



UNIVERSIDAD DE COLIMA
DIRECCIÓN GENERAL DE DESARROLLO DEL PERSONAL ACADÉMICO

Evaluación del desempeño docente usando Teoría de Respuesta al Ítem

Elaborado por
Dr. Osva Antonio Montesinos López

Teoría de respuesta al Ítem (IRT)

IRT es una técnica estadística muy popular en educación para la construcción y evaluación de cuestionarios, exámenes, instrumentos de medición con respuestas binarias (0,1) y ordinales (escala Likert).

IRT es útil para calibrar ítems en exámenes, cuestionarios y otro tipo de instrumentos de medición, así como para calcular la habilidad o actitud latente de cada individuo.

Teoría de respuesta al Ítem (IRT)

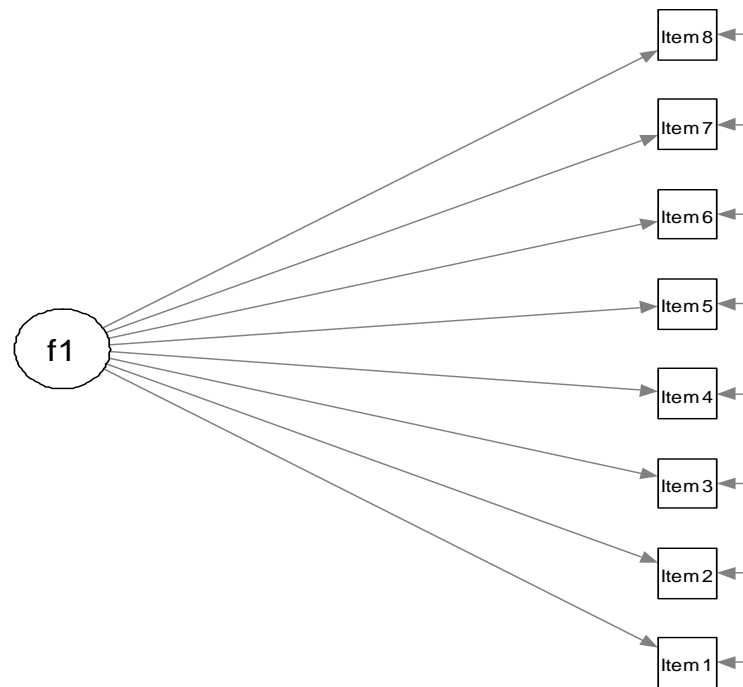


Figura 1. Cada individuo elige entre un número limitado de opciones de respuesta (binarias u ordinales) para cada ítem. **IRT liga estas respuestas a la variable latente.**

Origen de IRT

IRT fue propuesta originalmente en psicometría para medir la habilidad de los estudiantes usando un examen de varias preguntas.

Una ventaja de usar IRT es que asume que la **dificultad** de cada ítem es diferente, incluye un parámetro de **capacidad discriminativa** y modela la distribución original (no asume normalidad) de las respuestas, por lo cual se han propuesto varios modelos para IRT.

Usos de IRT

Es utilizado para obtener scores en muchas pruebas e instrumentos. Ejemplos:

Graduate Record Examination (GRE),
Graduate Management Admission Test (GMAT),
Scholastic Aptitude Test (SAT) (Skrondal y Rabe-Hesketh S, 2004).

También, es utilizado en el área de la salud, para investigación clínica y para medir la calidad de vida.

Modelos IRT más comunes

Modelo logístico de un parámetro (1PL)

$$P(y_{ij} = 1 | \theta_i, \beta_j) = \frac{\exp(\theta_i + \beta_j)}{1 + \exp(\theta_i + \beta_j)} \quad (1)$$

θ_i representa (no observada) la habilidad del sujeto i y β_j el parámetro de facilidad del ítem j .

Modelo logístico de dos parámetros (2PL)

$$P(y_{ij} = 1 | \theta_i, \alpha_j, \beta_j) = \frac{\exp[\alpha_j(\theta_i + \beta_j)]}{1 + \exp[\alpha_j(\theta_i + \beta_j)]} \quad (2)$$

α_j es el parámetro de discriminación para ítem j .

