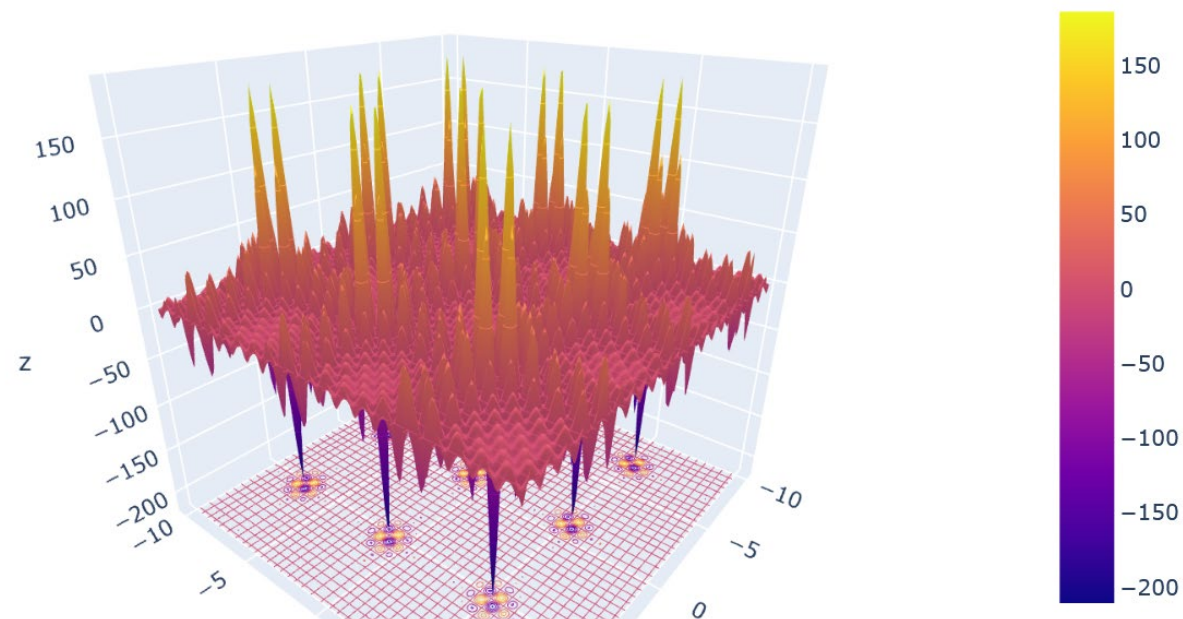
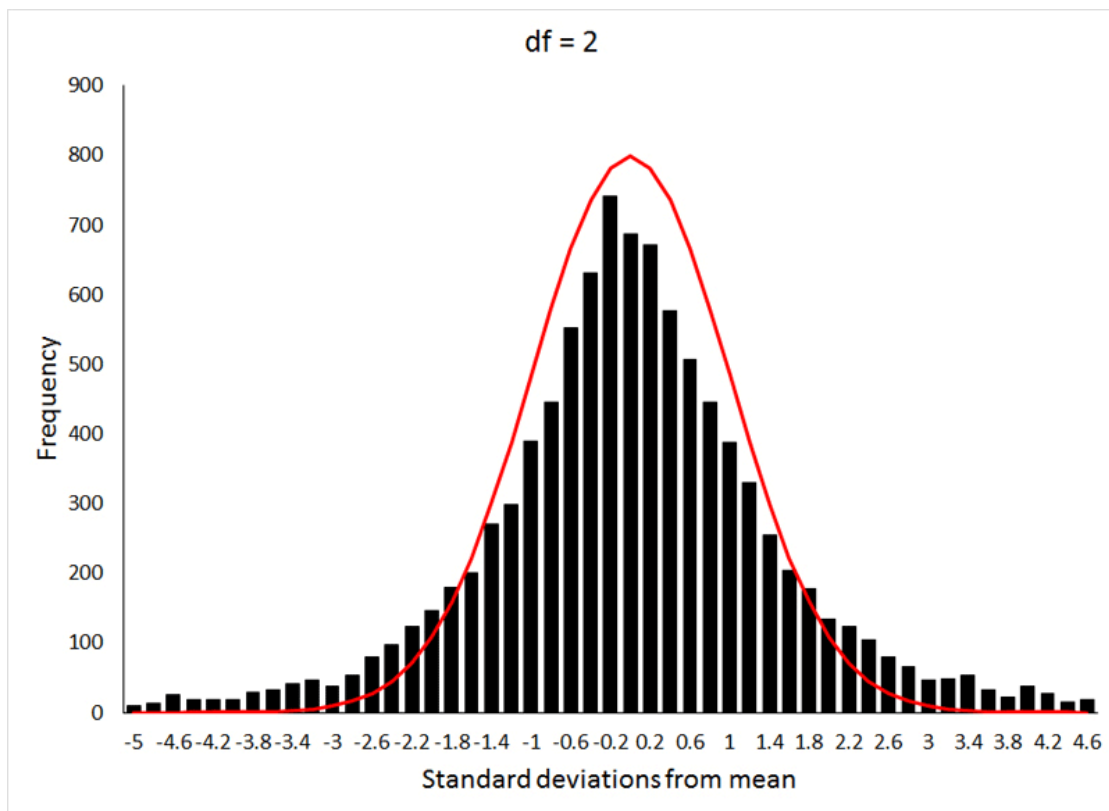


# Clase 10: Teoría de respuesta al ítem

Héctor Nájera



# Expectativa



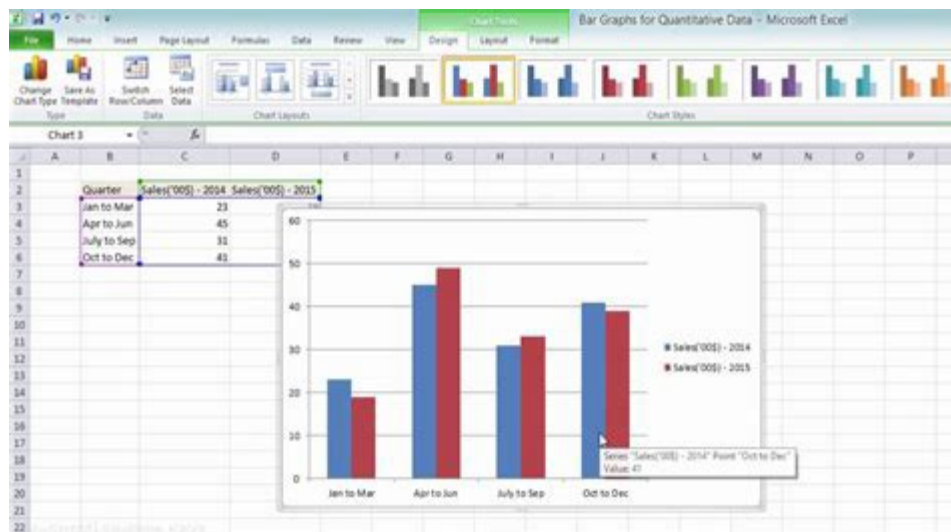
# Realidad

Warning: Maximum number of iterations has been exceeded.  
Current function value: 0.545891  
Iterations: 35

## Results: Logit

```
=====
Model:                Logit                No. Iterations:      35.0000
Dependent Variable:    y                    Pseudo R-squared:    0.212
Date:                 2018-09-10 12:16      AIC:                 55867.177
No. Observations:     51134                BIC:                 56044.021
Df Model:              19                   Log-Likelihood:      -27914.
Df Residuals:          51114                LL-Null:              -35443.
Converged:             0.0000                Scale:               1.0000
=====
```

	Coef.	Std.Err.	z	P> z	[0.025	0.975]
euribor3m	-0.4634	0.0091	-50.9471	0.0000	-0.4813	-0.445
job_blue-collar	-0.1736	0.0283	-6.1230	0.0000	-0.2291	-0.118
job_housemaid	-0.3260	0.0778	-4.1912	0.0000	-0.4784	-0.173
marital_unknown	0.7454	0.2253	3.3082	0.0009	0.3038	1.187
education_illiterate	1.3156	0.4373	3.0084	0.0026	0.4585	2.172
default_no	16.1521	5414.0744	0.0030	0.9976	-10595.2387	10627.542
default_unknown	15.8945	5414.0744	0.0029	0.9977	-10595.4963	10627.285
contact_cellular	-13.9393	5414.0744	-0.0026	0.9979	-10625.3302	10597.451
contact_telephone	-14.0065	5414.0744	-0.0026	0.9979	-10625.3973	10597.384
month_apr	-0.8356	0.0913	-9.1490	0.0000	-1.0145	-0.656
month_aug	-0.6882	0.0929	-7.4053	0.0000	-0.8703	-0.506
month_dec	-0.4233	0.1655	-2.5579	0.0105	-0.7477	-0.099
month_jul	-0.4056	0.0935	-4.3391	0.0000	-0.5889	-0.222
month_jun	-0.4817	0.0917	-5.2550	0.0000	-0.6614	-0.302
month_mar	0.6638	0.1229	5.3989	0.0000	0.4228	0.904
month_may	-1.4752	0.0874	-16.8815	0.0000	-1.6465	-1.303
month_nov	-0.8298	0.0942	-8.8085	0.0000	-1.0144	-0.645
month_oct	0.5065	0.1175	4.3111	0.0000	0.2762	0.736
poutcome_failure	-0.5000	0.0363	-13.7706	0.0000	-0.5711	-0.428
poutcome_success	1.5788	0.0618	25.5313	0.0000	1.4576	1.700



