2: Porción de la interfaz gráfica del simulador de evaluación adaptable computarizada. Se muestra un ejemplo de distribución de dificultades en el depósito de ítems, indicando algunos índices estadísticos como el índice de vecino más cercano, el índice de variación estándar y el índice de densidad.

Por supuesto, la interfaz gráfica del simulador proporciona los medios para definir diferentes distribuciones de dificultad, condición indispensable para enriquecer los diversos contextos de experimentación posibles. La estructura del depósito de ítems resulta ser un factor bastante importante en los resultados que se obtienen en aplicaciones reales, según reportes realizados por otros autores [referencia].

La construcción de un algoritmo destinado a dar solución al problema de construcción de un depósito de ítems es, por lo tanto, relativamente directa. Más aún si se cuenta con un lenguaje de programación que contenga funcionalidades mínimas de generación de números aleatorios con distribución uniforme.

3.3.2. Definición de los sujetos de prueba

También se pueden integrar más de un sujeto de prueba por cada uno de los experimentos. Como se ha señalado previamente, la lógica de construcción de experimentos se basa en la definición de un depósito de ítems, un cierto número de sujetos de prueba, con sus correspondientes habilidades reales a ser predichas por la simulación, las habilidades iniciales de estos sujetos de prueba y la definición apropiada de condiciones de administración de las pruebas en cada uno de los experimentos.

Con los datos de habilidad real y de semilla o valor inicial de habilidad, se lleva a cabo una combinación (producto cartesiano) de los valores dados, lo que aumenta el número de experimentos posibles. Al combinar también estos valores con los dados para la administración de las pruebas, el número de experimentos aumenta aún más.

Por ejemplo, si se definen m sujetos de prueba, a través de la inserción de un igual número de habilidades reales, entonces también debe haber m habilidades iniciales o semillas, lo que da un total de m^2 combinaciones. Por otro lado, si se definen n vecindades de selección, junto con un igual número de cardinalidades en las mismas, entonces se tendrán m^2n^2 combinaciones. Finalmente, si se definen k vecindades de prueba, entonces se tendrá un total de m^2n^2k experimentos posibles en cada corrida de simulación.

Durante la definición de los sujetos de prueba, también se integra el modelo psicométrico a emplear, cuyo valor por defecto para esta investigación es el denominado modelo de Rasch o modelo logístico de un solo parámetro. Este modelo está descrito por la ecuación logística siguiente,

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

que proporciona la probabilidad de que el sujeto de prueba responda correctamente a un ítem con dificultad μ, sabiendo de antemano que dicho sujeto cuenta con habilidad real θ.

Cabe aclarar que este modelo se emplea para simular la respuesta que el sujeto de prueba otorga a un ítem con dificultad μ, asumiendo que el sujeto tiene una habilidad real θ. Basándose solamente en estas respuestas, y las dificultades de los ítems correspondientes , el sistema tiene que deducir que la habilidad real del sujeto es aproximadamente igual a θ, con cierto grado de precisión.

El algoritmo que implementa este comportamiento utiliza el concepto de distribución uniforme y el resultado numérico que arroja el modelo de Rasch. El conjunto de valores que puede tomar $P(X = 1|\theta, \mu)$ está definido por el intervalo unitario $(0, 1)$. Es claro que mientras más grande sea el valor de habilidad θ con respecto a la dificultad μ, entonces mayor será la probabilidad de que se responda correctamente a la pregunta del ítem. Esto lo ilustra gráficamente Figura 3.3.

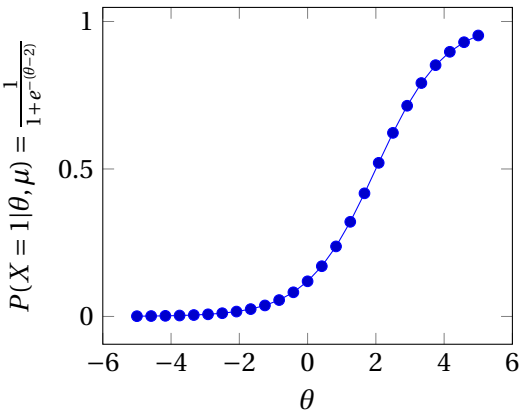


Figura 3.3: Modelo psicométrico logístico de un solo parámetro como función de la habilidad θ. En este caso, se asume que la dificultad del ítem es igual a 2. Obsérvese que la dificultad es un parámetro; es decir, valores diferentes proporcionan curvas diferentes, aunque basadas en el modelo 1PL. Las unidades de habilidad y dificultad son *logits*.

Suponiendo que r es el valor de la probabilidad, este será un punto que divide en dos segmentos al intervalo unitario, uno de ellos con longitud r (referido al extremo inferior del intervalo unitario) y el otro con longitud $1 - r$ (referido al extremo superior del intervalo unitario). Mientras más grande sea r, entonces mayor probabilidad se tendrá de responder correctamente la pregunta del ítem con dificultad μ, y menor será la probabilidad de responder incorrectamente $(1 - r)$.

Aleatoriamente, y de manera uniforme, se selecciona un punto del intervalo unitario $(0,1)$, sea s este punto. Dependiendo del valor r, el valor de s tendrá mayor probabilidad de caer o no en el primero segmento. Si es un elemento de él, entonces se dice que la respuesta es correcta; si no lo es, entonces la respuesta es incorrecta. La simulación de las evaluaciones suponen que los ítems son dicotómicos; es decir, solamente aceptan una respuesta que puede ser cierta o falsa. Basándose en esta definición de la respuesta del sujeto de prueba, Algoritmo 1 implementa el proceso de respuesta completo.

Algoritmo 1 Implementación de respuesta de sujeto de prueba, que cuenta con habilidad real θ^*

Entrada: Sujeto de prueba responde honestamente a cada ítem...
Salida: Una respuesta al ítem estadísticamente aceptable...

1: **procedimiento** RESPUESTA($\theta^*, \mu^{real}_{nuevo}$) ▷ Respuesta al ítem...
2: $a \leftarrow b$
3: Calcular probabilidad de respuesta correcta por parte sujeto de prueba...

$$P(X = 1|\theta^*, \mu^{real}_{nuevo}) = \frac{1}{1 + e^{-(\theta^* - \mu^{real}_{nuevo})}}$$

4: $r \leftarrow P(X = 1|\theta^*, \mu^{real}_{nuevo})$
5: Seleccionar aleatoria y uniformemente, número s en intervalo $(0,1)$...
6: **si** $r < s$ **entonces** ▷ Se tiene respuesta falso, si $r < s$...
7: $respuesta \leftarrow 0$
8: **en otro caso**
9: $respuesta \leftarrow 1$
10: **fin si**
11: **regresa** $respuesta$ ▷ *respuesta* es lo que responde sujeto de prueba...
12: **fin procedimiento**

Por supuesto, existen otras formas diferentes de obtener el patrón de respuesta y sus correspondientes dificultades. Una de ellas consiste simplemente en que el usuario del simulador proporcione estos datos, otra consiste en llevar a cabo un experimento real, cuidando en este caso el contexto de las pruebas. Sin pérdida de generalidad, solamente se instrumenta en este simulador el caso ya descrito a profundidad.

Lo que es importante resaltar aquí es que, finalmente, el simulador necesita que internamente se estén alimentando dificultades de ítems y respuestas a las preguntas de éstos, sin que el módulo de administración y el módulo de emisión de resultados sepa de dónde o cómo provienen estos datos. Ellos se encargan de

determinar los valores estimados de las variables de interés. En este sentido, las extensiones que pueden hacerse acerca de la forma de alimentar estos datos al sistema, pertenecen a un trabajo futuro.

3.3.3. Definición del proceso de selección de ítem

La definición del proceso de selección de un ítem (realmente la definición de una dificultad, para los propósitos de este trabajo), hace uso de las estimaciones que se tengan sobre la habilidad real del sujeto de prueba. Inicialmente, este estimado de habilidad se da a través de una semilla, cuya procedencia no es de interés conocer en este momento.

Con este estimado inicial de habilidad del sujeto de prueba, se da inicio el proceso de realización de esta última, presentando en forma recurrente el siguiente ítem, cuya selección no es de ninguna manera aleatoria, sino basada en criterios dados por el concepto de información de Fisher, o posiblemente algunas de sus múltiples variantes. En esta etapa del diseño del simulador, solamente se instrumenta internamente el criterio de Fisher, considerando el modelo psicométrico 1PL.

La idea de la información de Fisher radica en determinar qué ítems (dificultades) dentro del depósito proporcionan mayor información acerca de la habilidad real del sujeto de prueba, sabiendo que se cuenta con un estimado de ésta. Esto es lo que está detrás del concepto de máxima información de Fisher. Por supuesto, la expresión de esta información depende en gran medida del modelo psicométrico que se esté empleando. Recuérdese que, en este trabajo, se ha adoptado el modelo 1PL por su simplicidad de manejo y significado.

Es conocido que la información de Fisher obtenida al emplear el modelo 1PL, se hace máxima cuando la dificultad μ coincide con la habilidad estimada θ. Esto significa que, dado el estimado de habilidad θ, se busca en el depósito de ítems (dificultades) áquel que tenga asociada una dificultad cercana a este estimado de habilidad. Algoritmo 2 bosqueja la implementación que se ha hecho en este trabajo.

Debe tenerse en mente que, dado el valor del estimado de habilidad, no necesariamente se encuentra una dificultad que coincida con él. El estimado puede ser cualquier valor real (recuérdese, por ejemplo, el valor inicial de habilidad, la semilla), pero en el depósito solamente se encuentra un conjunto discreto finito de valores de dificultad. Una vez encontrado este valor, entonces se procede a mostrar al sujeto de prueba el correspondiente ítem, repitiendo este proceso hasta que se cumpla la condición de parada que se haya declarado previamente para la prueba. Esta condición de parada está íntimamente relacionada con la forma

Algoritmo 2 Esquema del algoritmo de selección de ítems. En este trabajo, la dificultad seleccionada dentro del depósito es la más cercana a θ_0. Se emplea al modelo 1PL.

Entrada: Estimación θ_0 de habilidad real del evaluado...

Salida: La complejidad μ de próximo ítem a ser presentado en la evaluación. El reactivo con esta complejidad da la máxima información acerca de la habilidad real del evaluado...

1: **procedimiento** SELECT(θ_0) ▷ Dificultad ítem maximiza información Fisher...
2: Dado conjunto de ítems (depósito), selecciona aquel cuya complejidad maximiza información de Fisher con estimación de habilidad θ_0. Sea esta complejidad μ...
3: **regresa** μ ▷ μ es la dificultad que maximiza información de Fisher...
4: **fin procedimiento**

en que se estima el valor de habilidad en cada iteración, y ambos temas son analizados en las próximas dos secciones de este capítulo.

3.3.4. Definición del proceso de estimación de habilidad

Este trabajo adopta el criterio de máxima verosimilitud, dejando para extensiones futuras del simulador otros criterios diferentes, por ejemplo Kullback-Leibler (criterios entrópico) o Maximum a Posteriori (MAP, basado en la construcción de un kernel).

La idea detrás del concepto de función de verosimilitud radica en construirla por medio de la secuencia de dificultades y respuestas dadas por el sujeto de prueba, a lo largo de esta última. Ya se ha mencionado que la prueba comienza con un valor inicial del estimado de habilidad, y que este estimado se emplea para seleccionar el primer ítem (dificultad) del depósito y presentárselo al sujeto de prueba. Ya se ha comentado también, con mayor detalle, este proceso de selección en Subsección 3.3.3.

Al maximizar la función de verosimilitud, se obtiene un problema que consiste en determinar los ceros de una función de la habilidad θ, parametrizada por las dificultades y las respuestas dadas hasta el instante, o punto, en que se encuentre la realización de la prueba. Por supuesto, la estructura de esta función depende del modelo psicométrico que se esté utilizando.

Resulta sumamente importante conservar los estimados de habilidad que se obtienen de esta forma, ellos son empleados en cada una de las iteraciones que definen la prueba para detectar el momento en que ella termina. En este sentido,

Algoritmo 3 Estimación de habilidad del sujeto de prueba.

Entrada: Patrón respuestas y dificultades obtenidas etapas previas de prueba...
Salida: Estimado de habilidad del sujeto de prueba...

1: **procedimiento** ESTIMATE(R, μ)▷ Estimado de habilidad de sujeto de prueba...
2: Construir función dependiente de habilidad θ, parametrizada con R y μ...

$$f(\theta_i, R, \mu) = \sum_{k=1}^{n} (r_{ik} - P(\theta_i|\mu_{ik}))$$

3: Determinar las raíces de la función anterior, sea ésta θ^*...
4: **regresa** θ^* ▷ θ^* es el estimado de habilidad siguiente...
5: **fin procedimiento**

el criterio de terminación se basa en la convergencia del estimado de habilidad hacia algún valor que, se asume, representa la habilidad real del sujeto de prueba.

3.3.5. Definición del proceso de terminación de una prueba

Por supuesto, existen criterios de terminación de una prueba que consideran un tiempo máximo de realización de la misma, la presentación de un número máximo de ítems (dificultades), los cambios en estimador de habilidad, a lo largo de un cierto número de presentaciones consecutivas más recientes de ítems (dificultades), y combinaciones posibles de todos estos casos.

El criterio adoptado en este trabajo considera solamente la convergencia de los estimados de habilidad, quedando las otras opciones para futuras extensiones del simulador. En este sentido, el tipo de convergencia que se asume es de tipo Cauchy, tal y como se define más adelante.

Se selecciona el criterio de Cauchy para la convergencia del estimado de habilidad debido a su sencillez y a que una definición usual de convergencia asume que se conoce el punto límite de la secuencia cosa que, por supuesto, no es el caso aquí, ya que se trata de determinar una aproximación al valore de habilidad real del sujeto de prueba. La siguiente, es una definición usual de convergencia para una secuencia,

Definición 3.3.1. (Convergencia) La sucesión $\theta(n)$ converge a θ^* cuando se cumple lo siguiente: $\forall \epsilon > 0, \exists K$ tal que $|\theta(n) - \theta^*| < \epsilon \forall\, n \geq K$.

Como se ha dicho, el inconveniente de esta definición radica en que se asume conocido el valor del límite θ^* de la secuencia. La siguiente definición prescinde de este conocimiento y solamente considera la comparación entre las iteraciones más recientes durante la prueba,

Definición 3.3.2. (Convergencia de Cauchy) La sucesión $\theta(n)$ converge en forma de Cauchy hacia una habilidad θ^* si, y solamente si, se cumple lo siguiente: $\forall \epsilon > 0, \exists K$ tal que $|\theta(n) - \theta(m)| < \epsilon$ siempre que $n, m > K$.

Intuitivamente, esta definición considera que la secuencia converge hacia un límite, no dice cuál, siempre que para cualquier vecindad que se dé sea posible encontrar que todos los puntos de la secuencia, definidos a partir de un cierto entero positivo K conocido, se encuentran dentro de la vecindad.

Por supuesto, si el límite existe, la convergencia de Cauchy permite determinar con buena precisión el valor de éste. Esto depende del radio ϵ de la vecindad. Esto se oye muy bien, pero existe todavía un gran problema; a saber, la secuencia de habilidades que se genera durante una prueba no sigue necesariamente un patrón tan uniforme. Ya se ha comentado que la estructura de la función a la que se le calcula su raíz (la habilidad) está parametrizada por las dificultades y las respuestas. Como se sabe, la selección de éstas tiene un comportamiento aleatorio, lo que significa que cualquier comportamiento uniforme que pueda tener la secuencia de valores estimados de habilidad, está sujeto a variaciones en algún momento, haciendo aparentemente inútil esperar que los valores siempre se encuentren dentro de una vecindad previamente definida.

Afortunadamente, se puede demostrar que este comportamiento no es tan frecuente a lo largo de una prueba y que la convergencia se comporta bastante bien casi seguramente, tal y como lo demuestra el siguiente teorema cuya demostración se debe a los autores [Chang and Ying, 2009],

Teorema 3.3.1. *Supóngase que $\{\hat{\theta}_k\}$ es la secuencia de estimadores especificados por las etapas del modelo de Rasch. Entonces, cuando $n \longrightarrow +\infty$, $\hat{\theta}_n \longrightarrow \theta$ casi seguramente*[1] *(a.s., almost surely) and $\sqrt{n/4}(\hat{\theta}_n - \theta) \longrightarrow N(0,1)$. Además, $4I_n(\hat{\theta})/n \longrightarrow 1$ a.s., donde $I_n(\theta) = \sum_{i=1}^{n} exp(\theta - b_i)/(1 + exp(\theta - b_i))^2$ es la información de Fisher observada.*

Este teorema sugiere además un procedimiento de validación del funcionamiento correcto del simulador. Basados en este teorema, la terminación de una prueba se lleva a cabo proporcionando una vecindad de radio ϵ y verificando que dentro de ella se tenga un cierto número de estimados de habilidad previamente definido, con la condición de que estos estimados sean los más recientes en forma consecutiva (uno tras otro). Este número se denomina la cardinalidad de la vecindad.

[1] Decir que un evento ocurre ***casi seguramente*** significa que ocurre con probabilidad aproximadamente igual a 1.

Algoritmo 4 Terminación de una prueba, basada en convergencia de secuencia de habilidades

Entrada: Valor de estimado de habilidad, radio de vecindad de selección, cardinalidad de vecindad de selección, estimados de habilidad más recientes contenidos en vecindad de selección...

Salida: Cierto o falso que vecindad de selección contiene la cardinalidad deseada de vecindad de prueba...

1: **procedimiento** TESTSTOP($\theta^*, \epsilon_s, B_{\epsilon_s}(\cdot), |B_{\epsilon_s}|$) ▷ Cierto o falso que vecindad contiene ya la cardinalidad deseada, incluyendo a θ^*...
2: Integrar nueva habilidad a vecindad...
3: **si** cardinalidad deseada se logra **entonces**
4: $end \leftarrow verdadero$
5: **en otro caso**
6: $end \leftarrow falso$
7: **fin si**
8: **regresa** end ▷ end confirma o no el logro de la cardinalidad deseada...
9: **fin procedimiento**

Algoritmo 5 Esquema del algoritmo de evaluación

Entrada: Estimación inicial θ_0 de habilidad real del sujeto de prueba...

Salida: La habilidad real del sujeto de prueba...

1: **procedimiento** EVALUATION(θ_0) ▷ Estimado habilidad real sujeto de prueba...
2: **mientras** No se satisfaga algún criterio de terminación **haz**
3: Responder a ítem obtenido en proceso de selección...
4: Estimar habilidad determinando raíces bajo historial de respuestas a ítems previos. Sea θ esta habilidad...
5: Select(θ)
6: **fin mientras**
7: **regresa** θ ▷ θ es el estimado de habilidad...
8: **fin procedimiento**

3.3.6. Algoritmo completo de evaluación adaptable computarizada

3.4. Diseño experimental

Responder al problema que se plantea en este trabajo, requiere de una propuesta de diseño experimental, en donde se especifiquen los diversos factores que intervienen en el proceso de simulación de una evaluación adaptable basada en computadoras. Esta propuesta es también importante porque debe incluir la definición de las unidades experimentales y la forma en que se verifica y valida el funcionamiento del simulador.