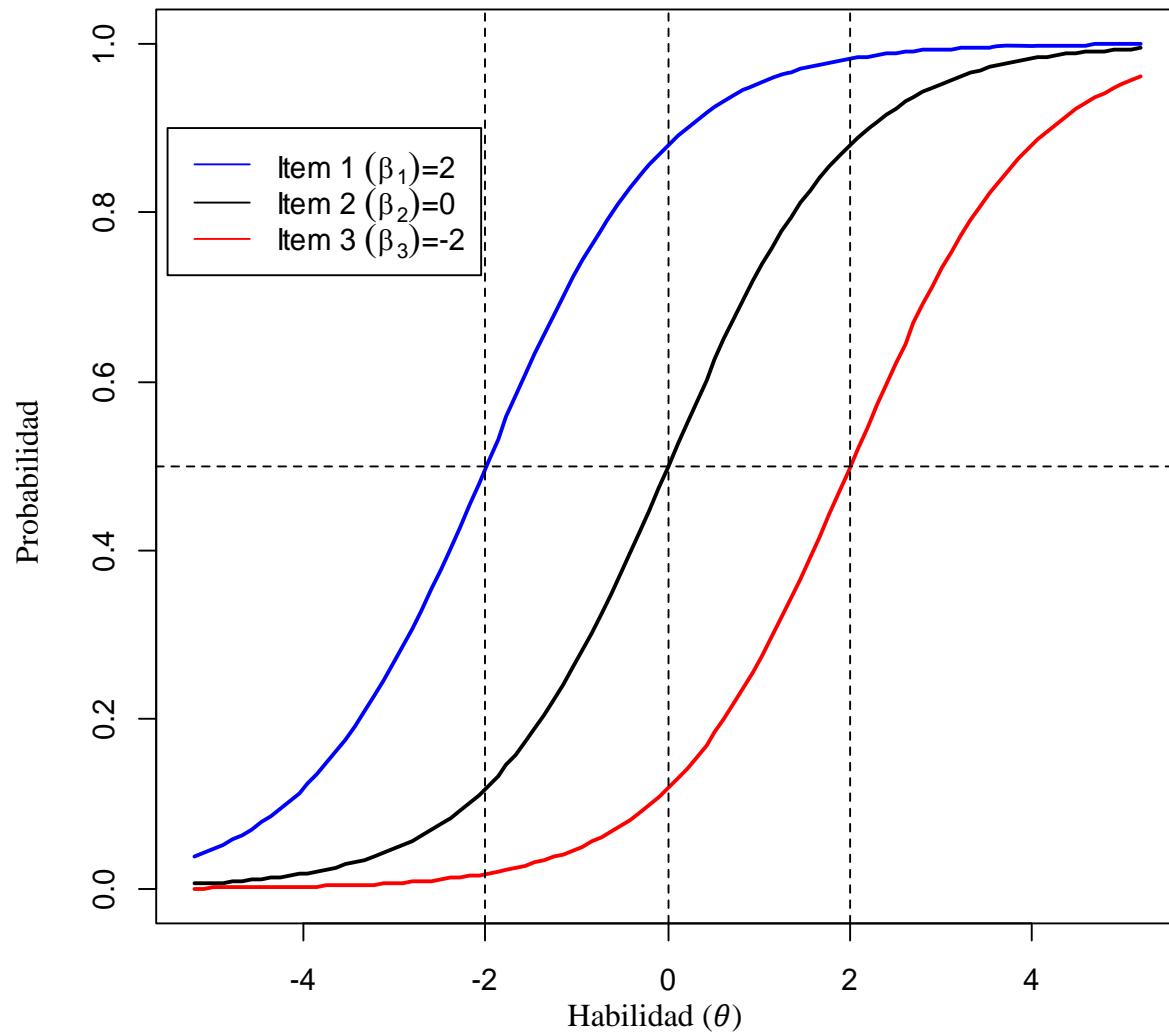


Figura 2. Curvas característica del ítem (ICCs por sus siglas en ingles) de tres ítems para el modelo **1PL** con diferentes valores de dificultad, se muestra la probabilidad de una respuesta correcta en función a la habilidad θ .

