The Examiner: Automatic Generation of "Good" Exams

Francisco J. Torres-Rojas
Escuela de Computación
Instituto Tecnológico de Costa Rica
torresrojas@gmail.com

***Abstract*—As educators, we must design, prepare, proctor and grade hundreds of exams during their careers. From this overwhelming task, we collect little or none objective evidence about the quality of the exams themselves. Thus, at most there is an intuitive learning about what characterizes a good or a bad exam. It is very likely that we blindly repeat in our exams rights and wrongs of the past. There exist metrics about the quality of an exam, and even metrics about the quality of each of the individual items in the exam. Using actual college courses, our research found experimental evidence that proves that it is possible to predict with great accuracy, parting from historical statistical data, the quality metrics that an exam will show even before applying it to a standard group of college students. With this result, we built an automatic system that generates "good" exams from an item bank enriched with statistical information from previous exams. Besides, powerful tools for analysis and controlled adjustment of each exam and each item were developed.**

***Index Terms*—education, evaluation, exams, IRT.**

***Resumen*—Las personas dedicadas a la educación deben diseñar, preparar, ejecutar y revisar cientos de exámenes durante sus carreras. De esta tarea abrumadora se recolecta poca o ninguna evidencia objetiva de la calidad de cada examen por lo que hay si acaso un aprendizaje intuitivo de qué caracteriza a un buen examen o a un mal examen. Es muy probable que se repitan errores o aciertos del pasado a ciegas. Existen métricas de la calidad de un examen e inclusive de cada uno de los ítemes que lo forman. Se encontró de manera experimental en cursos universitarios reales que, usando información estadística previa, es posible predecir con bastante exactitud las características de calidad que un examen tendrá aún antes de que se aplique a un grupo normal de estudiantes. Con este resultado se construyó un sistema automatizado para la generación automática de "buenos" exámenes armados partiendo de un banco de ítemes enriquecido con información histórica de su uso en evaluaciones previas. Además, se diseñaron poderosas herramientas para un análisis detallado y un ajuste controlado de cada examen y de cada ítem.**

I. Introducción

Como educadores tenemos la misión fundamental de transmitir conocimiento a otras personas. En una gran mayoría de los casos, existen requisitos legales y administrativos que nos exigen que midamos y que certifiquemos oficialmente la cantidad de este conocimiento que haya sido efectivamente asimilado por cada estudiante.

Sin embargo, pese a nuestra buenas intenciones, esta *cantidad de conocimiento* es un ejemplo clásico de lo que se conoce en estadística, psicología y otros campos como una **variable latente** ([9], [11], [20], [23], [30], [51]). Estas son variables inobservables que por su naturaleza no permiten una medición directa, así que solo pueden ser estimadas o inferidas de manera indirecta a través de un modelo matemático definido sobre otras variables que sí pueden ser observadas y medidas directamente. Otros ejemplos de variables latentes son calidad de vida, confianza en los negocios, moral del personal, felicidad de un país, o inclusive el coeficiente intelectual [46].

En un curso típico hay alguna mezcla de tareas, proyectos, asignaciones, exámenes y otras actividades que son evaluadas y medidas individualmente para luego ser combinadas con una fórmula ponderada cuyo resultado se redondea o se clasifica con ciertas reglas para producir una nota final, con la que pretendemos haber medido el conocimiento adquirido por cada persona que haya llevado el curso [32].

Con este contexto en mente, cada examen hecho en un curso puede ser entendido como una sonda o muestra estadística que recupera pistas para ayudarnos a inferir la variable latente del conocimiento adquirido por cada estudiante [11].

Si nos dedicamos a la docencia, tarde o temprano nos toca diseñar, preparar, ejecutar y revisar docenas de exámenes al año, que se acumulan en cientos o incluso miles en una carrera normal. Hay quienes disfrutan esta tarea, pero es más corriente que sea considerada una labor tediosa y abrumadora. Hay un alto consumo de horas tanto en la preparación como en la revisión de los exámenes. Esto puede precipitar en un cierto descuido de los docentes buscando atajos como reciclar exámenes completos de semestres recientes, que usualmente ya están en manos de las personas a ser evaluadas[1], o incluir, sin saber, preguntas exageradamente difíciles o fáciles. Además, siempre está el fantasma de las preguntas mal planteadas.

Creemos, porque así lo sentimos, que algunos de los exámenes que hicimos estuvieron bien y que otros fueron un desastre. A veces, con suerte, podemos identificar algunas preguntas de un examen que fueron muy buenas porque parecen separar claramente a las personas con dominio de la materia de aquellas que no. Pero la mayor parte es simple y sencillamente intuición. Entonces, la pregunta es: ¿qué caracteriza a un buen examen?

[1]Jamás se debe entender que el propósito de un examen es que a los estudiantes les vaya mal, pero tener a mano el examen *a priori* es un obvio fraude.

Hay estadísticas básicas tales como la media, la moda, la nota mínima, la nota máxima y la desviación estándar, que dan cierta luz respecto a la calidad de un examen. Un histograma de las notas siempre da información útil. Sin embargo, la triste realidad es que, en la mayoría de los casos, aparte de las notas obtenidas en cada examen que son registradas en alguna hoja de cálculo, es poco lo que aprendemos respecto al instrumento en sí. Así que muy probablemente repetiremos errores o aciertos a ciegas.

Podemos señalar los siguientes problemas en la mecánica usual del diseño y revisión de exámenes:

- Se desperdicia la oportunidad de un análisis detallado de la amplia información subyacente en cada examen.
- No hay un análisis a nivel de cada uno de los ítemes o reactivos que forman un examen.
- No se establecen relaciones con datos históricos previos (los cuales en todo caso usualmente no existen).
- No hay un mecanismo objetivo que nos advierta de preguntas mal diseñadas o mal planteadas.
- No hay manera de identificar objetivamente buenas preguntas que discriminen de manera efectiva.
- No se asimilan lecciones que nos permitan diseñar mejores exámenes en el futuro.

Se exploraron las propiedades estadísticas de ciertos parámetros básicos de un examen, usando como insumo información recolectada en múltiples cursos ofrecidos en distintos semestres de Computación. La hipótesis fue:

Si se cuenta con una base de datos de ítemes con información estadística histórica, se pueden predecir las características estadísticas de un examen construido con un subconjunto aleatorio de ítemes tomados de dicha base de datos.

Encontramos que esta hipótesis era verdadera. Esto trae como consecuencia la factibilidad de automatizar el proceso de generación de "buenos exámenes", es decir exámenes que satisfagan ciertos parámetros establecidos de calidad. Por ejemplo, se puede solicitar la generación automática de un examen donde se espera que la media de las notas sea de 67.5, y donde el promedio de los r_{pb} o coeficientes de discriminación asociados a los ítemes sea de al menos 0.15; o se puede solicitar la generación automática de un examen cuya consistencia interna (α de Cronbach [8]) sea mayor o igual a 0.80. El producto de software conocido como "*The Examiner*" descrito en detalle en este documento es un prototipo bastante funcional que muestra la validez del resultado.

En la Sección II, se revisan los antecedentes de esta investigación. Los parámetros estadísticos asociados a exámenes e ítemes son explorados en la Sección III. La Sección IV presenta a *The Examiner*, cuyas características técnicas son descritas en la Sección V. El proceso de generación de un examen es explicado en la Sección VI. La Sección VII ahonda en el análisis de un examen ya aplicado y la Sección VIII menciona la fase de actualización y generación de reportes. Finalmente, las conclusiones y el trabajo futuro pueden ser encontrados en la Sección IX.

II. Antecedentes

Este trabajo se coloca en el contexto de Teoría de Respuesta al Ítem o *Item Response Theory* (IRT), también conocida como teoría del rasgo latente (TRL) o teoría de respuesta al reactivo (TRR).