La verificación y validación del simulador resultan particularmente importantes. En este sentido, se emplean pruebas estadísticas que sustentan comportamientos razonables del simulador. En particular, las pruebas estadísticas χ^2 y F se emplean para confirmar hipótesis nulas basadas en igualdades de medias y valores esperados de varianzas.

3.4.1. Especificación del problema

El problema que se considera en este trabajo es el relacionado con la tasa de exposición de dificultades. Lo que se propone es estudiar esta tasa de exposición en términos de diversas variables entre las que se encuentran incluidas los radios ϵ_s y ϵ_p. El trabajo experimental consiste en determinar algunas de las causas posibles de altas tasas de exposición de dificultades.

Los factores posibles incluyen los radios antes dichos, la cardinalidad de una de estas vecindades, y la estructura del banco de ítems. Una primera aproximación acerca de cuál de estos factores influye más en la tasa de exposición, considera a la cardinalidad y radio de la vecindad relacionada con el proceso de convergencia.

3.4.2. Variable a ser medida

Las variables a ser medidas incluyen el radio de ambas vecindades, la cardinalidad de una de ellas, los índices que definen la estructura del banco de ítems (densidad, agrupamiento y aleatoriedad), los intervalos y subintervalos de valores de la estructura del banco de ítems, así como las diferentes condiciones de definición del evaluado y las políticas de administración de las pruebas.

Debe quedar claro, que algunas de las variables anteriores se mantendrán fijas durante el desarrollo de un experimento, mientras que otras serán realmente variables. En particular, las variables dependientes serán definidas por la tasa de exposición de las dificultades o, si se desea, el número total de iteraciones que se realizan en una prueba hasta llegar a satisfacer el pseudo criterio de Cauchy.

Algoritmo 6 Descripción completa del proceso de evaluación adaptable computarizada

Entrada: Haber contestado incorrectamenta al menos un ítem y correctamente al menos un ítem...

1: **procedimiento** CAT(θ_0)

2: $\theta_{nuevo} \leftarrow \theta_0$

3: **mientras** No es cierto que existe una secuencia de N habilidades consecutivas recientes $\theta_{k_1}, \ldots, \theta_{k_N}$ tal que $|\theta_{k_i} - \theta_{k_j}| < \epsilon_s, \forall i, j \in \{1, \ldots, N\}$. ¡Ojo, estas habilidades son las determinadas por las raíces de

$$\sum_{k=1}^{nuevo} \left(r_k - P(\theta | \mu_k^{real}) \right) = 0!$$

haz

4: Determinar qué dificultad es más informativa (en 1PL y usando criterio de máxima información de Fisher $\mu_{nuevo} = \theta_{nuevo}$)

5: Ir a depósito de ítems y, usando ϵ_p, determinar qué ítems en el depósito, con dificultad μ, están en la vecindad $|\mu - \mu_{nuevo}| < \epsilon_p$

6: Seleccionar aleatoriamente (asumiendo distribución uniforme) uno de los ítems con dificultad en la vecindad $|\mu - \mu_{nuevo}| < \epsilon_p$. Sea μ_{nuevo}^{real} la dificultad del ítem seleccionado

7: Presentar este ítem al sujeto de prueba

8: Hacer que sujeto de prueba responda a este ítem, asumiendo que él tiene habilidad real θ^*. En modelo 1PL, la probabilidad de responder correctamente es

$$P(\theta^*) = \frac{}{1 + e^{-(\theta^* - \mu_{nuevo}^{real})}}$$

Sea $r_{nuevo} \in \{0, 1\}$ esta respuesta

9: Sustituir respuesta r_{nuevo} y dificultad μ_{nuevo}^{real} en ecuación

$$\sum_{k=1}^{nuevo} \left(r_k - P\left(\theta | \mu_k^{real}\right) \right) = 0$$

y determinar raíz habilidad nueva θ_{nuevo}

10: **fin mientras**

11: **fin procedimiento**

Los factores que definen a las variables independientes son los que se refieren a los radios de vecindad y a la cardinalidad de una de estas vecindades. Asimismo, se incluyen como variables independientes a los índices que definen la estructura del banco de ítems. Factores relacionados con la habilidad real y estimación de la habilidad inicial del sujeto de prueba, se mantienen fijos durante la realización de un experimento, entendiendo que un experimento está definido por varias pruebas. Lo mismo ocurre en lo que respecta a variables que definen el número de ítems en el banco de ítems, y los intervalos de valores de dificultad. Estas características de la estructura del banco de ítems son independientes entre sí.

3.4.3. Factores y niveles a ser usados en el experimento

Los factores diversos deben ser controlables y el diseño dirá exactamente qué combinaciones de estos factores se deben correr y en qué orden. Todo esto debe estar dictado por la necesidad de corroborar el impacto que tienen sobre la tasa de exposición, o el número de iteraciones para lograr satisfacer el pseudo criterio de Cauchy, las variables independientes antes mencionadas.

Debe quedar claro que esta combinación de factores es dada en términos de las etiquetas que los identifican, mientras que sus valores correspondientes definen los niveles deseados de combinación. Por supuesto, existen otros factores que no se pueden controlar como lo son las respuestas que proporciona el sujeto de prueba. También se incluyen los valores específicos de dificultad dentro del banco de ítems. Con un diagrama causa–efecto, o bien, por medio de una especificación funcional, es relativamente fácil identificar la forma en que interactúan los factores con las variables, tal y como lo indica el diagrama causa–efecto mostrado por Figura 3.4.

Como se observa, el radio de la vecindad de selección depende del factor definido por la severidad del instructor. La severidad del instructor determina el grado con que el instructor está dispuesto a aceptar que el estimado de habilidad actual no es muy diferente de un conjunto de estimados de habilidad previos más recientes. Mientras más severo sea el instructor, menor será la distancia que debe existir entre estos estimados. Es decir, el radio de vecindad de selección ϵ_s es inversamente proporcional a la severidad del instructor. Un instructor muy severo exige mayor precisión en el estimado de habilidad.

Por otro lado, la cardinalidad de la vecindad de selección B_{ϵ_s} depende del factor definido por la tolerancia del instructor. La tolerancia del instructor determina el grado con que el instructor está dispuesto a aceptar que es suficiente que el estimado de habilidad actual no es muy diferente de un número previamente definido de estimados de habilidad previos más recientes. Mientras más tolerante es el instructor, menor será este número. Es decir, la cardinalidad de la vecindad de

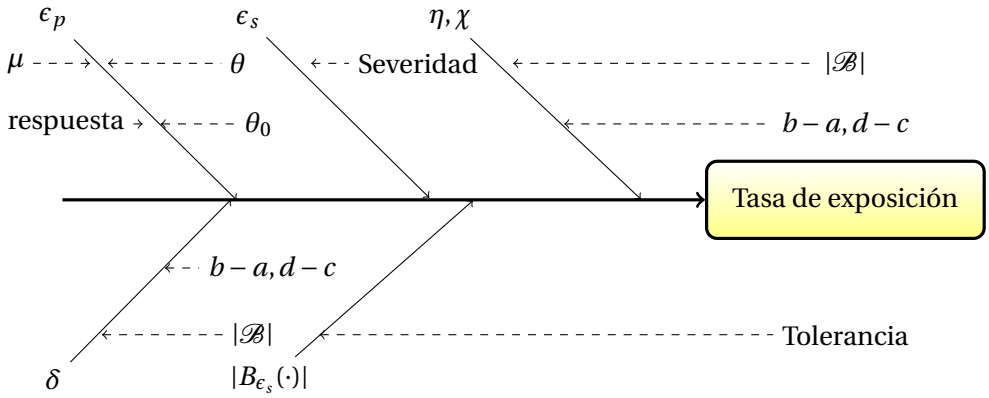


Figura 3.4: Diagrama causa–efecto en donde se muestran los factores principales que afectan la tasa de exposición de dificultades en un banco de ítems. Los factores de severidad y tolerancia se refieren al instructor y, tal como se comenta en el documento, estos están representados por el radio de vecindad de selección y la cardinalidad de esta vecindad, respectivamente.

prueba es inversamente proporcional a la tolerancia del instructor. Un instructor muy tolerante exige una muy baja cardinalidad.

También se observa en la Figura 3.4 que los índices con que se define la estructura del banco de ítems dependen de los factores determinados por la cardinalidad del banco de ítems (dificultades) y la longitud del intervalo y subintervalos ($(c,d) \subseteq (a,b)$).

Finalmente, factores tales como habilidad real (θ), estimado inicial de habilidad (θ_0), respuesta y dificultad de ítem (μ) afectan directamente al radio de la vecindad de prueba ϵ_p.

3.4.4. Definición del espacio de inferencia para el problema

No todos los factores citados en la Sección 3.4.3 servirán como variables. Algunos de ellos se fijan durante el desarrollo del experimento. Otro dato importante es que resulta imposible, por la naturaleza misma de un proceso de evaluación, fijar algunos de los factores secundarios que se muestran en Figura 3.4. Un ejemplo de ello es la respuesta que proporciona el sujeto de prueba.

Todos los factores que se consideran en una experimentación son los que se indican en Figura 3.4, aunque algunos se incluyen de una forma indirecta como son los casos de la Severidad y Tolerancia del instructor. Los factores que se man-

tienen variables y controlables incluyen solamente al radio ϵ_s de la vecindad de selección, $B_{\epsilon_s}(\cdot)$, y la cardinalidad $|B_{\epsilon_s}(\cdot)|$ de esta última, aunque una de ellas se mantiene fija en un experimento mientras la otra se considera variable.

Factores como la respuesta del sujeto de prueba se dejan variar de manera natural, descritos por el modelo logístico de un solo parámetro, aunque de una forma aleatoria. Es claro que un comportamiento de esta naturaleza se manifestará en las diferencias de resultados arrojados por el simulador. La dificultad μ del ítem es también un factor que tiene un comportamiento similar a este. La dificultad se escoge del banco de ítems $\mathscr{B}$ siguiendo parcialmente un proceso aleatorio que utiliza la vecindad de prueba definida por el radio ϵ_p y el estimado actual de habilidad.

No basta con definir los factores, también es necesario especificar los niveles de tratamiento de cada uno de ellos. Los resultados experimentales que se obtengan se aplicarán solamente al nivel de tratamiento específico, definiendo de esta forma el espacio de inferencia. La política seguida consiste en no fijar demasiados factores en el experimento, evitando que el espacio de inferencia se haga demasiado pequeño, pero al mismo tiempo evitando también que demasiados factores varíen libremente y que ello conduzca a resultados experimentales poco significativos.

3.4.5. Selección de los materiales con que se experimenta

El material experimental con que se trabaja incluye a los sujetos de prueba en una primera instancia. Estos sujetos se definen en términos de su habilidad real, supuestamente desconocida por el simulador, o bien por un conjunto de respuestas asociadas a un conjunto de ítems para los que se asume conocida su correspondiente dificultad.

También se incluye como material experimental al banco de ítems, el cual está descrito por los índices de agrupamiento, estructura y densidad, así como por los intervalos de valores de dificultad. En este sentido, cada valor de dificultad en el banco de ítems es en realidad parte del material experimental.

La experimentación está interesada en conocer los efectos que tienen en la precisión de la estimación de la habilidad del sujeto de prueba factores tales como los radios de vecindad de prueba y selección, y la cardinalidad de esta última vecindad, suponiendo que las características que definen al banco de ítems se mantienen constantes.

Asimismo, la experimentación se interesa en conocer los efectos que tienen sobre la tasa de exposición de cada ítem en el banco, factores tales como el radio de la vecindad de selección y su cardinalidad. A falta de una medición directa de

esta tasa de exposición por ítem, se propone también incluir el número de iteraciones que una prueba necesita para terminarse como un indicador de esta tasa de exposición. Aunque este indicador sería solamente a nivel global, en lugar de local. También aquí se asume que las características que definen al banco de ítems se mantienen fijas durante la experimentación.

Por supuesto, se puede proponer una experimentación en la que las características que definen al banco de ítems no sean constantes, pero esto se deja como una posible futura actividad de investigación. En cada una de las situaciones anteriores, los niveles de factor estarán definidos por los valores apropiados para cada una de las variables y los factores que se mantienen fijos.

3.4.6. Bosquejo del diseño

Una vez que se ha definido qué factores se van a considerar como variables y cuáles como fijos, resta definir la forma exacta en que se correrán los experimentos. Para evitar influencias que factores desconocidos puedan tener sobre los resultados experimentales, lo cual podría manifestarse en resultados experimentales sesgados, se aplican criterios combinatorios a las unidades experimentales. Así, se asignan aleatoriamente combinaciones de tratamiento y órdenes de corrida o ubicación de estas combinaciones.

Diseño del experimento considerando un solo factor

En una primera aproximación para plantear el diseño del experimento, se considera como factor importante el radio ϵ_s de la vecindad de selección $B_{\epsilon_s}(\cdot)$, mientras que las unidades experimentales son definidas por los sujetos de prueba. Se supone que existe un total de I diferentes niveles de tratamiento del factor de interés; es decir, I valores diferentes de radios de vecindad ϵ_s.

Cada nivel del factor de interés se aplicará a J diferentes sujetos de prueba, lo que da un total de J unidades experimentales diferentes. Si estos sujetos son también diferentes de un nivel a otro pero en igual número, entonces se tendrán IJ unidades experimentales diferentes en el experimento. Por cada nivel del factor de interés (ϵ_s), se empleará el mismo número de unidades experimentales o sujetos de prueba, lo que implica que en este trabajo el experimento se considera balanceado. Es necesario remarcar que este balanceo es por el igual número de sujetos de prueba por nivel, pero que estos sujetos de prueba son diferentes intraniveles e interniveles.

Se asumirá que las variables dependientes del radio de vecindad de selección son la precisión del estimado de habilidad o el número de iteraciones necesarias

para que se cumpla el pseudo criterio de Cauchy. Una de estas variables se denota por y_{ij}, y representa la respuesta medida asociada a la unidad experimental j (el sujeto de prueba j) considerando el nivel i del único factor (el radio de vecindad de selección).

Se desea determinar si los promedios de tratamiento por unidad experimental difieren o no. En este caso, los promedios se refieren a los promedios de errores de estimación de habilidad. Para asegurar la validez de la prueba estadística, deben seleccionarse aleatoriamente las IJ unidades experimentales (sujetos de prueba), asignar aleatoriamente las unidades a los niveles de tratamiento, y aleatoriamente correr el experimento.

Tabla 3.1 es una muestra de la forma en que se ordenan las combinaciones aleatorias y las órdenes de corrida. La primera columna especifica los niveles de tratamiento de la variable de interés que, en este caso, corresponde al radio ϵ_s de la vecindad de selección. Las filas definidas por la segunda hasta la última columnas, definen las unidades experimentales por cada nivel de tratamiento del factor de interés. Por supuesto, cada tabla como esta se acompaña de información relacionada con los otros factores que se mantienen fijos.

Tabla 3.1: Estructura de la definición de combinaciones aleatorias de unidades experimentales y sus órdenes de corrida.

Nivel de factor ϵ_s	Sujeto de prueba (θ)					
$\epsilon_s(1)$	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1J}
$\epsilon_s(2)$	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2J}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\epsilon_s(i)$	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{iJ}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\epsilon_s(I)$	x_{I1}	x_{I2}	$\cdots$	x_{Ij}	$\cdots$	x_{IJ}

Este método de experimentación que considera un solo factor como variable, mientras los demás se mantienen fijos, se denomina One Variable At a Time (OVAT) o One Factor At a Time (OFAT) y ha sido duramente criticado por algunos autores [Antony, 2003], argumentando que ignora efectos cruzados entre dos o más variables, lo que conduce a conclusiones óptimas de funcionamiento que no son apropiadas, sobre todo en lo que respecta a su generalidad.

Sin embargo, según algunos autores [Frey and Jugulum, 2006], bajo ciertas condiciones experimentales el método OFAT puede resultar más apropiado que el método factorial en el diseño de experimentos. Debido a que en este trabajo de tesis el principal enfoque consiste en determinar algunas de las condiciones bajo

las cuales se puede mejorar el funcionamiento de un evaluador adaptable computarizado real, más que obtener un modelo completo del funcionamiento de un sistema de esas características, se ha optado por un método OFAT para especificar el diseño experimental. Adicionalmente, parece natural llevar a cabo inicialmente una primera aproximación de este tipo, antes de adentrarse en la aplicación de métodos factoriales para el diseño experimental, enfoque que será completamente necesario conforme se agregen funcionalidades más complejas al sistema.

Para seleccionar aleatoriamente los IJ sujetos de prueba a partir de N unidades posibles, etiquétense las unidades potenciales desde 1 hasta N. Posteriormente, en forma aleatoria uniforme se seleccionan las IJ unidades experimentales. Estas son las unidades que se emplean en el experimento. Se debe tener cuidado en incluir unidades experimentales (sujetos de prueba) descritas por un amplio abaníco de valores de habilidad real.

A continuación, se asignan en forma aleatoria uniforme a las IJ unidades, seleccionadas como se describe arriba, los niveles de tratamiento especificados por los valores de vecindad previamente definidos. Nótese que ambos procesos de selección se llevan a cabo sin reemplazo.

Finalmente, para lograr un orden aleatorio en el orden de corrida del experimento, se usa de nueva cuenta una selección aleatoria unifirme, y sin reemplazo, de cada una de las IJ unidades experimentales ya obtenidas.

Aunque esto pudiera parecer estar realizando mucho trabajo innecesario, es la única forma de protegerse contra sesgos u otros factores desconocidos que pudieran afectar los resultados experimentales. En realidad, el proceso de selección aleatoria que se ha descrito es un buen método para asegurar la validez de los resultados experimentales. Este es un método de aleatorización completa, tal como se describe en referencia [Korenzen and Anderson, 1993].