# Modelos de medición con categorías

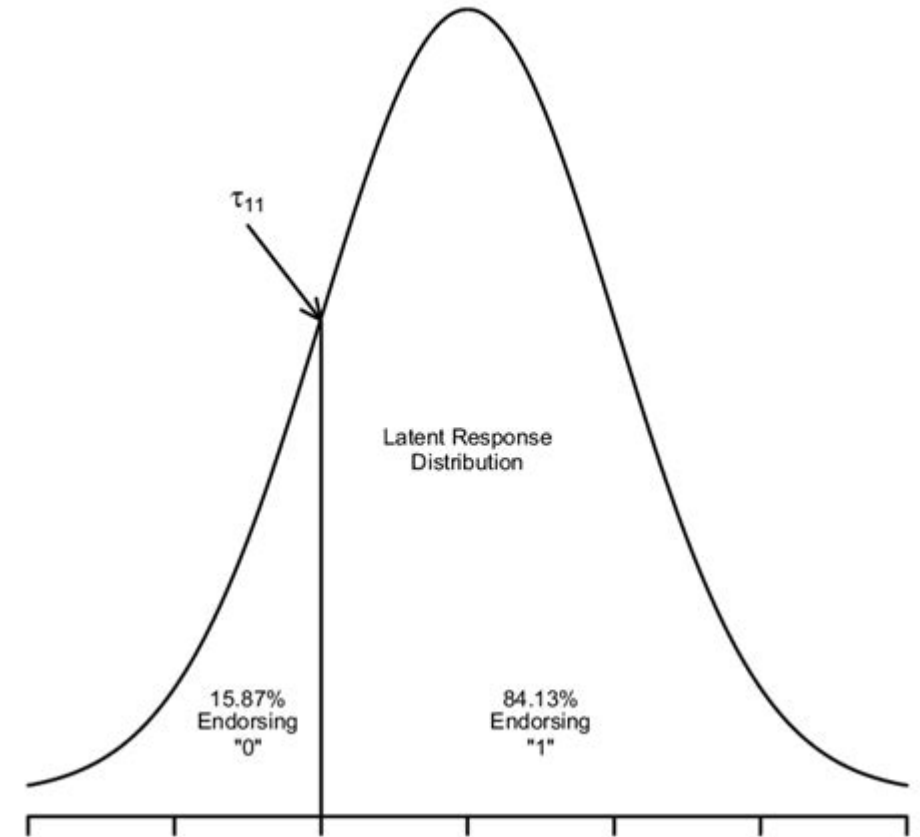
- Sabiduría popular: Si dicotomizas una variable pierdes información.
- Sabiduría informada: No hagas regresión lineal cuando la dependiente es categórica
- ¿Cómo podemos abordar el problema de trabajar con categorías en medición?
- Vimos que AFC permite tratar con variables categóricas, pero no es muy claro cómo hacer interpretaciones que vayan más allá de la varianza atribuible al factor

# Repensando las categorías

Tratamiento clásico. **Categorías absolutas/predefinidas**:  
¿Sexo?, Grupos de edad, estado de nacimiento, religión.

Tratamiento moderno. No todas las categorías son absolutas. **Podemos pensar en un continuo**, en el que hay un valor crítico que separa a los grupos de interés

¿Qué **valor crítico** en el continuo produce una buena separación?



Nivel crítico de **separación o creación** de grupos

# Dos modelos de medición complementarios

Las respuestas categóricas como manifestaciones de cierto fenómeno (medición)

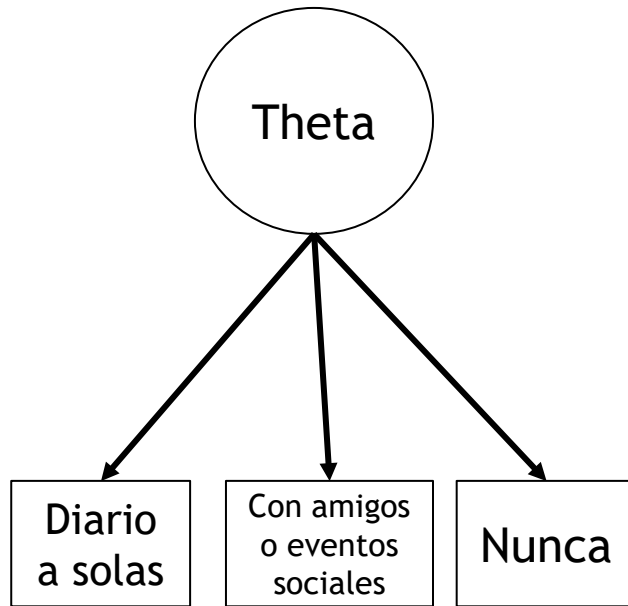
- Análisis factorial
- Teoría de respuesta al ítem

Las respuestas categóricas como manifestaciones de pertenencia a cierto grupo (clasificación)

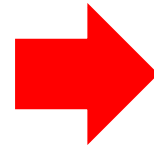
- Análisis de mezclas
- Clases latentes



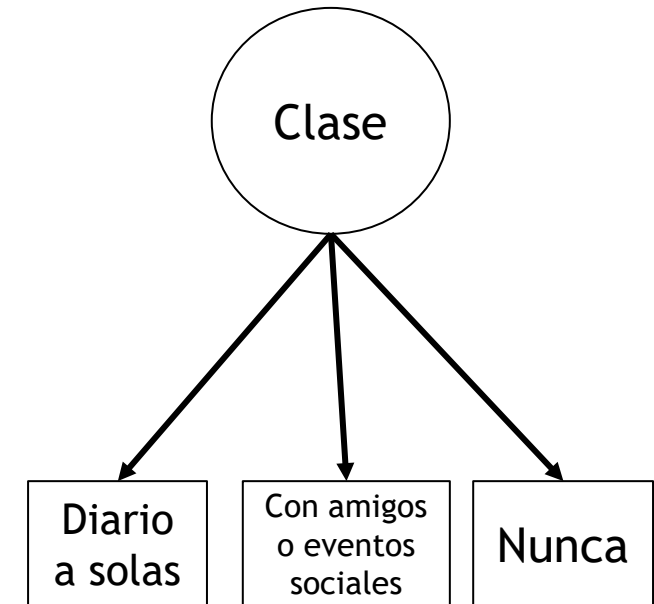
# Dos modelos de medición complementarios



Cambios en el nivel subyacente de alcoholismo explica la variabilidad de las respuestas



*Buena clasificación requiere buena medición*

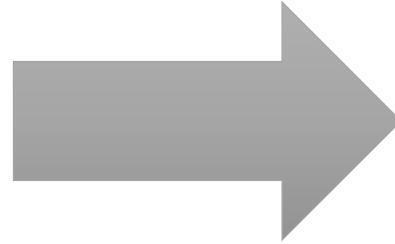


Pertenencia al grupo de alcohólicos explica la probabilidad de respuesta

# Estadísticas sociales y variables categóricas



Variables categóricas



1	0	1	0	1	1	1	1
2	1	0	1	0	0	0	1
3	0	1	1	1	1	0	0
4	0	0	0	0	0	1	1
5	0	0	1	0	0	1	0
6	1	1	1	1	0	1	0
7	0	0	0	0	0	1	0
8	1	1	1	1	0	0	1
9	0	0	0	0	0	0	0
10	1	0	0	1	0	0	0
11	0	0	0	1	1	0	0
12	1	1	0	0	1	1	1
13	1	0	0	1	0	1	1
14	1	0	0	1	0	1	1
15	1	0	1	1	0	1	0
16	0	1	0	1	0	0	1
17	1	0	1	1	1	0	1
18	1	0	0	0	0	1	0
19	1	1	0	1	0	0	0

¿Por qué observamos patrones de respuesta: 1 (Sí) ; 0 (No)?



# Item response modelling (Edinburgh)

PSYCHOMETRIKA—VOL. 7, NO. 1  
FEBRUARY, 1942

## ITEM SELECTION BY THE CONSTANT PROCESS\*

GEORGE A. FERGUSON

DEPARTMENT OF EDUCATIONAL RESEARCH, UNIVERSITY OF TORONTO

This paper relates the constant process used in psychophysics to the problem of item selection. Each test item may be described in terms of a limen, which is an index of the point at which an item discriminates, and the standard deviation of the limen, which is an index of the spread of the distribution. The limen and standard deviation may be related to the item characteristics curve. The description of a test in terms of a limen and standard deviation is a simple and concise way of describing a test.

PSYCHOMETRIKA—VOL. 9, NO. 1  
MARCH, 1944

## THE APPLICATION OF PROBIT ANALYSIS TO THE RESULTS OF MENTAL TESTS

D. J. FINNEY

ROTHAMSTED EXPERIMENTAL STATION

The application of the Müller-Urban constant process to item selection, as considered in a recent paper in this journal, is shown to be closely analogous to a method now in general use for the analysis

**XXIII.—On Problems connected with Item Selection and Test Construction.** By **D. N. Lawley**, Moray House, University of Edinburgh. *Communicated by Professor GODFREY H. THOMSON.*

(MS. received December 16, 1942. Read March 1, 1943.)

