

# Uso de pesos muestrales en la calibración de los parámetros de los ítems en evaluaciones a gran escala

Víctor H. Cervantes - Álvaro Uzaheta

Subdirección de Estadística

Instituto Colombiano para la Evaluación de la Educación - ICFES

*vcervantes@icfes.gov.co - auzaheta@icfes.gov.co*

II Seminario Internacional de Investigación sobre la Calidad de la Educación

3 de Noviembre de 2011

# Uso de pesos muestrales en calibración de ítems

## 1 Background

- Modelos de respuesta al ítem
- Evaluaciones en educación

## 2 Implementación de los pesos muestrales

- Estudios internacionales
- Modelos multinivel

## 3 Estudio de simulación

- Objetivos
- Diseño
- Variables de respuesta
- Resultados

## 4 Discusión

- Discusión
- Recomendaciones
- Limitaciones
- Investigaciones futuras

- Los modelos de respuesta al ítem datan de finales de los años 50 y comienzos de los 60
- Estos modelos proponen un modelo de medida que relaciona la magnitud de una habilidad o atributo individual no observable con ciertas características del ítem
- Modelo de Rasch

$$P(U_{ij}|\theta_j, b_i) = \frac{1}{1 + e^{-(\theta_j - b_i)}} \quad (1)$$

- Modelo de respuesta al ítem logístico de tres parámetros

$$P(U_{ij}|\theta_j, (a_i, b_i, c_i)) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}} \quad (2)$$

- Los modelos multinivel consideran la estructura jerárquica de los datos
- El modelo de Rasch puede verse como un modelo multinivel (e.g., Kamata, 2001)

|         |                        |
|---------|------------------------|
| Nivel 1 | Respuestas a los ítems |
| Nivel 2 | Individuos             |

- Otros modelos TRI pueden verse como modelos multinivel no lineales generalizados.

## Muestreo en evaluaciones en educación

- Los diseños de muestreo en evaluaciones educacionales tienen una estructura jerárquica natural
- Estos diseños suelen constar de:
  - Varios estratos
  - Varias etapas
  - Selección de grupos (p.e. clases) en lugar de individuos
- En los últimos años, estos diseños han sido denominados como diseños de muestras complejas

Al ajustar un modelo TRI y calibrar los ítems se encuentra que

- algunas evaluaciones emplean los pesos del diseño muestral: NAEP - TIMSS(2007) - SABER(2009).
- algunas no los usan: PISA (2003, 2006).

# Uso de pesos muestrales en calibración de ítems

## 1 Background

Modelos de respuesta al ítem  
Evaluaciones en educación

## 2 Implementación de los pesos muestrales

Estudios internacionales  
Modelos multinivel

## 3 Estudio de simulación

Objetivos  
Diseño  
Variables de respuesta  
Resultados

## 4 Discusión

Discusión  
Recomendaciones  
Limitaciones  
Investigaciones futuras

- Hay un diseño muestral para cada país.
- La calibración se hace partir de una submuestra de cada país (500 estudiantes por muestreo aleatorio simple)
- Esta submuestra es usada para la calibración.
- No se reporta uso alguno de los pesos.



- Hay un diseño muestral para cada país.
- De cada país se emplea una submuestra de 1000 estudiantes de cada país (TIMSS 2003) o la muestra completa (TIMSS 2007).
- La submuestra de cada país es reponderada para la calibración.
- La reponderación busca que cada país “pese igual” que los demás.

## Reescalamiento de pesos en modelos multinivel

- Algunos autores han sugerido que en los modelos multinivel los pesos muestrales deben reescalarsen (e.g. Asparouhov, 2006; Carle, 2009)
- Algunos de los métodos son:
  - Reescalar de forma que la suma de pesos sea igual al tamaño de muestra.
  - Reescalar de forma que la suma de pesos dentro de cada agrupación sea igual al número de unidades de muestreo.

# Uso de pesos muestrales en calibración de ítems

## 1 Background

Modelos de respuesta al ítem  
Evaluaciones en educación

## 2 Implementación de los pesos muestrales

Estudios internacionales  
Modelos multinivel

## 3 Estudio de simulación

Objetivos  
Diseño  
Variables de respuesta  
Resultados

## 4 Discusión

Discusión  
Recomendaciones  
Limitaciones  
Investigaciones futuras

- Identificar entre los esquemas de ponderación considerados el más apropiados para la calibración de los parámetros de los ítems
- Evaluar los métodos de estimación disponibles en los paquetes para calibración empleados regularmente en evaluaciones internacionales cuando se emplean estos esquemas de ponderación.