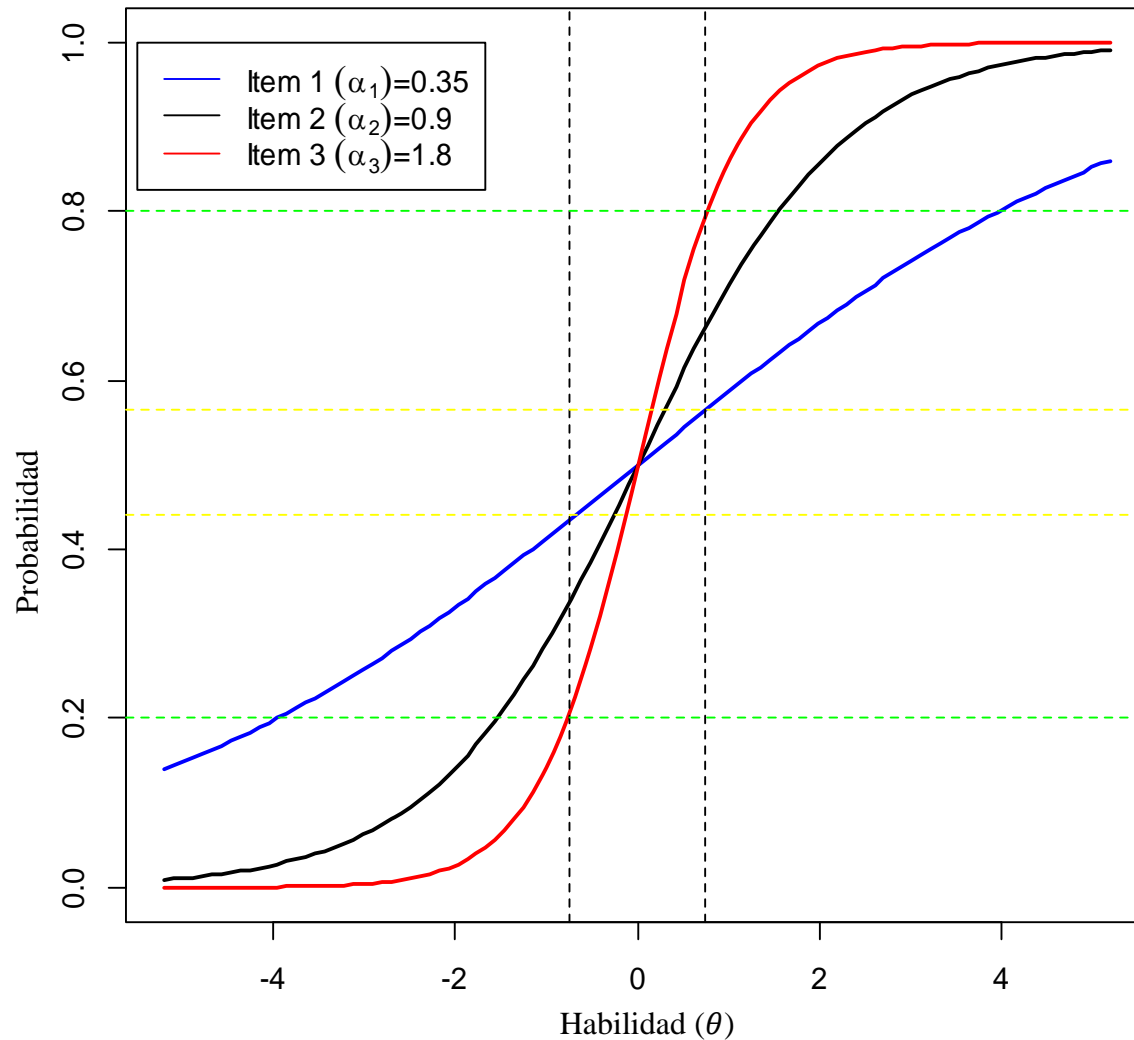


Figura 3. Curvas característica del ítem (ICCs por sus siglas en ingles) de tres ítems con diferentes valores de α_j , y la misma de dificultad ($\beta_1 = \beta_2 = \beta_3 = 0$), se muestra la probabilidad de una respuesta correcta en función de la habilidad θ .

Modelo Graded Response (Samejina, 1969)

$$P(y_{ij} \leq c \mid \alpha_j, \beta_{jc}, \theta_i) = \frac{\exp\{\alpha_j(\theta_i - \beta_{jc})\}}{1 + \exp\{\alpha_j(\theta_i - \beta_{jc})\}} \quad (3)$$

y_{ij} respuesta para el i -ésimo individuo para el ítem j .

c es una de las 5 opciones de respuesta ($c=0,1,2,3,4$).

θ_i es la variable latente (habilidad) para el individuo i ; $\theta_i \in (-\infty, \infty)$.

β_{jc} representa un parámetro de dificultad para el ítem j en la categoría c .

α_j representa un parámetro de capacidad discriminativa del ítem j .

$$P(y_{ij} \leq c \mid \alpha_j, \beta_{jc}, \theta_i) = f(y_{ij} \mid \boldsymbol{\vartheta}, \theta_i).$$

$$P(y_{ij} = c | \boldsymbol{\vartheta}, \theta_i) = g(\eta_{jc}) - g(\eta_{jc+1})$$

$$\text{Donde, } \eta_{jc} = \alpha_j(\theta_i - \beta_{jc}), g(\eta) = \frac{\exp\{\eta\}}{1 + \exp\{\eta\}}$$

La implementación del modelo fue realizada en el paquete ITM (Rizopoulos, 2006) de R.

El IRT confirmatorio fue implementado en MPLUS

6.

(Muthén and Muthén, 2006).

Estimación usando el método de Máxima Verosimilitud

Sea $\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{ij}, \dots, y_{Ij})$, con $i = 1, \dots, I$ y $j = 1, \dots, n$, el vector de I respuestas observadas para j th individuo con habilidad θ_j . Bajo independencia local, la verosimilitud condicional de observar el vector de respuestas \mathbf{y}_j para el individuo j es:

$$L_j(\boldsymbol{\vartheta} | \theta_j) = \prod_{i=1}^I P(y_{ij} | \theta_j, \boldsymbol{\vartheta}) = \prod_{i=1}^I f(y_{ij} | \theta_j, \boldsymbol{\vartheta})$$

$\boldsymbol{\vartheta}$ = vector de parametros de dificultad y discriminación que se desan
estimar

Para obtener la contribución independiente del individuo j para la función de verosimilitud, los efectos aleatorios a nivel individual (variables latentes) son tomados en cuenta:

$$L_j(\boldsymbol{\vartheta}) = \int_{-\infty}^{\infty} \prod_{i=1}^I f(y_{ij} | \theta_j, \boldsymbol{\vartheta}) \varphi(\theta_j) d\theta_j$$

Donde $\varphi(\theta_j)$ es una distribución normal estándar. Por último la verosimilitud marginal (incondicional) es el producto de las verosimilitudes individuales.

$$L = \prod_{j=1}^n L_j(\boldsymbol{\vartheta})$$

Adaptative Gauss-Hermite quadrature (Pinheiro and Bates, 2000) fue usado para maximizar L y estimar los parametros de dificultad y discriminación. La estimación de la habilidad se realiza con el método Empirical Bayes.

