NOTAS
DE
CLASE

*Educación Matemática*
*Vol. 11 No. 3 Diciembre 1999*
*pp. 104-119*

# Valoración de los ejercicios en las pruebas de rendimiento escolar[1]

**Fecha de recepción:** Noviembre, 1998

**Horacio Félix Attorresi, María Silvia Galibert y María Ester Aguerri**
Instituto de Investigaciones, Facultad de Psicología
Universidad de Buenos Aires, Buenos Aires, Argentina
hattorre@psi.uba.ar, galibert@psi.uba.ar,
maguerri@psi.uba.ar

**Resumen.** *El presente trabajo trata sobre el modo en que conviene ponderar los ítemes en la elaboración de la calificación para un mejor aprovechamiento de la información obtenida con una prueba de rendimiento. Se explica la metodología después de introducir algunos conceptos elementales y se ilustra con un ejemplo de aplicación a datos reales. La metodología que se propone se fundamenta en uno de los modelos de la Teoría de Respuesta al Ítem (TRI), que constituye un enfoque psicométrico más moderno que la Teoría Clásica de Tests (TCT). El propósito de este trabajo es explotar las ventajas de una y otra y usar la relación entre ellas para obtener una buena ponderación de los ítemes de una prueba. La metodología consiste en utilizar los pesos óptimos que provee la TRI y calcularlos aproximadamente usando coeficientes definidos en el marco de la TCT.*

**Abstract.** *This paper describes the most convenient method for item weighing to be applied when a score system is devised, in order to attain a better use of the information collected with a performance test. The methodology is explained after introducing some basic concepts and an example of application to actual data is provided. The methodology advanced is based on one of the Item response Theory (IRT) models, which is a more up-to-date psychometric approach than the Classical Tests Theory (CTT). The objective of this work is to take advantage of the best aspects of both theories and to use their relationship so as to reach a good weighing of test items. The methodology consists in using the optimum weights provided by the IRT and to calculate them using coefficients defined within the frame of the CTT.*

## Introducción

Una etapa importante en el proceso de enseñanza - aprendizaje es la evaluación del rendimiento. Sin duda el "buen criterio" del profesor es insustituible en la elección de

[1] La investigación que se presenta en este artículo fue realizada con subsidios de la Universidad de Buenos Aires (UBACyT TP02), del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET PIP4423) y de la Agencia Nacional de Promoción Científica y Tecnológica (PICT04-04704).

los contenidos por evaluar, en el modo de formulación de las preguntas y en la manera de asignar una calificación en función de las respuestas de los alumnos; pero algunas técnicas que proporciona la psicometría pueden mejorar considerablemente la calidad del proceso de evaluación. Ha de tenerse conciencia de que dicha evaluación es el resultado de una "medición" cuya calidad depende no sólo de la bondad del instrumento —la prueba— sino también del aprovechamiento de la información obtenida de la misma a la hora de asignar las calificaciones.

La construcción de buenos instrumentos de medición requiere de dos etapas: la elaboración propiamente dicha, cuyos criterios de diseño pueden hallarse en Ebel (1977), Mehrens y Lehmann (1982), y el análisis posterior de la misma a partir de los resultados observados una vez que fue administrada a un conjunto de sujetos. El análisis de la prueba se lleva a cabo con auxilio de la estadística y el más sencillo, accesible aún para legos en la materia, consiste en examinar algunos índices básicos que provee la Teoría Clásica de Tests (TCT). Ésta se desarrolló sobre la base del modelo lineal de puntuaciones de Spearman (1904) y ha alcanzado gran difusión. Al respecto pueden consultarse Cortada de Kohan (1968), Muñiz (1992) y Santisteban Requena (1990).

Una vez que la prueba es administrada, el profesor debe decidir qué calificación asignar a cada alumno según su desempeño. Un criterio usual y simple es clasificar cada respuesta como correcta o incorrecta y obtener la calificación como el número de respuestas correctas transformada en la escala vigente de calificaciones; por ejemplo, si la prueba consiste de veinte ejercicios y la escala de calificaciones es de 0 a 10, al alumno que tiene 14 respuestas correctas se lo califica con 7. Este criterio otorga a cada ejercicio o ítem el mismo peso: todos valen igual. Pero si las preguntas tienen distintos grados de dificultad el profesor puede considerar razonable asignarles puntajes diferentes y darles más "peso" en el puntaje total a unos ítemes que a otros. Los profesores suelen asignar estos pesos con un criterio que depende fundamentalmente de su experiencia, por cierto valiosa, pero no totalmente objetiva.