1. IN constructing tests designed to measure mental ability it has been a common practice to use a fairly large number of questions or items each



# Variables latentes y probabilidad de respuesta

1	0	1	0	1	1	1	1
2	1	0	1	0	0	0	1
3	0	1	1	1	1	0	0
4	0	0	0	0	0	1	1
5	0	0	1	0	0	1	0
6	1	1	1	1	0	1	0
7	0	0	0	0	0	1	0
8	1	1	1	1	0	0	1
9	0	0	0	0	0	0	0
10	1	0	0	1	0	0	0
11	0	0	0	1	1	0	0
12	1	1	0	0	1	1	1
13	1	0	0	1	0	1	1
14	1	0	0	1	0	1	1
15	1	0	1	1	0	1	0
16	0	1	0	1	0	0	1
17	1	0	1	1	1	0	1
18	1	0	0	0	0	1	0
19	1	1	0	1	0	0	0

$\theta$

Si las respuestas son manifestaciones de cierto fenómeno ( $\theta$ ), podemos pensar que la matriz observada de respuestas **refleja la probabilidad** de registrar 1 o 0

Por ejemplo, si  $\theta$  es el nivel latente de precariedad laboral, esperaríamos que personas con niveles distintos tuvieran **probabilidades diferentes** de respuesta a cierta pregunta

Incluso podemos pensar que su respuesta es condicional a qué tan **severa (difícil)** es la pregunta. Por ejemplo, no es lo mismo “tener un salario” a “contar con seguridad social completa”.

Podríamos pensar también que la probabilidad de respuesta depende de qué tan bien una pregunta distingue (**discrimina**) entre sujetos con valores diferentes. Por ejemplo, la seguridad social puede hacer un mejor trabajo en distinguir a dos personas con niveles distintos de  $\theta$  que tener contrato

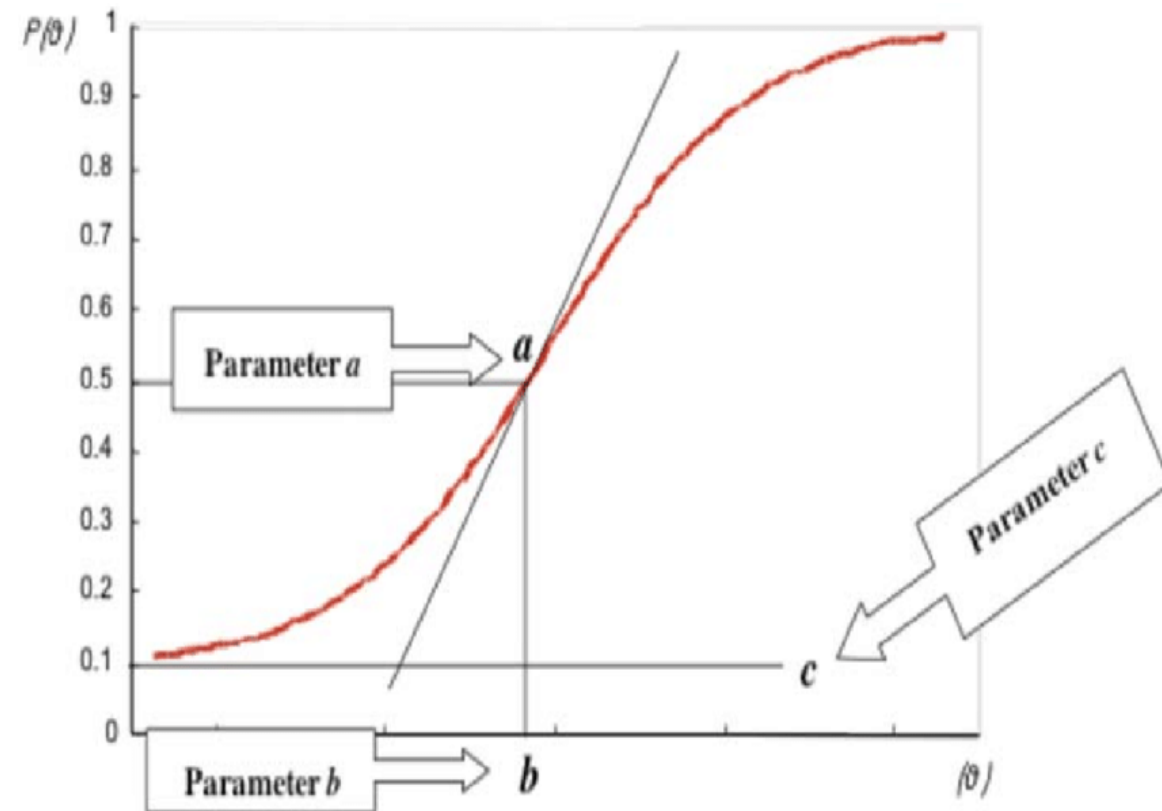
# Teoría de respuesta al ítem

- Parte de que las unidades de análisis tienen un valor latente respecto al fenómeno de interés:  $\theta$  (Es estructural)
- Las respuestas (score observados) son una probabilidad condicional a  $\theta$  y otros posibles parámetros
  - Severidad / Dificultad
  - Discriminación
  - Suerte
- Estos tres son sujetos a perturbaciones

# TRI o IRT en inglés

$$P(X_{ij} = 1 | \theta, a, b_i) = \frac{e^{Da(\theta_j - b_i)}}{1 + e^{Da(\theta_j - b_i)}} \quad (14.1)$$

where  $P$  means probability,  $X = 1$  means that the item is correct (or the response is 1), and the symbol  $|$  means “conditional on.” So we read  $P(X_{ij} = 1 | \theta_j, a, b_i)$  as: The probability of a correct response to item  $i$  by examinee  $j$  conditional on examinee  $j$ ’s ability and the difficulty ( $b_i$ ) of the item. The wording “conditional on” simply means that the probability of a correct response can (and usually does) depend on the examinee’s ability level and the item difficulty. The expression  $P(X_{ij} = 1 | \theta_j, a, b_i)$  is often shortened to  $P(\theta)$ , which will be used in the remainder of the model description.



# TRI de un parámetro

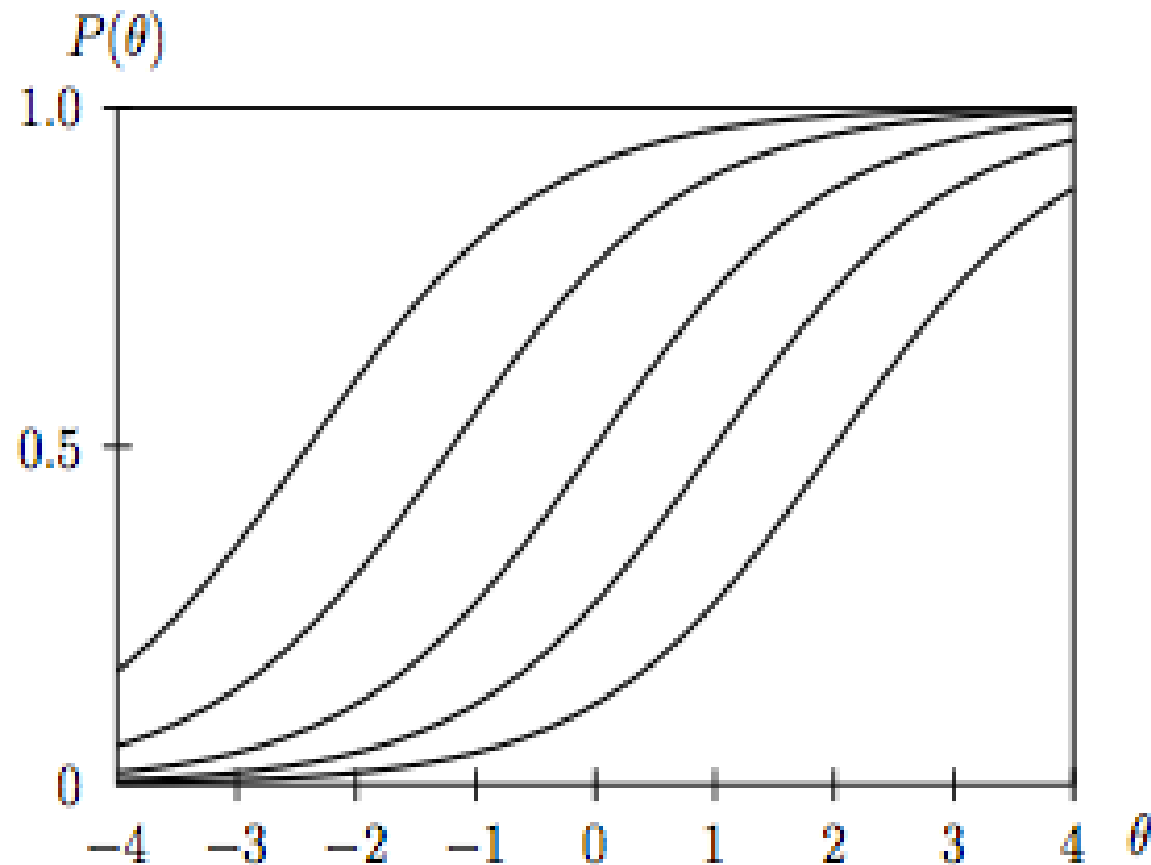
El TRI de un parámetro sólo usa “b”  
(Severidad / Dificultad)

Asume que todos los indicadores  
“Discriminan” igual.