La IRT es usada para la construcción de exámenes y pruebas psicológicas. Describe la relación entre un conjunto de datos obtenidos en un proceso de medición (las respuestas a los ítemes de una prueba) con determinadas variables latentes, como por ejemplo el dominio de la materia de un estudiante, o los rasgos de personalidad de los sujetos a quienes se ha administrado. El tipo de variables o propiedades latentes que se pretende medir pueden ser cuantitativas o cualitativas ([32], [37], [41]).

Las ideas fundamentales de IRT pueden ser rastreadas a los trabajos de Louis Leon Thurstone en 1912 y de Alfred Binet en 1905 ([15], [26], [34]), pero se establece como disciplina académica entre las décadas de 1950 y 1960, con los trabajos del psicometrista estadounidense Frederick M. Lord trabajando para el Educational Testing Service (ETS) [33], el matemático danés Georg Rasch [43], y el sociólogo austriaco Paul Lazarsfeld ([27], [30]). A pesar de los sólidos desarrollos teóricos iniciales, el uso de la IRT no se extendió hasta las décadas de 1970 y 1980, cuando la proliferación de computadores personales permitió un acceso barato al procesamiento de datos requerido.

El libro clásico por excelencia de IRT es el de Lord [33], pero varios de sus desarrollos ya resultan obsoletos. Hambleton et al. dan un buen resumen de las técnicas principales de IRT en [20] y en [23]. También Hambleton junto con Van der Linden presentan un manual muy práctico de varios modelos de IRT pero que requiere conocimientos más avanzados del tema en [49]. Una introducción muy accesible a IRT se encuentra en el libro de Embretson [10], pero como el mismo título lo indica está muy orientada a psicólogos. Otro libro introductorio a IRT muy útil es el de Baker [5]. El mismo Baker explora varios modelos alternos de IRT presentando explicaciones detalladas de algoritmos que pueden ser usados para estimar parámetros de un ítem en [6].

De Boeck da una introducción a IRT orientada a investigadores y estudiantes de postgrado en [7]. Rafael Jaime de Ayala nos ofrece una versión más actualizada de los conceptos de IRT en [3]. Fox discute un interesantísimo enfoque bayesiano para el modelaje de IRT en [16]. Nering y Ostini son editores de un manual comprehensivo de los modelos de IRT más utilizados [38].

Existen diversos paquetes de software estadístico que pueden ser usados para análisis de datos desde el punto de vista de IRT. Por ejemplo, SPSS ([31], [47]), SAS ([31], [44]), R [42], o hasta el mismo Minitab [36] pueden calcular varios de los parámetros requeridos. Eventualmente es posible calcular parámetros tales como el α de Cronbach usando la hoja de cálculo EXCEL [13]. Por su naturaleza general, estos paquetes requieren la preparación de datos en el formato que cada uno maneja y no ofrecen ningún manejo de una base de datos

histórica de ítemes usados lo cual los hacía inconvenientes para nuestro proyecto de investigación.

Similarmente, existen otros softwares comerciales o libres con distintas capacidades para cálculo de parámetros de IRT, pero ninguno da el enfoque integral (datos históricos, generación de exámenes, análisis estadístico) que ofrece *The Examiner*. A continuación mencionamos a los más significativos:

- **BILOG-MG**: producto de *Scientific Software International* que tienen interfaces gráficas y múltiples capacidades de análisis de IRT. [45].
- **exMIRT**: paquete de software para análisis de ítemes y calificación de exámenes [14].
- **ICL**: lenguaje para estimaciones de parámetros de IRT escrito por Bradley Hanson y disponible en forma gratuita [24].
- **jMetrik**: software libre escrito por Patrick Meyer de la Universidad de Virginia para IRT [35].

Hay disponibles distintos generadores de exámenes tales como el trabajo de Grün y Zeilis con R [18], o una gran variedad de paquetes para LaTeX que automatizan parte del trabajo de edición ([50], [28]).

En cualquier caso, ninguno de los anteriores hace una análisis de parámetros de IRT. Tampoco ninguno muestra el alcance ni la sofisticación de *The Examiner*, cuyos objetivos no terminan con la generación de un examen, sino que incluyen el registro de datos históricos, la construcción de modelos predictivos del comportamiento de los exámenes, y los análisis estadísticos asociados.

III. Parámetros

A continuación describimos los parámetros calculados y utilizados en esta investigación:

III-A. Dificultad de un ítem (p_i)

Se interpreta como la probabilidad de que un estudiante conteste correctamente el ítem i. Por lo tanto, p_i es un número real entre 0 y 1.0, el cual es más bajo entre más difícil sea una pregunta. Inicialmente, este dato se obtiene de una estimación de la persona que crea la pregunta. Posteriormente, la estimación es reemplazada con los datos históricos, dividiendo la cantidad de estudiantes que contestan correctamente esta pregunta entre la cantidad total de estudiantes que la contestan.

III-B. Varianza y desviación estándar de las respuestas

Para los propósitos de *The Examiner*, únicamente se usarán preguntas de selección única. Estas sólo pueden estar correctas o incorrectas. Con esto en mente, el lector atento notará que al ser las preguntas de naturaleza dicotómica (i.e., correctas o incorrectas) la varianza del ítem i se puede calcular como:

$$\sigma_i^2 = p_i(1 - p_i) \tag{1}$$

donde p_i es la dificultad del ítem i definida previamente. La desviación estándar σ_i del ítem i sería la raíz cuadrada de esta cantidad. Entonces:

$$\sigma_i = \sqrt{p_i(1 - p_i)} \tag{2}$$

III-C. Coeficiente biserial puntual (r_{pb})

Esta cantidad se puede interpretar como la capacidad de discriminación de un ítem. Esencialmente, el r_{pb} del ítem i es el coeficiente de correlación entre tener bueno o malo dicho ítem y la nota final que la estudiante obtiene en el examen.

Al ser un factor de correlación, el r_{pb} es un número real entre -1.0 y +1.0. Entre más cercano esté este valor a +1.0 significa que la pregunta discrimina muy bien entre estudiantes que conocen la materia y aquellos que no, es decir que las personas bien preparadas tienden a contestarla bien y las que tengan deficiencias tienden a contestarla mal. Por otro lado, valores negativos del r_{pb} significan que la pregunta muestra una curiosa discriminación inversa: las personas que sacan mala nota en el examen la contestan bien y las de buena nota en el examen la tienen mala (usualmente esto es indicador de que la pregunta está mal planteada o que la opción indicada como correcta está equivocada).

Finalmente, si el r_{pb} vale exactamente 0.0 significa que la pregunta no discrimina entre estudiantes que conozcan la materia o no. Este último caso se da cuando todas las personas evaluadas contestan correctamente la pregunta, o aún peor cuando todas la contestan mal. Ambas situaciones son indeseables porque significan que la pregunta no contribuye a detectar el nivel de conocimiento real.

Supongamos que n estudiantes hicieron un examen y cada uno obtuvo una nota final denotada como X_i, siendo $\bar{X}$ el promedio general de notas. La varianza s_n^2 de estas notas se calcula como:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3}$$

Por tanto, la desviación estándar s_n es:

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{4}$$

Supongamos además, que el ítem i fue contestado correctamente por n_1 personas, e incorrectamente por $n_0 = n - n_1$ personas. Si sólo consideramos los exámenes de los que lo contestaron correctamente, denotaremos como M_1 el promedio de las notas de estos exámenes. Similarmente, denotaremos como M_0 el promedio de las notas de los exámenes de las personas que contestaron mal el ítem i.