El presente trabajo trata sobre el modo en que conviene valorar cada ejercicio para un mejor aprovechamiento de la información obtenida con la prueba. Se explica la metodología después de introducir algunos conceptos elementales y se ilustra con un ejemplo de aplicación a datos reales. La metodología que se propone se fundamenta en uno de los modelos de la Teoría de Respuesta al Ítem (TRI). Esta teoría constituye un enfoque psicométrico más moderno que la TCT. Sobre la base de trabajos pioneros, fue desarrollada por separado por Birnbaum (1968) y Rasch (1960). No es el propósito de este trabajo ahondar en la comparación de ambas teorías sino explotar las ventajas de una y de otra, más aún, la relación entre ellas, para obtener una conveniente valoración de cada ejercicio de una prueba. Si bien la TRI es desde el punto de vista teórico superior a la TCT y posibilita interesantes aplicaciones que detalla Lord (1980), en contrapartida requiere de un trabajo computacional mayor que hace imprescindible el uso de softwares específicos —fuera del alcance de la mayoría de los profesores— y de un gran número de sujetos a los cuales administrar la prueba. Sin embargo, bajo ciertas condiciones que serán señaladas oportunamente, ambas teorías pueden relacio-narse y así aproximar ciertos resultados de la TRI, difíciles de obtener, mediante la TCT. La metodología consistirá, por tanto, en utilizar los pesos óptimos que provee la TRI y calcularlos aproximadamente usando coeficientes definidos en el marco de la TCT.

## Modelos psicométricos

**Modelo Lineal de Puntuaciones.** Introducido por Spearman (1904), es la base de la Teoría Clásica de Tests. Su formulación es:

$$X = V + \varepsilon$$

donde

X es el puntaje observado de un individuo, elegido al azar de una población, al administrarle un test.
V es el puntaje verdadero que le correspondería a dicho individuo en el test si la medición se realizara sin error.
$\varepsilon$ es el error de medición.

Fijado el sujeto, su puntaje verdadero es un valor fijo pero tanto el puntaje observado como el error de medición varían aleatoriamente.

## Definición

Se dice que dos conjuntos de puntuaciones X y X' son medidas paralelas si se cumple que

$$X = V + \varepsilon \quad y \quad X' = V' + \varepsilon' \quad con \quad V = V' \quad y \quad Var(\varepsilon) = Var(\varepsilon')$$

## Supuestos

$E(\varepsilon/V) = 0$
$Var(\varepsilon/V) = Var(\varepsilon)$
$Cov(\varepsilon, \varepsilon') = 0$ para toda medida paralela X'

Donde E(/), Var(/) y Cov son la esperanza condicional, varianza condicional y covarianza matemáticas respectivamente.

## Inferencias acerca de las puntuaciones verdaderas

Si a los supuestos generales del modelo lineal se agrega el de la distribución normal bivariada de sus componentes X y V, pueden inferirse las puntuaciones verdaderas a partir de las observadas mediante la siguiente ecuación de regresión lineal:

$$V = E(V/X) = \rho_{XV}\frac{\sigma_V}{\sigma_X}(X - \mu_X) + \mu_V \qquad [1]$$

donde $\hat{V}$ es el puntaje inferido mediante la ecuación de regresión, es la correlación lineal entre los puntajes observados y los verdaderos, $\mu_x$, $\mu_v$, $\rho_x$ y $\rho_v$ son sus correspondientes medias y desviaciones estándar.

## Índices característicos de los ítemes que componen un test

Son parámetros característicos de los ítemes que se utilizan en la etapa de construcción del test llamada "análisis de ítemes". Sobre la base de estos resúmenes se seleccionan los mejores ítemes para integrar un test a partir de pruebas piloto.

## Índice de dificultad del ítem i ($p_i$)

Es la frecuencia relativa de respuesta correcta al ítem i. La interpretación es obvia, ya que estima la proporción de sujetos en la población que es capaz de contestar correctamente al ítem. Ebel (1977) y Hurtado de Mendoza (1982) recomiendan llamarlo índice de facilidad.

## Coeficientes biserial puntual ($r_{iX}$), QEX y biserial ($r_{iX}$)

El coeficiente biserial puntual $r_{iX}$ es el coeficiente de correlación de Pearson[2] entre el ítem i (dicotómico) y el puntaje total en la prueba que lo contiene. Cuanto más alta es la correlación mejor representa el ítem a la prueba. Sin embargo, esta correlación se ve favorecida por la presencia del ítem en la prueba; para corregir esta influencia y tener una medida más real de su representatividad se calcula el índice QEX (Question Evaluation Index) que es el coeficiente de correlación Biserial Puntual entre el ítem y los puntajes totales de la prueba excluyendo al ítem.

El coeficiente biserial es la correlación de Pearson entre un ítem que no es dicotómico por naturaleza sino una dicotomización de una variable continua con distribución normal. Se relaciona con el biserial puntual mediante la siguiente expresión:

$$r_{iX} = r_{iX}\sqrt{p_i(1 - p_i)}/y$$

donde $y$ es la ordenada correspondiente al valor de la puntuación típica en la curva normal que deja por debajo un área igual a $1 - p_i$.

Una manera sencilla de calcularlo es mediante la fórmula

$$r_{iX} = \frac{(\overline{X}_a - \overline{X}) \times p_i/y}{S_x} \qquad [2]$$

donde:

- $\overline{X}_a$ es el promedio de la cantidad de respuestas correctas de los individuos que contestaron bien al ítem.
- $\overline{X}$ es el promedio de la cantidad de respuestas correctas de los sujetos.
- $p_i$ es el índice de dificultad definido por la TCT: el número de sujetos que respondieron correctamente al ítem dividido el total de sujetos.
- En la tabla del apéndice se encuentra el valor de $p/y$ conociendo $p$. $y$ es la ordenada correspondiente a la abscisa $z$ en una distribución normal estándar que deja debajo de sí un área igual a $1 - p_i$.