IRT Confirmatorio

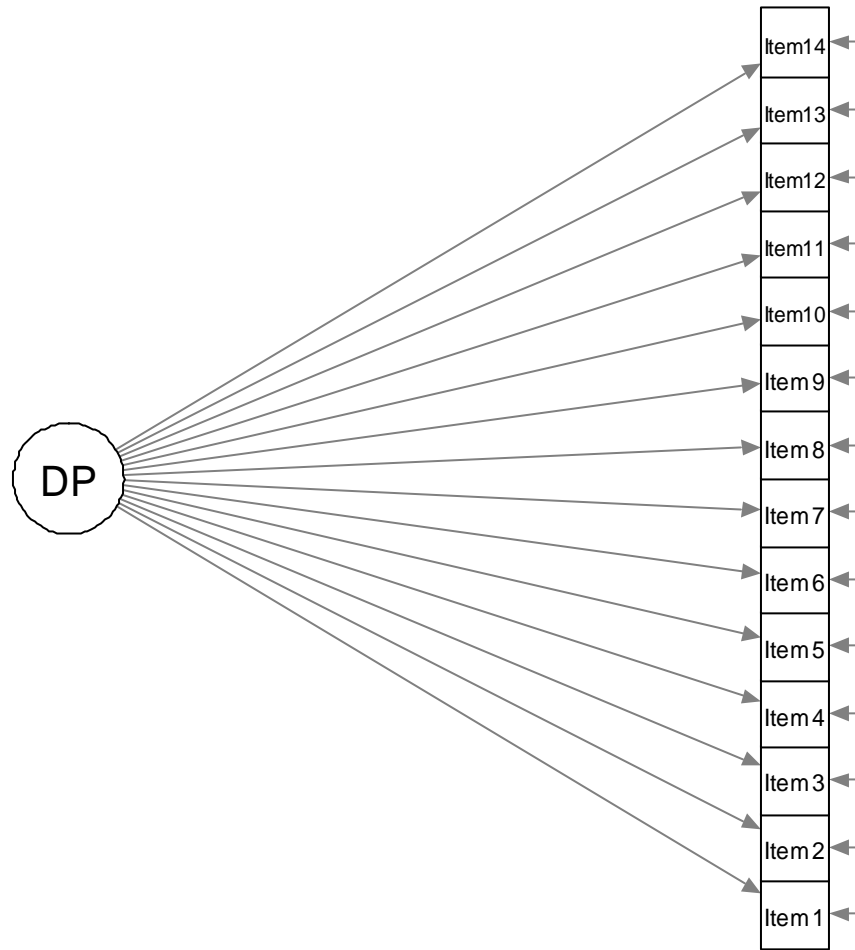


Figura 4. Modelo hipotético para medir el desempeño del profesor.

IRT Confirmatorio

| Semestre | RMSEA | CFI | TLI |
|--------------------------|-----------------|-----------------|-----------------|
| ENE13-JULIO13 | 0.071 | 0.981 | 0.978 |
| AGO13-ENE14 | 0.071 | 0.980 | 0.976 |
| Critero a cumplir | <0.08 | >0.95 | >0.95 |

RMSEA = Root Mean Sqaure Error of Approximation

CFI=Comparative Fit Index

TLI=Tucker Lewis Index

Estimaciones de los parámetros de dificultad y discriminación bajo el modelo IRT.
Evaluación ene13-jul13

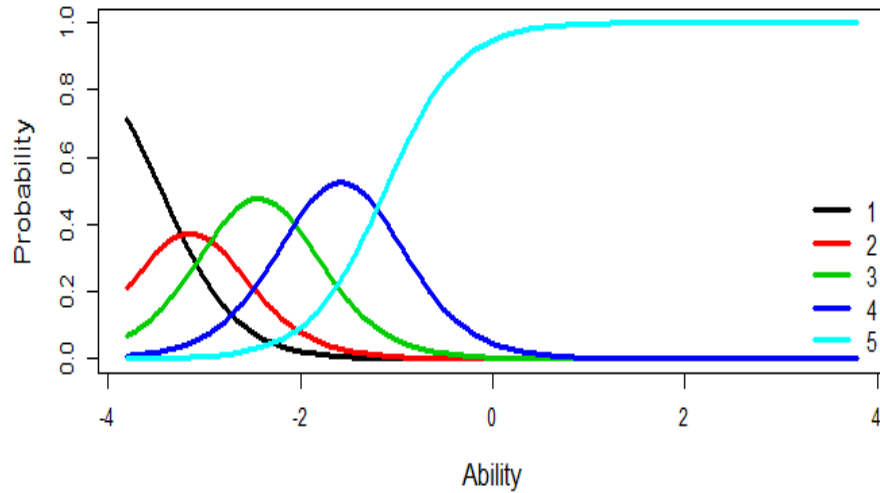
| Ítem | β_{j1} | β_{j2} | β_{j3} | β_{j4} | α_j |
|--------|--------------|--------------|--------------|--------------|------------|
| Ítem1 | -3.517 | -2.877 | -2.029 | -1.124 | 2.64 |
| Ítem2 | -3.198 | -2.54 | -1.649 | -0.721 | 2.791 |
| Ítem3 | -3.25 | -2.67 | -1.86 | -0.974 | 2.755 |
| Ítem4 | -3.289 | -2.643 | -1.807 | -0.973 | 2.884 |
| Ítem5 | -3.125 | -2.559 | -1.795 | -0.963 | 3.265 |
| Ítem6 | -3.07 | -2.49 | -1.694 | -0.832 | 3.251 |
| Ítem7 | -2.95 | -2.409 | -1.649 | -0.788 | 3.61 |
| Ítem8 | -2.894 | -2.353 | -1.618 | -0.818 | 3.48 |
| Ítem9 | -2.79 | -2.198 | -1.438 | -0.663 | 1.548 |
| Ítem10 | -3.259 | -2.691 | -1.928 | -0.998 | 2.78 |
| Ítem11 | -3.326 | -2.765 | -2 | -1.12 | 2.577 |
| Ítem12 | -2.974 | -2.386 | -1.614 | -0.75 | 3.064 |
| Ítem13 | -3.061 | -2.503 | -1.699 | -0.828 | 2.35 |
| Ítem14 | -2.953 | -2.458 | -1.757 | -0.969 | 2.084 |