Entonces el coeficiente biserial puntual o r_{pb} del ítem i, se calcula con la fórmula:

$$r_{pb} = \frac{M_1 - M_o}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{5}$$

Nótese que:

$$p_i = \frac{n_1}{n}$$

y que:

$$(1 - p_i) = \frac{n_0}{n}$$

Entonces la ecuación 5 se puede reescribir como:

Tabla 1: Interpretación de α de Cronbach

α de Cronbach	Consistencia Interna
$\alpha \geq 0{,}9$	Excelente
$0{,}8 \leq \alpha < 0{,}9$	Buena
$0{,}7 \leq \alpha < 0{,}8$	Aceptable
$0{,}6 \leq \alpha < 0{,}7$	Cuestionable
$0{,}5 \leq \alpha < 0{,}6$	Pobre
$\alpha < 0{,}5$	Inaceptable

$$r_{pb} = \frac{M_1 - M_o}{s_n}\sqrt{p_i(1-p_i)} \tag{6}$$

y reemplazando 2 en 6:

$$r_{pb} = (M_1 - M_o)\frac{\sigma_i}{s_n} \tag{7}$$

III-D. α de Cronbach

Cuando tratamos de medir una variable latente no directamente observable, por ejemplo, el conocimiento adquirido, en un grupo de personas, podemos aplicar un examen con n ítemes o reactivos al grupo. Ya que suponemos que los n ítemes están relacionados con la variable latente inobservable de interés, los n ítemes debieran mostrar un elevado nivel de correlación entre ellos, puesto que cada una está por premisa correlacionado con la variable latente ([8], [9], [11]). Esto es lo mismo que decir que los n ítemes son consistentes entre ellos.

El α de Cronbach permite cuantificar el nivel de fiabilidad o **consistencia interna** de una escala de medida para la magnitud inobservable construida a partir de las n variables observadas. Esta métrica fue propuesta originalmente por Lee Cronbach [8], pero hay antecedentes en [22] y en [19].

Por construcción, es una cantidad que se mueve entre 0.0 y 1.0. Valores cercanos al 1.0 son preferidos. Por ejemplo, para un examen universitario el α de Cronbach debe ser mayor o igual a 0.75, considerándose que valores superiores a 0.9 son excelentes indicadores, mientras que valores menores a 0.5 hacen inútil al instrumento. La tabla 1 muestra la intepretación normalmente aceptada de esta métrica [44].

Supongamos que se aplica un examen de K ítemes a n personas. La ecuación 3 calcula s_n^2 la varianza de las notas del examen, y la ecuación 1 calcula la varianza σ_i^2 del i-ésimo ítem del examen. Entonces, el α de Cronbach de este examen se calcula con la fórmula:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_i^2}{s_n^2}\right) \tag{8}$$

III-E. Efecto individual (positivo o negativo) de cada ítem en el α de Cronbach

Inicialmente se calcula el α de Cronbach para todo el examen tal como se describió en la ecuación 8. Ahora, se recalcula pero omitiendo el i-ésimo ítem, como si el examen sólo tuviera los otros $K-1$ ítemes.

Al calcular la diferencia (positiva o negativa) entre ambos valores del α de Cronbach podemos estimar la contribución individual del ítem i a la consistencia interna del examen completo. Así, si esta diferencia es positiva esto nos indica que el i-ésimo ítem reduce la consistencia interna del examen (i.e., es inconsistente con las otras preguntas del examen), y si la diferencia es negativa implica que al retirar el i-ésimo ítem, el examen se volvió menos consistente internamente.

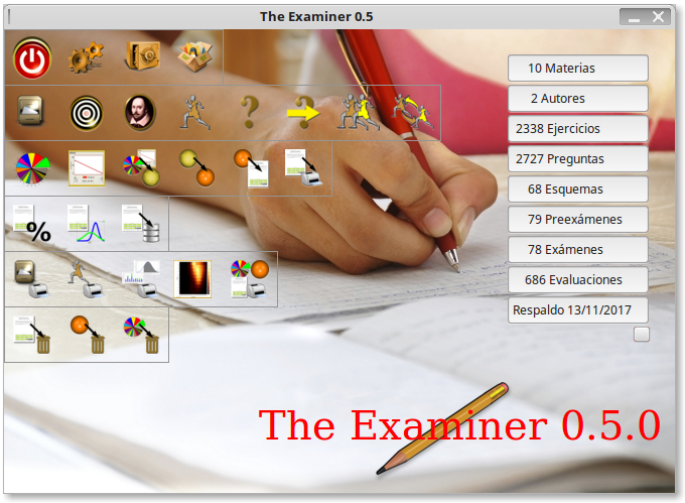


Figura 1: *The Examiner*

IV. *The Examiner*

Se siguió una metodología experimental donde se generaron exámenes de selección única que fueron aplicados en cursos ofrecidos en distintos semestres en la carrera de Ingeniería en Computación de la Escuela de Computación del Instituto Tecnológico de Costa Rica [25]. Se conservó una base de datos con información histórica del comportamiento de cada examen y de cada ítem utilizado para su análisis estadístico. Todo el proceso fue automatizado con un sistema computacional llamado *The Examiner*, que facilita la preparación, evaluación y análisis de exámenes. La Figura 1 muestra un ejemplo de la interfaz principal.

The Examiner maneja exámenes de selección única con 5 opciones (una verdadera y cuatro distractores). Se espera incluir otros tipos de preguntas en el futuro. Sin embargo, contrario a leyendas urbanas corrientes entre muchos estudiantes y algunos profesores, existe abundante evidencia científica de que las notas obtenidas en este tipo de exámenes se correlacionan fuertemente con las notas que se obtienen en, por ejemplo, exámenes con preguntas de desarrollo ([12], [21], [39]). Hay que agregar además ventajas inherentes a los exámenes de selección única tales como no requerir que las personas evaluadas tengan que escribir - muy posiblemente a mano - largos textos, la objetividad y facilidad de revisión, las posibilidades de *data mining*, y otras más.

El software permite que el usuario construya y mantenga una base de datos de preguntas, que se enriquece cuando son resueltas por estudiantes en exámenes reales. Para cada pregunta se obtienen y se conservan datos tales como:

- Fecha de creación.

- Fecha del uso más reciente en un examen.
- Persona creadora del ítem.
- p_i: dificultad de la pregunta. Este dato es inicialmente estimado por la persona que diseña el ítem, y luego reemplazada por datos reales de dificultad cuando la pregunta es resuelta.
- r_{pb}: Coeficiente biserial puntual de la pregunta completa.
- Coeficiente biserial puntual de cada opción (correcta y distractores) del ítem.
- Varianza de las notas de los exámenes donde aparece esta pregunta.
- Varianza de las notas de los exámenes donde aparece esta pregunta y fue contestada correctamente.
- Varianza de las notas de los exámenes donde aparece esta pregunta y fue contestada incorrectamente.
- Cantidad de estudiantes que han contestado esta pregunta
- Otras métricas estadísticas

Usando la información acumulada, *The Examiner* puede predecir una serie de propiedades estadísticas de exámenes armados automáticamente con subconjuntos de ítemes tomados de la base de datos, aún antes de ser aplicados. El objetivo de estas predicciones es generar automáticamente "buenos exámenes", i.e., exámenes que sean apropiados y justos desde el punto de vista de todas las entidades involucradas (centros de estudios, estudiantes y docentes).

Después que un examen es ejecutado y revisado, el software realiza un detallado análisis ítem por ítem que podrá revelar errores puntuales en algunos de ellos. *The Examiner* tiene la capacidad de hacer una variedad de ajustes inmediatos al examen actual y usar esta información para mejorar iterativamente exámenes futuros.

V. Características Técnicas

The Examiner es un software libre que puede ser estudiado, copiado y distribuido sin ninguna restricción. De momento, ejecuta únicamente en el sistema operativo Linux. Fue desarrollado como parte de los experimentos de esta investigación.