[2] El coeficiente de correleación de Pearson es una medida entre –1 y 1 para cuantificar el grado de relación lineal entre dos variables.

- $S_x$ es el desvío estándar de la cantidad de respuestas correctas, cuya fórmula es

$$S_x = \sqrt{\frac{\Sigma X_1^2 - nX^2}{n}} \qquad [3]$$

donde:

- $X_i$ es la cantidad de respuestas correctas del individuo i.
- $X$ es el promedio de la cantidad de respuestas correctas de los sujetos.

El coeficiente biserial varía entre –1 y 1. Cuanto más alta es la correlación mejor representa el ítem a la prueba.

## Índice de discriminación ($ID_i$)

Es de esperar que un ítem bien construido sea contestado correctamente con mayor frecuencia por los sujetos de puntajes superiores que por los de puntajes inferiores. Por ello se define

$$ID_i = (S\text{-}I)/T$$

donde S e I son el número de sujetos que contestan correctamente el ítem i correspondientes a los grupos de sujetos con puntajes superiores e inferiores respectivamente y T es la cantidad de sujetos de cada grupo. Este índice varía entre -1 y 1. Un ítem se acepta cuando el índice de discriminación es positivo y es mejor cuanto más cercano a uno. Existen distintos índices de discriminación según la regla empleada para determinar la cantidad T. D'Agostino & Cureton (1975) demostraron que el 21,5% maximiza el índice de discriminación.

- **Modelo Logístico deTres Parámetros.** Es uno de los modelos de la Teoría de Respuesta al Ítem. Su formulación es:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}}$$

donde

$\theta$ es el rasgo latente que se desea medir con el ítem i, por ejemplo, el nivel de conocimientos alcanzados por un estudiante en un determinado tema. El origen y unidad de la escala de $q$ son arbitrarios; es decir, están indeterminados.

$P_i(\theta)$: *función característica o de respuesta al ítem i.* Es la probabilidad de contestar correctamente el ítem i para un nivel dado de $\theta$. Su gráfica se denomina *curva característica del ítem.*

$b_i$ : *índice de dificultad del ítem i.* Coincide con el valor $\theta$ necesario para tener probabilidad $0{,}5+c_i/2$ de contestar correctamente el ítem i.

$a_i$: *índice de discriminación del ítem i.* Es proporcional a la pendiente de la recta tangente en el punto de inflexión de la curva característica del ítem, salvo el factor $1\text{-}c_i$.

$c_i$: *índice por azar del ítem i.* Es la probabilidad de contestar correctamente el ítem i cuando el nivel de habilidad $\theta$ tiende a $-\infty$. (Valor asintótico de $P(\theta)$).

*D*: es un factor de escala. Para $D = 1,7$, las curvas logísticas de los modelos de uno y dos parámetros difieren de las correspondientes al modelos de de ojiva normal[3] en menos de 0,01 para todos los valores de $\theta$.

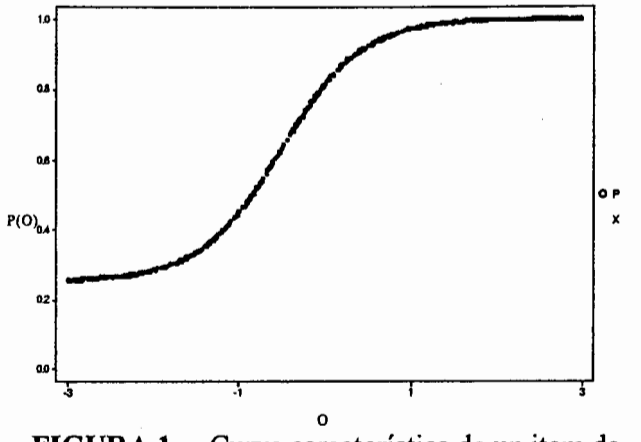


**FIGURA 1.** Curva característica de un item de parámetros a = 1,25 b = –0,5 c = 0,25

Cuando $c_i = 0$ se tiene el modelo de dos parámetros. En este caso el índice de dificultad $b_i$ corresponde al nivel de $\theta$ necesario para tener una probabilidad de 0,5 de contestar correctamente el ítem y $a_i$ resulta proporcional a la pendiente de la recta tangente en el punto de inflexión.

Si además todos los índices de discriminación coinciden, pueden considerarse iguales a 1, con lo que resulta el modelo de un parámetro o modelo de Rasch (1960).

Por rasgo latente se entiende una variable que por su carácter abstracto no es directamente observable; por ejemplo, algún tipo de capacidad como la habilidad matemática, la comprensión verbal, la inteligencia general o, como en la presente aplicación, el nivel de conocimientos (rendimiento) alcanzado por un estudiante. Un ítem será más difícil que otro si se requiere de un mayor nivel de habilidad o conocimiento para tener la misma probabilidad de responderlo correctamente; de allí que *b* exprese el índice de dificultad del ítem, ya que puede interpretarse como el nivel de rendimiento requerido para tener 0,5 de probabilidad de respuesta correcta en los modelos de uno o dos parámetros y $0,5 + c/2$ en el modelo de tres parámetros.