Estimaciones de los parámetros de dificultad y
discriminación bajo el modelo IRT.
Evaluación ago13-ene14

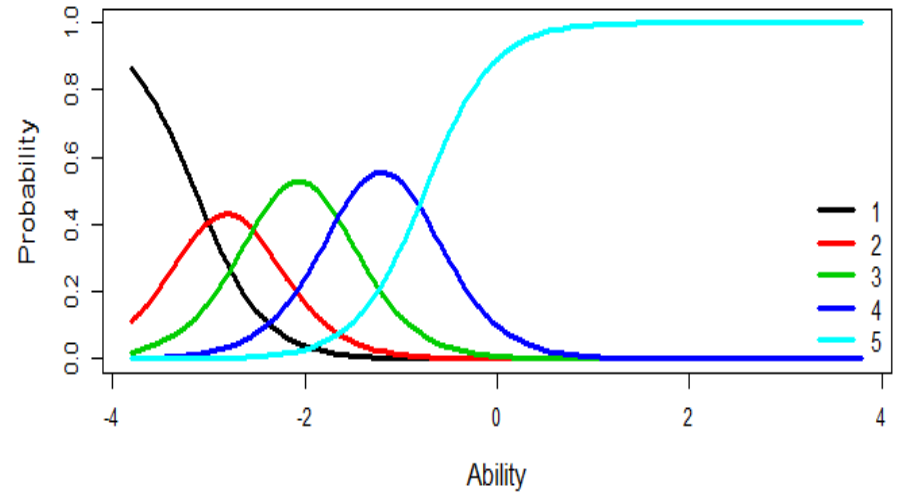
| Ítem | β_{j1} | β_{j2} | β_{j3} | β_{j4} | α_j |
|--------|--------------|--------------|--------------|--------------|------------|
| Ítem1 | -3.586 | -2.926 | -2.044 | -1.152 | 2.588 |
| Ítem2 | -3.127 | -2.476 | -1.617 | -0.722 | 2.77 |
| Ítem3 | -3.299 | -2.684 | -1.884 | -1.022 | 2.719 |
| Ítem4 | -3.206 | -2.561 | -1.769 | -0.962 | 2.927 |
| Ítem5 | -3.108 | -2.537 | -1.786 | -0.981 | 3.256 |
| Ítem6 | -3.066 | -2.483 | -1.695 | -0.857 | 3.277 |
| Ítem7 | -2.957 | -2.397 | -1.651 | -0.807 | 3.613 |
| Ítem8 | -2.83 | -2.304 | -1.609 | -0.841 | 3.482 |
| Ítem9 | -2.756 | -2.171 | -1.452 | -0.704 | 1.511 |
| Ítem10 | -3.3 | -2.738 | -1.953 | -1.027 | 2.694 |
| Ítem11 | -3.37 | -2.796 | -2.03 | -1.149 | 2.511 |
| Ítem12 | -2.955 | -2.366 | -1.602 | -0.761 | 3.046 |
| Ítem13 | -3.009 | -2.446 | -1.674 | -0.819 | 2.319 |
| Ítem14 | -2.914 | -2.429 | -1.758 | -0.979 | 2.084 |

Probabilidad de elección de cada una de las diferentes opciones de respuesta con respecto a la habilidad del individuo.

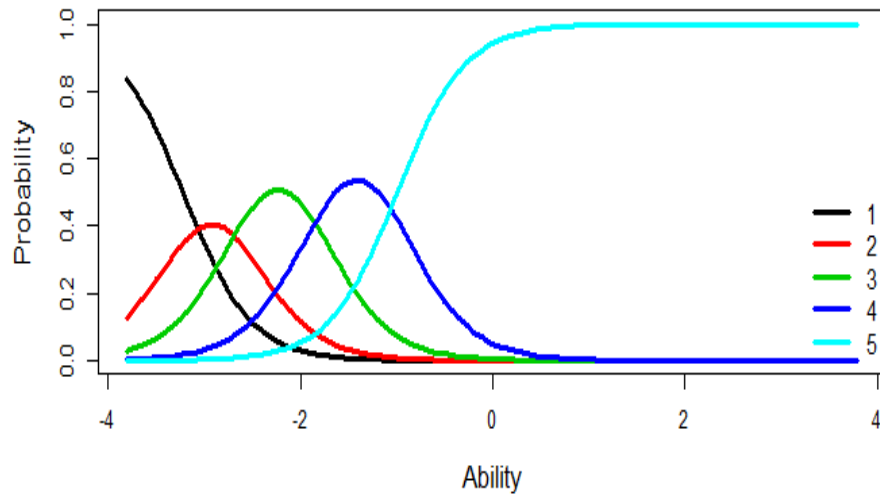
Item Response Category Characteristic Curves
Item: calif1