Este sistema está formado por:

1. 49 mil líneas de código escrito en Lenguaje C y repartidas en 29 programas.
2. 102614 líneas de XML para describir las interfaces usando `glade`. Toda la interacción gráfica es hecha a través de `gtk`.
3. 7 megabytes repatidos en 165 imágenes para formar la iconografía, botones y decoración del sistema.

La información es mantenida internamente en una base de datos relacional con 17 tablas. Se utilizó `Postgres 9.3` [40]. Para la síntesis de todos los reportes y documentos (incluyendo muy especialmente a los exámenes) del sistema, *The Examiner* genera código LaTeX ([17], [28], [50]) que es convertido a archivos PDF de manera relativamente transparente para el usuario. En algunos módulos, estos reportes son en realidad presentaciones `Beamer` [50] que el usuario puede parametrizar. Diversos gráficos estadísticos son preparados con `gnuplot`.

Se le ha puesto especial énfasis a la facilidad de uso, por lo que el diseño busca que las interfaces sean intuitivas y amigables. La mayor parte del funcionamiento es bastante automática, logrando que uno se pueda enfocar más en redactar y pulir propiamente las preguntas, sin tener que preocuparse de detalles de carpintería en la construcción, revisión y análisis de exámenes. Al mismo tiempo, muchas características funcionales son parametrizables y hay posibilidades de refinamiento e intervención manual en puntos clave del proceso [29].

The Examiner hereda todo el poder de LaTeX para generar documentos y exámenes de gran calidad y belleza tipográfica ([17], [28], [50]). Gráficos, imágenes, tablas, fórmulas matemáticas, distintos tipos de letras, y muchas otras facilidades están disponibles para la preparación de preguntas. Además, *The Examiner* puede generar múltiples versiones de un mismo examen, alterando el orden de las preguntas y de las opciones dentro de las mismas, para desalentar intentos de fraude.

VI. Generación de Exámenes

La primera tarea a la que los usuarios de *The Examiner* se tienen que avocar es la entrada de datos. Esto consume tiempo, sobre todo cuando se está usando el software por primera vez, pero se compensa con creces en las fases posteriores. Entre otras cosas, se deben ingresar primero materias, autores y profesores al sistema. Después se ingresan ejercicios, preguntas y ligas entre ejercicios de diferentes materias. Un ejercicio contiene una o más preguntas interrelacionadas que muy posiblemente comparten información común. Toda pregunta pertenece a un ejercicio. La combinación de estas dos labores tiene como objetivo construir la base de datos de preguntas con la que se generarán exámenes en otros módulos del sistema. La Figura 2 muestra los entes involucrados en este paso.

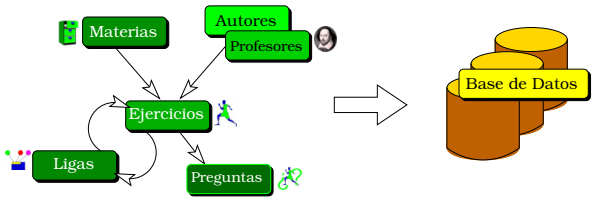


Figura 2: Entrada de Datos

Suponiendo que ya existen suficientes ejercicios y preguntas en la base de datos, como primer paso para generar un examen con *The Examiner*, hay que definir un esquema de examen donde se especifica la cantidad de preguntas de cada tema y subtema de una misma materia que aparecerán en dicho examen. La Figura 3 ilustra gráficamente este concepto. Un esquema no dice cuáles preguntas estarán en un examen, solamente cuántas preguntas de cada tema y subtema.

En el siguiente paso del proceso se selecciona un subconjunto de preguntas de la base de datos que sigue estrictamente las proporciones indicadas en un esquema preparado con anterioridad. Hay un sorteo para escoger aleatoriamente estas preguntas. Cada pregunta candidata a ser escogida tiene una cantidad de "boletos" asignada según sean las características

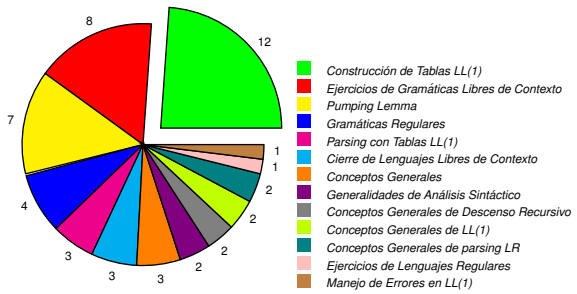


Figura 3: Esquema de Examen

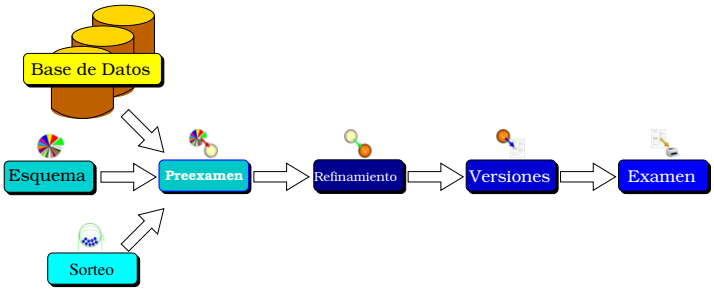


Figura 4: Proceso de Generación de Exámenes

que se deseen estimular (e.g., darle más boletos a preguntas difíciles, o darle más boletos a preguntas nuevas o que hace mucho tiempo no se usan, etc.).

Esta colección no ordenada de preguntas seleccionadas aleatoriamente bajo la guía de un esquema se conoce como un preexamen. De desearlo, el usuario puede refinar manualmente un preexamen, eliminando o agregando preguntas de manera controlada. Algo muy interesante es que usando la información histórica asociada a las preguntas seleccionadas para un preexamen, el software ya puede hacer predicciones respecto a las propiedades estadísticas de este futuro examen.

La diferencia fundamental entre un preexamen y un examen, es que este último está dotado de un orden. Así, cuando se genera un examen a partir de un preexamen se establece el orden de los ejercicios dentro del examen, el orden de las preguntas dentro de un ejercicio, e inclusive el orden de las opciones dentro de una pregunta. *The Examiner* da la capacidad al usuario de preparar simultáneamente múltiples versiones de un mismo examen donde, para propósitos de seguridad, los órdenes de las preguntas y de las opciones sean diferentes para cada versión.

La última fase del proceso agrega a todas las versiones de un examen la información administrativa necesaria para la aplicación al estudiantado. Esto incluye fecha, institución, profesor, instrucciones, juramento[2], hoja de respuestas, apéndices y otros detalles menores. La Figura 4 ilustra estos pasos y la Figura 5 muestra la portada típica de un examen y una página del enunciado.

VII. Análisis de un Examen

Por la naturaleza obvia de las preguntas de selección única, las respuestas respectivas solo pueden ser una de las 5 opciones existentes (i.e., A, B , C, D, E), un carácter inválido o un espacio en blanco. Esto hace que sea muy fácil una primera pasada casi automática de revisión de los exámenes. La entrada de estos datos es un proceso rápido y hasta entretenido. Se calculan las notas preliminares con un modelo básico de porcentaje de preguntas correctas que resulta suficiente para arrancar las fases posteriores de análisis.

Tomando un grupo completo de exámenes ya revisado, *The Examiner* hace un análisis detallado donde calcula una serie de métricas y estadísticas. Algunas, tales como media, varianza, desviación estándar, α de Cronbach, promedio de los r_{pb}, etc., por su carácter global permiten establecer criterios respecto al comportamiento general del examen.