Cuanto mayor sea *a*, más variará la probabilidad de respuesta correcta por unidad de cambio en el nivel de rendimiento $\theta$, lo que le da sentido a su interpretación como índice de discriminación.

Finalmente, *c* refleja la probabilidad de contestar correctamente por azar; por cuanto a niveles nulos de rendimiento ($-\infty$) corresponde alguna probabilidad $c > 0$ de respuesta correcta.

[3] Modelo de Ojiva Normal de Tres Parámetros:

## Supuestos

### Independencia local

Significa que la probabilidad de una determinada respuesta correcta al ítem para un valor dado de $\theta$ coincide con la probabilidad de dicha respuesta al ítem para el mismo valor de $\theta$ y de respuesta a cualquier subconjunto de ítemes $i_1, i_2, \ldots, i_k$. En lenguaje matemático,

Si $U_i = u_i$, $i = 1, 2, \ldots, n$, es la variable Bernoulli asociada a la respuesta al ítem $i$, entonces $P(U_i = u_i/\theta) = P(U_i = u_i/\theta, u_{i_1}, \ldots u_{i_k})$ $(i \neq i_1, \ldots, i_k)$ o, equivalentemente, $P(U_{i_1} = u_{i_1}, \ldots, U_{i_k} = uik/\theta) = P(U_{i_1} = u_{i_1}/\theta). \ldots .P(U_{i_k} = u_{i_k}/\theta)$ $(i_j \neq i_k$ si $j \neq k)$

### Unidimensionalidad

La probabilidad de contestar correctamente a un ítem depende sólo de un factor, que es el rasgo latente. Por tanto, dicha probabilidad queda determinada para cada sujeto por su nivel del rasgo latente, nivel que se desea estimar por el proceso de medición.

Lo contrario de la unidimensionalidad es el "sesgo" de un ítem; esto es, que de dos sujetos con el mismo nivel del rasgo latente, uno tenga más probabilidad de responder correctamente que otro debido a que hay otro factor implicado.

Lord (1980) afirma que la unidimensionalidad no es un supuesto adicional a la independencia local, sino que ésta última se deriva necesariamente de aquélla.

## Funciones de Información

La TRI permite obtener las funciones de información de los puntajes en una prueba y de cada ítem. La función de información del ítem i es

$$I_i(\theta) = P'^2_i(\theta) / P_i(\theta)Q_i(\theta)$$

donde $P'_i(\theta)$ es la derivada de la función característica $P_i(\theta)$ y $Q_i(\theta)$ es $1\text{-}P_i(\theta)$.

La función de información del puntaje total en la prueba es la suma de las funciones de información de los ítemes que la componen: $\Sigma P'^2_i(\theta)/ P_i(\theta)Q_i(\theta)$.

Ellas indican para qué niveles de rendimiento q de los sujetos, el puntaje asignado constituye una medición más precisa. Por ejemplo; si la función de información del puntaje en una prueba tuviera la forma de la figura 2, indicaría que medirá con mayor precisión a los sujetos de rendimiento medio.

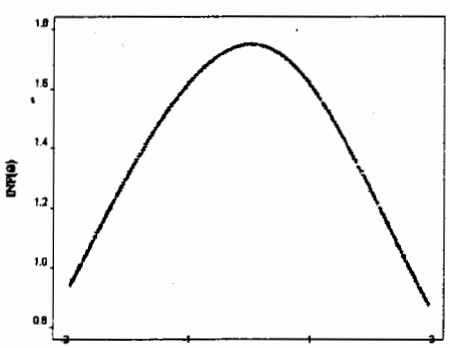


**FIGURA 2.** Función de información de un ítem

## Aplicaciones de la Teoría de Respuesta al Ítem

En los modelos de la Teoría de Respuesta al Ítem se caracteriza a cada ítem en función de sus parámetros propios independientemente de cómo se distribuya el rasgo latente en la población de sujetos que los contestan y de cuáles sean las características del resto de los ítemes que componen el test mientras sean unidimensionales; es decir, mientras midan todos al mismo rasgo en una sola dimensión. Del mismo modo, la medida $q$ característica de cada sujeto no depende de los parámetros de los ítemes ni de las medidas de los otros sujetos de la población. Es precisamente esta independencia entre el instrumento de medición y los sujetos de la misma lo que hace a la diferencia esencial entre los enfoques de la TCT y de la TRI y constituye su principal ventaja ya que, si se satisfacen los supuestos, se sigue que:

- Fijada una escala para $\theta$, salvo fluctuaciones de muestreo, se obtiene la misma estimación del rendimiento de un sujeto cuando es medido por diferentes subconjuntos de ítemes, aún cuando éstos difirieran por ejemplo, en sus índices de dificultad. En la TCT, en cambio, la medida del sujeto es el puntaje verdadero V en un determinado test, la cual será más alta si el test es fácil que si es difícil.
- La estimación del rendimiento de un sujeto depende sólo de sus respuestas a los ítemes y no de las características del nivel de rendimiento en la población de respondentes, como ocurre en la TCT cuando se infieren los puntajes verdaderos mediante la expresión [1], que depende de la media y desvío estándar de los puntajes verdaderos en la población.
- A su vez, los parámetros de los ítemes tampoco dependen de los rendimientos de los sujetos que los responden como ocurre en la TCT donde, por ejemplo, el índice de dificultad —proporción de respuestas correctas al ítem— variaría entre dos poblaciones de estudiantes que, en promedio, difirieran en su desempeño.