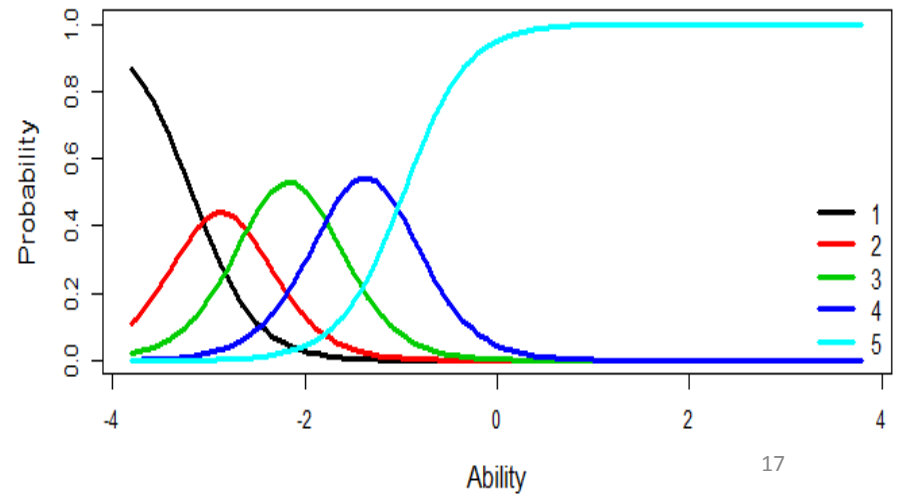
Item Response Category Characteristic Curves
Item: calif2



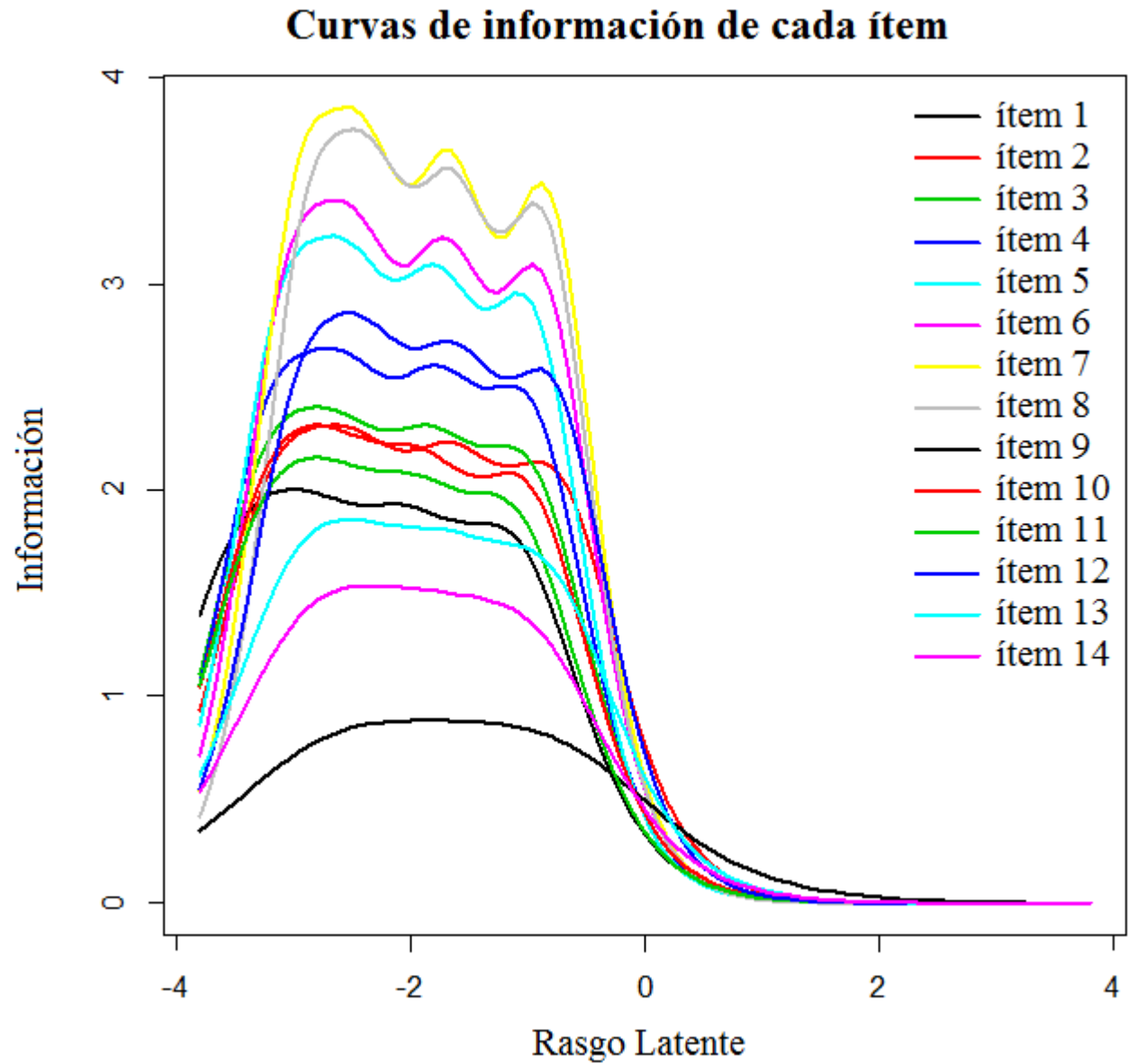
Item Response Category Characteristic Curves
Item: calif3