Como ejemplo, la Figura 6 muestra las primeras dos páginas del análisis que hace el software de un examen parcial real de 59 preguntas aplicado a 17 estudiantes de un curso real de Biología Molecular Computacional. Se pueden notar, entre otras cosas, como la predicción de la media de notas (65.873) fue casi idéntica a la media real obtenida (67.398); que el α de Cronbach fue de 0.951, lo que de acuerdo a la Tabla 1 indica una excelente consistencia interna; y que el r_{pb} promedio fue de 0.283, lo que indica que el examen discrima de manera excelente entre las personas de alto rendimiento y las de bajo rendimiento. En otras palabras, el análisis nos indica que este fue un muy **buen** examen.

En el gráfico de la parte inferior de la Figura 6 se cruzan la dificultad p de las preguntas con el r_{pb} y se muestra la frecuencia o cantidad de preguntas como la altura de dicho gráfico (o en versión aplastada en la parte inferior como un "gráfico de calor"). Un buen examen se caracteriza por tener una "montaña" en el lado derecho de este gráfico donde está la discriminación positiva, y que se acomoda en la parte superior o inferior del eje p de acuerdo a qué tan bien o mal les fue a quienes tomaron el examen[3].

En la que puede ser la parte más interesante del análisis, se estudia ítem por ítem del examen. Para cada uno, se presentan gráficamente la cantidad de estudiantes que seleccionó cada una de las 5 opciones y se calculan métricas tales como dificultad, uso de distractores, índice de discriminación (r_{pb}), y efecto individual en el α de Cronbach. Estos valores son calificados visualmente con una serie de banderas correspondientes a las características buenas (banderas verdes, azules y celestes) o a las características consideradas problemáticas (banderas rojas y amarillas).

La parte izquierda de la Figura 7 muestra un ejemplo de

[2]Hay evidencia científica fuerte de que al someter a las personas a un juramento firmado en el que se comprometen a no hacer trampas, la cantidad de fraude académico se reduce sensiblemente ([1], [2]).

[3]Viendo el gráfico de calor reconocemos una buena curva como una "ballena" sumergiéndose hacia la derecha. Entre más profundo esté esta ballena, más difícil fue el examen.

Investigación de Operaciones
Primer Examen Parcial - 12/4/2016

Nombre : _____ Carnet: _____

Instrucciones: Sólo se tomará en cuenta lo que aparezca en la caja de respuestas abajo, independientemente de cualquier anotación hecha en el enunciado interno del examen. Use letras mayúsculas (**A**, **B**, **C**, **D**, **E**) para contestar. Se considera como incorrecta cualquier casilla vacía, ilegible, ambigua o con una letra diferente a las permitidas. No hay penalidad adicional por dar una respuesta equivocada. Si desea cambiar respuestas ya escritas, táchelas y escriba las respuestas que desee a un lado de la caja, junto a una nota explicativa y su firma.

_____

Las respuestas de este examen serán resultado de mis decisiones individuales. No usaré, recibiré, ni ofreceré ayuda no autorizada. No copiaré de otros exámenes, ni permitiré que nadie copie parte alguna de este examen. No realizaré ninguna trampa ni procedimiento deshonesto. ***Juro por mi honor que todo lo anterior es cierto****.*

_____
Firma

Respuestas

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54						

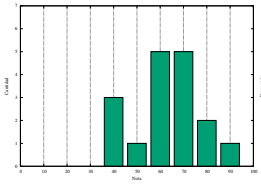
Correctas: _____ de **54** | Porcentaje: _____ | Ajuste: _____ de _____ | **Nota**: _____

Análisis de Segundo Examen Parcial
Biología Molecular Computacional (06/06/2015)

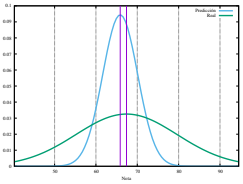
Instituto Tecnológico de Costa Rica	
Escuela de Ingeniería en Computación	
Ingeniería en Computación	
Materia	Biología Molecular Computacional
Profesor	Francisco J. Torres-Rojas
Descripción	Segundo Examen Parcial
Fecha	06/06/2015
Código de Examen	00009
Versiones	6
Preguntas	59
Estudiantes	17
Media	67.40
Nota mínima	45.76
Nota mínima ajustada	51.72
Nota máxima	91.53
Nota máxima ajustada	96.55
Desviación Estándar	12.2570
α **de Cronbach**	0.9510
r_{pb} **promedio**	0.2830

El examen muestra una **excelente consistencia interna** (α de Cronbach = **0.951000**). Diversos ítemes que miden la misma característica muestran un comportamiento bastante similar.

El examen muestra una buena discriminación promedio (**0.283000**). Distingue aceptablemente entre estudiantes de alto rendimiento y bajo rendimiento.



Histograma de Notas



Distribución Normal (real y predicción)

21. Al usar el árbol binario de búsqueda mostrado en la **Figura 2** se visitan en promedio 2.1 nodos para encontrar el dato requerido. ¿Cuál de las siguientes opciones podría estar mostrando las estadísticas de consultas por ciudad?

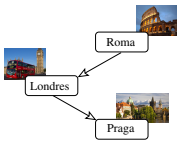


Figura 2

(a) Londres: 11315, Praga: 13200, Roma: 6821
(b) Roma: 1400, Praga: 400, Londres:200
(c) Roma: 657, Londres: 422, Praga: 118
(d) Praga: 2264, Londres: 7924, Roma: 1132
(e) Praga:40, Roma:67, Londres:83

22. El efecto donde una opción poco atractiva es ofrecida al público para hacer más atractiva otra opción que posiblemente no hubiera sido escogida sin la presencia de la poco atractiva fue estudiado por:

(a) Dan Ariely
(b) Dietrich Bonhoeffer
(c) Werner Heissenberg
(d) B. F. Skinner
(e) A. C. Doyle

23. Si se hace un alineamiento global de 2 hileras idénticas, ¿Cuáles de las siguientes afirmaciones serían **falsas**?

I. La ruta en la tabla quedará en la diagonal.
II. La tabla será cuadrada.
III. La cantidad de *gaps* en el alineamiento será 0.

(a) I y III.
(b) Todas son falsas.
(c) I y II.
(d) Todas son verdaderas.
(e) Sólo la III.

24. ¿Cuál de los 3 *white papers* (Robinson, Pringle y Crowder) asignados como lectura para este curso menciona al Método Símplex y las contribuciones de Dantzig al campo?

(a) Los 3 lo mencionan
(b) Harlan Crowder
(c) Lew Pringle
(d) Randy Robinson
(e) Ninguno de los artículos menciona ni al Símplex, ni a Dantzig.

	C_1	C_2	C_3	C_4
C_1	0	6	∞	3
C_2	2	0	∞	15
C_3	∞	7	0	∞
C_4	8	7	∞	0

Tabla 3: Distancias Directas entre Ciudades

	C_1	C_2	C_3	C_4
C_1	0	6	∞	3
C_2	2	0	∞	5
C_3	9	7	0	∞
C_4	7	8	∞	0

Tabla 4

25. Considere la ecuación recurrente

$$T(n) = T(\frac{n}{2}) + c$$

con $T(1) = c$.
Si $c = 1$, esta ecuación recurrente se reduce a

(a) 1
(b) $log_2 n + 1$
(c) $log_e n$
(d) n
(e) $2^n + 1$

Las preguntas 26 y 27 requieren la siguiente información:
La **Tabla 3** da las distancias directas entre 4 ciudades.