Este enfoque permite, pues, interesantes aplicaciones que detalla Lord (1980); entre ellas:

- Elaboración de bancos de ítemes. Un banco de ítemes es un conjunto de ítemes calibrados; esto es, ítemes cuyos parámetros ya han sido estimados. Disponer de un banco de ítemes permite seleccionar subconjuntos de ítemes para elaborar tests con características prefijadas (p.ej. tests para obtener estimaciones más precisas de la habilidad en determinados rangos de la escala) y elaborar tests adaptativos o "a la medida de los sujetos", eligiendo para cada sujeto los ítemes que arrojarán mediciones más precisas de su rasgo. Estas aplicaciones se basan en las *funciones de información* de los ítemes y del test.
- Estudio del Funcionamiento Diferencial del Ítem. El sesgo de los tests es un problema clásico y se refiere a la posibilidad de que ciertos sujetos se vean desfavorecidos con respectos a otros al ser medidos en una determinada habilidad, no por razón de su capacidad sino de su pertenencia a cierta población, por ejemplo, por su sexo, edad, raza, etc. La formulación de los modelos de la TRI mediante la función característica permite dar una definición precisa del sesgo: un ítem es sesgado (o

tiene un funcionamiento diferencial) si su función característica no es la misma entre poblaciones; y se proponen diversos métodos estadísticos para abordar este estudio. Eliminando los ítemes con DIF es posible satisfacer la invarianza pretendida por la TRI.

## Condiciones de aplicabilidad de la metodología

- Las pruebas deben consistir de ítemes dicotómicos: cada respuesta es clasificada como satisfactoria o no satisfactoria. En el contexto de las pruebas de rendimiento escolar, los ítemes son las cuestiones que el alumno debe resolver: problemas, ejercicios, preguntas, etc. Las preguntas pueden ser de respuesta abierta o, si se trata de pruebas de elección múltiple, deberán elegirse por lo menos cuatro opciones a fin de que la probabilidad de responder bien por azar sea pequeña para no sobreestimar el nivel, especialmente de aquellos sujetos de menores rendimientos. En otras palabras, se requiere que los datos se adecuen al modelo de dos parámetros.
- Los ítemes deben ser en lo posible independientes unos de otros: la probabilidad que tienen los individuos de un determinado nivel de rendimiento de contestar bien un ítem no aumenta ni disminuye si responden bien a otro ítem. Por ello deben ser evitados aquellos ítemes cuyas respuestas puedan ser deducidas del conocimiento de otros ítemes o que tengan similar estrategia específica de resolución.
- El nivel de conocimiento alcanzado por los sujetos se distribuye normalmente. Esto requiere que la mayoría de los alumnos haya alcanzado un rendimiento próximo al rendimiento medio y una minoría rendimientos excesivamente altos o bajos.
- Unidimensionalidad: la performance en el ítem debe depender sólo del rasgo que se desea medir; en el presente caso el rendimiento, y no de otros factores. Es decir; si dos sujetos tienen el mismo nivel de rendimiento, deben tener la misma probabilidad de contestar bien el ítem; aun cuando pertenezcan a poblaciones muy diferentes. Por ejemplo: Supongamos que para medir la comprensión lectora se construye un ítem basado en un relato de una escena que transcurre en el campo. Posiblemente un niño de la ciudad tenga menor probabilidad de contestarlo correctamente que un niño del campo, aun cuando ambos tengan la misma capacidad de comprensión lectora. Esto ocurriría porque el éxito en la resolución no depende sólo de la capacidad de comprensión lectora que se desea medir sino de otro factor: la familiaridad con el ambiente en el que transcurre la escena. Este tipo de situaciones suelen darse en variables de difícil acceso a la observación directa como son los distintos tipos de habilidades: razonamiento lógico, comprensión lectora, capacidad para discriminar relaciones, etc. En las pruebas de rendimiento no es tan problemático lograr la unidimensionalidad porque el nivel de conocimientos alcanzado es una variable que puede observarse directamente: la cantidad de conocimientos alcanzados.

## Pesos óptimos de los ítemes en el puntaje total

El concepto de función de información proporciona un criterio para definir qué se entiende por "pesos óptimos": serán óptimos aquéllos que la maximicen; en otras palabras, serán óptimos los pesos que arrojen mediciones lo más precisas posibles.

Lord (1980) demuestra, en una de las aplicaciones de los modelos de la TRI, que los pesos óptimos de los ítemes, bajo las condiciones antes mencionadas, los constituyen los coeficientes de discriminación de la TRI. El puntaje de cada alumno resultará, pues, de sumar los índices de discriminación de los ítemes que ha contestado correctamente.

Cuando dichos índices no pueden ser calculados, ya por carecerse del software específico, ya porque la muestra es pequeña, dichos índices pueden aproximarse mediante la siguiente expresión:

$$a_i = \frac{r_{iX}}{\sqrt{1 - r_{iX}^2}} \qquad [4]$$

donde $r_{iX}$ es el coeficiente biserial del ítem i.