3.4.7. Desarrollo del modelo matemático

Supóngase que y_{ij} representa la observación del error de estimación de habilidad producido por el tratamiento con nivel i del radio $\epsilon_s(i)$ de la vecindad de selección $B_{\epsilon_s(i)}(\cdot)$ aplicado al sujeto de prueba j (unidad experimental j). En lo que sigue, se asume que las observaciones y_{ij} provienen del siguiente modelo matemático,

$$y_{ij} = \mu + A_i + e_{ij} \tag{3.1}$$

donde μ es la media del proceso completo, A_i es el efecto diferencial debido al nivel de tratamiento i del factor fijo A_i, y e_{ij} es una componente de error aleatorio. Debe recordarse que por cada par i, j se lleva a cabo una prueba. Al sumar sobre

todos los niveles de tratamiento, se obtiene lo siguiente,

$$\sum_{i=1}^{I} y_{ij} = I\mu + \sum_{i=1}^{I} A_i + \sum_{i=1}^{I} e_{ij}$$

En la definición de los niveles de tratamiento de los sujetos de prueba, se sugiere se clasifiquen éstos de acuerdo a las siguientes etiquetas,

1. 0 = imposible de resolver (+4 logits).
2. 1 = con mucha dificultad (+2 logits).
3. 2 = con alguna dificultad (0 logits).
4. 3 = con muy poca dificultad (−2 logits).
5. 4 = sin ninguna dificultad (−4 logits).

aunque un refinamiento en la partición permite acceder a más combinaciones posibles de sujeto de prueba y los niveles de tratamiento para el factor variable. Se opta por agregar habilidades con saltos de +0.5 logits, partiendo de −4 logits hasta llegar a +4 logits, que son los valores extremos que definen, respectivamente, a sujetos de prueba con muy alta habilidad y muy baja habilidad. Esto da un total de 17 posibles sujetos de prueba, sin considerar todavía los niveles de tratamiento correspondientes a los radios de vecindad de selección.

Debe recordarse que la habilidad real es el único factor que describe los niveles de tratamiento del sujeto de prueba. Por otro lado, se mantienen fijos otros factores como, por ejemplo, el estimado inicial de la habilidad. En este sentido, existen estudios en los que estas habilidades iniciales se clasifican como alta (+4 logits), media (0 logits) y baja (−4 logits) [Burston and Burston, 1995].

Por otro lado, se estima que una terminación aceptable de una prueba debe emplear un número de iteraciones dentro del rango de valores 20 a 40; es decir, una prueba debe presentar un número de ítems dentro de este intervalo de valores. Pruebas preliminares con el simulador, sugieren que los valores del radio de vecindad de selección deben estar ubicados dentro del intervalo de valores [0.05, 2] logits. Los valores más pequeños de este intervalo deben estar asociados con cardinalidades pequeñas (2 dificultades cercanas más recientes) y los valores grandes con cardinalidades pequeñas o grandes (2 a 4 dificultades cercanas más recientes).

Una vez definidos los niveles de tratamiento y los niveles en que se fijan algunos de los factores, lo siguiente resume la forma en que se construye el diseño experimental en este trabajo,

1. Definir primero N, el número de sujetos de prueba. Las N potenciales unidades experimentales solamente están definidas por el valor real de habilidad. No se incluye el estimado inicial de habilidad porque este es un factor que es controlable y no es intrínseco al sujeto de prueba (unidad experimental).

   Las unidades experimentales posibles se definen con las habilidades reales (en logits) mostradas por Tabla 3.2, considerando también los niveles de tratamiento para el factor variable que, en este caso, corresponde al radio de vecindad de selección,

Tabla 3.2: Unidades experimentales en segunda columna que definen al sujeto de prueba con base en su habilidad real, la cual es desconocida por el evaluador adaptable. También se muestran los niveles de tratamiento propuestos para el factor variable representado por el radio de vecindad de selección.

Etiqueta	Habilidad (logits)	Nivel de factor ϵ_s (logits)
1	−4.0	0.05
2	−3.5	0.06
3	−3.0	0.07
4	−2.5	0.08
5	−2.0	0.09
6	−1.5	0.10
7	−1.0	0.11
8	−0.5	0.12
9	0.0	0.13
10	+0.5	0.14
11	+1.0	0.15
12	+1.5	0.16
13	+2.0	0.17
14	+2.5	0.18
15	+3.0	0.19
16	+3.5	0.20
17	+4.0	

   Por cada nivel de tratamiento definido para el factor variable ($\epsilon_s(i)$), se tomarán solamente tres unidades experimentales. Así, $N = 17$ es el total de las unidades potenciales, mientras que $J = 3$ es el conjunto de unidades experimentales a escoger por cada nivel de tratamiento.

2. Definir después I, el número de radios de vecindad. El número de niveles de tratamiento I se tomará aleatoriamente del conjunto posible de radios de vecindad de selección (en logits) mostrados en la tercera columna de Tabla 3.2, lo que da un total posible $M = 16$. Solamente se tomarán $I = 4$ niveles de tratamiento para el factor definido por el radio de vecindad de selección.

3. Definir ahora los N valores de habilidad real que definen a los sujetos de prueba. Ya está hecho arriba en punto 1.

4. Definir ahora los I niveles de tratamiento. Ya está hecho arriba en punto 2.

5. Seleccionar aleatoriamente los IJ sujetos de prueba, a partir de formar IN unidades experimentales. Se debe contar con $IJ = 12$ unidades experimentales, 3 por cada uno de los cuatro niveles de tratamiento. Si por cada uno de los M niveles de tratamiento se asignaran N unidades experimentales, en total se tendrían $MN = 17 \times 16 = 114 + 160 = 274$ unidades experimentales (cada una con su nivel de tratamiento ya asignado). De estas 274 unidades experimentales con su nivel de tratamiento ya asignado, se tienen que seleccionar solamente 12, tomando en cuenta que por cada nivel de tratamiento (que consiste de solamente cuatro valores), se tienen que asignar solamente tres sujetos de prueba. Nótese que por cada nivel de tratamiento los sujetos de prueba son diferentes y que, posiblemente, estos sujetos de prueba son diferentes de un nivel a otro.

6. Asignar aleatoriamente a cada uno de los IJ unidades experimentales uno de los I niveles de tratamiento. Esto ya se hizo en el paso 5 previo.

7. Seleccionar aleatoriamente cada una de las IJ unidades experimentales obtenidas a través de los pasos previos. Esto ya se hizo en el paso 5.

Por ejemplo, se podrían obtener las propuestas de unidades experimentales (con nivel de tratamiento incluido) que muestra Tabla 3.4 usando la instrucción MatLab siguiente y cuyos resultados son mostrados por Tabla 3.3,

[floor(16 * rand(4, 1)) + 1 floor(17 * rand(4, 3)) + 1]

Tabla 3.3: Diseño experimental considerando ϵ_s como factor variable.

2	9	1	10
5	2	16	5
6	5	13	8
7	14	9	17

La primer columna de Tabla 3.3 se refiere al índice de los niveles de tratamiento para el factor ϵ_s según Tabla 3.2, mientras que las columnas 2 hasta 4 especifican los índices de las unidades experimentales (sujetos de pruebas) que les corresponden según Tabla 3.2. Tabla 3.4 muestra el mapeo de estos índices hacia los correspondientes valores de los factores.

Tabla 3.4: Mapeo de índices en Tabla 3.3 hacia los correspondientes valores de los niveles de tratamiento de los factores de entrada y variable.

Nivel de factor ϵ_s	Sujeto de prueba (θ)		
+0.06	0.00	+0.05	+0.50
+0.09	−3.50	+3.50	−2.00
+0.10	−2.00	+2.00	−0.50
+0.11	+2.50	0.00	+4.00

La interpretación adecuada del contenido de Tabla 3.4 determina el desarrollo experimental. Por ejemplo, los primeros tres experimentos que se llevan a cabo, son aquellos en los que el radio de vecindad de selección es igual a +0.06 logits y junto con él se asocian tres diferentes sujetos de prueba (unidades experimentales) correspondientes a los valores $0.00, 0.01, 0.50$ logits.

Por supuesto, también se deben especificar los factores que se han fijado. Ellos corresponden a la cardinalidad de la vecindad de selección que, como se indicado líneas arriba, tiene un valor fijo en el conjunto $\{2,3,4\}$, otro al estimado inicial de habilidad de los sujetos de prueba y otros factores más asociados a la estructura del banco de ítems y el radio de la vecindad de prueba. Tabla 3.5 muestra los resultados de un experimento, junto con las condiciones o espacio de inferencia dentro del cual son válidos.

Este espacio de inferencia está determinado por un banco de ítems (dificultades) con 200 dificultades, una habilidad inicial igual a -1, un intervalo y subintervalo definidos por $(-4,4)$, una densidad de puntos igual a +0.99427, un índice de vecino igual a +1.0879 y su correspondiente índice de Gauss igual a +1.5136. La cardinalidad de la vecindad de selección es 4, mientras que el radio de vecindad de prueba es +0.01 logits.

El análisis de los datos y las conclusiones, se llevan a cabo dentro del Capítulo C. Los resultados son analizados bajo pruebas estadísticas con las que se verifica y valida el funcionamiento del simulador. Esta verificación y validación sustentan la respuesta que se da al problema de investigación.

Tabla 3.5: Resultados experimentales en donde se han fijado los factores Cardinalidad de banco de ítems = 200, habilidad inicial = -1, $(a,b) = (c,d) = (-4,4)$, densidad = +0.99427, Gauss = +1.5136, índice de vecino = +1.0879, $|B_{\epsilon_s}(\cdot)| = 4$, $\epsilon_p = +0.01$. Se aplica solamente una prueba por experimento.

ϵ_s	Sujeto	Prueba	ϵ_p	$\lvert B_{\epsilon_s}(\cdot)\rvert$	Iter	Tasa Conv.	θ_0	θ_e	θ_d	error
0.06	0.00	1	0.01	4	25	10.6228	-1	-0.244776	-0.293915	0.200751
0.06	0.05	1	0.01	4	31	0.0622047	-1	0.0912504	0.132466	0.451673
0.06	0.5	1	0.01	4	38	7.61313	-1	0.426396	0.445904	0.0457447
0.09	-3.5	1	0.01	4	23	2.60532	-1	-3.88339	-3.85458	0.0074182
0.09	3.5	1	0.01	4	28	-31.9235	-1	3.03181	3.08288	0.0168441
0.09	-2	1	0.01	4	24	0.998779	-1	-2.00115	-1.99167	0.00474031
0.1	-2	1	0.01	4	27	-0.244097	-1	-1.80972	-1.79944	0.00568085
0.1	2	1	0.01	4	19	1.13321	-1	1.95044	1.94951	0.000480007
0.1	-0.5	1	0.01	4	17	1.03302	-1	0.20854	0.200361	0.0392188
0.11	2.5	1	0.01	4	19	1.35758	-1	2.59558	2.5886	0.00269194
0.11	0	1	0.01	4	20	0.55657	-1	0.313688	0.327073	0.0426693
0.11	4	1	0.01	4	23	-1.54825	-1	4.17478	4	0.0418661

3.4.8. Validación y verificación de la simulación

Los conceptos de verificación y validación son muy importantes en el desarrollo de modelos, principalmente modelos implementados bajo una simulación. La verificación, según algunos autores, consiste en determinar que un programa de simulación por computadora se desempeña como se desea, lo que significa que está más relacionada con el proceso de depurado de las componentes software que definen al simulador [Law, 2003, Kleijnen, 1995]. Por otro lado, la validación, según los mismos autores, es un proceso con el que se determina si un modelo de simulación es una representación exacta del sistema real que se simula.

La validación del simulador es sin duda la componente más difícil de analizar. Sin embargo, se puede esbozar un procedimiento que parcialmente valide el funcionamiento del simulador. En primer lugar, los resultados de la simulación deben ser razonables. Si los resultados son consistentes con la forma en que se percibe que el sistema debe operar, entonces se obtiene lo que se llama validez aparente (*face validity*) [Law, 2003]. La verficación y validez del funcionamiento del simulador, serán parte de la evidencia sobre la cual se soportan los resultados que responden al problema planteado en esta investigación.

Para la verificación, se pueden analizar cada uno de los siguientes puntos, los cuales pueden ser útiles para verificar que el simulador está produciendo valores que coinciden, con cierto nivel de significancia, con valores estadísticos esperados (media, varianza, etc.). Se llevan a cabo pruebas estadísticas sobre la precisión con que se predice la habilidad real del sujeto de prueba.

Esta precisión está dada por el error relativo existente entre el estimado de habilidad y la habilidad correspondiente dentro del banco de ítems. En este sentido, se plantea la hipótesis nula de que la media de cada una de las muestras experimentales es la misma e igual a cero. La hipótesis alternativa es que al menos dos medias no son iguales. Las pruebas estadísticas que se consideran son las llamadas prueba χ^2 y prueba F.

También se aplican estas mismas pruebas estadísticas a las desviaciones sistemáticas existentes entre los estadísticos observados, como por ejemplo la habilidad, y su valor esperado que, para los experimentos llevados a cabo, se asume que es la habilidad real, factor que se asume conocido de antemano (aunque no por el simulador de evaluación que lo predice).

Por supuesto, se puede pensar en otras evidencias de verificación del funcionamiento del simulador, aunque buscarlas y encontrarlas se sale de los propósitos iniciales de este trabajo. Por ejemplo, en lugar de aplicar pruebas estadísticas sobre la media y la varianza, se puede probar la distribución completa de la variable aleatoria. En este sentido, se puede aplicar una prueba de bondad de ajuste tal como, por ejemplo, las pruebas χ^2 y Kolmogorov-Smirnov [Kleijnen, 1995].

Por otro lado, la prueba t de Student asume respuestas de simulación distribuidas en forma normal e independientes, con media y varianzas desconocidas. Para estimar esta varianza desconocida, se pueden particionar las corridas de simulación en subcorridas y calcular la salida o respuestas en cada una de ellas, el promedio de cada una de las subcorridas y el promedio de los promedios de cada una de las subcorridas (que es igual al promedio de la corrida completa, considerándola como compuesta de las subcorridas), lo que conduce a la prueba estadística t de Student o bien a la prueba Chi cuadrado o bien la prueba F [Kleijnen, 1995].

Finalmente, cabe agregar que la validación hace uso de datos reales, de pruebas estadísticas que comparan los datos simulados con los datos reales (pruebas t de Student, pruebas gráficas y pruebas de Schruben-Turing). También emplea procedimientos estadísticos (basados en análisis de regresión) para probar que las respuestas reales y simuladas se encuentran positivamente correlacionadas y que, posiblemente, tengan las mismas medias [Kleijnen, 1995].

El gran problema de la validación, para el tipo de modelo de simulación que aquí se propone, es que es muy difícil encontrar datos reales característicos de un proceso de evaluación. Es decir, datos reales en donde se muestre, paso a paso, el tipo de respuesta que proporciona el sujeto de prueba, su correspondiente estimado de habilidad, la correspondiente dificultad del ítem, etc. Sin embargo, en este trabajo se hace un intento de encontrar una validación con estas características, acudiendo a resultados parciales reportados por otros autores, y agregando algunas suposiciones [Magis and Raîche, 2012]. Como complemento a esto, se emplea

el Teorema 3.3.1 como evidencia para la validación exitosa del funcionamiento del simulador.

Finalmente, en posibles trabajos futuros, sobre todo cuando el simulador integre otras funcionalidades, se estaría en posibilidad de validar el funcionamiento del simulador acudiendo a la muy conocida prueba de Turing, tal y como lo sugiere el autor [Kleijnen, 1995].

CAPÍTULO

4 Resultados de simulación

El diseño experimental desarrollado a lo largo del Capítulo 3 se emplea para determinar los diferentes resultados que son de interés para este estudio. Inicialmente, se obtienen aquellos resultados que son de interés para la verificación del funcionamiento del simulador. Estos resultados incluyen áquellos mostrados por Tabla 3.5, en donde el factor variable es el radio ϵ_s de la vecindad de selección.

Posteriormente, se procede a la validación del funcionamiento del simulador propuesto. Para ello, se hace uso del resultado planteado por Teorema 3.3.1 y de datos aportados por otros autores, los cuales se consideran suficientemente completos para emplearlos en la prueba de validación.

Finalmente, se llevan a cabo los experimentos asociados a la respuesta que se da al problema planteado en este estudio, y que es en relación con los efectos que tienen sobre la tasa de exposición, o sobre el número de iteraciones necesarias para llegar al estimado final de habilidad, factores tales como los radios de vecindad de selección y de vecindad prueba.

4.1. Verificación

Una suposición razonable acerca del buen funcionamiento de las componentes software que definen al simulador es que, tomando como factor variable el radio ϵ_s de la vecindad de selección $B_{\epsilon_s}(\cdot)$ y manteniendo fijos a los otros factores, la distribución de las diferencias existentes entre el estimado de habilidad y áquella otorgada por el banco de ítems, debe ser de tal manera que las medias de población por nivel de tratamiento del factor variable, sean exactamente las mismas.

Lo anteriormente es ciertamente correcto cuando al proceso al que se somete cada sujeto de prueba en un experimento, se le otorga un suficiente número de iteraciones (basadas en la cardinalidad de la vecindad de selección) con las que es posible decidir que finalmente se ha llegado al valor de habilidad más preciso posible. Por el diseño experimental propuesto, la cardinalidad está entre los factores que se mantienen fijos, así que todo sujeto de prueba durante el proceso de prueba se encuentra bajo las mismas condiciones en este sentido.

En lo anterior, también juega un papel importante, por supuesto, la estructura del banco de ítems mismo. Sin embargo, esta se mantiene fija a lo largo de una experimentación. Pero es claro que una estructura con una amplia dispersión podría conducir también a que las medias de población no sean necesariamente las mismas.

Así, en este sentido se tiene la hipótesis nula de que las medias de las distribuciones de diferencias existentes entre el estimado de habilidad y áquella otorgada por el banco de ítems, debe ser de tal forma que las medias de población por nivel de tratamiento del factor variable, son exactamente las mismas. La hipótesis alternativa es que ellas no lo son y que, por lo menos, existen dos que no son iguales.

A lo anterior, debe agregarse que la verdad o falsedad de estas hipótesis depende de la estructura del banco de ítems y de la cardinalidad de la vecindad de selección. Por ello, primeramente se desarrollan dos experimentos orientados a confirmar o invalidar la hipótesis nula. Cada uno de estos experimentos se diferencia en la forma de la estructura del banco de ítems solamente.

4.1.1. Primer experimento

Para el primer experimento, se toman en consideración los resultados mostrados por Tabla 3.5, en donde se observa que existen cuatro niveles de tratamiento para el factor variable ϵ_s. Para verificar la validez de la hipótesis nula, se aplica la prueba estadística F, tal y como se muestra a continuación.

La media por cada nivel de tratamiento y la media total son dadas por Tabla 4.1. Los valores en esta tabla se emplean para determinar el valor F observado, tal y como lo muestra Tabla 4.2. Debe recordarse que la igualdad de medias que predica la prueba F surge del hecho de que se asume se cumple el modelo matemático representado por Ecuación 3.1.

En este modelo, la media es la media de cada nivel de tratamiento y los términos A_i, e_{ij} representan, respectivamente, intervariabilidad entre niveles de tratamiento e intravariabilidad en cada uno de los niveles.