26. La distancia mínima entre la ciudad **2** y la ciudad **4** es:

(a) 4
(b) 6
(c) ∞
(d) 5
(e) 7

27. La tabla final de distancias óptimas entre ciudades es:

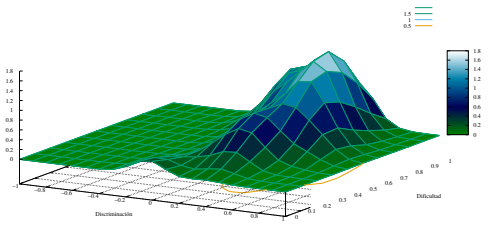
(a) Ver **Tabla 4**
(b) Ver **Tabla 5**
(c) Ver **Tabla 6**
(d) Ver **Tabla 7**
(e) Ver **Tabla 8**

	C_1	C_2	C_3	C_4
C_1	0	6	∞	3
C_2	2	0	∞	5
C_3	9	7	0	12
C_4	8	7	∞	0

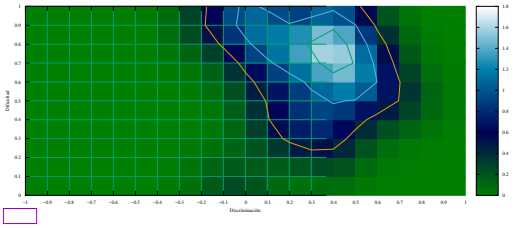
Tabla 5

	C_1	C_2	C_3	C_4
C_1	0	6	∞	3
C_2	2	0	9	6
C_3	∞	9	0	12
C_4	8	7	∞	0

Tabla 6



Cruce entre Coeficiente de Discriminación (r_{pb}) y Dificultad (p)



Líneas de Contorno del Cruce entre Coeficiente de Discriminación (r_{pb}) y Dificultad (p)

Figura 5: Portada y Página de Examen

Figura 6: Análisis de Características Generales del Examen

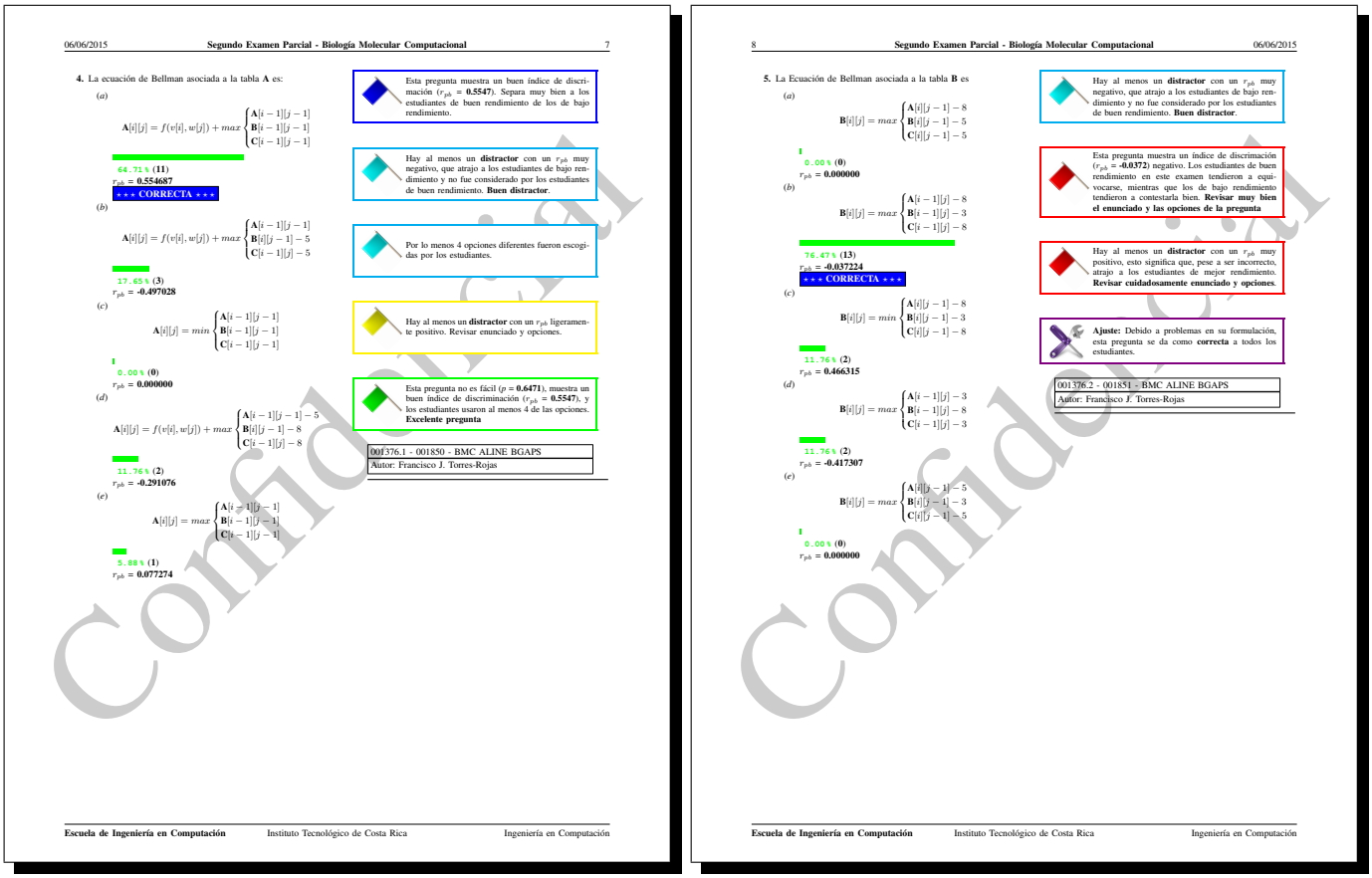


Figura 7: Análisis de Ítemes

un ítem que al ser analizado se consideró excelente (bandera verde). Este ítem muestra una dificultad p de 0.6471, lo cual indica que no era fácil. El r_{pb} del ítem es 0.5609, por lo que discrimina de manera excelente. Al menos 4 de las opciones fueron consideradas, lo que significa que los distractores estuvieron bien planteados, máxime cuando uno de ellos muestra un r_{pb} muy negativo (lo cual es un excelente indicador para un distractor). La única preocupación (bandera amarilla) viene del hecho de que haya un distractor con un r_{pb} ligeramente positivo.

Por otro lado, la parte derecha de la Figura 7 presenta un ítem problemático. El r_{pb} es negativo (lo que significa que los que lo contestaron bien sacaron mala nota, y los que lo contestaron mal tienen buena nota). Además, hay un distractor con un r_{pb} muy positivo. Posiblemente este distractor (que es una respuesta incorrecta) fue escogido por las dos personas de máxima nota en el examen. De hecho la pregunta mostró tantos problemas, que el profesor respectivo decidió hacer un ajuste a esta pregunta, dándola como buena a todas las personas.

Usando esta información, el docente puede hacer una diversidad de ajustes a las preguntas insatisfactorias (e.g., eliminar una pregunta, darla como buena a todas las personas evaluadas, considerarla crédito extra, etc.). La facilidad para hacer ajustes a un examen es una más de las funcionalidades de *The Examiner*. Estos cambios serán tomados en cuenta para reevaluar los exámenes automáticamente.

El sistema genera una serie de reportes asociados a este análisis e inclusive prepara automáticamente una presentación `Beamer` [50] con todos los detalles de lo analizado que resulta muy apropiada para ser mostrada a las personas evaluadas u otros entes interesados. La práctica con *The Examiner* nos ha enseñado que hay una respuesta muy positiva cuando se muestra este análisis detallado del examen, pues se comprende que hay interés en hacer una evaluación justa. Además, convierte la revisión del examen en una oportunidad extra de reforzamiento del aprendizaje.