También se le conoce como **modelo Rasch**  
(Aunque la gente que sigue Rasch NO  
estarían muy de acuerdo).

Rasch quería separar cualquier relación  
que pudiera tener el “examinado” del  
“test”.

IRT ve la interacción entre ambos



## Explicación de pizarrón

EMSA: **Por falta de recursos** en los últimos tres meses

1. Tuvieron poca variedad de alimentos
2. **Se quedaron todo un día sin comida**
3. Los adultos no tuvieron alguna de las tres comidas



# Rasch vs TRI de un parámetro



Food and Agriculture Organization  
of the United Nations

Lo importante es qué tanto se parecen  
los datos a la teoría

El modelo es correcto y los datos  
deben ajustarse

## The Food Insecurity Experience Scale:

Measuring food insecurity through people's experiences

**I**n September 2015, **the 193 Member States of the United Nations adopted** the 2030 Agenda for Sustainable Development to succeed the Millennium Development Goals. The UN Statistical Commission (UNSC) agreed in March 2016 on a global indicator framework comprising 230 indicators to monitor targets and measure progress towards achievement of the 17 new Sustainable Development Goals (SDGs).

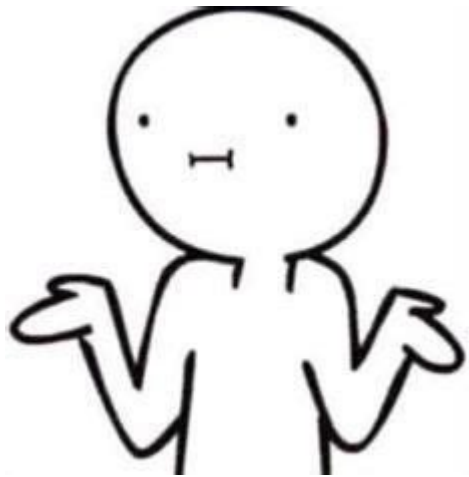
The UNSC created the Inter Agency and Expert Group on SDG indicators (IAEG-SDG), **whose members are the chief statisticians (or their delegates) of 28 countries**, elected on a rotating basis and representing all UN regions. The group includes regional and **international agencies as observers** with no deliberating power but who play an important advisory role.

IAEG-SDG has **appointed FAO to serve as custodian of 21 indicators**. FAO's responsibilities include maintaining the methodology needed to properly compile these indicators, providing technical support to countries, receiving data/indicators from countries to maintain a global database, and producing regional and global aggregates to report to the UN Department of Economic and Social Affairs (UNDESA).



# Rasch vs TRI de un parámetro

Resto de la comunidad estadística: Los modelos son imperfectos y lo mejor que podemos hacer es evaluar si resultan en una buena representación del mundo observado



Rasch: Los datos deben ajustar al modelo

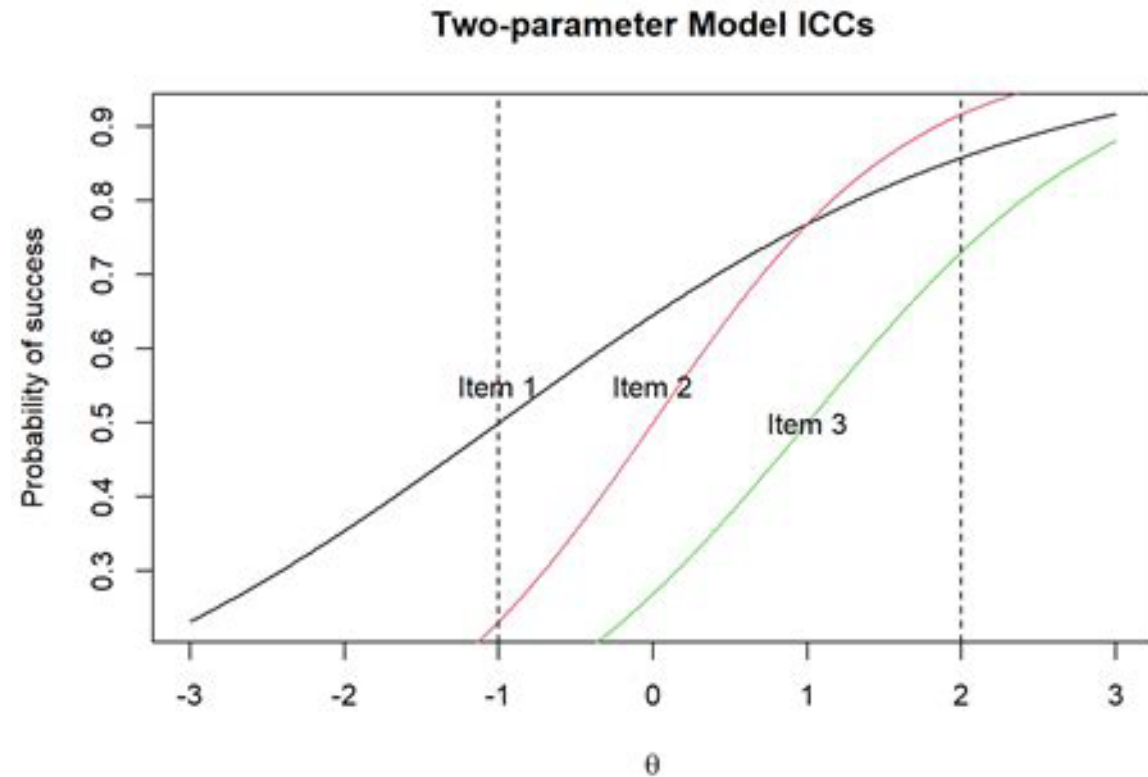
TRI: Los modelos deben ajustar a los datos



# TRI dos parámetros:

Severidad: Eje de las x

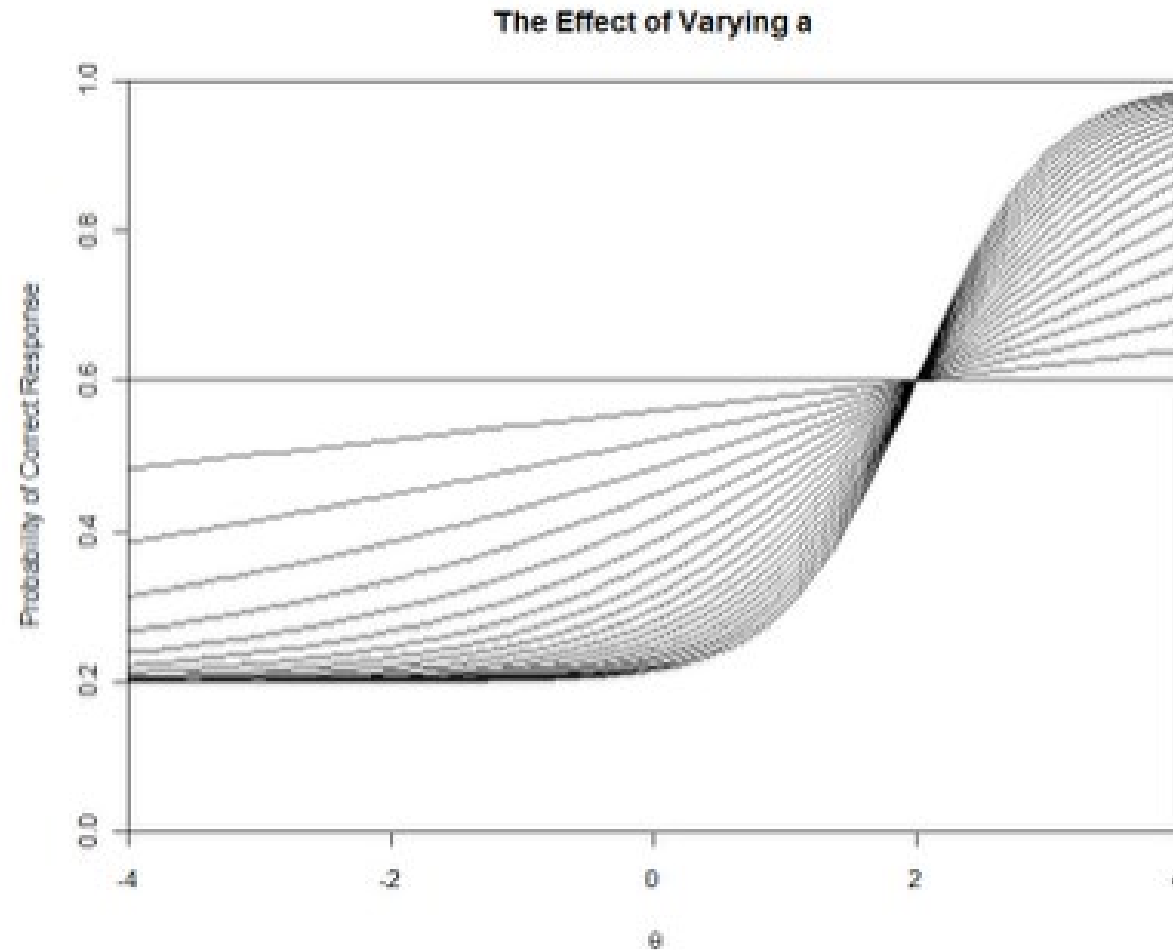
Discriminación: Pendiente



# La importancia de la pendiente


¿Qué curva nos hablaría de un mal indicador?