Tabla 4.1: Medias parcial y total como datos para prueba F

Nivel factor ϵ_s	Media
0.06	0.2327229
0.09	0.00966753666666667
0.10	0.0151265523333333
0.11	0.02907578
Media total	0.07164819225

Tabla 4.2: Cálculos ANOVA para determinar el valor de F observado

Fuente	gl	SS	MS	F
A_i	3	0.104381307620823	0.034793769206941	3.23920229225608
$\epsilon_{j(i)}$	8	0.0859	0.0107	

La prueba estadística t-Student es útil para determinar significativamente que la media de la población tiene un valor igual a cero. Así, considerando ahora que la hipótesis nula es que la media de la población es cero, se tiene lo siguiente. Para aproximar la varianza de la población, se trabaja con la versión no sesgada de la varianza de la muestra, tal y como se describe en Apéndice B.

Considerando de nueva cuenta los resultados mostrados por Tabla 3.5, los cuales representan a la muestra experimental, se tiene entonces que $\sigma^2 = 0.0173$ para una muestra de tamaño $n = 12$. Así, la desviación estándar tiene un valor aproximado de $\sigma = 0.1315$. La media de la muestra tiene un valor $\hat{\mu} = 0.0716$, mientras que el número de grados de libertad es igual a $\nu = 11$. El valor observado de la variable t de Student es entonces el siguiente

$$\begin{aligned} t_{obs} &= \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma} \\ &= \sqrt{12}\cdot\frac{0.0716-0}{0.1315} \\ &= 1.8862 \end{aligned}$$

Con un nivel de significancia de $\alpha = 0.05$, una consulta a la tabla de la distribución t de Student correspondiente (11 grados de libertad) proporciona un valor crítico $t_{crit} = \pm 2.201$ (two tails test). Por lo tanto, t_{obs} cae fuera de la región crítica (dentro de la cual se rechaza la hipótesis nula). Así, la media de la población es igual a cero, dentro del espacio de inferencia especificado.

4.1.2. Segundo experimento

Considerando una nueva propuesta de unidades experimentales para ser específico, 4 unidades de tratamiento y 5 unidades experimentales, mediante el uso de la instrucción en Matlab [floor(16*rand(4,1))+1 floor(17*rand(4,5))+1] se obtiene la Tabla 4.3:

Tabla 4.3: Diseño experimental considerando ϵ_s como factor variable.

12	11	16	12	15	2
13	2	5	4	16	4
5	2	9	8	9	14
11	8	3	11	2	4

Tabla 4.4: Intervalo (-4,4), subrango(-1,1) No. de items 50 índice de... 0.29888, Desviación:-5.9475 e índice de densidad:0.2476

ϵ_s	Sujeto	Prueba	ϵ_p	$\lvert B_{\epsilon_s}(\cdot)\rvert$	Inter	Tasa Conver.	θ_0	θ_e	θ_d	error
0.09	-3.5	1	0.01	4	20	-1.89809	-1	-3.8696	-4	0.0336888
0.09	0.0	1	0.01	4	20	0.744879	-1	0.539107	0.531034	0.0149752
0.09	-0.5	1	0.01	4	18	0.92186	-1	-0.737843	-0.746026	0.0110905
0.09	0.0	1	0.01	4	28	0.891051	-1	0.31885	0.310956	0.024757
0.09	+2.5	1	0.01	4	30	0.950831	-1	1.90363	0.941186	0.617008
0.15	-0.5	1	0.01	4	13	0.886477	-1	-0.737223	-0.746026	0.0119416
0.15	-3.0	1	0.01	4	19	0.924183	-1	-2.87885	-4	0.389442
0.15	+1.0	1	0.01	4	21	0.337701	-1	0.688336	0.698259	0.0144157
0.15	-3.5	1	0.01	4	24	0.0952724	-1	-3.50683	-4	0.140631
0.15	-2.5	1	0.01	4	18	0.918147	-1	-2.34512	-0.936334	0.600731
0.16	+1.0	1	0.01	4	19	0.353424	-1	1.18839	0.941186	0.208015
0.16	+3.5	1	0.01	4	17	-1.86536	-1	3.83208	4	0.0438204
0.16	+1.5	1	0.01	4	20	-10.5267	-1	1.45336	0.941186	0.352406
0.16	+3.0	1	0.01	4	20	0.916639	-1	3.31018	4	0.208395
0.16	-3.5	1	0.01	4	18	0.917763	-1	-2.92111	-4	0.369344
0.17	-3.5	1	0.01	4	15	-5.24414	-1	-3.597	-4	0.112939
0.17	-2.0	1	0.01	4	16	-7.4158	-1	-1.41413	-0.936334	0.337873
0.17	+2.5	1	0.01	4	9	18.1534	-1	2.38433	0.941186	0.605262
0.17	+3.5	1	0.01	4	18	-10.9016	-1	3.40804	4	0.173694
0.17	-2.5	1	0.01	4	17	0.909116	-1	-3.18972	-4	0.254027

4.2. Tasa de exposición

En el ámbito de la evaluación adaptable por computadora, el estudio de la tasa de exposición se ha desarrollado con el objetivo de mantener la seguridad

de los bancos de reactivos. Las investigaciones en este campo se han realizado con el mecanismo de selección de preguntas más frecuentemente empleado: el que busca entre los reactivos no presentados en un banco aquel que máximiza la función de información de Fisher para el nivel de rasgo estimado en ese momento.

¿Cómo se mide la tasa de exposición?

La regla de selección de pregunta que se pretende utilizar se basa únicamente en un valor puntual.

$$r_i = \frac{m_i}{p} = \frac{\text{N° de veces que se ha usado el reactivo}}{\text{N° de evaluaciones que se aplicó}}$$

4.3. Validación

La validación del simulador se ha hecho en relación con el comportamiento de convergencia de la habilidad previa, aunque algunas otras opciones, o pruebas complementarias, pueden ser utilizadas. Por ejemplo, el Teorema 3.3.1 proporciona algunas herramientas útiles en este sentido dicho, y demostrado por otros autores en este campo [Chang and Ying, 2009].

La figura 4.1 ilustra la convergencia de la habilidad estimada para el valor real de la habilidad de algunas condiciones de prueba. Particularmente, la estimación se prevé con una precisión de 0.25 % en error relativo. Puede observarse que las habilidades estimadas, y los valores de las dificultades dadas por el banco de ítems, se acumulan en torno a un vecindad bidimensional definido por estas variables.

Por otro lado, la Figura4.2 ilustra la forma en que el radio de la vecindad de selección afecta el número de iteraciones para llegar al valor verdadero de habilidad. Las mismas condiciones experimentales se mantienen para cada experimento, representado por el número de iteraciones en función del radio de vecindad.

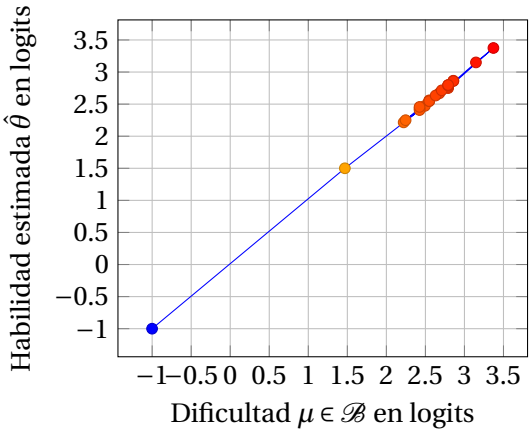


Figura 4.1: Habilidad estimada $\hat{\theta}$ (como lo predijo la simulación) versus selección de las dificultad del ítem $\mu \in \mathscr{B}$. Las condiciones de las predicciones asumen una habilidad real desconocida con un valor de 3 logits y una habilidad inicial con un valor de -1 logits. La habilidad estimada, después de 23 iteraciones (ítems presentados ó, equivalencias, en test con 23 ítems), con un valor de 2.79809 logits, mientras que la última dificultad del ítem seleccionado (elemento de la base de datos $\mathscr{B}$) tiene un valor de 2.79107 logits, con un valor relativo del 0.25 %. Por otra parte la estructura del del banco de ítems y las condiciones de administración del ítem son, respectivamente, definidas por un índice de vecindad de 0.94633 y un índice de densidad de 0.98772, una vecindad de pseudo Cauchy de 0.1 con cardinalidad 5 y una prueba de 0.01. La distribución de dificultades en el banco de ítems es una característica intrínseca aceptable y no es debido a los efectos aleatorios como se indica por el valor del del índice Gaussiano de -1.134. El banco de ítems tiene 300 dificultades y el intervalo y subintervalo están dados por $(-4,+4)$.

4.3.1. Validación por Teorema en referencia [Chang and Ying, 2009]

Dados los datos originales X arrojados por el simulador, existen cuatro histogramas que comparar; a saber,

1. En la figura 4.3, es el histograma donde se muestran resultados de subexposición y sobre exposicion de los datos experimentales de la frecuencia de uso de ítems del banco, considerando ϵ_s variable, y los siguientes datos: ϵ_p, $B_{\epsilon_p}(\theta^*)$, habilidades reales $\{-2,0,2\}$, θ_0 y 10 pruebas, constantes.

2. Las figuras: 4.4 y 4.5 muestran resultados de subexposición y sobre exposicion de los datos experimentales de la frecuencia de uso de ítems del ban-

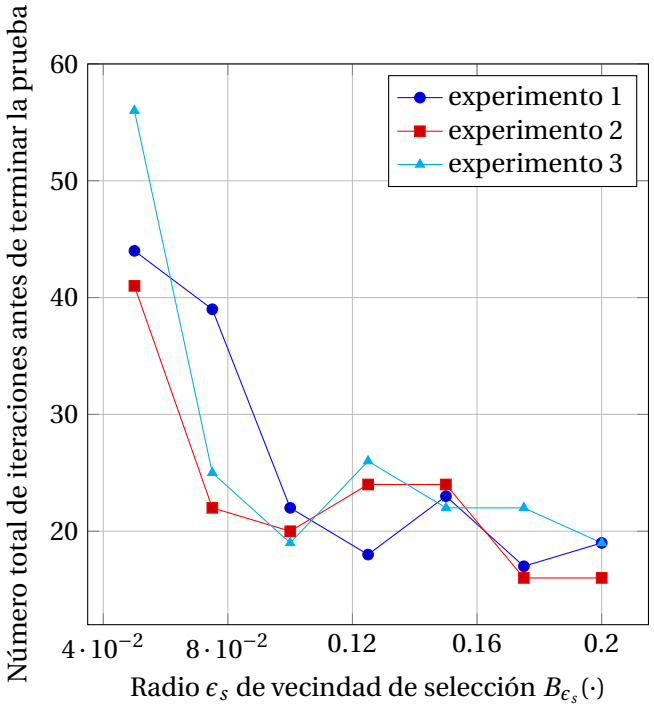


Figura 4.2: Resultados de la simulación para el número total de iteraciones antes de terminar la preueba frente al radio ϵ_s de vecindad de selección $B_{\epsilon_s}(\cdot)$.Cada punto en un experimento representa una prueba bajo las condiciones especificadas de simulación. Para radios pequeños ϵ_s, por tanto, más alto es el número de iteraciones y, por lo tanto, menor la posibilidad de utilizar más a menudo las deficultades en el banco de ítem. Por otro lado, para un radio mayor ϵ_s, menor es el número de iteraciones y, por lo tanto, menor la posibilidad de utilizar más a menudo las dificultades en el banco del ítem. Las condiciones experimentales definen un banco con 100 dificultades, con un índice de vecino más cercano de 0.8979, estructura aleatoria de -1.2363, índice densidad de 0.9856, vecindad de prueba de 0.01, cardinalidad de Cauchy de 5, habilidad inicial de -1 logits y una habilidad real supuesta de $+1$ logits. Los puntos máximo y mínimo del intervalo y subintervalo de dificultades son $(-4,+4)$ logits.

co, considerando ϵ_p variable, y los siguientes datos: ϵ_s, $B_{\epsilon_p}(\theta^*)$, habilidades reales $\{-2,0,2\}$, θ_0 y 10 pruebas, constantes.

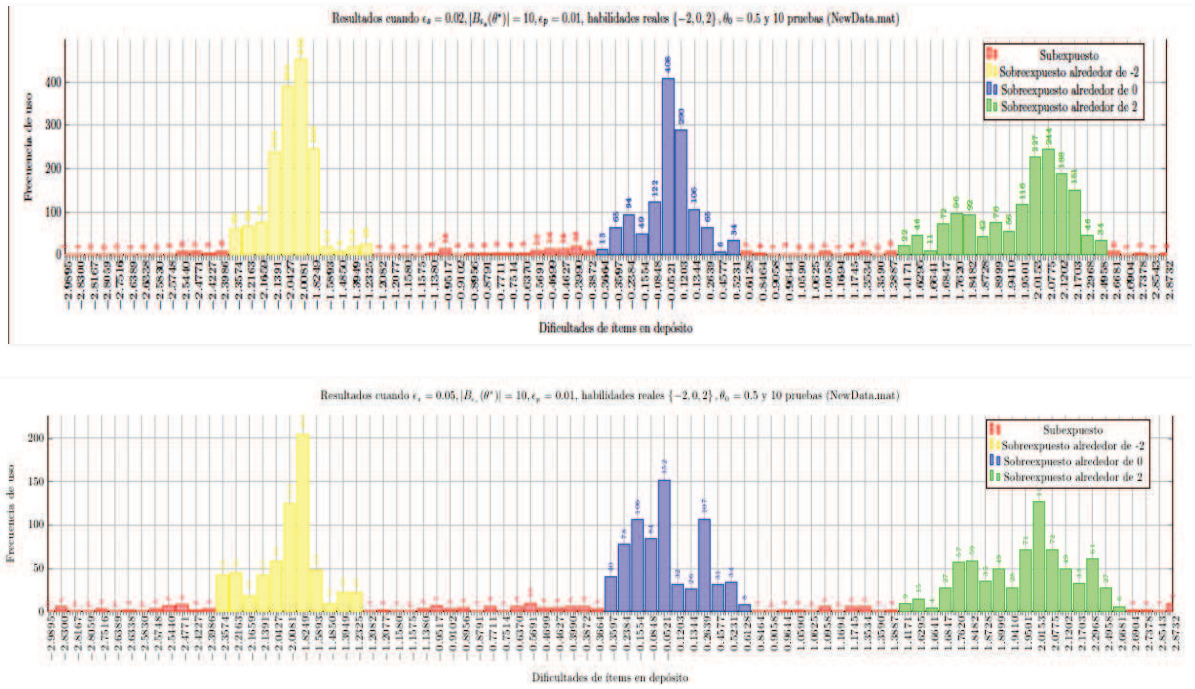


Figura 4.3: Comportamiento de uso de los ítems en relación con la dificultad de los ítems en el banco

El punto importante a recordar que las dos primeras distribuciones, cuando se aplica en ellas la prueba de bondad de ajuste χ^2, no proporcionan una aceptación de la hipótesis nula, la cual establece que los datos se distribuyen normalmente con media cero y desviación estándar uno.

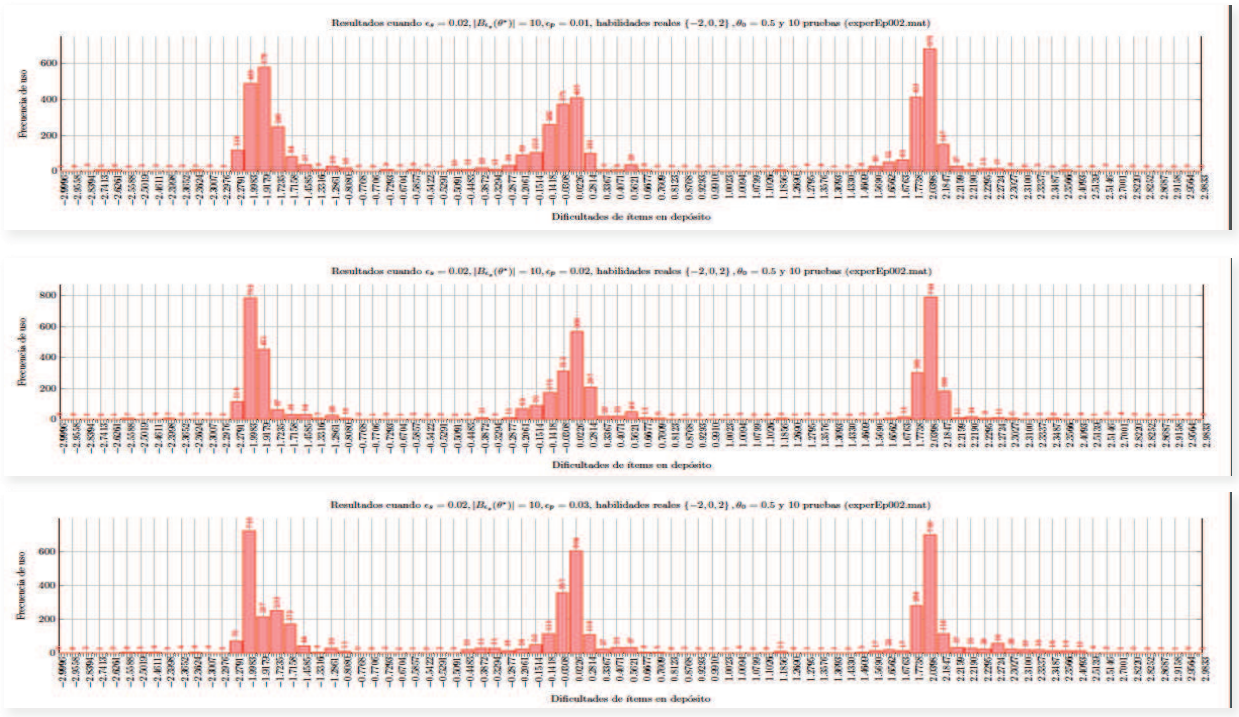


Figura 4.4: Valores experimentales del comportamiento de la subexposición y sobrexposición de los ítems en el banco, considerando ϵ_p variable y los demás datos constantes