VIII. Actualización y Reportes

Finalmente, la información estadística recolectada en la ejecución, revisión y ajuste de un examen debe ser llevada

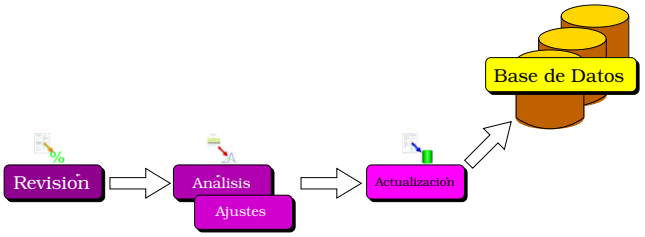


Figura 8: Actualización de Base de Datos

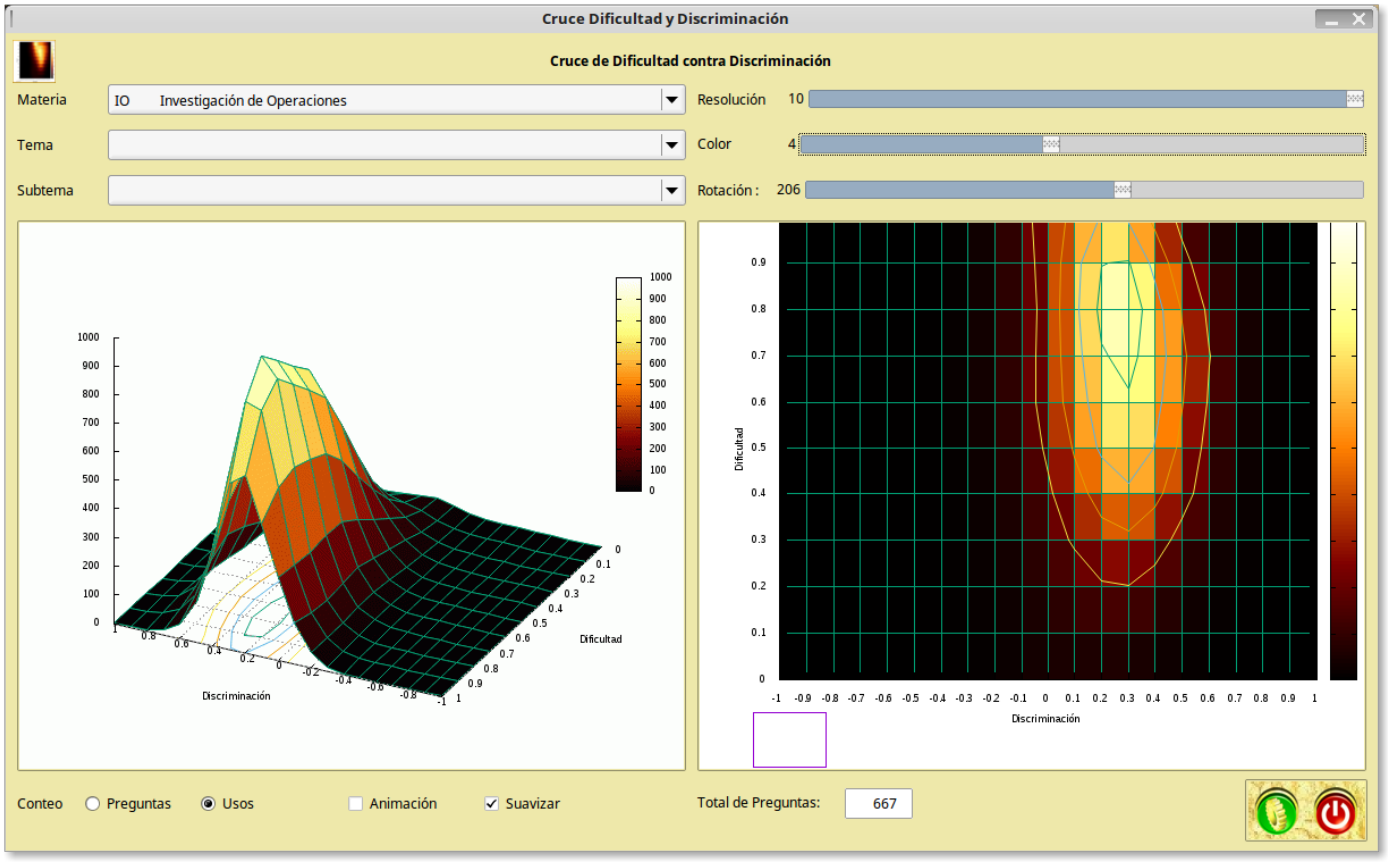


Figura 9: Análisis de la Base de Datos. Hay 667 ítemes de Investigación de Operaciones con información estadística histórica.

de vuelta a la base de datos para mejorar procesos futuros de preparación de exámenes e incrementar el poder predictivo de la herramienta. Inclusive, se lleva un registro de la última fecha en la que se usó un ítem para controlar que las preguntas no se repitan de manera muy seguida en exámenes. Por tanto, hay una fase final de actualización que se ejecuta después de haber completado los ajustes y análisis necesarios. Este proceso se ilustra en la Figura 8.

Aparte del procesamiento de un examen individual, la base de datos construida se puede analizar con una variedad de herramientas que muestran lo descubierto respecto a cada pregunta. El potencial para actividades de *data mining* es evidente. Por ejemplo, la Figura 9 muestra un análisis gráfico del comportamiento de la base de datos de preguntas del curso de Investigación de Operaciones. Hay capacidad para consultar visualmente sobre el gráfico, tocando cualquier entrada del mapa de calor, para ver ejemplos de preguntas en cada categoría.

Hay una amplia variedad de reportes (ya sea como archivos PDF o de forma interactiva) que produce *The Examiner*. Ninguno de ellos genera cambios en la información existente, por lo que pueden ser solicitados tantas veces como se desee La Figura 10 muestra gráficamente esta funcionalidad.

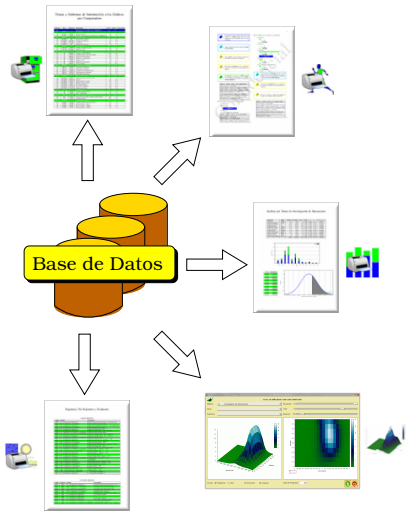


Figura 10: Generación de Reportes

IX. Conclusiones

Si se cuenta con una base de datos de ítemes con información estadística histórica se pueden predecir las características estadísticas de un examen construido con una combinación aleatoria de ítemes tomados de dicha base de datos. Los experimentos realizados presentaron predicciones de gran exactitud

y precisión. Como era de esperar, se encontró que la calidad de las predicciones mejora conforme se tenga más información histórica acumulada. Por lo tanto, es factible automatizar el proceso de generación de exámenes de calidad.

También se encontraron otros resultados indirectos muy interesantes, como el hecho de que preguntas en un examen que vengan directamente de ejemplos mostrados en clase o de pruebas cortas realizadas antes del examen tienen un r_{pb} muy alto. Originalmente, se especulaba que estas preguntas no discriminarían efectivamente porque se esperaba que todo el mundo las contestara correctamente. Sin embargo, los datos indican que las personas distraidas no notarán que la pregunta es la misma de un ejemplo reciente visto en clases.

Se espera poder continuar el desarrollo de *The Examiner* hasta convertirlo en un producto completo. Hay potencial para usar capacidades de *computer vision* para automatizar aún más la fase de revisión ([4], [48]). Hay que difundirlo entre las comunidades docentes nacionales e internacionales. En particular, *The Examiner* tendría un uso muy efectivo en cursos colegiados con números altos de estudiantes.