¿Qué quisiéramos en un buen indicador?



## Explicación de pizarrón

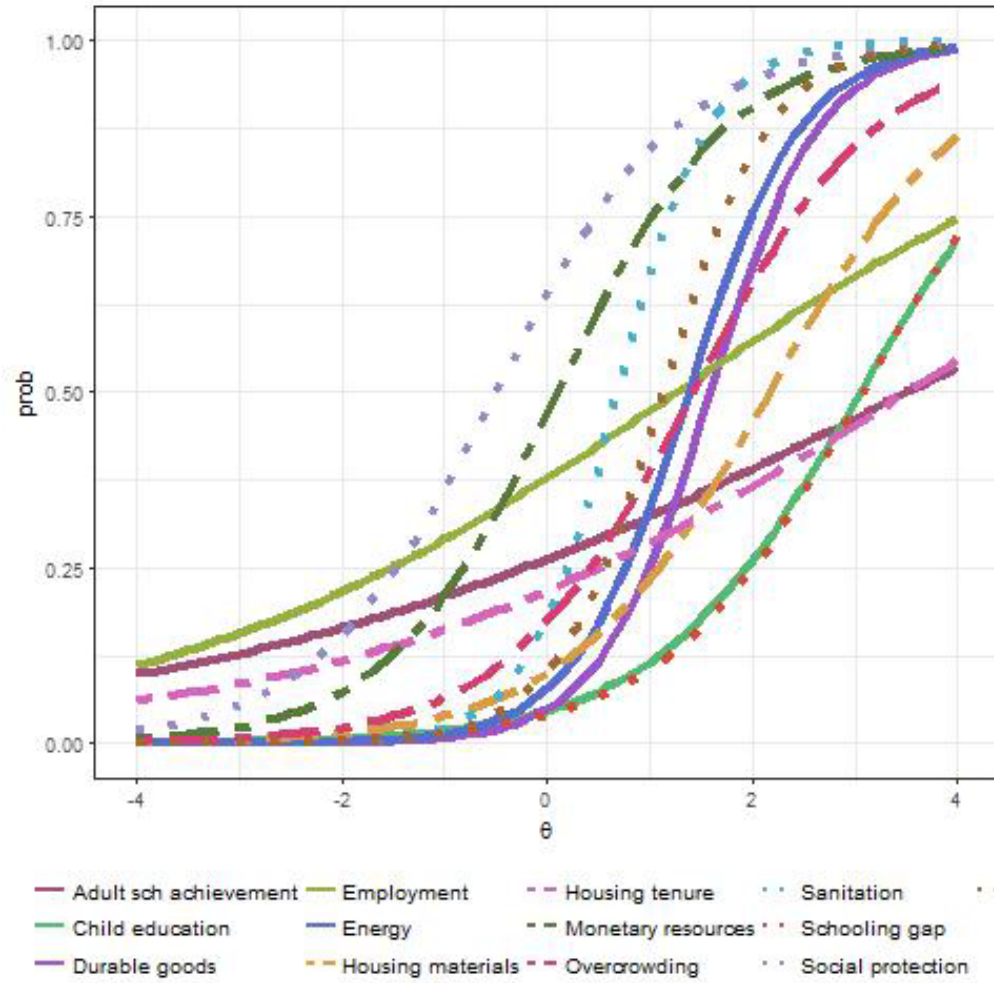
EMSA: **Por falta de recursos** en los últimos tres meses

- 
1. Tuvieron poca variedad de alimentos
  2. **Se quedaron todo un día sin comida**
  3. Los adultos no tuvieron alguna de las tres comidas
  4. Los adultos no tomaron cerveza

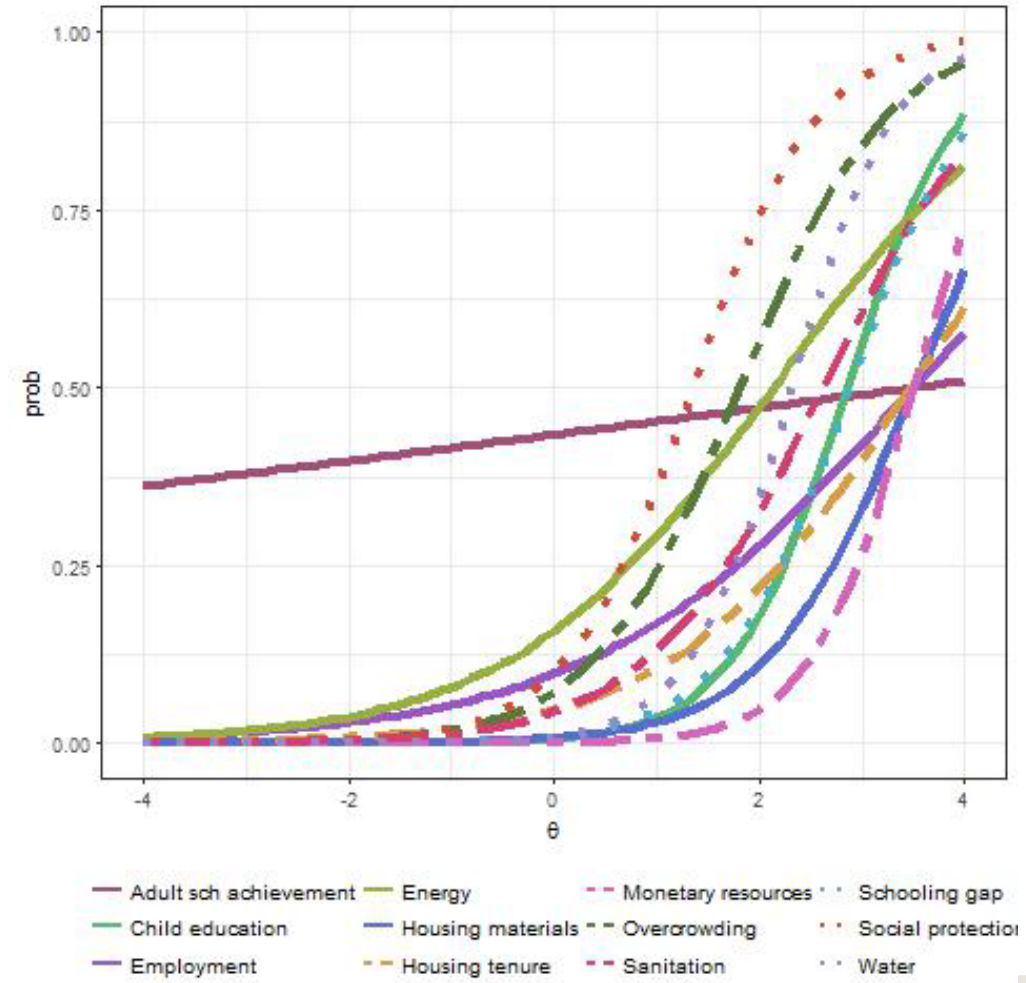
Inseguridad alimentaria



# Ejemplo

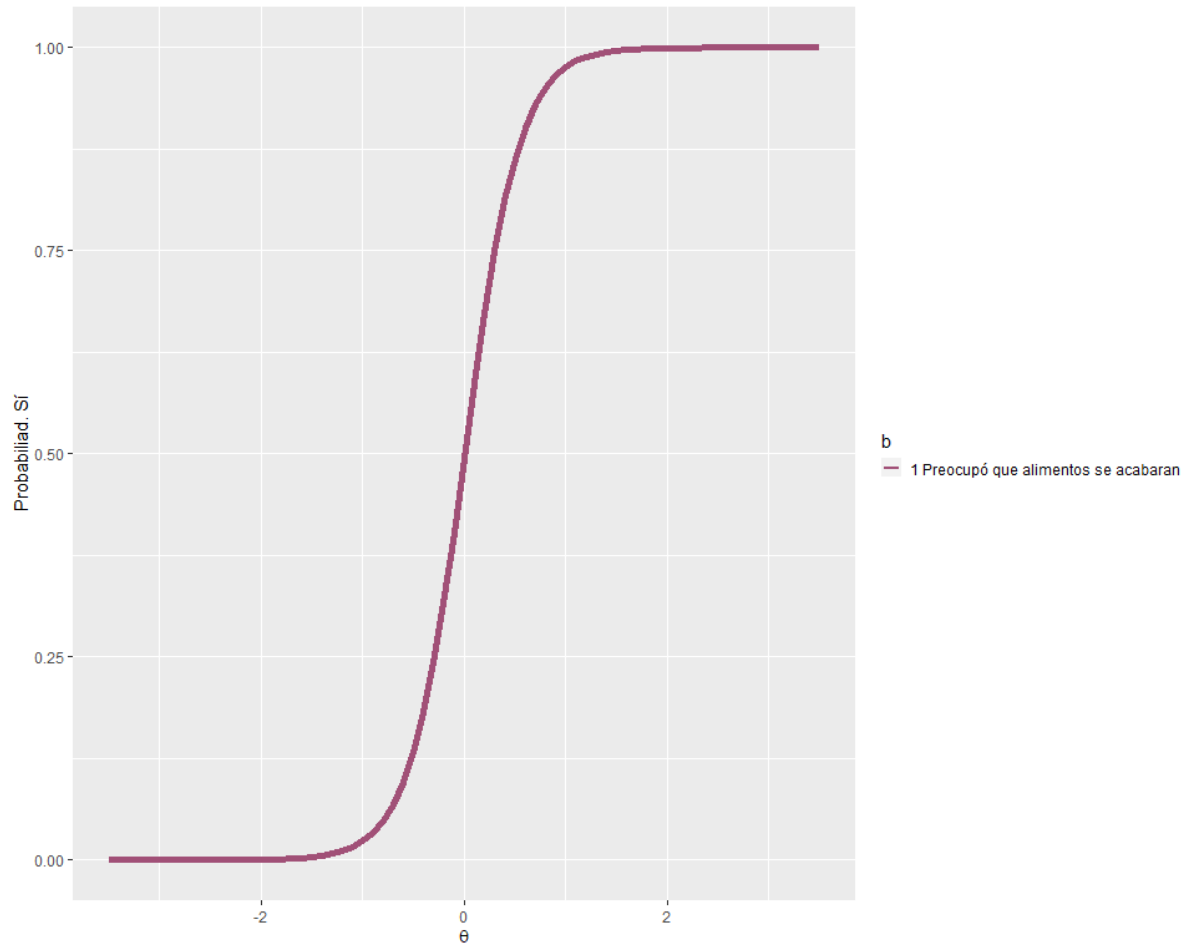


México (2012)