- Longitud de prueba:  $k = 25$  ítems
- Parámetros de los ítems:  $b \sim N(0, 1.3)$  y  $a \sim \log N(0, 0.04)$
- Habilidades en la población:  $\theta = \alpha_j + \varepsilon_i, \varepsilon \sim N(-0.4, 0.6)$ ,  $\alpha$  tomado del promedio de la sede-jornada en Saber 5º y 9º, 2009, reescalados a una media de  $-0.4$  y varianza de  $0.4$ .
- CCI = 0.25
- Réplicas por condición:  $R = 200$

## Variables manipuladas y niveles analizados

- Diseño de muestra

D1 Saber 5º y 9º, 2009: diseño estratificado por conglomerados en tres etapas

1ª etapa Estratos definidos por los grados ofrecidos y por algunas entidades territoriales.

Selección aleatoria simple de instituciones educativas en cada estrato

2ª etapa Selección de todas las sedes-jornadas de cada institución seleccionada.

3ª etapa ~ dos tercios de los estudiantes de cada sede-jornada presentan cada prueba

D2 PIRLS 2011 (COL): Estratos por zona (Rural-Urbana), sector (Privado-Público) y grado. Instituciones seleccionadas por PPT sistemática

Una sede-jornada es seleccionada en la segunda etapa.

## Variables manipuladas y niveles analizados

- Diseño de muestra

- D3 Diseño para pruebas piloto ICFES: Estratos por tamaños de ciudades

- Se seleccionan ciudades PPT en la primera etapa

- Sedes-jornadas seleccionadas PPT en la segunda etapa

- D4 Saber 5º y 9º, 2011: diseño estratificado por conglomerados en tres etapas

- 1ª etapa Estratos definidos por los grados ofrecidos y por algunas entidades territoriales, zona y sector. Estratos implícitos por desempeño en 2009 y por tamaño.

- Selección sistemática de instituciones educativas en cada estrato

- 2ª etapa Selección de todas las sedes-jornadas de cada institución seleccionada.

- 3ª etapa ~ la mitad de los estudiantes de cada sede-jornada presentan cada prueba

## Variables manipuladas y niveles analizados

- Esquema de ponderación

W0 Sin pesos

W1 Ponderado - pesos obtenidos directamente por el diseño de muestra.

W2 Ponderado - pesos reescalados para que su suma sea igual al tamaño de muestra.

W3 Ponderado - pesos reescalados por etapa para que su suma sea igual al número de unidades seleccionadas en la misma.

W4 Ponderado - pesos reescalados como en W3 en todas las etapas con excepción de la primera.



## Variables manipuladas y niveles analizados

- Método de estimación de los parámetros de los ítems

**EWS** Máxima verosimilitud conjunta (JML)

**EWB** JML con corrección de sesgo

**EMB** Máxima verosimilitud marginal

**EML** Mínimos cuadrados reponderados penalizados

S Sesgo

$$\frac{1}{R} \sum_{r=1}^R (\hat{b}_{ir} - b_i) \quad (3)$$

SP Sesgo promedio:

$$\frac{1}{kR} \sum_{i=1}^k \sum_{r=1}^R (\hat{b}_{ir} - b_i) \quad (4)$$

ECM Error cuadrático medio promedio:

$$\frac{1}{kR} \sum_{i=1}^k \sum_{r=1}^R (\hat{b}_{ir} - b_i)^2 \quad (5)$$

EE Error estándar promedio:

$$\frac{1}{k} \sum_{i=1}^k \sqrt{\frac{1}{R-1} \sum_{r=1}^R (\hat{b}_{ir} - \bar{\hat{b}}_i)^2} \quad (6)$$

ME Efecto de (mala) especificación promedio:

$$\left[ \prod_{i=1}^k \prod_{r=1}^R \left( \frac{se_{b_i}^2}{\hat{se}_{b_{ir}}^2} \right) \right]^{\frac{1}{kR}} \quad (7)$$