REFERENCIAS

[1] D. Ariely, "*Predictably Irrational: The Hidden Forces that Shape our Decisions*", Harper Perennial, New York, U.S.A., 2009.

[2] D. Ariely, "*The (Honest) Truth about Dishonesty*", Harper Perennial, New York, U.S.A., 2012.

[3] R. J. de Ayala , "*The Theory and Practice of Item Response Theory*", The Guilford Press, New York, U.S.A., 2009.

[4] D. L. Baggio, D. M. Escrivá, N. Mahmood, R. Shilkrot, S. Emami, K. Levgen, J. Saragih, "*Mastering OpenCV with Practical Computer Vision Projects*", PACKT Publishing, Birmigham, U. K., 2012.

[5] F. B. Baker, "*The Basics of Item Response Theory*". ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD., 2001.

[6] F. B. Baker, S-H. Kim, "*Item Response Theory Parameter Estimation Techniques*", Second Edition, Revised and Expanded, Marcel Dekker, Inc., CRC Press, Boca Raton, U.S.A., 2004.

[7] P. De Boeck, M. Wilson, "*Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*", New York: Springer. 2004.

[8] L. Cronbach, "*Coefficient alpha and the internal structure of tests*", Psychometrika, 16 (3): 297-334, 1951.

[9] C. DeMars, "*Item Response Theory Understanding Statistics Measurement*", Oxfor University Press, New York, U.S.A., 2010.

[10] S. Embretson, S. Reise, "*Item response theory for psychologists*", Mahwah, NJ: Erlbaum. 2000.

[11] R. F. DeVellis, "*Scale Development Theory and Applications*", Third Edition, SAGE, U.S.A., 2012.

[12] R. L. Ebel, D. A. Frisbie, "*Essentials of educational measurement*". Englewood Cliffs: Prentice-Hall; 1991.

[13] C. Elvin, "*Test Item Analysis Using Microsoft Excel Spreadsheet Program*", `http://www.eflclub.com/elvin/publications/2003/itemanalysis.html`, 2003.

[14] FlexMIRT, `http://flexmirt.vpgcentral.com/`, 2016.

[15] R. Foschi, E. Cicciola, "*Politics and naturalism in the 20th century psychology of Alfred Binet*", History of psychology 9 (4): 26789, Nov. 2006.

[16] J-P. Fox, "*Bayesian Item Response Modeling: Theory and Applications*", Springer, New York, U.S.A., 2010.

[17] G. Grätzer, "*First Steps in LATEX*", Birkhäuser & Springer-Verlag, Boston, U.S.A., 1999.

[18] B. Grün, A. Zeileis, "*Automatic Generation of Exams in R*", Journal of Statisticak Software, Vol. 29, Issue 10, Feb 2009.

[19] L. Guttman, "*A Basis for Analyzing test-retest reliability*", Psychometrika, 10 (4): 255-282, 1945.

[20] R. K. Hambleton, H. Swaminathan, H. J. Rogers, "*Fundamentals of Item Response Theory*", Newbury Park, CA: Sage Press, 2009.

[21] T. Hogan, "*Relationship between Free-Response and Choice-Type Tests of Achivement: A Review of the Literature*", ERIC Document Reproduction SErvice No. ED 224811.

[22] C. Hoyt, "*Test Reliability Estimated by Analysis of Variance*", Psychometrika, 6 (3): 153-160, 1941.

[23] R. K. Hambleton, H. Swaminathan, "*Item Response Theory: Principles and Applications*", Kluwer-Nihjoff Publishing, Norwell, U.S.A., 2010.

[24] B. Hanson, "*IRT Command Language (ICL)*", `http://www.b-a-h.com/software/irt/icl/index.html`, 2002.

[25] G. Hernández, L. M. Rodríguez, M. G. Antón, E. J. Muñoz-Martínez, G. Duval, "*Filosofía de la Experiencia y Ciencia Experimental*", Fondo de Cultura Económica, México, 2003.

[26] P. Horst, "*L. L. Thurstone and the Science of Human Behavior*", Science 122 (3183): 125960, 1955.

[27] H. Jebek, "*Paul Lazarsfeld The Founder of Modern Empirical Sociology: A Research Biography*", International Journal of Public Opinion Research 13:229-244, 2001.

[28] H. Kopka, P. W. Daly, "*A Guide to LaTeX*", 3rd ed. Harlow, England: Addison-Wesley, 1999.

[29] J. Lazar, J. H. Feng, H. Hochheiser, "*Research Methods in Human-Computer Interaction*", Wiley Publications, Glasgow, Great Britain, 2010.

[30] P.F. Lazarsfeld, N. W. Henry, "*Latent Structure Analysis*", Boston: Houghton Mifin, 1968.

[31] R. Levesque, "*SPSS Programming and Data Management: A Guide for SPSS and SAS Users*", Fourth Edition, SPSS Inc., Chicago Ill. 2007.

[32] John Michael Linacre, "*Diseño de mejores pruebas, utilizando la Técnica de Rasch*", Ponencia Magistral de III Foro Nacional de Evaluación Educativa, 29 de octubre de 1998, Veracruz, México, MESA Memo # 68, 1998.

[33] F. M. Lord, "*Applications of item response theory to practical testing problems*", Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 1980.

[34] O. Martin, "*Psychological measurement from Binet to Thurstone, (1900-1930)*", Revue de synthese (4): 45793, 1997.

[35] P. Meyer, `http://www.itemanalysis.com/`, 2007.

[36] Minitab Inc., "*Minitab Statistical Software*", `https://www.minitab.com/en-us/`, 2016.

[37] National Council on Measurement in Education `http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorI`

[38] M. Nering, R. Ostini (eds.), "*Handbook of Polytomous Item Response Theory Models*", Routledge Taylor & Francis Group, New York, U.S.A., 2010.

[39] D. G. Patterson, "*Do new and old type examinations measure different mental functions?*", School and Society, 24, 246-248.

[40] Postgres, "*The world's most advanced open source database*", `https://www.postgresql.org/`, 2017.

[41] G. Prieto y A. R. Delgado, "*Análisis de un test mediante el modelo de Rasch*", Psicothema 2003, vol. 15 n 1, pp. 94-100, ISSN 0214 - 9915 CODEN PSOTEG.

[42] R Foundation, "*The R Project for Statistical Computing*", `https://www.r-project.org/`, 2016.

[43] G. Rasch, "*Probabilistic models for some intelligence and attainment tests*". Copenhagen, Danish Institute for Educational Research, The University of Chicago Press. 1980.

[44] C. F. Sheu, C. T. Chen, Y. H. Su, W. C. Wang, "*Using SAS PROC NLMIXED to t item response theory models*", PubMed, `http://www.ncbi.nlm.nih.gov/pubmed/16171193`, 2005.

[45] Scientic Software International, `http://www.ssicentral.com/irt/index.html`, 2003.

[46] D. Shenk, "*The Genius in All of Us New Insights into Genetics, Talent, and IQ*", Anchor Books, New York, U.S.A., 2010.

[47] SPSS Inc., "*SPSS 15.0 Command Syntax Reference*", Chicago Ill. 2006.

[48] R. Szeliski, "*Computer Vision Algorithms and Applications*", Springer-Verlag, Londond, England, 2011.

[49] W. J. Van der Linden, R. K. Hambleton, R.K., "*Handbook of modern item response theory*", New York: Springer. 1997.

[50] M. R. C. van Dongen, "*LATEX and Friends*", X.media.publishing Springer, Berlin, Alemania, 2012.

[51] M. Wilson, "*Constructing Measures An Item Response Modeling Approach*", Psychology Press Taylor & Francis Group, New York, U.S.A., 2005.