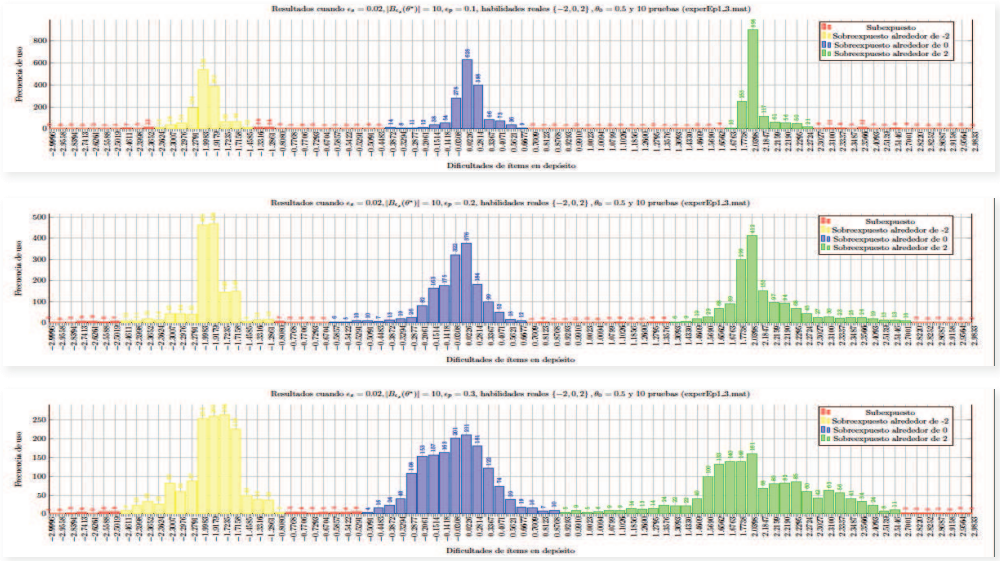


Figura 4.5: Valores experimentales del comportamiento de la subexposición y sobrexposición de los ítems en el banco, considerando ϵ_p variable y los demás datos constantes

CAPÍTULO

5 Conclusiones y trabajo futuro

5.1. Conclusiones

Según los resultados obtenidos de acuerdo al comportamiento en la administración del depósito de ítems, cuando el radio de vecindad de seleccin y la cardinalidad de esta vecindad se incluyen como variables de interś, podemos observar que este comportamiento es muy interesante. El control de estas variables tiene un impacto sobre los valores de la tasa de exposiciń del ítem. Hay un comportamiento regulado de la tasa de exposición del ítem en función del radio de vecindad de la selección. La tasa de exposición del ítem es inversamente proporcional al radio. Radio pequeño tiende a proporcionar valores grandes en el número total de iteraciones antes de terminar una evaluación. La cardinalidad de la vecindad de selección tiene también un fuerte impacto en la exposición del ítem. Un valor alto de cardinalidad aumenta el número de iteraciones antes de terminar una prueba. El radio y la cardinalidad de vecindad de selección representan, de cierta manera y metafóricamente hablando, la severidad del instructor. La cardinalidad de la vecindad de selección representa el número de elementos secuenciales que el evaluado respondió de manera correcta, o incorrecto, respuestas dadas antes de que el instructor decide terminar la prueba y, por otro lado, el radio significa la tolerancia que el instructor asigna para que la habilidad estimada representar el valor de habilidad real. En la etapa actual de la investigación presentada en este documento, la estadística de orden parece ser la herramienta natural para describir la estructura de un banco de ítems. La estructura del Banco de ítems está dado principalmente por la distribución de las dificultades y, de forma natural, la estadística de orden introduce el concepto de densidad del ítem, lo que se refiere al número de ítems por dificultad. Por otro lado, todavía hay algunas otras preguntas muy

importantes de la investigación que necesitan ser tratadas en la obra. Labor futura que debe realizarse para estudiar a la relación entre la estructura de un banco de ítems, representado por el índice de vecino más cercano, la varianza aleatoria estándar del índice de curva normal y el índice de densidad y la precisión de la habilidad estimada y la tasa de exposición del ítem. El problema de control de exposición del ítem en términos de teoría de colas o acciones de control necesita ser resuelto.

Además, la descripción o análisis de la estructura de un banco real de ítems a través de los distintos índices ya discutido, se ha iniciado el desarrollo de un verdadero sistema informático de evaluación adaptable llamado Ariya [Alcántara J, 2010, Barranco JA, 2010, T.Canales, 2012, Cano G, 2011]. El sistema está actualmente orientado a evaluar la habilidad de los evaluados para entender diferentes conceptos físicos dentro del tema de cinemática y el subtema de movimiento uniformemente acelerado. Este desarrollo se utilizará para discutir otro punto importante preocupados por la forma en que diferentes técnicas de tasa de control de exposición (Randomesque, Simpson-Hetter, etc.[Barrada JR. J Olea. V Ponsoda, 2006]) se comportan como una función de la estructura del Banco de ítems, junto con sus relaciones a las características de la vecindad $B_{\epsilon_s}(\cdot)$.

5.2. Trabajo futuro

1. Enriquecer el simulador, incluyéndole modelos psicométricos más complejos (2PL, 3PL y 4PL).
2. Incluir en el simulador otros métodos de selección, como por ejemplo Kullback-Leibler, etc.
3. Incluir en el simulador aspectos tales como la severidad del instructor, ítems multivaluados o modelos de crédito partial, , etc.
4. Considerar otras pruebas estadísticas como, por ejemplo, el índice de Ripley.
5. Análisis de tasas de exposición.
6. Considerar la administración de ítems como un problema de Teoría de Colas aplicado a control de inventarios.
7. Incluir opciones de estimación de habilidad diferentes al concepto de máxima verosimilitud que aquí se emplea (por ejemplo EAP o Expectation a posteriori).
8. Colectar datos reales, de tal forma que se puede llevar a cabo una validación más profunda del funcionamiento del simulador.

Bibliografía

[Alcántara J, 2010] Alcántara J, R. G. (2010). Esquema de seguridad para un ambiente de evaluación adaptable. Tesis de licenciatura en sistemas computacionales, Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo.

[Antony, 2003] Antony, J. (2003). *Design of Experiments for Engineers and Scientists*. Elsevier Science & Technology Books.

[Barhum et al., 2007] Barhum, K., Goldreich, O., and Shraibman, A. (2007). *On approximating the average distance between points*. Technical report partially supported by the Israel Science Foundation (grant no. 460/05), Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

[Barrada JR, 2004] Barrada JR, J Olea, V. P. (2004). Reglas de selección de ítems en tests adaptativos informatizados.*Metodología de las Ciencias del Comportamiento, Suplemento* 2004; volumen especial: 55-61. Technical report, Universidad Autónoma de Madrid.

[Barrada JR. J Olea. V Ponsoda, 2006] Barrada JR. J Olea. V Ponsoda, F. A. (2006). Incorporating randomness in the fisher information for improving item exposure control in cats. Reporte de investigación., Departamento de Psicología Social y Metodología, Facultad de Psicología, Universidad Autónoma de Madrid.

[Barranco JA, 2010] Barranco JA, R. L. (2010). Análisis, diseño e implementación de una base de datos para la administración de reactivos en un evaluador educativo. Tesis de licenciatura en sistemas computacionales, Centro de Investi-

gación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo.

[Benko, 2001] Benko, M. (2001). Testing the equality of means and variances across populations and implementation in xplore. Bsc. degree in statistic thesis, Wirtschaftwissenschaftliche Fakultat, Humboldt Universitat zu Berlin.

[Burgstaller and Pillichshammer, 2009] Burgstaller, B. and Pillichshammer, F. (2009). The average distance between two points. Bull. Austral. Math. Soc., 80(3):353–359.

[Burston and Burston, 1995] Burston, J. and Burston, M. M. (1995). Practical design and implementation considerations of a computer adaptive foreign language test: The monash/melbourne french cat. *The Computer Assisted Language Instruction Consortium Journal*, 13(1):26–46.

[Cano G, 2011] Cano G, M. P. (2011). Tecnología java y mysql como alternativa para la creación de ítems multimedia, dentro de un sistema de evaluación en línea. Tesis de licenciatura en sistemas computacionales, Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo.

[Chang and Ying, 2009] Chang, H. H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37(3):1466–1488.

[Cheng and Liou, 2002] Cheng, P. E. and Liou, M. (2002). Computerized adaptive testing using the nearest-neighbors criterion. Research Report supported by grant from the National Science Council, ROC, Institute of Statistical Science, Academia Sinica, ROC, Taipei 115, Taiwan.

[Choi, 2008] Choi, S. W. (circa 2008). Firestar: Computerized adaptive testing (cat) simulation program for polytomous irt models. Research Report Patient Outcomes Measurement Information System (PROMIS) U-01 AR 052177-04, Northwestern University Feinberg School of Medicine and National Institute of Health (NIH).

[Clark and Evans, 1954] Clark, P. J. and Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453.

[Frey and Jugulum, 2006] Frey, D. D. and Jugulum, R. (2006). The mechanisms by which adaptive one-factor-at-a-time experimentation leads to improvement. *Transactions of the ASME*, 128:1050–1060.

[Glas, 2005] Glas, C. A. W. (2005). The impact of item parameter estimation on computerized adaptive testing with item cloning. Computerized Testing Report 02-06, LSAC Research Report Series, University of Twente, Enschede, The Netherlands, Law School Admission Council.

[Horn, 2012] Horn, R. A. (2012). *Understanding the one way anova*. Northern Arizona University, Educational Psychology Department, Eastburn Education Building – Room 206J.

[Illowsky and Dean, 2008] Illowsky, B. and Dean, S. (2008). F distribution and anova. In *Collaborative Statistics*, chapter 13, pages 511–530. Connexions, Rice University, Houston, Texas.

[Kleijnen, 1995] Kleijnen, J. P. C. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82:145–162.

[Korenzen and Anderson, 1993] Korenzen, T. J. L. and Anderson, V. L. (1993). *Design of Experiments, A No-Name Approach*. Statistics: Textbooks and Monographs. Marcel Dekker, New York, USA.

[Lavery, 2004] Lavery, R. (2004). An animated guide: The logic of hypothesis testing and anova. In *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*, pages Paper 192–29.

[Law, 2003] Law, A. M. (2003). How to conduct a successful simulation study. In Chick, S., Sánchez, P. J., Ferrin, D., and Morrice, J., editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 66–70.

[Levine, 2004] Levine, N. (2004). *CrimeStat III Version 3.0, A Spatial Program for the Analysis of Crime Incident Locations, Operation Manual*. The National Institute of Justice, Washington, DC, USA.

[Loureiro and García, 2007] Loureiro, E. F. and García, P. (*c*irca 2007). Métodos estadísticos y valor p (p – Valor): Historia de una controversia. Notas dentro de la Universidad de Buenos Aires, Facultad de Ciencias Económicas, Instituto de Investigaciones en Administración, Contabilidad y Matemática, Sección de Investigaciones en Matemática (Estadística y Econometría).

[Magis and Raîche, 2012] Magis, D. and Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the r package catr. *Journal of Statistical Software*, 48(8):1–31.

[Olea and Ponsoda, 2003] Olea, J. and Ponsoda, V. (2003). *Tests Adaptativos Informatizados*. Aula Abierta. Ediciones Universidad Nacional de Educación a Distancia, San Romualdo, Madrid, España, first edition.

[Patelis, 2000] Patelis, T. (2000). An overview of computer-based testing. Research Notes RN–09, College Board Office of Research and Development.

[Raîche and Blais, 2005] Raîche, G. and Blais, J.-G. (2005). *SIMCAT 1.0 - A SAS Computer Program for Simulating Computer Adaptive Testing User's Guide*. Université du Québec à Montréal.

[T.Canales, 2012] T.Canales, C. (2012). *Reingeniería del Modelo de Datos a un Sistema de Evaluación por Computadora.* Reporte técnico de proyecto de estadía dentro del programa de ingeniería en software., Universidad Politécnica de Pachuca.

[Vokey and Allen, 2011] Vokey, J. R. and Allen, S. W. (2011). The general logic of anova. In *Thinking with Data*, chapter 15, pages 149–156. PsΨence Ink Book.

[Weingart and Selvin, 1995] Weingart, M. and Selvin, S. (1995). Nearest neighbor analysis in one dimension. Technical Report LBL-36888 UC-605, Information and Computing Sciences Division, Lawrence Berkeley Laboratory and Division of Biostatistics, School of Public Health, University of California, Berkeley, California 94720, USA.

APÉNDICE

A Densidad de puntos

La propiedad de densidad de dificultades en el depósito de ítems está caracterizada por índices de densidad, uno de los cuales toma en cuenta el concepto de distancia promedio entre puntos en un intervalo. Este problema consiste en asumir que se cuenta con un intervalo real cerrado $[a,b]$ y lo que se desea es determinar para cada punto $x \in [a,b]$ la distancia promedio de este punto con relación a los puntos $y \in [a,b]$ [Barhum et al., 2007, Burgstaller and Pillichshammer, 2009].

La justificación fundamental del porqué se quiere resolver este problema, consiste en construir una medida para el grado de densidad de un conjunto de puntos discretos dentro del intervalo $[a,b]$. En relación con la distribución de dificultades en el depósito de ítems, este índice será una referencia acerca de la bondad del depósito, en sustitución del número de ítems en él.

Al contar con una medida de la distancia promedio entre puntos del intervalo $[a,b]$, que se podría decir representa la situación ideal de contar con un depósito de ítems con una distribución infinita de valores de dificultad, es posible comparar con ella la distancia promedio entre los puntos discretos en el intervalo , construyendo así una medida de la densidad de los puntos discretos y, por lo tanto, la bondad del depósito de ítems.

Así, supóngase que $x, y \in [a,b]$. La distancia entre estos dos puntos está dada por el valor aboluto de su diferencia; es decir, $|x-y|$. Para el caso particular del punto x fijo, el promedio de sus distancias a los puntos $y \in [a,b]$ es dada por la expresión siguiente, asumiendo una distribución uniforme en las mismas

$$\int_a^b f(y)|x-y|dy$$

donde

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq y \leq b \\ 0 & \text{en caso contrario} \end{cases}$$

es la función de distribución uniforme sobre el intervalo $[a,b]$.

Por lo tanto el valor promedio de las distancias del punto x a cada uno de los puntos y dentro del intervalo $[a,b]$ es el siguiente

$$\begin{aligned} \frac{1}{b-a}\int_a^x (x-y)\,dy - \frac{1}{b-a}\int_x^b (x-y)\,dy &= \frac{1}{b-a}\left[x(x-a) - \frac{1}{2}(x^2-a^2) - x(b-x)\right. \\ &\quad \left. + \frac{1}{2}(b^2-x^2)\right] \\ &= \frac{1}{b-a}\left[x^2 - (a+b)x + \frac{1}{2}(a^2+b^2)\right] \end{aligned} \quad \text{(A.1)}$$

Por ejemplo, para $x = a$ o $x = b$, el promedio de las distancias desde x a cada uno de los puntos en el intervalo $[a,b]$ es simplemente $\frac{1}{2}(b-a)$ (basta sustituir en Ecuación (A.1) el valor de x correspondiente); mientras que el promedio de las distancias del punto medio del intervalo, $x = \frac{b+a}{2}$, es sencillamente $\frac{b-a}{4}$.

Sin embargo, la cuestión es que interesa conocer el promedio de las distancias entre cada uno de los puntos del intervalo $[a,b]$, promedio que por supuesto no debe depender de algún punto específico dentro de este intervalo. Este valor se obtiene directamente de la Ecuación (A.1) al integrar de la siguiente forma

$$\begin{aligned} \frac{1}{b-a}\int_a^b f(x)\left(x^2 - (a+b)x + \frac{1}{2}(a^2+b^2)\right)dx &= \frac{1}{(b-a)^2}\left[\frac{1}{3}(b^3-a^3)\right. \\ &\quad - \frac{1}{2}(a+b)(b^2-a^2) \\ &\quad \left. + \frac{1}{2}(a^2+b^2)(b-a)\right] \\ &= \frac{1}{3}(b-a) \end{aligned} \quad \text{(A.2)}$$

Por consiguiente, un depósito de ítems ideal debe estar definido por una distribución de dificultades cuya densidad tenga el valor

$$\frac{1}{3}(\mu_{max} - \mu_{min})$$

en donde μ_{max} y μ_{min} son los valores de dificultad máximo y mínimo en el depósito.

Es de suponer que si el promedio de distancias entre cada una de las dificultades es aproximadamente igual a la densidad ideal dada por Ecuación (A.2), entonces las dificultades se encuentran demasiado agrupadas, situación que debe ser altamente preferible ya que con ello se debe garantizar una mayor precisión en el valor del estimado de habilidad.

A.1. Otra versión

Considérese el experimento de seleccionar aleatoriamente y de manera uniforme n puntos $x_1, x_2, \ldots, x_n$ del intervalo (a, b). Estos n puntos son posteriormente ordenados en forma creciente $x_{i_1}, x_{i_2}, \ldots, x_{i_n}, i_k \in \{1, 2, \ldots, n\}$, donde x_{i_k} representa un valor posible de la variable aleatoria X_k. La cuestión consiste en determinar algunos de los estadísticos importantes de cada una de las n variables aleatorias X_k como, por ejemplo, la media y la varianza.

Por ejemplo, considérese el caso del intervalo $(-4, 3)$ y que solamente se consideran dos variables aleatorias X_1, X_2. Se puede demostrar que el valor medio de X_1 es $-\frac{5}{3}$. Para probar experimentalmente que esto es así, basta seleccionar aleatoriamente dos puntos del intervalo $[-4, 3]$, ordenarlos en forma creciente y repetir todo este proceso tantas veces como se desee, anotando en cada caso los pares de valores. La media de la primera columna de los valores así obtenidos debe aproximarse al valor $-\frac{5}{3}$. Este caso se conoce como estadística de orden 2.

A continuación se analiza la estadística de orden n, en la que se toman n variables aleatorias con $n \geq 2$. El análisis se lleva a cabo sobre un intervalo general (a, b) siguiendo una metodología similar a la que se emplea en referencia [Weingart and Selvin, 1995], que solamente consideró el intervalo unitario $(0, 1)$.

Entonces, en lo que sigue, se asume que $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas (i.i.d.), cada una con densidad $f(x)$ (función densidad de probabilidad) y distribución $F(x)$ (función acumulativa de distribución). Como se sabe, la relación existente entre $f(x)$ y $F(x)$ es

$$f(x) = \frac{d}{dx} F(x)$$

Definición A.1.1. (Estadísticas de orden) Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas, cada una con densidad $f(x)$ y distribución $F(x)$. Supóngase que $x_1, x_2, \ldots, x_n$ son los valores adquiridos durante la realización de un experimento $\mathscr{E}$ y que estos valores se ordenan en forma creciente $x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}$, donde $(\pi(1), \pi(2), \ldots, \pi(n))$ es la permutación que permite este ordenamiento particular. Se dice entonces que $x_{\pi(k)}$ es el valor que adquiere

la variable aleatoria X_k durante la realización del experimento $\mathscr{E}$ y el conjunto de valores $x_1, x_2, \ldots, x_n$ se denomina estadísticas de orden de las $X_k's$. □

Nótese que la permutación π que permite el ordenamiento no es única, depende del conjunto de valores particular a ordenar. De hecho existen $n!$ posibles permutaciones, una de las cuales se aplica para un caso específico de valores experimentales. Dada una permutación fija π, existen infinitos resultados experimentales a los que se les puede aplicar, para obtener un subconjunto de resultados experimentales posibles de la estadística de orden de las X_k. Los resultados adicionales posibles son dados por las permutaciones restantes.