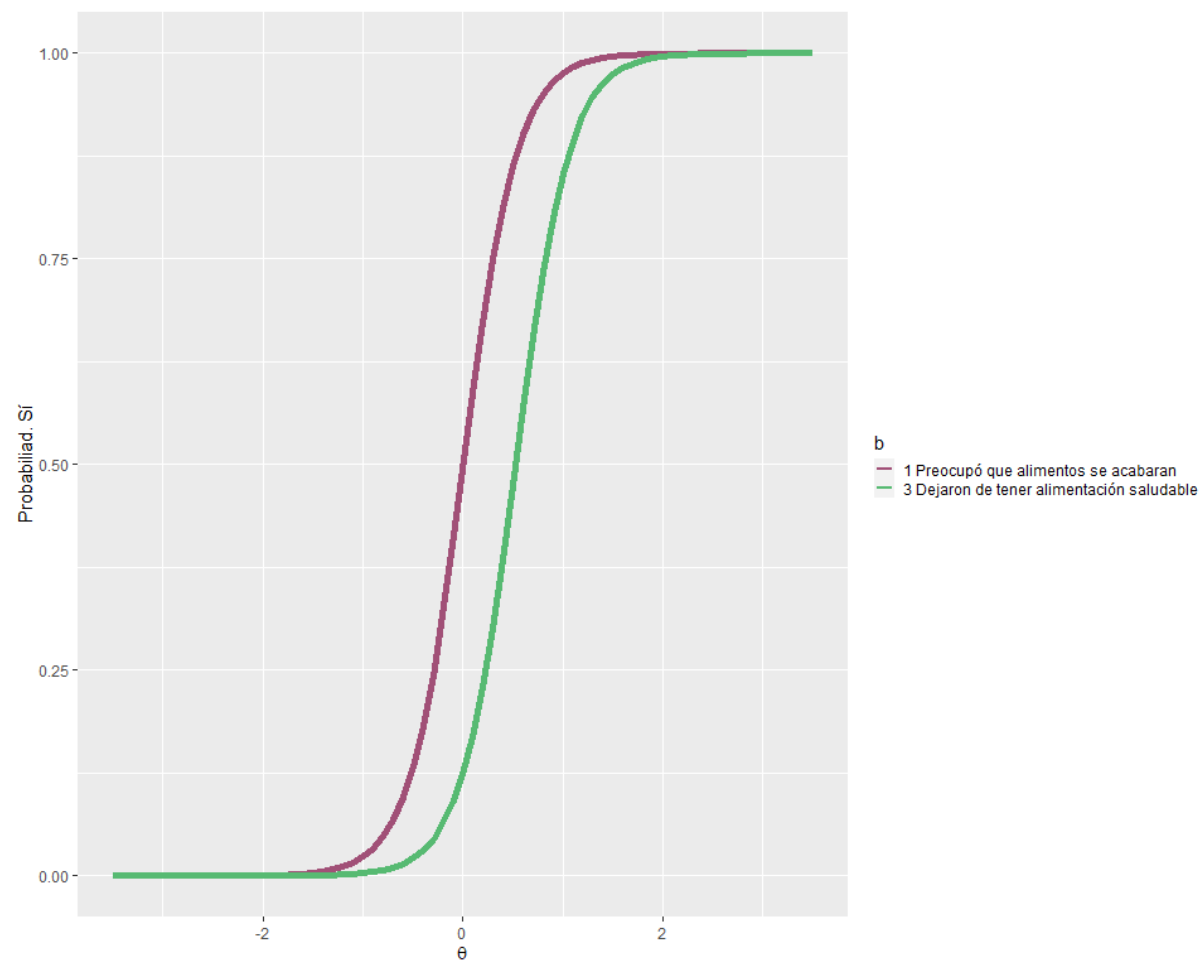


Argentina 2012

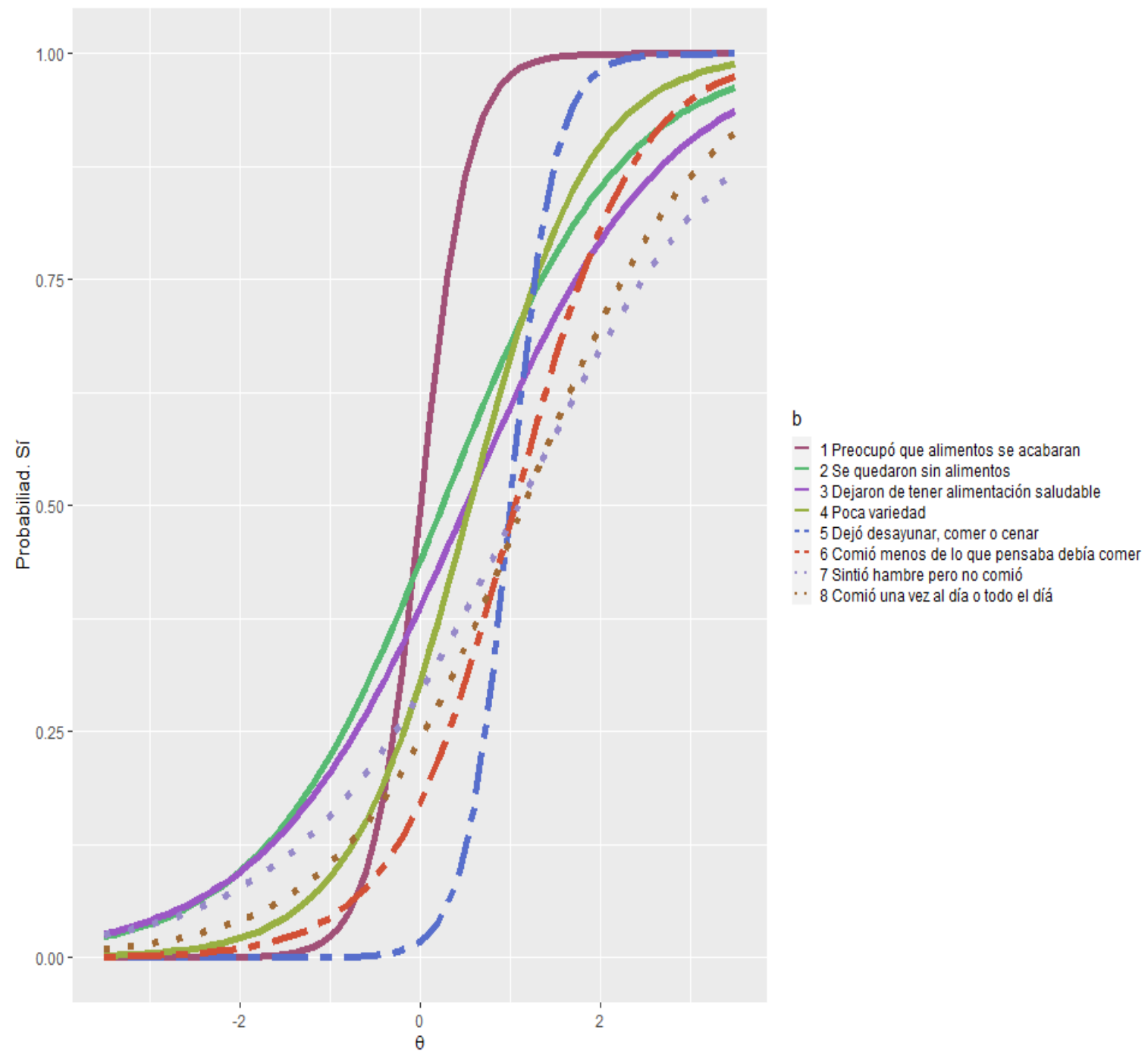
# Ejemplo: Modelo de un parámetro (severidad)



Curvas características. Modelo de un parámetro (Rasch). Adultos



Curvas características. Modelo de un parámetro (Rasch).Adultos



# ¿Uno o dos parámetros?

- La ventaja de TRI, al igual que en CFA, es que podemos comparar modelos usando estadísticos globales de ajuste
  - TLI, CFI, BIC, AIC, BICn
  - Podemos saber qué modelo hacer un mejor ajuste de los datos.