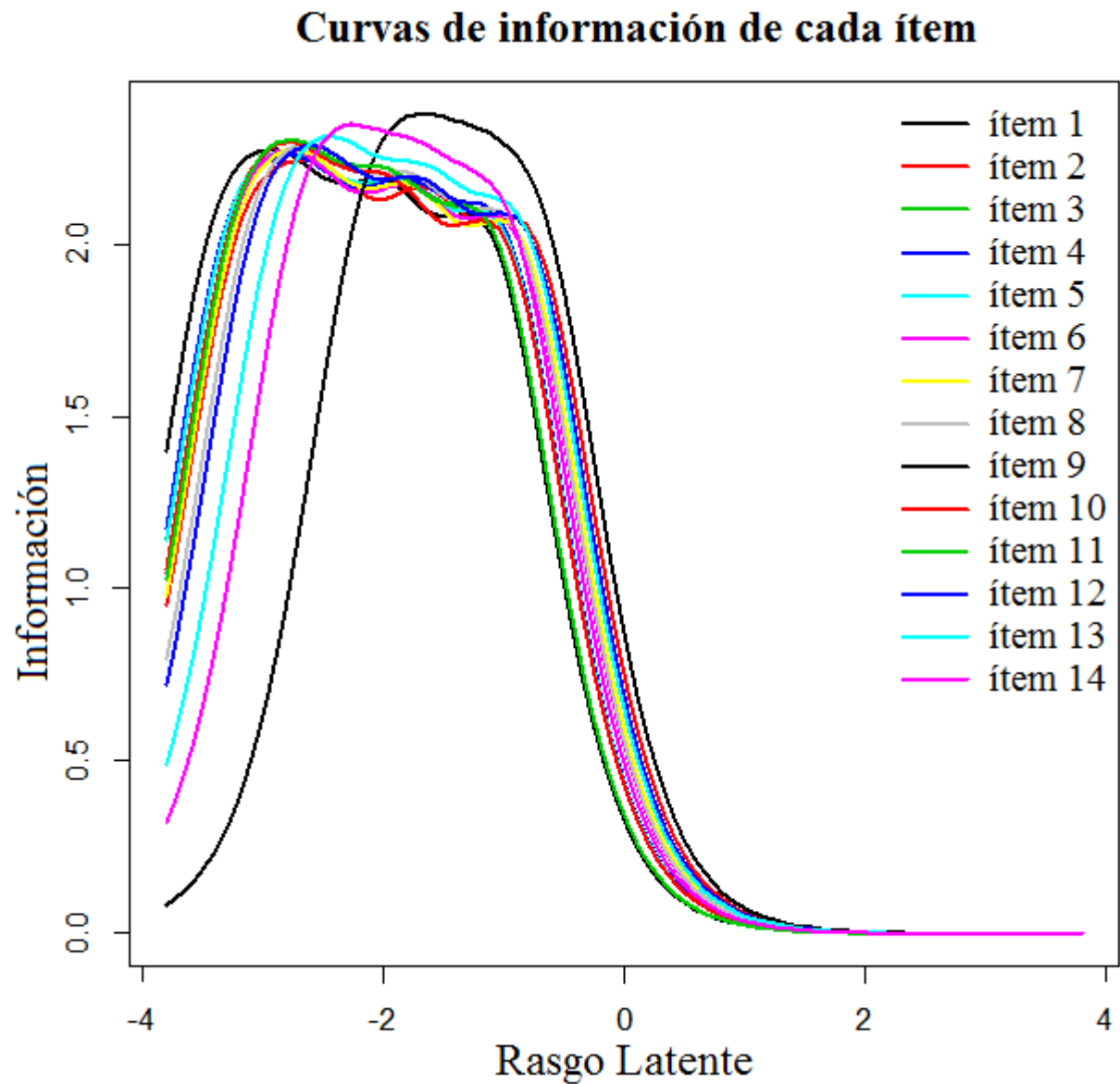
Item Response Category Characteristic Curves
Item: calif4



Información proporcionada por cada ítem tomando en cuenta ambos parámetros (dificultad y discriminación).



Información proporcionada por cada ítem ignorando el parámetro de discriminación.



Habilidad en el intervalo de 0 a 100

El puntaje con IRT se construye con las estimaciones de θ_i (habilidad del profesor i); $\theta_i \in (-\infty, \infty)$. Para expresarlo en valores en el intervalo de 0 a 100.

$$IRT_i = \left(\frac{\theta_i - \theta_{\min}}{\theta_{\max} - \theta_{\min}} \right) \cdot 100 \quad (4)$$