Dado un número real arbitrario x, la probabilidad de que el valor de X_n, sea menor que x es igual a la probabilidad de que $X_i \leq x \forall i$, lo cual se expresa como sigue

$$\prod_{i=1}^{n} P\{X_i \leq x\}$$

por cuestiones de independencia. Pero como $P\{X_i \leq x\}$ es la función de distribución $F(x) \forall i$ se cumple

$$F_{Y_n} = [F(x)]^n$$

y

$$f_{Y_n} = n\,[F(x)]^{n-1} f(x)$$

donde se conviene en identificar como Y_k la variable aleatoria que asocia al orden $k-$ésimo más pequeño de la secuencia de variables original $X_k's$.

En lugar de repetir nuevamente todos los pasos que se han hecho para el caso del intervalo $(0,1)$, es más conveniente definir una variable nueva en términos de la variable x que toma valores en el intervalo (a,b). Partiendo del hecho de que

$$a \leq x \leq b$$

se obtiene

$$0 \leq \frac{x-a}{b-a} \leq 1$$

y así, el cambio de variable

$$y = \frac{x-b}{b-a}$$

coloca a y en el intervalo $(0,1)$.

El valor esperado de X_k en la estadística de orden es

$$E(Y_k) = \frac{k}{n+1}$$

por lo que el valor esperado en el intervalo (a,b) es simplemente

$$E(Y_k) = a + \frac{k}{n+1}(b-a) \forall k = 1,2,\ldots,n$$

Similarmente, la varianza de X_k en la estadística de orden es

$$\sigma^2(Y_k) = \frac{1}{n+2}\frac{i}{n+1}\left(1 - \frac{i}{n+1}\right)(b-a)^2 \forall k = 1,2,\ldots,n$$

En el análisis de la distribución de las dificultades en el depósito de ítems, es importante considerar el valor promedio de las distancias entre cada valor de dificultad. Cuando se considera al vecino más cercano, este promedio está dado por la siguiente ecuación si se asume que se tienen n valores de dificultad en el intervalo $[a,b]$

$$\hat{d} = \frac{1}{n}\left[abs(x_2 - x_1) + \sum_{i=2}^{n-1} min(x_i - x_{i-1}, x_{i+1} - x_i) + abs(x_n - x_{n-1}) \right]$$

El valor promedio de esta distancia promedio, para el caso en que el intervalo es $(0,1)$ es

$$\frac{n+2}{2n(n+1)}$$

por lo que la ecuación que da el promedio de las distancias promedios en el intervalo (a,b) es simplemente

$$\langle \hat{d} \rangle = \frac{n+2}{2n(n+1)}(b-a) \forall n \geq 2$$

o bien

$$\langle \hat{d} \rangle = 0.5\left(1 + \frac{1}{n+1}\right)\frac{1}{\rho} \forall n \geq 2$$

donde ρ indica la densidad de puntos por unidad de longitud que, en el caso de dificultades de ítems en el depósito, representa el número de ítems por cada dificultad en el mismo.

Nótese que para $n \gg 1$ se tiene que

$$\langle \hat{d} \rangle \approx 0.5\frac{1}{\rho}$$

ecuación en la que se expresa claramente que la distancia promedio entre puntos depende en forma inversamente proporcional a la densidad, la que a su vez

depende inversamente de la longitud del intervalo que contiene a los puntos y directamente del número de puntos en el mismo.

Cualquiera de estas dos últimas ecuaciones será de utilidad para describir el grado de agrupamiento de las dificultades en el depósito de ítems. Con ello, se podrá verificar si se encuentran en grupos, con distribución regular o completamente dispersas.

Otro estadístico que será de utilidad en el análisis del comportamiento del depósito de ítems es la varianza de la distancia promedio $\hat{d}$. Esta varianza es dada por la ecuación

$$\sigma^2(\hat{d}) = \frac{1}{n^2}\frac{2n^2+17n+12}{12(n+1)^2(n+2)}(b-a)^2 \approx \frac{0.1667}{n\rho^2}, n \gg 1$$

Las pruebas estadísticas para determinar el grado de agrupamiento de los puntos en el intervalo (a,b) se basan en el índice de vecino más cercano η y en el índice de variación estándar de la curva normal λ, los cuales se definen, respectivamente, como sigue [Clark and Evans, 1954]

$$\eta = \frac{\hat{d}}{\langle \hat{d} \rangle} \approx 2\rho\hat{d}$$

y

$$c = \frac{\hat{d} - \langle \hat{d} \rangle}{\sigma(\hat{d})} \approx 3n\rho(\eta - 1), n \gg 1$$

Si para una observación dada resulta que $\eta < 1$, entonces significa que se tiene un agrupamiento significativo, mientras que si $\eta > 1$, entonces se tiene una dispersión de los puntos en el intervalo. Para el caso $r = 1$, la distribución se puede considerar completamente obtenida por azar.

Sin embargo, todavía puede quedar la duda de que la observación obtenida haya sido lograda por azar (bien sea que $\eta \gg 1$ o $\eta << 1$). En este caso, se busca entonces un procedimiento que permita determinar si la separación que tiene η de 1 no es casual (es decir, por azar), sino que se debe a que ésta es realmente la estructura que tienen los datos. La prueba estadística c resuelve este problema de la siguiente manera.

Si resulta que

$$|c \pm 2.27| \leq 0.31$$

entonces la probabilidad de que las observaciones se desvíen aleatoriamente del hecho de que sucedan al azar es menor al 5 %. Es decir, el agrupamiento es consistente con una probabilidad del 95 % de que esto no haya ocurrido al azar. Por otro

lado, si

$$2.58 < |c|$$

entonces la probabilidad de que las observaciones se desvíen aleatoriamente del hecho de que ocurran al azar es menor al 1 %. Es decir, el agrupamiento es consistente con una probabilidad del 99 % de que esto no haya ocurrido al azar.

APÉNDICE

B Prueba estadística F

La prueba estadística F se basa en la distribución que sigue una variable aleatoria F definida en términos de varianzas intramuestras e intermuestras. Esta variable aleatoria tiene una distribución que depende de lo que se conoce como grados de libertad de las muestras intra e inter. Estos grados de libertad dependen del tamaño de las muestras.

Conocidos los grados de libertad y el valor de la variable F, es posible construir una prueba estadística que, usando también el concepto de $p-$Valor, permita determinar la probabilidad de desechar una hipótesis nula previamente establecida en términos de las medias de cada uno de los grupos de muestras, implicando con ello una decisión incorrecta.

Todo lo anterior resulta importante, ya que conduce directamente al concepto de Análisis de Variabilidad o Análisis de Varianza (ANOVA por ANalysis Of VAriability o ANalysis Of VAriance), método con el cual es posible determinar si un conjunto de muestras proceden de una misma distribución. Esto es útil para llevar a cabo pruebas de hipótesis.

B.1. Introducción

Supóngase que se tienen una serie de muestreos aleatorios con distribución normal cuya media es μ y varianza σ^2. Si estos muestreos representan a las variables aleatorias $X_1, X_2, \ldots, X_k$, entonces el muestreo correspondiente a la variable aleatoria X_i se denotará por la secuencia de valores $x_{i1}, x_{i2}, \ldots, x_{ir}, \forall i = 1, 2, \ldots, k$ y un entero positivo r común a todas las muestras.

Si la variable aleatoria X_i tiene media μ_i, entonces una aproximación a esta media de la población es la media de la muestra asociada a X_i; es decir,

$$\hat{\mu}_i = \frac{1}{r}\sum_{j=1}^{r} x_{ij}, \forall\, i = 1, 2, \ldots, k$$

Similarmente, si la varianza de la población es σ_i, entonces una estimación de esta varianza es dada por la ecuación siguiente,

$$\begin{aligned}
\hat{\sigma}_i^2 &= \frac{1}{r}\sum_{j=1}^{r}(x_{ij} - \hat{\mu}_i)^2 \\
&= \frac{1}{r}\left[\sum_{j=1}^{r} x_{ij}^2 - 2\sum_{j=1}^{r} x_{ij}\hat{\mu}_i + \sum_{j=1}^{r}\hat{\mu}_i^2\right] \\
&= \frac{1}{r}\left[\sum_{j=1}^{r} x_{ij}^2 - 2r\hat{\mu}_i^2 + r\hat{\mu}_i^2\right] \\
&= \frac{1}{r}\left[\sum_{j=1}^{r} x_{ij}^2 - r\hat{\mu}_i^2\right]
\end{aligned}$$

Sin embargo, si esta estimación se considera realmente una estimación de la varianza de la muestra, entonces el valor esperado de ella en toda la población es

dada por la ecuación,

$$
\begin{aligned}
E(\hat{\sigma}_i^2) &= E\left(\frac{1}{r}\sum_{j=1}^{r} x_{ij}^2 - r\hat{\mu}_i^2\right) \\
&= \frac{1}{r}\left[\sum_{j=1}^{r} E\left(x_{ij}^2\right) - rE\left(\hat{\mu}_i^2\right)\right] \\
&= \frac{1}{r}\left[\sum_{j=1}^{r} E\left((x_{ij}-\mu+\mu)^2\right) - rE\left((\hat{\mu}_i-\mu+\mu)^2\right)\right] \\
&= \frac{1}{r}\left[\sum_{j=1}^{r}\left[E\left((x_{ij}-\mu)^2\right) - 2\mu E(x_{ij}-\mu) + E(\mu^2)\right]\right] - \\
&\quad -\left[E((\hat{\mu}-\mu)^2) - 2\mu E(\hat{\mu}-\mu) + E(\mu^2)\right] \\
&= \frac{1}{r}\left[r\sigma^2 + r\mu^2\right] - \left[E((\hat{\mu}-\mu)^2) + \mu^2\right] \\
&= \sigma^2 - E((\hat{\mu}-\mu)^2) \\
&= \sigma^2 - E\left(\left(\frac{1}{r}\sum_{j=1}^{r} x_{ij} - \mu\right)^2\right) \\
&= \sigma^2 - \frac{1}{r^2}E\left(\left(\sum_{j=1}^{r}(x_{ij}-\mu)\right)^2\right) \\
&= \sigma^2 - \frac{1}{r^2}E\left(\left(\sum_{j=1}^{r}(x_{ij}-\mu)\right)\left(\sum_{k=1}^{r}(x_{ik}-\mu)\right)\right) \\
&= \sigma^2 - \frac{1}{r^2}E\left(\sum_{j=1}^{r}\sum_{k=1}^{r}(x_{ij}-\mu)(x_{ik}-\mu)\right) \\
&= \sigma^2 - \frac{1}{r^2}\sum_{j=1}^{r}\sum_{k=1}^{r}E((x_{ij}-\mu)(x_{ik}-\mu)) \\
&= \sigma^2 - \frac{1}{r^2}\sum_{j=1}^{r}E((x_{ij}-\mu)^2) \\
&= \sigma^2 - \frac{1}{r^2}r\sigma^2 \\
&= \frac{r-1}{r}\sigma^2
\end{aligned}
$$

en donde se ha supuesto que μ es la media de la población y que las variables que definen la muestra son estadísticamente independientes. Es claro que el estimador de la varianza así definido es un estimador sesgado que, para valores grandes de r, coincide exactamente con la varianza de la población, σ. Se puede definir un estimador no sesgado si se toma $\hat{\sigma}^2_{i,nosesgado}$ como

$$\hat{\sigma}^2_{i,nosesgado} = \frac{r}{r-1}\sigma_i^2$$

y en este caso se tiene

$$E\left(\sigma^2_{i,nosesgado}\right) = \sigma$$

B.2. Distribución Chi cuadrada

La distribución Gaussiana es la distribución más conocida. Sin embargo, las pruebas estadísticas también emplean otro tipo de distribuciones, entre las cuales se encuentra la denominada distribución Chi Cuadrada, comúnmente denotada por χ^2.

Esta distribución se emplea frecuentemente para llevar a cabo pruebas estadísticas sobre la varianza de una población. Esto resulta de una forma natural a partir de la misma definición de la variable aleatoria que representa a la función χ^2. Supóngase que se tiene una población (no una muestra de la población) cuyos elementos son descritos por la variable aleatoria X, cuyos valores se encuentran normalmente distribuidos con media (de la población) μ y varianza (de la población) σ^2. Es decir, la variable aleatoria X tiene una distribución normal $\mathcal{N}(\mu,\sigma^2)$. Se define una nueva variable aleatoria como sigue,

$$Z^2 = \frac{(X-\mu)^2}{\sigma^2}$$

Esta variable aleatoria tiene una función densidad de probabilidad denominada Chi cuadrada, y se dice que tiene un grado de libertad. Sin embargo, es posible definir más variables aleatorias a través de la suma siguiente,

$$\chi_\nu = \sum_{i=1}^{\nu} \frac{(X_i-\mu)^2}{\sigma^2}$$

siendo ν los grados de libertad. Cuando se determina la variable Chi cuadrada para una muestra, no para una población como se ha hecho hasta ahora, los grados

de libertad dependen del tamaño de la muestra menos la unidad. Esto es así debido a que se emplea una estimación de la media de la población a través de la media de la muestra. Nótese que la varianza de la población se encuentra entonces relacionada con la varianza de la muestra a través de la siguiente ecuación,

$$\begin{aligned} \chi^2_{r-1} &= \sum_{i=1}^{r} \frac{(X_i - \hat{X})}{\sigma^2} \\ &= \frac{(r-1)\hat{\sigma}^2}{\sigma^2} \end{aligned}$$

donde $\hat{\sigma}^2$ es la versión no sesgada de la varianza de la muestra.

Así, la distribución de la variable aleatoria Chi cuadrada es útil para llevar a cabo pruebas estadísticas en las que se tiene que tomar una decisión acerca de la varianza de una población. Para ello, es necesario contar con una muestra de interés y el nivel de significancia que se desea (valor α).

El $p-$Valor es la probabilidad que permite declarar la significancia de una prueba estadística. El término "significancia de una prueba" indica que la probabilidad es suficientemente pequeña como para rechazar la hipótesis nula (probabilidad de rechazar la hipótesis nula siendo una decisión incorrecta, si esta probabilidad es muy pequeña, entonces se rechaza la hipótesis nula ya que es poco probable que se tome una decisión incorrecta; es decir, es más probable que sea correcta) [Loureiro and García, 2007]. En otras palabras, generalmente el $p-$Valor busca ser una evidencia en contra de la hipótesis nula.

¿Qué se quiere decir por un $p-$Valor suficientemente pequeño, de tal forma que permita rechazar la hipótesis nula y adoptar la hipótesis alternativa? Generalmente, los valores de $p-$Valor menores a 0.05 se consideran como pequeños. Un valor de 0.05 indica que si un experimento se lleva a cabo veinte veces, diecinueve de los resultados experimentales soportarán la evidencia a favor de la hipótesis nula, pero una de ellos será en contra. Pero lo interesante de este resultado radica en el hecho de que una diferencia tan grande como esta hubiera surgido por casualidad si realmente hubiera sido completamente cierta la hipótesis nula. Es decir, esta diferencia surge porque la hipótesis nula no se está cumpliendo del todo. Mientras más veces se repita el experimento y resulte que al menos uno de sus resultados contradice la hipótesis nula de tal forma que resulte por pura casualidad, entonces más altamente significativo se vuelve el $p-$Valor asociado.

En términos de la función densidad de probabilidad, el $p-$Valor está dado por el área bajo la curva que describe la distribución en el intervalo definido por el valor de la variable aleatoria resultante de la prueba estadística correspondiente hasta infinito y/o hasta menos infinito.

Considerando un sola cola de la distribución (lado derecho), si este valor resultante de la variable aleatoria debido a la prueba estadística es mayor que el valor de la variable aleatoria asociado al nivel de significancia α seleccionado, entonces se desecha la hipótesis nula (esto implica que el $p-$Valor es menor que el nivel de significancia α seleccionado).

Considerando la cola izquierda de la distribución, si este valor resultante de la variable aleatoria debido a la prueba estadística es menor que el valor de la variable aleatoria asociado al nivel de significancia α seleccionado, entonces se desecha la hipótesis nula (esto implica que el $p-$Valor es menor que el nivel de significancia seleccionado).

Cada nivel de tratamiento define una muestra sobre unidades experimentales. El resultado del muestreo depende de un parámetro abstracto θ, que pertenece a algún espacio paramétrico desconocido Θ. El valor real del parámetro frecuentemente se desconoce, solamente se sabe de una clase de valores posibles para θ, denótese esta clase como el espacio paramétrico Θ. Sin embargo, se puede construir un conjunto de dos hipótesis acerca de este parámetro (equivalentemente, dividir el espacio paramétrico en dos subespacios):

Definición B.2.1. (Hipótesis nula)[Benko, 2001] La hipótesis nula es una suposición acerca del parámetro θ, la cual se desea demostrar que es verdadera

$$H_0 : \theta \in \omega, \text{ donde } \omega \subseteq \Theta$$

La situación se encuentra completamente especificada solamente cuando se conoce que existen otras alternativas para θ además de los valores en ω. Esto se conoce como a hipótesis alternativa. Uno de los ejemplos más comunes es la hipótesis alternativa que es complementaria a la hipótesis nula; a saber,

$$H_1 : \theta \in \Theta - \omega$$

N.B. En el trabajo que nos ocupa, los parámetros pueden ser la media o la varianza, y la hipótesis nula se puede referir, por ejemplo, a que la media de cada una de las muestras es la misma, o a que la varianza de cada una de las muestras es la misma. Nótese que se puede hablar de un espacio de parámetros de varias dimensiones. Por ejemplo, puede ser el espacio de parámetros N dimensional definido por vectores cuyas componentes están definidas por N medias. □

B.2.1. Hipótesis bidireccional versus hipótesis unidireccional

En lo que sigue, se asumirá implícitamente un parámetro unidimensional, una hipótesis de un solo punto ($\omega \in \Omega$) y $\Theta \subseteq R$. Esta suposición divide nuestra situación abstracta a dos tipos básicos de Hipótesis:

1. Hipótesis bidireccional ($\Theta = R$): La hipótesis nula

$$H_0 : \theta = \theta_0$$

contra la hipótesis alternativa

$$H_1 : \theta \neq \theta_0$$

donde $\theta_0 \in R$.

2. Hipótesis unidireccional ($\Theta \subseteq R$), en este tipo se distinguen dos casos:

a) $\Theta = \{\theta \geq \theta_0 | \theta, \theta_0 \in R\}$ con correspondiente hipótesis:

$$H_0 : \theta = \theta_0$$

contra la alternativa

$$H_1 : \theta \geq \theta_0$$

b) $\Theta = \{\theta \leq \theta_0 | \theta, \theta_0 \in R\}$ con correspondiente hipótesis:

$$H_0 : \theta = \theta_0$$

contra la alternativa

$$H_1 : \theta \leq \theta_0$$

Definición B.2.2. (Prueba de H_0 contra H_1) Probar H_0 contra H_1 es un proceso de decisión que se basa en el muestreo $X_1, X_2, \ldots, X_n$, el cual conduce al rechazo o no rechazo de H_0.