## Procedimiento para el cálculo

Las fórmulas que anteceden están basadas en el cálculo de promedios y desvíos estándar que actualmente se llevan a cabo con calculadoras de bolsillo o con softwares estadísticos de fácil manipulación. Con todo, ejemplificaremos su cálculo manual sobre las calificaciones de 34 alumnos obtenidas en una prueba de elección múltiple de 14 ítemes.

Conviene presentar los resultados de la evaluación en una matriz de unos y ceros, donde los unos significan que los ítemes fueron bien respondidos y cero lo contrario. Cada fila de la matriz es el patrón de respuestas de un sujeto. El primer sujeto contestó correctamente las preguntas 1, 2, 3, 5, 6, 8, 9, 13 y 14, es decir, tiene 9 respuestas correctas.

| | Ítemes | | | | | | | | | | | | | | Total de respuestas |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | correctas por sujeto |
| Sujeto 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 9 |
| Sujeto 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 10 |
| Sujeto 3 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 9 |
| Sujeto 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 13 |
| Sujeto 5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 9 |
| Sujeto 6 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 9 |
| Sujeto 7 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 9 |
| Sujeto 8 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 10 |
| Sujeto 9 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 4 |
| Sujeto 10 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5 |
| Sujeto 11 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 7 |
| Sujeto 12 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 12 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sujeto 13 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 8 |
| Sujeto 14 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 7 |
| Sujeto 15 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 9 |
| Sujeto 16 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 6 |
| Sujeto 17 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 11 |
| Sujeto 18 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 6 |
| Sujeto 19 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 10 |
| Sujeto 20 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 7 |
| Sujeto 21 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 8 |
| Sujeto 22 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 5 |
| Sujeto 23 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 12 |
| Sujeto 24 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 8 |
| Sujeto 25 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 8 |
| Sujeto 26 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 5 |
| Sujeto 27 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 11 |
| Sujeto 28 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 12 |
| Sujeto 29 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 |
| Sujeto 30 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 6 |
| Sujeto 31 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 8 |
| Sujeto 32 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 6 |
| Sujeto 33 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 7 |
| Sujeto 34 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 7 |
| Cantidad de aciertos al ítem. | 25 | 22 | 24 | 17 | 23 | 26 | 19 | 8 | 19 | 21 | 7 | 21 | 15 | 29 | |
| Indice $p_i$ | .735 | .647 | .706 | .500 | .676 | .765 | .559 | .235 | .559 | .618 | .206 | .618 | .441 | .853 | |

## Cálculo del promedio de la cantidad de respuestas correctas: $X$

Se suman todos los valores de la columna del total de respuestas correctas por sujeto y se lo divide por el número de sujetos.

$$S\,X_i = 9 + 10 + 9 + 13 + 9 + 9 + 9 + \ldots + 6 + 7 + 7 = 276$$

$$X = 276 / 34 = 8.118$$

## Cálculo del desvío estándar de la cantidad de respuestas correctas: $S_x$

Se suman los cuadrados de los valores de la columna del total de respuestas correctas y se reemplaza en la fórmula [3].

$$\Sigma X_i^2 = 9^2 + 10^2 + 9^2 + 13^2 + 9^2 + \ldots + 6^2 + 7^2 + 7^2 = 2438$$

$$S_x = \sqrt{\frac{2438 - 34 \times 8.118^2}{34}} = 2.409$$

Se ejemplificará el cálculo del peso óptimo aproximado para el primer ítem.

Los individuos que contestaron bien el ítem 1 tienen un 1 en la primera columna: son los sujetos 1, 3, 4, 5, 7, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 27, 28, 32, 33 y 34.

Promediando la cantidad de respuestas correctas de este grupo se obtiene

$$\overline{X}_a = (9 + 9 + 13 + ... + 7 + 7) / 25 = 215/25 = 8.600$$

La proporción de respuestas correctas para el ítem 1 es

$$p_1 = 25 / 34 = 0.7353 \cong 0.74$$

Según la tabla del apéndice resulta que $p/y = 2.281$

Reemplazando estas cantidades en la fórmula [2] se obtiene el coeficiente de correlación biserial para el ítem 1.

$$r_{1X} = \frac{(8.600 - 8.118) \times 2.281}{2.409}$$

Sustituyendo en la fórmula [4] se obtiene el índice de discriminación aproximado de la TRI para el ítem 1, el que constituirá su peso.

Operando de la misma manera se obtienen los pesos para los otros ítemes. Éstos son:

| Ítem | Peso |
|---|---|
| 1 | 0.512 |
| 2 | 0.592 |
| 3 | 1.165 |
| 4 | 0.356 |
| 5 | 0.420 |
| 6 | 1.055 |
| 7 | 0.514 |
| 8 | 0.294 |
| 9 | 0.775 |
| 10 | 0.741 |
| 11 | 0.094 |
| 12 | 0.878 |
| 13 | 0.578 |
| 14 | 0.576 |
| Peso total | 8.558 |

El puntaje de un sujeto es la suma de los pesos de los ítemes que contestó correctamente. Así, como el sujeto 1 contestó correctamente los ítemes 1, 2, 3, 5, 6, 8, 9, 13 y 14, le corresponde un puntaje igual a

$$0.512 + 0.592 + 1.165 + 0.420 + 1.055 + 0.294 + 0.775 + 0.578 + 0.576 = 5.971$$

Observemos que el sujeto 5 también tiene 9 respuestas correctas y sin embargo el puntaje que le corresponde (6.656) es diferente ya que su patrón de respuestas lo es.