# Supuestos básicos de TRI

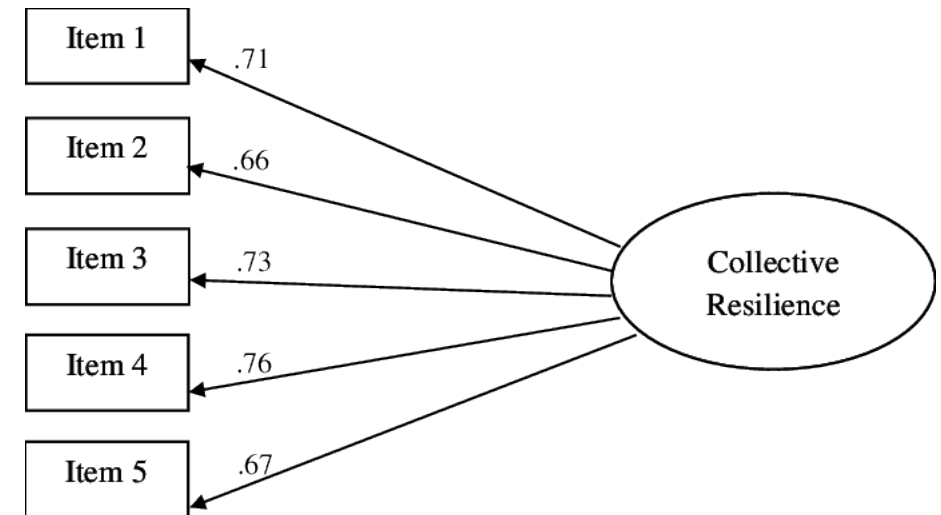
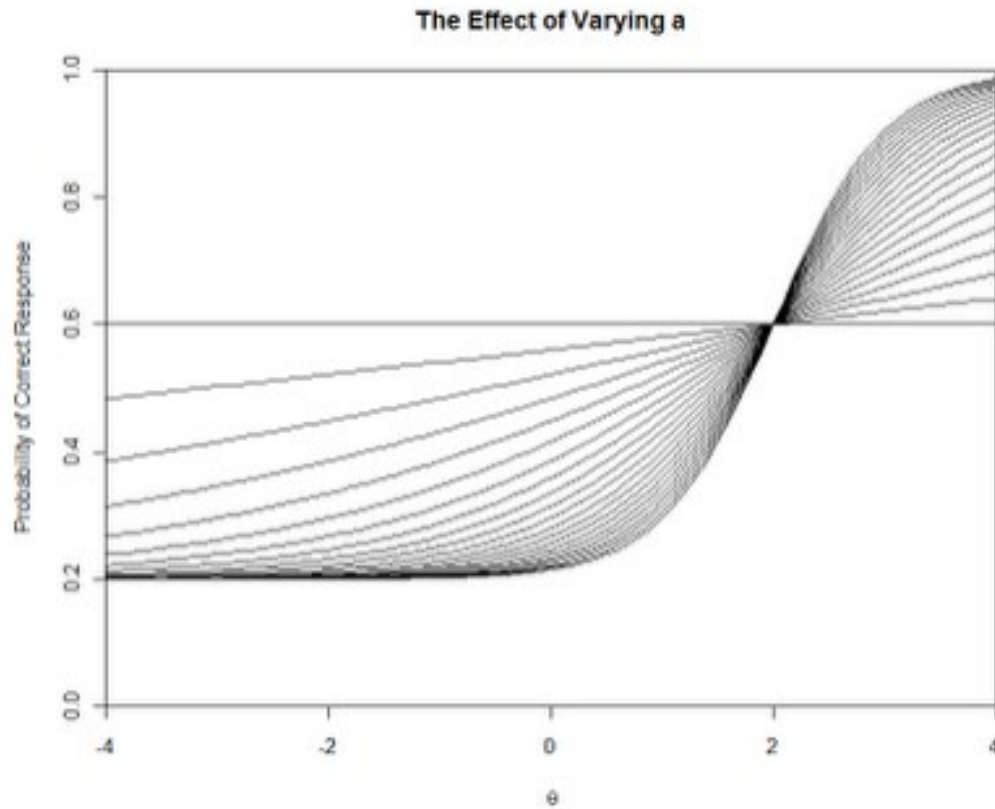
- Dimensionalidad: Están pensados para modelos unidimensionales, aunque pueden acomodar modelos multidimensionales
- Independencia local: La variación se debe al factor y la correlación entre ítems se debe al factor
- Forma funcional: Relación entre el factor y las respuestas

# Relación entre TRI y análisis factorial

- Parecen tener varios aspectos en común
- ¿Cuál puede ser la relación entre TRI y análisis factorial?
- Piensen en los parámetros de un modelo unidimensional
  - Discriminación - - - Carga factorial
- También pueden aplicarse a variables nominales y ordinales
  - Polytomous response theory

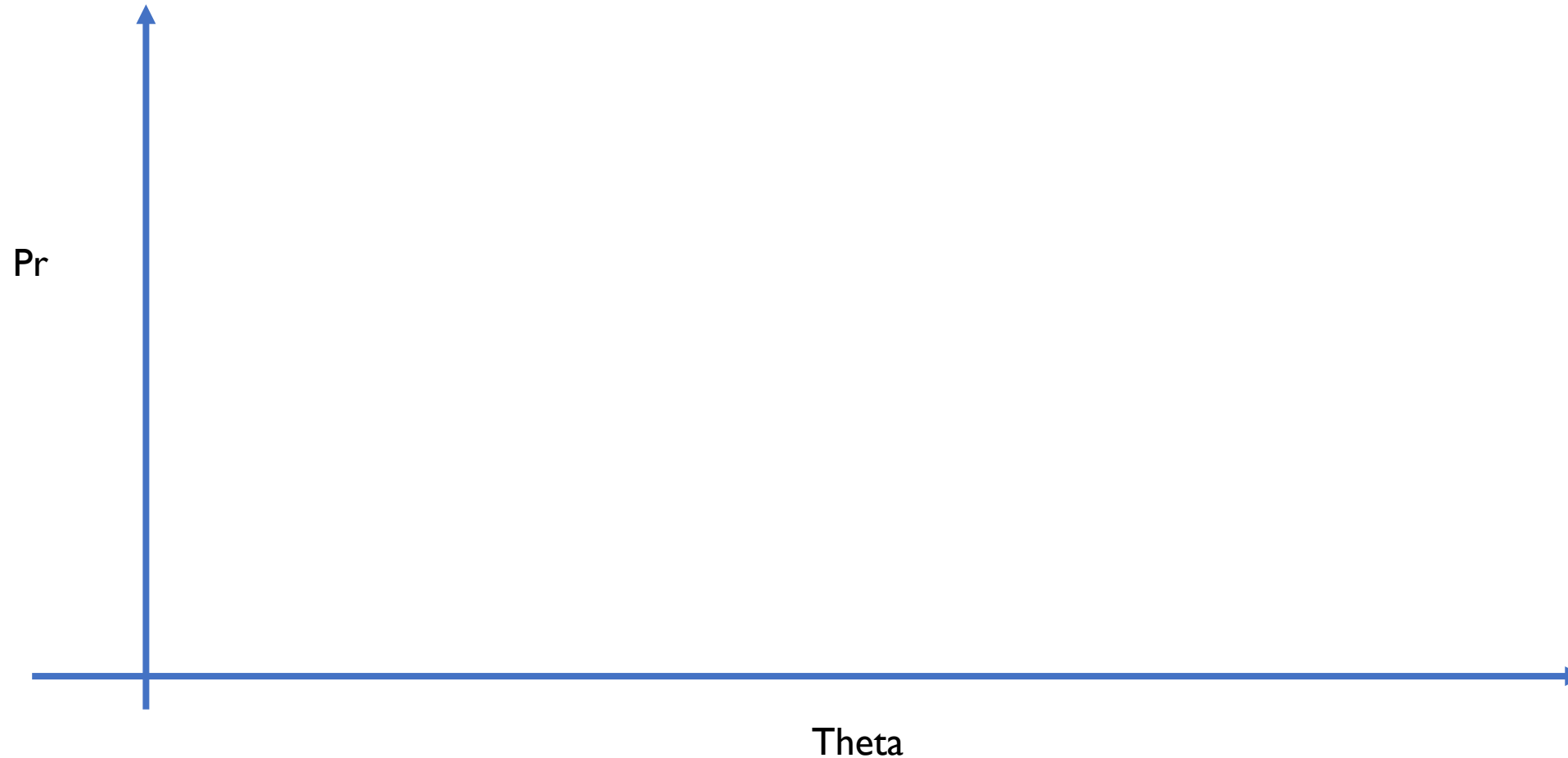


# Parámetro de discriminación ~ carga factorial



Son equivalentes. La diferencia es la parametrización (Logit v métrica del factor)

- Si las cargas factoriales son bajas ¿Cómo lucirían las curvas?