Después de la prueba, puede ocurrir una de cuatro situaciones posibles,

1. H_0 es verdadera y se decide no rechazar H_0, lo cual es una decisión correcta.

2. H_0 es verdadera, pero se decide rechazar H_0, lo cual es una decisión incorrecta.

3. H_1 es correcta (por lo tanto, H_0 es incorrecta), pero se decide no rechazar H_0, lo cual es una decisión incorrecta.

4. H_1 es correcta (por lo tanto, H_0 es incorrecta) y se decide rechazar H_0, lo cual es una decisión correcta.

N.B. σ^2 es la varianza de las medias de las muestras [Illowsky and Dean, 2008]. Debe entenderse que la variables F es una variable aleatoria y que para diferentes muestreos se tendrán, por tanto, diferentes valores de F, asumiendo aún que se tienen los mismos grados de libertad pata todos los muestreos. Parece ser que el procedimiento para realizar la prueba F es el siguiente,

1. Determinar los grados de libertad del numerador y del denominador.
2. Determinar el valor de F a través del cociente de las varianzas.
3. Proporcionar el nivel de significancia que se desea.
4. Con los grados de libertad, el nivel de significancia deseado y el valor de la función F, determinar el $p-$valor.
5. Si el $p-$valor es mayor que el nivel de significancia deseado, entonces no se puede rechazar la hipótesis H_0.

Una referencia excelente es dada por [Horn, 2012]. En ella se explica de manera clara la forma en que se rechaza o acepta la hipótesis nula. Otra excelente referencia es [Lavery, 2004]. En esta referencia se afirma que el $p-$valor indica la probabilidad de tomar una decisión incorrecta cuando se rechaza la hipótesis nula H_0. Esto quiere decir que valores pequeños del $p-$valor sugieren rechazar la hipótesis nula sin temor a equivocarse casi seguramente.

$$r = \text{rechazar hipótesis nula}$$

$$s = \text{tomar una decisión incorrecta}$$

El $p-$Valor es la probabilidad de que $r \wedge s$, mientras que $1-(p-\text{Valor})$ es la probabilidad de que $\neg(r \wedge s) = \neg r \vee \neg s$; es decir, de que no se rechaze la hipótesis nula o no se tome una decisión incorrecta, o bien de que se acepte la hipótesis nula o se tome una decisión correcta.

El $p-$Valor proporciona la probabilidad de cometer un error de tipo 1, en donde se dice que se rechaza la hipótesis nula sabiendo que es verdadera (decisión incorrecta cuando se rechaza la hipótesis nula). Esto es más claro de si habla de la hipótesis alterna H_1. El $p-$Valor es la probabilidad de tomar una decisión incorrecta cuando se acepta (rechaza) la hipótesis $H_1(H_0)$.

Evento: La hipótesis nula es verdadera pero se rechaza creyendo que la hipótesis alterna es verdadera, lo cual es una decisión incorrecta. La probabilidad de que ocurra esta situación es dada por el $p-$Valor. Mientras más pequeño es el $p-$Valor menos propenso se está en cometer este error de Tipo 1 y, por lo tanto, más posibilidades existen de que se sostenga la hipótesis nula. Mientras más grande es

el p – Valor, más propenso se está en cometer este error de Tipo 1 y, por lo tanto, menos posibilidades existen de que se sostenga la hipótesis nula.

La probabilidad de rechazar la hipótesis nula implicando tomar una decisión incorrecta es dada por el p – Valor. Mientras más pequeño es el p – Valor, con mayor razón se debe rechazar la hipótesis nula ya que menos probable es que ocurra el evento "rechazar la hipótesis nula implicando tomar una decisión incorrecta". Mientras más grande es el p – Valor, con menor razón se debe rechazar la hipótesis nula ya que más probables es que ocurra el evento "rechazar la hipótesis nula implicando tomar una decisión incorrecta".

La clave está entonces en definir el evento "rechazar la hipótesis nula implicando tomar una decisión incorrecta"y definir el p – Valor como la probabilidad de que ocurra. El p – Valor indica la probabilidad de estar mal si se rechaza la hipótesis nula (error de Tipo 1).

correcto pero se rechaza y pvalor pequeño, entonces rechazar

correcto pero se rechaza y pvalor grande, entonces no se rechaza □

Comparing the difference between means to the variability within contestant distributions is the basis for ANOVA [Lavery, 2004].

[Vokey and Allen, 2011]

APÉNDICE

C Publicación

C.1. Artículo presentado en el CIIECC 2013

Como resultado de esta investigación se propuso el siguiente artículo publicado en la revista elsevier; se agrega documento para constancia de la asitencia a la 3ra. Conferencia Iberoaméricana de Ingeniería Electrónica y Ciencias Computacionales, en San Luís Potosí, México, 24-26 de Abril 2014. Además se muestra una gráfica que representa el interés por el artículo de algunos lectores de diferentes países.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 7 (2013) 273 – 281

Procedia Technology

2013 Iberoamerican Conference on Electronics Engineering and Computer Science

Order statistics and item bank analysis in computer adaptive testing

J.Suárez[a,*], A.Franco[a], R.A.Santos[b]

[a] *Systems and Information Technologies Research Center, Autonomous University of the State of Hidalgo, Pachuca - Tulancingo Avenue Km. 4.5 , Mineral de la reforma, Hidalgo, Mexico C.P. 42184*
[b] *Computer Science Department, Technological University of Xicotepec of Juárez, Puebla, Mexico*

Abstract

This paper addresses the problem of items exposure rate in computer adaptive testing and its relation with the structure of an item bank. An item is a structure defined by real and/or virtual components possibly containing text, image, audio and/or video elements, which are useful to build a context where a question is made about, and of diverse elements or mechanisms for information acquisition to provide an answer to this question. Every item has an associated difficulty that depends on how easy is to answer the question about the defined context. An item bank is a deposit of this kind of structures and, in this paper, the item bank structure is defined in terms of statistical indexes arising from the onedimensional Order Statistics Theory, namely, nearest neighbor index and the standard variate of normal curve; and another one from the concept of compactness of intervals of real numbers. In this sense, the work talks about items difficulty exposure rate assuming that the item bank is, in fact, defined by a finite discrete set of items difficulties. Therefore, the emphasis is given on the items difficulty and it is assumed that the number of items per difficulty in the item bank is unlimited. The experimental results are obtained through a simulation environment that takes into account the definition of the structure of an item bank, the definition of a testing subject and the definition of an item administration context. Therefore, the results are mainly experimental rather than theoretical, although the validation of the simulation environment is based on theoretical results of other authors in the field.

Selection and peer-review under responsibility of CIIECC 2013

Keywords: item bank, order statistics, computer adaptive testing, simulation, item difficulty

1. Introduction

Some Learning Management Systems (LMS's) [17] and Web site development systems [22] have a testing component to measure the degree of achievement of their users in specific knowledge areas (Blackboard,

☆This document is a collaborative effort of the Autonomous University of the State of Hidalgo and the Technological University of Xicotepec of Juarez, under the financial support of the Mexican National Council of Science and Technology and the State of Hidalgo's Government, FOMIX'08 97071 Project.

*Corresponding author

Email addresses: jsuarez@uaeh.edu.mx (J.Suárez), afranco@uaeh.edu.mx (A.Franco), rando_san@hotmail.com (R.A.Santos)

2212-0173 © 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CIIECC 2013
doi:10.1016/j.protcy.2013.04.034

Moodle and Hot Potatoe for example). This testing component defines a completely broad research field by itself, and includes different aspects related with the administration of items used in tests.

The item administration becomes rather simple or complex depending on the philosophy of testing at hand. Relatively simple procedures for item administration based on a philosophy of paper and pencil can be found but, when this philosophy of testing is based on concepts such as those introduced by Computer Adaptive Testing (CAT) [15, 19, 20], things become not so simple.

CAT is not just a very promising philosophy of testing but, actually, a reality based on the idea of adapting the test to the testee, instead of adapting the testee to the test, like in the paper and pencil case. There are different ways of implementing this type of evaluation. One of them is based on the Item Response Theory (IRT) [20]. Knowledge structure, rule space and factor analysis are other three different methods oriented to implement CAT philosophy, as well [12, 13, 18].

However, within CAT testing philosophy, there are still many problems embracing topics related with item bank administration, mainly the specification of an optimal item selection criterion. Item administration in real time is one of the main characteristics of an adaptive testing system because of its design and implementation. Several models of initial ability estimation, item selection, item's exposure rate control, and diverse psicometric models, among others, have been proposed.

The problems of item selection and item's exposure rate control are quite related, and they have a direct interaction with the item bank, which is a repository from where the items are extracted along a test processing. Given a context of testing, the item selection process consists in applying a map between the estimated ability of a testee and an item's difficulty, with the main intention of selecting the items with the best information about the real ability level of the testee. In this sense, if Θ defines the set of abilities and Δ is a collection of subsets of item's dificulties, then $f : \Theta \longmapsto \Delta$ defines the function f with a not so simple correspondence rule, as it will be seen later on. The function f behaves almost like an interval map does. Under ideal circumstances, the item bank should have an adequate amount of items (with an acceptable distribution of difficulties) to avoid item overexposure at any time.

Even though the domain Θ of f is a real interval, this does not happen for the Δ collection which, in a concrete application, its elements are finite numerable sets of difficulties. This situation is even more complicated, because the cardinality of these finite numerable sets can also be variable in some contexts of selection procedure [3].

If $\mathcal{B}$ denotes the item bank, then the finite cardinality of the repository (total number of different difficulties values, not the number of items with these difficulties!) is $|\mathcal{B}|$. In a real scenario, for every difficulty value μ_i in the repository there is a number m_i of items with this difficulty, so that the total number of items in the repository is given by the expression $\sum_{k=1}^{|\mathcal{B}|} m_k$.

Hence, metaphorically speaking, every difficulty labels a set of m_i items in some instant of time t, and the role of the function f consists in selecting the set of difficulties with the highest information about the real ability of the testee, considering that an estimate of this ability is already known. To simplify the analysis, hereafter the number of items per difficulty is assumed to be unlimited, and future work is addressed toward the case of a finite number of items per difficulty, considering this problem as an specific case of queue theory or stock control. Figure 1 illustrates the case discussed in this paper.

The topic research in this paper is concerned with the problem of studying the behavior of an item bank as a function of the radius ϵ_p from the testing neighborhood $B_{\epsilon_p}\left(f\left(\hat{\theta}\right)\right)$, and the radius ϵ_s from the neigborhood of selection $B_{\epsilon_s}(\cdot)$. Do they have an effect on the size of the difficulty exposure rate? Notice that, because of the context already proposed, instead of questioning about item exposure, the question asks about difficulty exposure. In fact, the contribution of this paper is twofold: to formulate useful criteria, based on Order Statistics Theory, to categorize the goodness of the structure of an item bank, and to remark that the characteristics that define the neighborhood $B_{\epsilon_s}(\cdot)$ directly affect the evaluation process and the item administration.

2. Order Statistics Theory and item bank description

The item bank structure has a strong impact on the precision of the estimated ability for a testee. For example, an item bank containing a small amount of items (difficulties) and with high dispersion on the

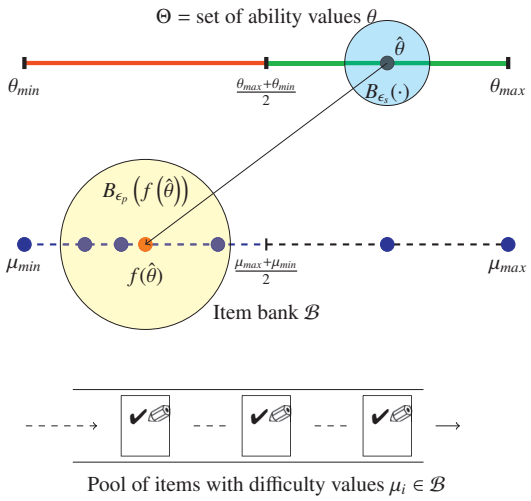


Fig. 1. Estimate ability values $\hat{\theta}$ are used to select the next difficulty inside the item bank $\mathcal{B}$. This selection considers items inside a test neighborhood $B_{\epsilon_p}\left(f\left(\hat{\theta}\right)\right)$. The number of difficulties inside $\mathcal{B}$ is a finite one, but the number of items per difficulty is unlimited. At the same time, there is a neighborhood $B_{\epsilon_s}(\cdot)$ whose elements, if any, are estimated abilities that satisfy the pseudo Cauchy criterion for convergence. The test continues while this stop criterion is not satisfied.

difficulty values, will almost surely produce estimated abilities beyond the acceptable values. Therefore, it is reasonable to say that the structure of the item bank is defined by the number of items, the distribution of their difficulties, or the number of items per difficulty, the interval of possible values for these difficulties, the type of item, etc.

The method of item (difficulty) selection is also affected by the structure of the item bank. It is clear that, no matter what method of item (difficulty) selection is chosen, an item bank with a quite disperse distribution of item difficulties will make that the exposure rate of these difficulties will increase quickly.

Some efforts have been made to solve these and many others problems associated with the item bank administration. For example, the simplest way of item selection suggests to choose the item with the higher information about the real ability of the testee. For some authors, this is a criterion based on a single value of the estimated ability [3, 16].

However, if the item bank $\mathcal{B}$ is considered as a neighborhood containing the whole set of item difficulties, the criterion of single ability estimate value can be seen as one criterion based on item selection over one interval or neighborhood. This observation suggests the possible construction of more general methods of selection based on neighborhoods of difficulties rather than just a single point. Indeed, several authors have studied selection rules with this kind of neighborhood [10]. These proposals are, in fact, critics to the item selection method based on a single value of the estimated ability .

However, one of the inconveniences of these proposals, even those based on Fisher information, is that high item (difficulties) exposure rates are obtained, or they tend to affect the precision of the estimated abilities, or they produce both effects at the same time [10]. Furthermore, the precision on the estimated ability is not just a function of the item bank size, as other authors said [16], but also a function of the distribution of the item difficulties.

With no doubt, the number of items is very important to get an item bank with acceptable performance, but the specification of this valor is not, in any way, a warranty for the item bank to have a sufficiently large number of difficulties with reasonable distribution, which is a *sine qua non* condition to obtain good

precisions of the estimated abilities, and to have an effective policy of item exposure control. For these reasons, it is very important to talk about number of items per difficulty, concept that appears in a natural way when the distribution of difficulties is considered as a clustering problem in the data mining field.

This is only one part of the whole story, because other kind of neighborhood must also be considered. Particularly, the kind of neighborhood associated with the way of finishing a test. One criterion is based on the idea of convergence of the estimated abilities, and further enquiry needs to be done in this sense. There are some interesting results of other authors explaining the ways in which this type of convergence occurs[9]. In fact, some of these results are used to validate the functioning of the simulator used in this paper.

There are statistical tools that can also be applied to study the phenomenon of item exposure rate, and Order Statistics Theory is one of them. There are different applications of the Order Statistics Theory in several branches of science, mainly in biology, geology, etc. [11]. The concept of nearest neighbor is also presented in this theory, but in a very different sense of the previous discussion [21]. Appendix A introduces in a greater detail some important aspects of Order Statistics Theory and the way this theory relates with the field of item bank analysis.

Therefore, the analysis of difficulties distribution might be seen as one dimensional clustering problem that can be described through very well known indexes in the data mining field, specifically those indexes defined by the nearest neighborg method. These indexes are useful to describe the degree of clustering of a set of points, which is precisely the situation in an item bank. Second order properties or local properties are a complement of first order or global properties associated to important patterns of distribution (mean, variance, mode, etc.) [2, 6, 11, 14] and the indexes are a very useful tool to specify these local properties.

It is quite interesting to search for these kind of indexes and how they can be useful for determining the goodness of the item bank structure. For example, the type of effects that a very disperse distribution of difficulties has on the estimated ability and the item exposure rate, even though the number of items per difficulty in the structure of the item bank is acceptable. This point of view remarks the important difference between the number of items and the distribution of difficulties, it makes evident that the problem of item exposure rate only embraces the control of item presentation for every difficulty, and that the problem of precision in the estimated ability is closely related with the distribution of difficulties in the item bank.

3. Algorithms

Algorithm StoppingConvergence implements the satisfiability of the pseudo Cauchy criterion of convergence and Algorithm 1 describes the part of the main procedure, where the function StoppingConvergence is called. There are, of course, many other procedures and functions still not mentioned, but with the same importance for the good functioning of the simulator and the implementation of the ideas already posed in previous sections of this paper.

The Computer Adaptive Testing process can only stop when one of the following three conditions is satisfied: the evaluation has used the maximum number of permitted items, or the evaluation has lasted for the maximum permitted time, or the estimated ability does not sufficiently change within a neighborhood previously defined. This paper only studies the effects of the third condition on the evaluation process, and Line 7 of Algorithm 1 makes a call to algorithm StoppingConvergence, looking for the satisfiability of the pseudo Cauchy criterion to stop. The pseudo Cauchy criterion is satisfied when the neighborhood $B_{\epsilon_s}(\cdot)$ reaches the desired cardinality $\left|B_{\epsilon_s}(\cdot)\right|$.

4. Simulation results

The simulation results are supported by a previous validation of the simulator. After validating the system, some conditions of testing are created to produce some results related with the difficulty exposure rate.

Algorithm 1 Sketch of main procedure to show the construction of the current set of estimated abilities (latest $|B_{\epsilon_s}(\cdot)|$ values of estimated abilities) to test the pseudo Cauchy convergence.