## Transformación a la escala de calificaciones

Una vez asignados los puntajes es preciso transformarlos a la escala usual de calificaciones. Para ello Cano de Becerra (1971) propone diferentes e interesantes criterios que toman en cuenta el desempeño general del grupo. Aquí ejemplificaremos con el criterio más sencillo que consiste en hacer corresponder el máximo puntaje posible -suma de todos los pesos- con la mayor nota en la escala de calificaciones y el mínimo puntaje -cero- con el mínimo de la escala.

Por ejemplo, si la escala de calificaciones va de cero a diez el problema se resuelve aplicando la sencilla regla de tres simple, se hace corresponder el peso total 8.558 con el diez. La constante de proporcionalidad es 10 / 8.558 = 1.168. Cada calificación se obtiene multiplicando el puntaje de cada sujeto por esta constante. Luego podrán redondearse a la precisión deseada.

Para el sujeto 1 la calificación será

$$5.971 \times 1.168 = 6.975$$

y para el sujeto5, $6.656 \times 1.168 = 7.778$

## Resultados

Se muestran en la siguiente tabla. Para cada sujeto se registra:

- En la primera columna su puntaje bruto, esto es, la cantidad de respuestas correctas (cada ítem vale lo mismo).
- En la segunda columna, el puntaje bruto expresado en la escala de calificaciones de 0 a 10.
- En la tercera columna, el puntaje pesado; es decir, la suma de los pesos de los ítemes que cada sujeto contestó correctamente.
- En la cuarta columna, el puntaje pesado expresado en la escala de calificaciones de 0 a 10.

| SUJETO | PUNTAJE BRUTO | PUNTAJE BRUTO ESCALADO DE 0 A 10 | PUNTAJE PESADO | PUNTAJE PESADO ESCALADO DE 0 A 10 |
|---|---|---|---|---|
| 1 | 9 | 6.429 | 5.971 | 6.975 |
| 2 | 10 | 7.143 | 6.631 | 7.747 |
| 3 | 9 | 6.429 | 6.458 | 7.546 |
| 4 | 13 | 9.286 | 8.264 | 9.655 |
| 5 | 9 | 6.429 | 6.657 | 7.778 |
| 6 | 9 | 6.429 | 6.322 | 7.387 |
| 7 | 9 | 6.429 | 6.023 | 7.037 |
| 8 | 10 | 7.143 | 7.237 | 8.455 |
| 9 | 4 | 2.857 | 2.341 | 2.735 |
| 10 | 5 | 3.571 | 3.891 | 4.546 |

| | | | | |
|---|---|---|---|---|
| 11 | 7 | 5.000 | 3.958 | 4.624 |
| 12 | 12 | 8.571 | 7.426 | 8.677 |
| 13 | 8 | 5.714 | 5.637 | 6.585 |
| 14 | 7 | 5.000 | 5.146 | 6.012 |
| 15 | 9 | 6.429 | 6.641 | 7.759 |
| 16 | 6 | 4.286 | 4.131 | 4.826 |
| 17 | 11 | 7.857 | 7.591 | 8.869 |
| 18 | 6 | 4.286 | 4.323 | 5.051 |
| 19 | 10 | 7.143 | 7.076 | 8.267 |
| 20 | 7 | 5.000 | 4.576 | 5.347 |
| 21 | 8 | 5.714 | 5.126 | 5.989 |
| 22 | 5 | 3.571 | 3.680 | 4.299 |
| 23 | 12 | 8.571 | 7.426 | 8.677 |
| 24 | 8 | 5.714 | 5.587 | 6.528 |
| 25 | 8 | 5.714 | 5.455 | 6.374 |
| 26 | 5 | 3.571 | 2.327 | 2.719 |
| 27 | 11 | 7.857 | 7.655 | 8.944 |
| 28 | 12 | 8.571 | 8.170 | 9.545 |
| 29 | 3 | 2.143 | 1.107 | 1.293 |
| 30 | 6 | 4.286 | 2.905 | 3.394 |
| 31 | 8 | 5.714 | 5.109 | 5.969 |
| 32 | 6 | 4.286 | 3.823 | 4.467 |
| 33 | 7 | 5.000 | 4.452 | 5.202 |
| 34 | 7 | 5.000 | 4.844 | 5.659 |

A manera ilustrativa se muestra a continuación cómo resultaron ordenados algunos de los sujetos según sus puntajes cuando éstos corresponden a pesos iguales y a pesos diferentes respectivamente.