IC Cobertura aproximada y esperada de los intervalos de confianza

$$\frac{1}{kR} \sum_{i=1}^k \sum_{r=1}^R I_{\{\hat{b}_{ir} \pm z_{1-\alpha/2} se_{b_{ir}}^*\}}(b_i) \quad (8)$$

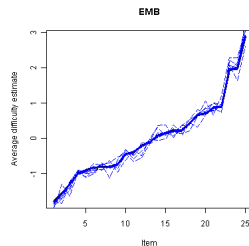
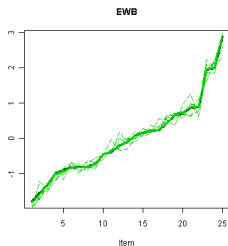
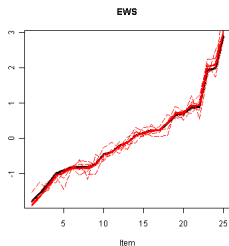


Figure: Sesgo

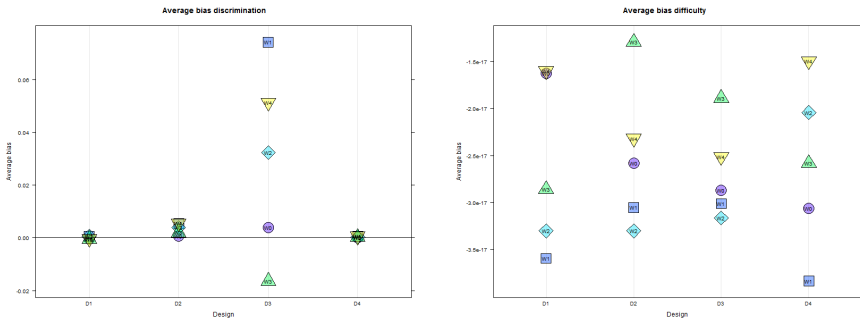


Figure: Sesgo promedio

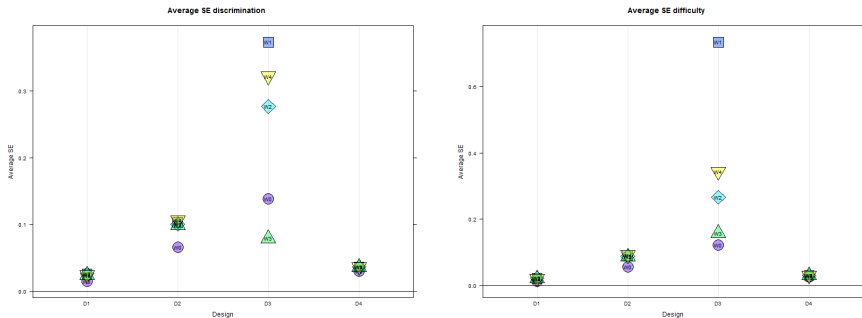


Figure: EE promedio

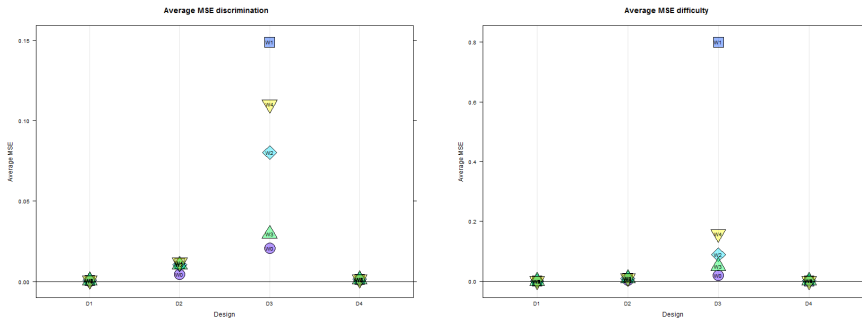


Figure: Average MSE

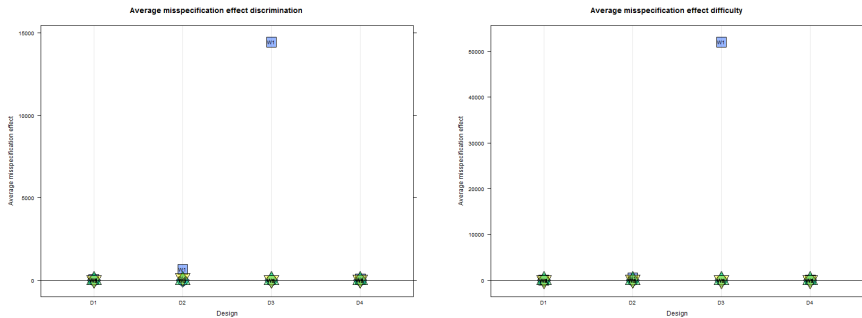


Figure: Error de especificación promedio



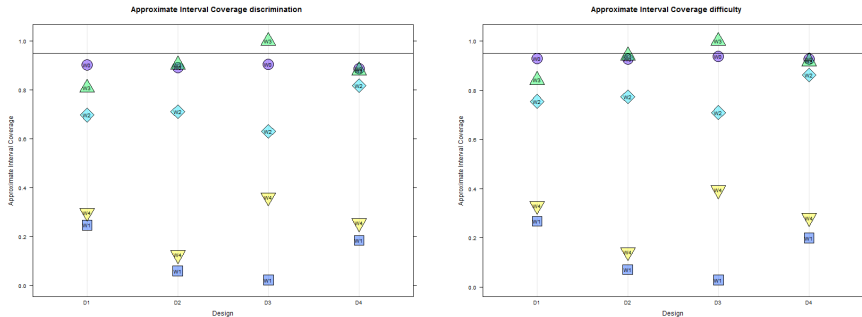


Figure: Cobertura de los intervalos de confianza

# Uso de pesos muestrales en calibración de ítems

## 1 Background

- Modelos de respuesta al ítem
- Evaluaciones en educación

## 2 Implementación de los pesos muestrales

- Estudios internacionales
- Modelos multinivel

## 3 Estudio de simulación

- Objetivos
- Diseño
- Variables de respuesta
- Resultados

## 4 Discusión

- Discusión
- Recomendaciones
- Limitaciones
- Investigaciones futuras

- La estimación por JML no es apropiada y debería evitarse (cf. Cyr and Davies, 2005).
- Incluso con la corrección de sesgo, la estimación por JML no parece apropiada.
- A partir de los resultados, no parece ser necesario el uso de los pesos muestrales para la calibración de los ítems.
- La ponderación mediante los paquetes analizados trata los pesos muestrales como pesos de frecuencia, lo cual produce errores estándar incorrectos y engañosos.

- Si no es posible no utilizar los pesos en la calibración de ítems de evaluaciones educativas con muestras complejas, se deberían reescalar de acuerdo con el diseño muestral y en todas las etapas.
- La información del peso total por individuo no es suficiente para realizar este reescalamiento de los pesos.