Require: Arguments of simulation
Ensure: Results of simulation
procedure MAIN(Arguments of simulation)
    ...
    $q \leftarrow 1$
    $Certificate \leftarrow$ false
    while $Certificate$ is false **do**
        Compute componente *qth* of current set of estimates of ability
        if current set of abilities has cardinality $|B_{\epsilon_s}(\cdot)|$ **then**
            $Certificate \leftarrow$STOPPINGCONVERGENCE(current set, ϵ_s, $|B_{\epsilon_s}(\cdot)|$) ▷ Certify current set
        end if
        ...
        if current set of abilities has cardinality $|B_{\epsilon_s}(\cdot)|$ and $Certificate$ is false **then**
            $q \leftarrow |B_{\epsilon_s}(\cdot)|$ ▷ Module $|B_{\epsilon_s}(\cdot)|$ definition of indexes of elements current set ability estimates
            Make left shift on current set abilities, leaving empty rightmost cell to include next estimate
        else
            $q \leftarrow q + 1$ ▷ Current estimated set does not have $|B_{\epsilon_s}(\cdot)|$ elements, yet. Still being filled
        end if
        ...
    end while
    ...
end procedure

4.1. Simulation validation

The validation of the simulator has been made considering the convergence behavior of the estimated abilities, although some other options, or complementary evidence, can be used. For example, the following theorem provides some useful tools in this sense and has been stated and proved by other authors in the field,

Theorem 1. *Let $\{\hat{\theta}_k\}$ be the sequential estimators specified by steps 1-3 for the Rasch model. Then, as $n \to \infty$, $\hat{\theta}_n \to \theta$ a.s. and $\sqrt{n/4}(\hat{\theta}_n - \theta) \to N(0,1)$. Furthermore, $4I_n(\hat{\theta})/n \to 1$ a.s., where $I_\theta = \sum_{i=1}^{n} exp(\theta - b_i)/(1 + exp(\theta - b_i))^2$ is the observed Fisher information [9].*

Figure 2 illustrates the convergence of the estimated ability to the real value of ability for some testing conditions. Particularly, the estimate is predicted with a precision of 0.25% in relative error. It can be seen that the estimated abilities, and those difficulty values given by the item bank, accumulate around a bidimensional neighborhood defined by these variables.

4.2. Selection neighborhood and item exposure

On the other hand, Figure 3 illustrates the way the radius of the neighborhood of selection affects the number of iterations to reach the true ability value. The same experimental conditions are hold for every experiment, as represented by the total number of iterations as a function of the neighborhood radius.

5. Conclusions and future work

According to the results obtained from the behavior of the item bank administration, when the radius and cardinality of the neighborhood of selection are included as variables of interest, we can observe that this behavior is very interesting. The control of these variables has an impact on the values of the item exposure rate. There is a regular behavior of the item exposure rate as a function of the radius of the neighborhood

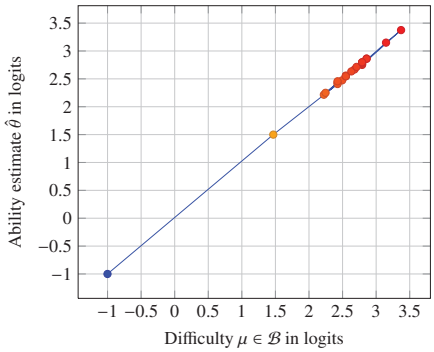


Fig. 2. Ability estimate $\hat{\theta}$ (as predicted by the simulation) versus selected item difficulty $\mu \in \mathcal{B}$. The conditions of predictions assume an unknown real ability with value of 3 logits and an initial ability with value of −1 logits.The predicted ability estimate, after 23 iterations (ítems presentations or, equivalently, a test with 23 ítems), has a value of 2.79809 logits, while the last selected item difficulty (element of the data bank $\mathcal{B}$) has the value 2.79107 logits, which gives a relative error of 0.25%. On the other hand the item bank structure and item administration conditions are, respectively, defined by a neighborhood index of 0.94633 and a density index of 0.98772, a pseudo Cauchy neighborhood of 0.1 with cardinality 5 and a test neighborhood of 0.01. The distribution of difficulties in the item bank is an acceptable intrinsic characteristic and it is not due to random effects as indicated by the Gaussian index value of −1.134. The item bank has 300 difficulties and the interval and subinterval are given by (−4, +4).

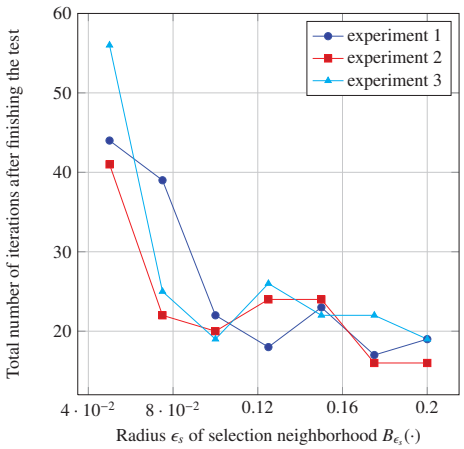


Fig. 3. Simulation results for total number of iterations before finishing a test versus the radius ϵ_s of the neighborhood of selection $B_{\epsilon_s}(\cdot)$. Every point in an experiment represents a test under the specified simulation conditions. For smaller radius ϵ_s, higher the total number of iterations and, therefore, higher the possibility of using more frequently the difficulties in the item bank. On the other hand, for higher radius ϵ_s, smaller the total number of iterations and, therefore, smaller the possibility of using more frequently the difficulties in the item bank. The experimental conditions define an item bank with 100 difficulties, with a nearest neighbor index of 0.8979, random structure of −1.2363, density index of 0.9856, test neighborhood of 0.01, Cauchy cardinality of 5, initial ability of −1 logits and a suppossed real ability of +1 logits. The minimum and maximum points of the interval and subinterval of difficulties are (−4, +4) logits.

of selection. The item exposure rate is inversely proportional to the radius. Small radius tend to provide large values on the total number of iterations before finishing a test. The cardinality of the neighborhood of selection has also a strong impact on the item exposure. A high value of cardinality increases the total number of iterations before finishing a test. The radius and cardinality of the neighborhood of selection represent, in a certain way and methaporically speaking, the severity of the instructor. The cardinality of the neighborhood of selection represents the number of sequential items that the testee has rightly, or wrongly, answered just before the instructor decides to finish the test and, on the other hand, the radius means the tolerance that the instructor assigns for the estimated ability to represent the real ability value. At the current stage of the research presented in this paper, Order Statistics seems to be the natural tool to locally describe the structure of an item bank. The structure of the item bank is mainly given by the distribution of difficulties and, in a natural way, Order Statistics introduces the concept of item density, which refers to the number of items per difficulty. On the other hand, there are still some other very important research questions that need to be addressed to the work. Future work should be conducted to study the relation between the structure of an item bank, represented by the nearest neighbor index, the standard variate of normal curve index and the density index, and the precision of the estimated ability and the item exposure rate. The problem of item exposure control in terms of queue theory or stock control needs to be solved.

Furthermore, the description or analysis of the structure of a real item bank, through the different indexes already discussed, has been started by the development of a real Computer Adaptive Testing System called Ariya [1, 5, 7, 8]. The system is currently oriented to evaluate the testee's ability to understand different physical concepts inside the topic of kinematic and the subtopic of uniformly accelerated motion. This development will be used to discuss another important point concerned with the way in which some different exposure rate control techniques (Randomesque, Simpson–Hetter, etc. [4]) behave as a function of the item bank structure, along with their relations to the characteristics of the neighborhood $B_{\epsilon_s}(\cdot)$.

Appendix A. Order Statistics and density indexes

Appendix A.1. Density index

The difficulty density property of the item bank is characterized by the density index, which is defined through the concept of average distance between points in a compact interval and the concept of average distance between a set of discrete points inside the same compact interval. Therefore, in the definition of the density index, it is assumed that there is a real open interval (a, b) and that it is required to compute, for every point $x \in (a, b)$, the average distance between this point and the set of points $y \in (a, b)$ [2, 6]. This average distance changes from one point to another and, for a point $x \in (a, b)$, it is given by

$$\frac{1}{b-a}\left[x^2 - (a+b)x + \frac{1}{2}(a^2+b^2)\right]$$

Next, it can be proved that the average of all these distances becomes $\frac{1}{3}(b-a)$. For a discrete finite set of difficulties, the mean of the average distances is computed as indicated by Definition Appendix A.1, and the density index is $\delta = 3\frac{\hat{D}}{b-a}$.

Definition Appendix A.1. *Let assume a random and one dimensional spatial distribution defined by n difficulties $x_1, x_2, \ldots, x_n$ with increasing order (it does not matter the form of the difficulties distribution at this moment). Let D_{ij} the distance between the point i and the point j, which is defined as follows, $D_{ij} = \left|x_i - x_j\right|, \forall i, j \in \{1, 2, \ldots, n\}$, then the average distance related with the difficulty i is given by the following equation,*

$$\hat{D}_i = \frac{1}{n}\sum_{j=1}^{n} D_{ij}, \forall i \in \{1, 2, \ldots, n\} \text{ and the mean of these averages becomes } \hat{D} = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i$$

The index of density of the set of difficulties describes how much dense the set becomes when compared with the density of the compact real interval (a, b). Notice that, in the general case, the set of discrete

difficulties can be strongly grouped inside a subinterval of the interval (a, b), but this fact does not necessarily imply a high density with respect the interval (a, b). The maximum value of the density index occurs when there are only two discrete difficulties and the value of one of them acquires the value a and the other one the value b. In this case, the density index has the value 1.5. The minimum value of the density index occurs when there are just two discrete difficulties, but they are close enough to make the average distance equal to zero, so the density index is equal to zero in this case. A density index close to 1 implies a high density value for the set of discrete difficulties.

Appendix A.2. Order Statistics

Let consider the experiment of randomly and uniformly selecting n difficulties $x_1, x_2, \ldots, x_n$ from the interval (a, b). Afterwards, these n difficulties are ordered in an increasing form $x_{i_1}, x_{i_2}, \ldots, x_{i_n}, i_k \in \{1, 2, \ldots, n\}$, where x_{i_k} represents a possible value of the random variable X_k. The main concern consists in determining some of the important statistics for every one of the n random variables such as, for example, the mean and the variance. This is called the order n statistics. In what follows, $n \geq 2$ and the analysis is made over the general interval (a, b) following the methodology from reference [21], where the analysis is only made over the unitary interval $(0, 1)$. It is relatively easy to obtain the results for order n statistics in the interval (a, b) by mean of the following change of variable $y = \frac{x-b}{b-a}$, where it is assumed that $a < x < b$. Therefore, $y \in (0, 1)$ and the results of reference [21] can be applied. The expected value of X_k in the order n statistics of the interval $(0, 1)$ is $E(Y_k) = \frac{k}{n+1}$, so that the expected value in the interval (a, b) is

$$E(Y_k) = a + \frac{k}{n+1}(b-a), \forall k = 1, 2, \ldots, n$$

Similarly, the variance of the order n statistics for the interval (a, b) is

$$\sigma^2(Y_k) = \frac{1}{n+2}\frac{k}{n+1}\left(1 - \frac{k}{n+1}\right)(b-a)^2, \forall k = 1, 2, \ldots, n$$

These results can be used to analyze theoretically the behavior of the average of the distances between every pair of difficulties in the item bank, when only closest neighbors are considered. In this case, the average is given by the following equation, under the assumption that there are n difficulty values in the interval (a, b),

$$\hat{d} = \frac{1}{n}\left[abs(x_2 - x_1) + \sum_{i=2}^{n-1} min(x_i - x_{i-1}, x_{i+1} - x_i) + abs(x_n - x_{n-1})\right]$$

The theoretical mean of these averages is given by the equation

$$\left\langle \hat{d} \right\rangle = 0.5\left(1 + \frac{1}{n+1}\right)\frac{1}{\rho}, \forall n \geq 2,$$

where ρ represents the number of items per difficulty or the item density. The variance of the averages is also given by

$$\sigma^2(\hat{d}) = \frac{1}{n^2}\frac{2n^2 + 17n + 12}{12(n+1)^2(n+2)}(b-a)^2, \forall n \geq 2$$

The statistic tests to determine the level of clustering of the difficulties in the interval (a, b) are given by the nearest neighbor index η and the standard variate of normal curve c, which are respectively defined as follows [11],

$$\eta = \frac{\hat{d}}{\left\langle \hat{d} \right\rangle} \text{ and } c = \frac{\hat{d} - \left\langle \hat{d} \right\rangle}{\sigma(\hat{d})}$$

References

[1] Alcántara J, RE García. *Esquema de Seguridad para un Ambiente de Evaluación Adaptable*. Tesis de Licenciatura en Sistemas Computacionales. Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo, Hidalgo, México; 2010.

[2] Barhum K, Goldreich O, Shraibman A. *On approximating the average distance between points*. Technical Report partially supported by the Israel Science Foundation grant No. 460/05. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel; March 2007.

[3] Barrada JR, J Olea, V Ponsoda. Reglas de selección de ítems en Tests Adaptativos Informatizados. *Metodología de las Ciencias del Comportamiento, Suplemento* 2004; Volumen Especial: 55-61.

[4] Barrada JR, J Olea, V Ponsoda, FJ Abad. *Incorporating Randomness in the Fisher Information for Improving Item Exposure Control in CATS*. Reporte de Investigación. Departamento de Psicología Social y Metodología, Facultad de Psicología, Universidad Autónoma de Madrid; 2006.

[5] Barranco JA, RA López. *Análisis, Diseño e Implementación de una Base de Datos para la Administración de Reactivos en un Evaluador Educativo*. Tesis de Licenciatura en Sistemas Computacionales. Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo, Hidalgo, México; 2010.

[6] Burgstaller B, F Pillichshammer. The average distance between two points. *Bull Austral Math Soc* 2009; 80(3): 353-359.

[7] Canales T. *Reingeniería del Modelo de Datos a un Sistema de Evaluación por Computadora*. Reporte Técnico de Proyecto de Estadía dentro del Programa de Ingeniería en Software. Universidad Politécnica de Pachuca, Hidalgo, México; 2012.

[8] Cano G, M Pedraza. *Tecnología Java y MySQL como Alternativa para la Creación de Ítems Multimedia, Dentro de un Sistema de Evaluación en Línea*. Tesis de Licenciatura en Sistemas Computacionales. Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo, Hidalgo, México; 2011.

[9] Chang HH, Z Ying. Nonlinear Sequential Designs for Logistic Item Response Theory Models with Applications to Computerized Adaptive Tests. *The Annals of Statistics* 2009; 37(3): 1466-1488.

[10] Cheng PE, Liou M. *Computerized adaptive testing using the nearest-neighbors criterion*. Research Report supported by grant from the National Science Council, ROC. Institute of Statistical Science, Academia Sinica, ROC, Taipei 115, Taiwan; 2002.

[11] Clark PJ, FC Evans. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology* October 1954; 35(4): 445-453.

[12] Doignon JP. *Probabilistic assessment of knowledges*. In: Albert D, editor. Knowledge Structures, New York: Springer Verlag; March 1994, Chapter 1, p. 1-52.

[13] Falmagne JC, Doignon JP. *Learning Spaces*. Interdisciplinary Applied Mathematics. Berlin, Heidelberg: Springer; 2011.

[14] Levine N. *CrimeStat III Version 3.0, A Spatial Program for the Analysis of Crime Incident Locations, Operation Manual*. Washington, DC, USA: The National Institute of Justice; 2004.

[15] McBride JR. *Research Antecedents of Applied Adaptive Testing*. In: Sands WA, Waters BK, McBride JR, editors. Computerized Adaptive Testing, From Inquiry to Operation, Washington, DC, USA: American Psychological Association; 1997, Chapter 5, p. 47-57.

[16] Olea J, Ponsoda V. *Tests Adaptativos Informatizados*. Primera Edición. San Romualdo, Madrid, España: Ediciones Universidad Nacional de Educación a Distancia; 2003.

[17] Randall B, Sweetin J, Steinbeiser D. *Learning Management System Feasability Study, Part II of the Open Source Collaborative Moodle Assessment Report*. Technical Report Version 1.3. North Carolina Community College System Office: Learning Technology Systems, North Carolina, USA; 2010.

[18] Tatsuoka KK. *Cognitive Assessment, An Introduction to the Rule Space Method*. Multivariate Application Series. Taylor & Francis Group, 270 Madison Avenue, New York, NY 10016: Routledge; 2009.

[19] Van der Linden WJ, Glas CAW, et al. *Computerized Adaptive Testing, Theory and Practice*. In: Van der Linden WJ, Glas CAW, editors. Computerized Adaptive Testing, Theory and Practice, New York, USA: ICO, Kluwer Academic Publishers; 2000.

[20] Van der Linden WJ, Hambleton RK, et al. *Handbook of Modern Item Response Theory*. In: Van der Linden WJ, Hambleton RK, editors. Handbook of Modern Item Response Theory, New York, USA: Springer Verlag; 1997.

[21] Weingart M, Selvin S. *Nearest Neighbor Analysis in One Dimension*. Technical Report LBL-36888 UC-605. Information and Computing Sciences Division, Lawrence Berkeley Laboratory and Division of Biostatistics, School of Public Health, University of California, Berkeley, California 94720, USA: February 1995.

[22] Wilde R. Technical Evaluation Report 33: Evaluating Authoring Tools. *The International Review of Research in Open and Distance Learning* August 2004; 5(2): 1-6 .

CIIECC 2013

Conferencia Iberoamericana de Ingeniería Electrónica y Ciencias Computacionales
San Luis Potosí, México
24-26 de Abril, 2013

La **Universidad Autónoma de San Luis Potosí**
a través de la
Facultad de Ciencias
otorga el presente:

RECONOCIMIENTO a:

J. Suárez, A. Franco ,
Randolfo Alberto Santos

Por su participación como:

Exponente(s) en las Sesiones Técnicas
Order statistics and item bank analysis in computer adaptive testing

dentro de los eventos llevados a cabo en la
3ra. Conferencia Iberoamericana de Ingeniería Electrónica y Ciencias Computacionales 2013
Efectuado durante los días 24 al 26 de Abril del 2013

Dr. José Martin Luna Rivera
Presidente del Comité Técnico

Dr. Enrique Stevens Navarro
Presidente del Comité Organizador

UASLP-FC

CIIECC
Iberoamerican Conference on Electronics Engineering and Computer Science

Article Usage Dashboard

Order Statistics and Item Bank Analysis in Computer Adaptive Testing

Sua'rez, J.; Franco, A.; Santos, R.A.

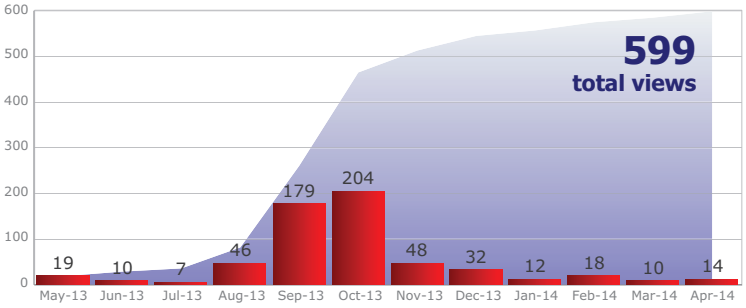
Procedia Technology, Volume(s) 7, 10-May-2013, Pages 273-281

View Article

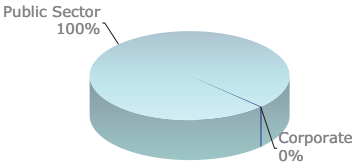
Views by geography

Top countries	Rank	Views	Pct
United States	1	519	87%
Iran, Islamic Republic of	2	13	2%
India	3	11	2%
China	4	10	2%
Mexico	5	9	2%

Trend and cumulative views



Corporate versus Public Sector



 Terms and conditions Support