**Pesos Iguales**

| Lugar en el ranking | Sujetos |
|---|---|
| 1 | 4 |
| 2 | 12, 23, 28 |
| 3 | 17, 27 |
| 4 | 2, 8, 19 |
| 5 | 1, 3, 5, 6, 7, 15 |

**Pesos diferentes**

| Lugar en el ranking | Sujetos |
|---|---|
| 1 | 4 |
| 2 | 28 |
| 3 | 27 |
| 4 | 17 |
| 5 | 12, 23 |
| 6 | 8 |
| 7 | 19 |

| | |
|---|---|
| 8 | 5 |
| 9 | 15 |
| 10 | 2 |
| 11 | 3 |
| 12 | 6 |
| 13 | 7 |
| 14 | 1 |

Si se ordenaran los sujetos con respecto a su calificación considerando pesos iguales resultarían varios grupos de sujetos "indistinguibles" como, por ejemplo, los sujetos 12, 23 y 28 en la segunda posición, 17 y 27 en la tercera, 2,8 y 19 en la cuarta, 1, 3, 5, 6, 7 y 15 en la quinta posición, etc. La ponderación de los ítemes, en cambio, permitió no sólo discriminar entre dichos sujetos sino también una corrección en el ranking. Así, el sujeto 27 pasó a estar por delante de los sujetos 12 y 23.

## Discusión

Esta metodología ofrece una manera objetiva de asignar los pesos a los ítemes de modo que el puntaje total rescata no sólo la cantidad de respuestas correctas sino su calidad; en otras palabras, no sólo tiene en cuenta cuántas sino cuáles. Esto permite discriminar mejor a los sujetos en función de su patrón de respuestas y obtener mediciones más precisas.

Puesto que el procedimiento demanda un trabajo computacional mayor que la manera usual de asignar calificaciones, es razonable plantearse hasta qué punto la ganancia en precisión justifica tal esfuerzo. No puede darse una respuesta en términos absolutos. En primer lugar dependerá del objetivo de la evaluación. Si es muy importante el aspecto competitivo -por ejemplo, una evaluación para otorgar becas a los mejores- deberá extremarse la precisión con el fin de discriminar y jerarquizar de la manera más justa posible a los individuos de mejores rendimientos. En segundo lugar, dependerá de las características propias de los ítemes; si todos ellos tienen una similar potencia discriminatoria, la calificación asignada por pesos óptimos no diferirá mucho de la que considera los pesos iguales. Finalmente, como la ganancia en precisión que aporta esta metodología, se pierde en parte por efecto del redondeo, no debe usarse una escala de calificaciones restringida a pocos valores enteros. Se recomienda el uso de una escala de 0 a 1000.

Por otra parte, el docente puede ir elaborando su propio banco de ítemes; esto es, ir acopiando la experiencia que obtiene de las sucesivas administraciones de los mismos. Si los grupos a los que se fue administrando la prueba guardan una cierta homogeneidad, como podría ser el caso de sucesivas promociones de un determinado curso y si los ítemes están bien construidos y son suficientemente representativos del contenido, éstos suelen tener un comportamiento similar entre los diferentes grupos en lo que hace a su poder de discriminación, por lo que no será necesario calcular los pesos cada vez, sino que podrán utilizarse los ya calculados. Es prudente, sin embargo, revisarlos cada tanto a fin de constatar que siguen vigentes o bien reactualizarlos.

El trabajo computacional se evita completamente si el docente aprende a trabajar (lo que sería muy deseable dada la importancia actual de la informática) con algún software estadístico sencillo, como por ejemplo el Statistix. El trabajo se reduce, entonces, a sólo cargar los datos y a efectuar algunas transformaciones sencillas.

## Referencias

**Birnbaum, A. (1968).** "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability" en Lord, F.M. y Novick, M.R., *Statistical Theories of Mental Test Scores. Reading*, Mass., Addison-Wesley.

**Cano de Becerra, F. (1971).** "Elementos de Estadística al Servicio de la Evaluación del Rendimiento". *Revista de Psicología*, 16 (1/2): 61-77.

**Cortada de Kohan, N. (1968).** *Manual para la Construcción de Tests Objetivos de Rendimiento.* Ed. Paidós, Buenos Aires.

**D'Agostino, R.B. & Cureton, E.E. (1975).** The 27 Percent Rule Revisited, *Educational and Psychological Measurement*, 35, 1975, pp. 45-50.

**Ebel, Rober L. (1977).** *Fundamentos de la Medición Educacional.* Ed. Guadalupe. Buenos Aires.

**Hurtado de Mendoza, Ma. (1982).** *Pruebas de Rendimiento Académico y Objetivos de la Instrucción.* Ed. Diana, México.

**Lord, F. (1980).** *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, N.J. Erlbaum.

**Mehrens, W. y Lehmann , I. (1982).** *Medición y Evaluación en la Educación y en la Psicología.* Ed. CECSA, México.

**Muñiz, J. (1992).** *Teoría Clásica de los Tests.* Ed. Pirámide, Madrid.

**Rasch, G. (1960).** *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

**Santisteban Requena, C. (1990).** *Psicometría.* Ed. Norma, Madrid.

**Spearman, C. (1904).** "The Proof and Measurement of Association betwen two Things". *American Journal of Psychology*, 15, 23 - 35.

[2] El coeficiente de correlación de Pearson es una medida entre -1 y 1 para cuantificar el grado de relación lineal entre dos variables.