- Los paquetes comerciales para TRI solo permiten considerar dos niveles, pero los datos en evaluaciones educativas generalmente comprenden más niveles.
- Los resultados pueden aplicarse directamente en los casos en los que todos los individuos responden al mismo conjunto de ítems o cuando se emplean diseños de cuadernillos o formas de prueba de bloques incompletos balanceados (BIB, por su sigla en inglés).
- Esperamos que los resultados sean similares en los casos de diseños de bloques no balanceados.
- Limitaciones de memoria y de velocidad dificultan el uso de lme4.

- Explorar aproximaciones de estimación y de ponderación alternativas (e.g. evaluar el método de estimación condicional, evaluar la aproximación basada en el diseño para la ponderación propuesta por Cohen et al. (2008) cuando/si llega a estar disponible, etc.).
- Evaluar una aproximación completamente multinivel que tenga en cuenta los demás niveles del diseño.

**Gracias !!**

Víctor H. Cervantes  
[vcervantes@icfes.gov.co](mailto:vcervantes@icfes.gov.co)

- Akihito Kamata. Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1):79–93, 2001.
- Tihomir Asparouhov. General multi-level modeling with sampling weights. *Communications in statistics: Theory and Methods*, 35(3):439–460, 2006.
- Adam C. Carle. Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9:49–61, 2009.
- André Cyr and Alexander Davies. Item response theory and latent variable modeling for surveys with complex sampling design: The case of the national longitudinal survey of children and youth in canada. Paper presented at the Federal Committee on Statistical Methodology Conference in Arlington, Virginia, EEUU, 2005.
- Jon Cohen, Tsze Chan, Tao Jiang, and Mary Seburn. Consistent estimation of Rasch item parameters and their standard errors under complex sample designs. *Applied Psychological Measurement*, 32(4):289–310, 2008.