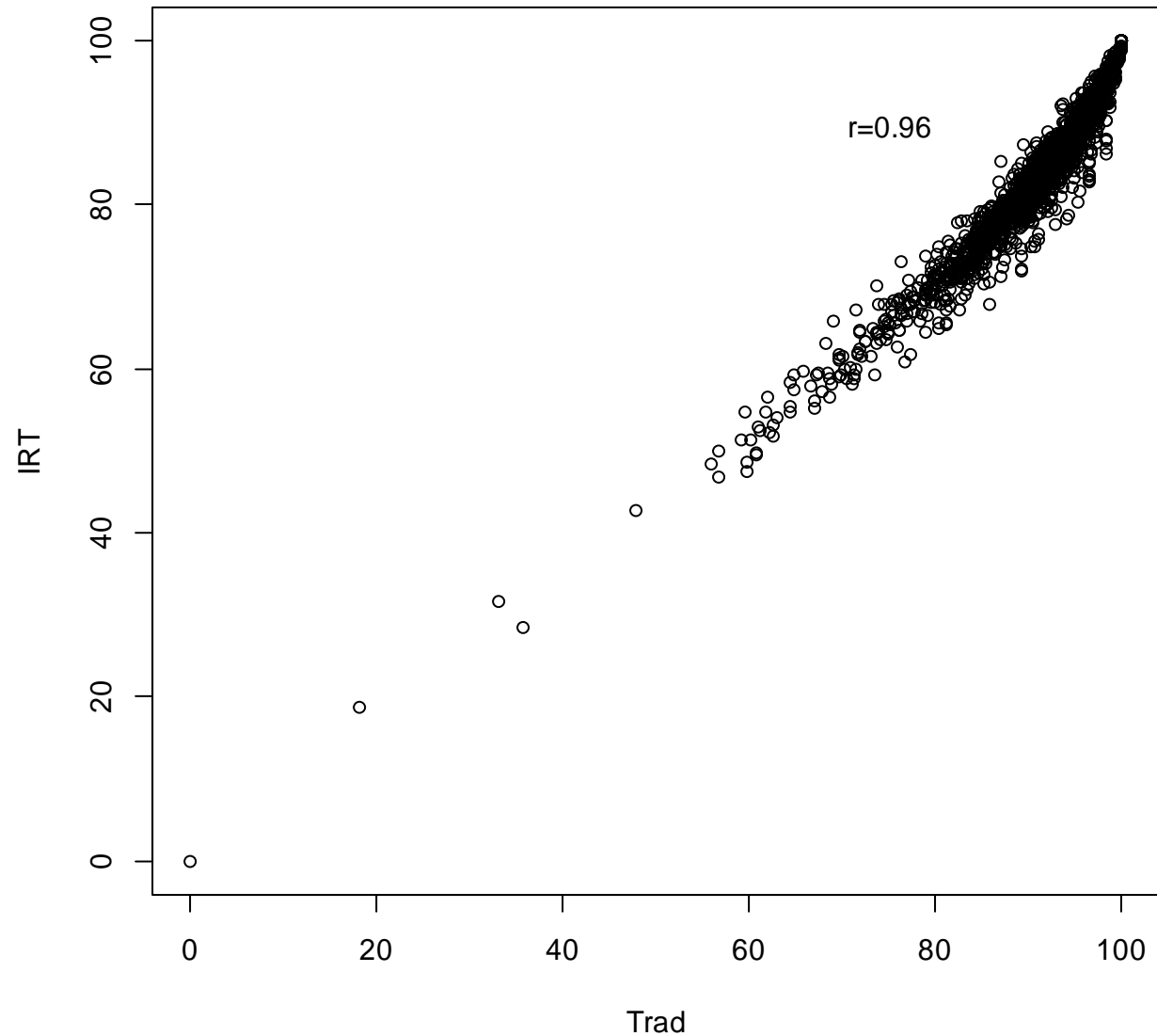
Donde: θ_{\min} es el índice de desempeño mínimo y θ_{\max} el índice de desempeño máximo de entre los individuos que conforman la muestra.

Distribución de profesores por intervalo de calificación con el modelo tradicional y el IRT.

| Índice Intervalo | Frecuencia | | Proporción | |
|---------------------|---------------|---------------|---------------|---------------|
| | Tradicional | IRT | Tradicional | IRT |
| (30,35] | 0 | 1 | 0.000 | 0.001 |
| (35,40] | 0 | 0 | 0.000 | 0.000 |
| (40,45] | 1 | 1 | 0.001 | 0.001 |
| (45,50] | 1 | 7 | 0.001 | 0.004 |
| (50,55] | 4 | 9 | 0.002 | 0.005 |
| (55,60] | 5 | 15 | 0.003 | 0.008 |
| (60,65] | 6 | 40 | 0.003 | 0.023 |
| (65,70] | 12 | 68 | 0.007 | 0.038 |
| (70,75] | 28 | 151 | 0.016 | 0.085 |
| (75,80] | 67 | 249 | 0.038 | 0.141 |
| (80,85] | 141 | 398 | 0.080 | 0.225 |
| (85,90] | 321 | 466 | 0.181 | 0.263 |
| (90,95] | 669 | 282 | 0.378 | 0.159 |
| (95,100] | 514 | 82 | 0.291 | 0.046 |
| Total | 1769 | 1769 | 1 | 1 |
| Promedio | 90.915 | 83.033 | 90.915 | 83.033 |

Correlación del tradicional vs el IRT ene13-ago13

Se aprecia que la correlación entre el desempeño calculado en forma tradicional con el nuevo método (IRT) es 0.96, lo cual indica que ambos métodos son muy similares.



Comentarios Finales

IRT está basado en un modelo probabilístico formal y toma en cuenta el nivel de facilidad y capacidad discriminativa de cada ítem, la correlación entre ítems y la distribución original de los datos.

La propuesta de utilizar IRT produce resultados más justos, presuponiendo unidimensionalidad en la variable latente (Desempeño del profesor). Si bien el cálculo resulta más complejo, es preferible a la forma empleada hasta ahora porque le asigna pesos distintos a cada ítem.

Sin embargo, se requiere un estudio a mayor profundidad para poder tomar en cuenta en la estimación del desempeño del profesor variables como área del conocimiento, cantidad de alumnos, entre otras.

Referencias

Johnson, R. A., Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (Sixth ed.). Prentice Hall.

Muthén LK, Muthén BO. *Mplus: The Comprehensive Modeling Program for Applied Researchers. User's Guide*. Los Angeles, CA: Muthén & Muthén; 1998–2006.

Samejima F. (1969). Estimation of Latent Ability using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement*, 34.

Skrondal A, Rabe-Hesketh S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall, Boca Raton, FL.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25.