# Evaluación de discriminación y dificultad

- Discriminación baja es definitivamente indeseable. Qué bajo es muy bajo: .9 (~.4 con cargas factoriales estandarizadas).
- Un modelo TRI o factorial con bajas cargas factoriales es un mal modelo.
- Severidad. Depende del estudio, pero extremos  $>+3$  o  $<-3$  desviaciones estándar difícilmente va a ser útil



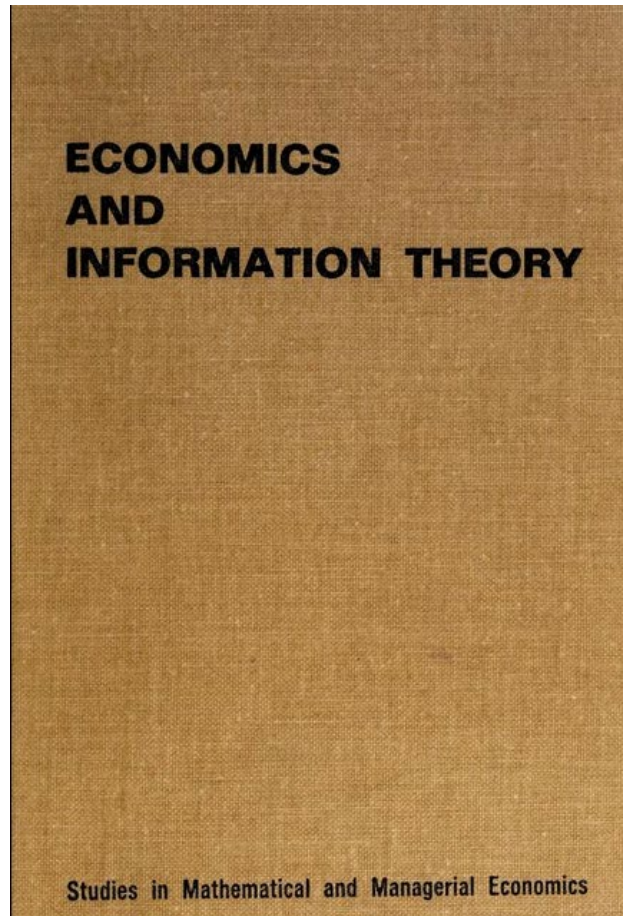
# TRI y teoría de la información

- Claude Shannon 1948
  - Información: Todo aquello de un mensaje que no nos sorprende -señal-
  - Entropía: Medida de la cantidad de información una vez que conozco el resultado

TRI: Respuestas aleatorias tienen poca información (alta entropía). Si la discriminación me dice la probabilidad de respuesta puedo deducir la información



# Teoría de la información y desigualdad



Producción de medidas de desigualdad

Descomposición comparable bajo  
índices de Theil

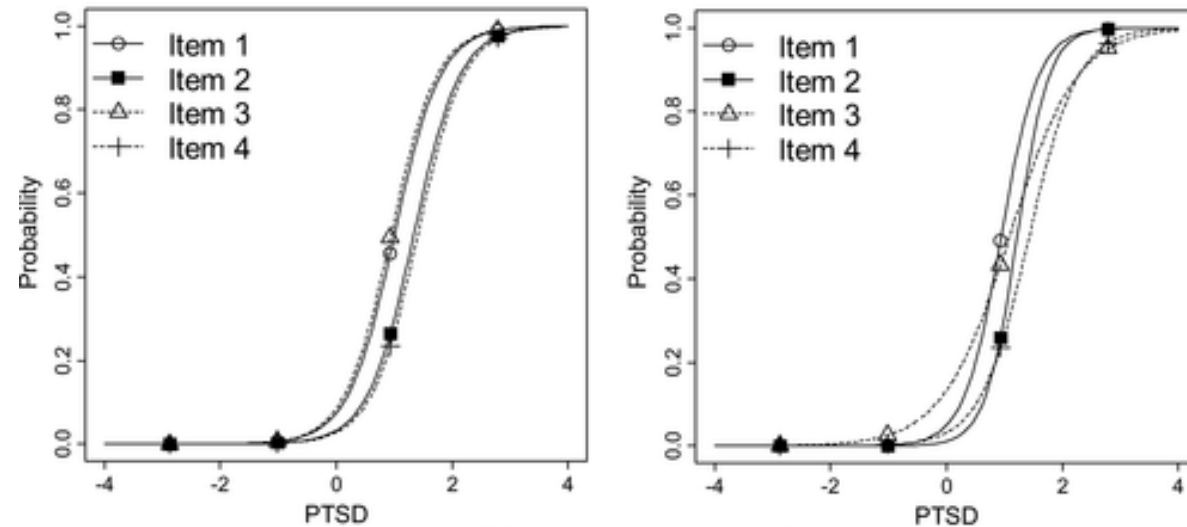


# ICC and information curves

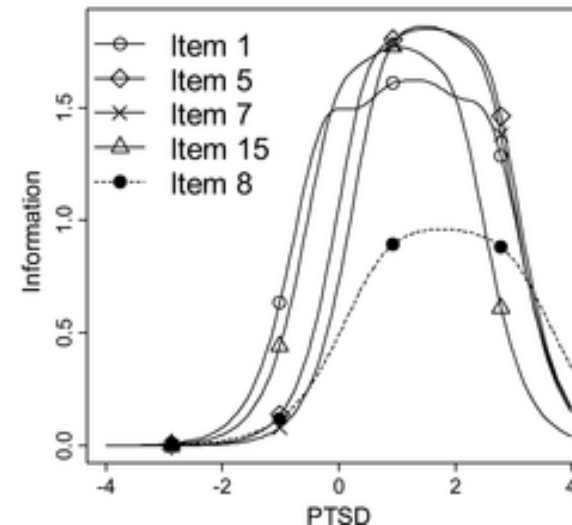
Las curvas de información me ayudan a visualizar “el rango” de la variable latente para el cuál tengo más información

Esto es útil para evaluar confiabilidad relativa

PC-PTSD Item Characteristic Curves

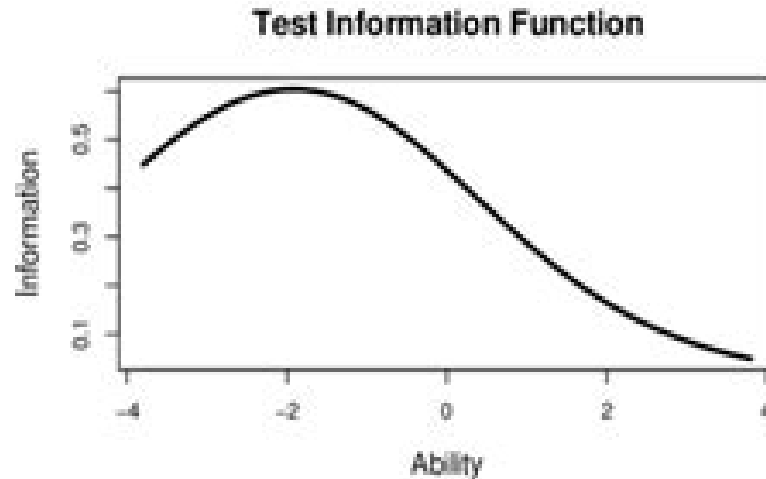
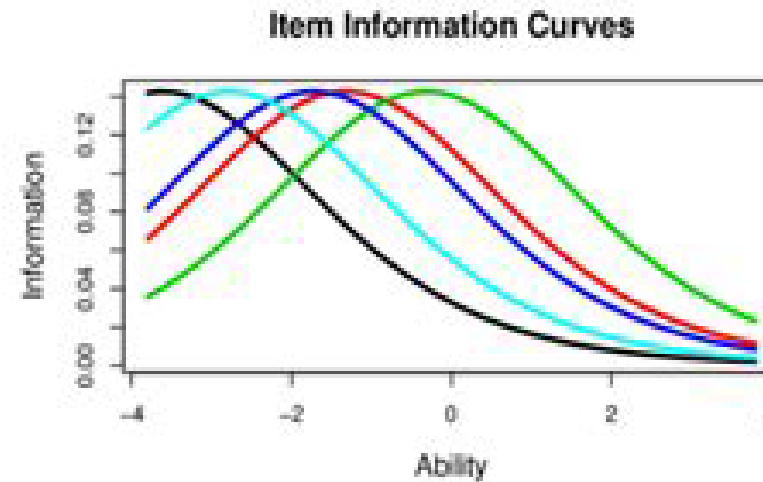
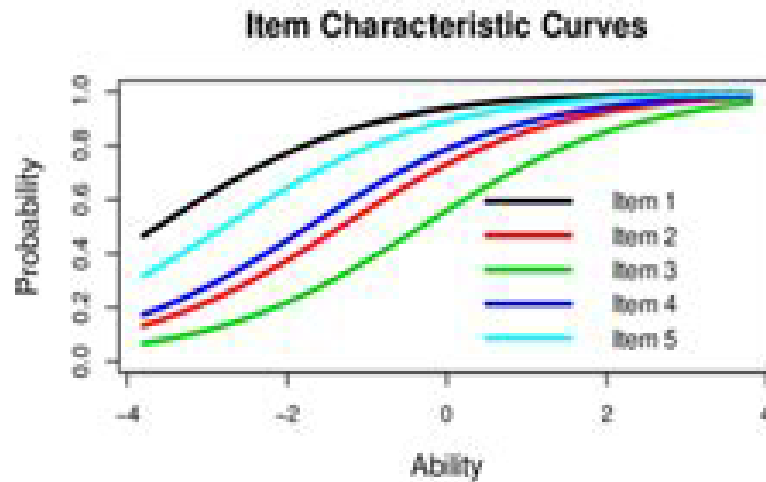


PCL Item Information Curves





# Curvas en TRI



Total Information: 3.763

Information in  $(-4, 0)$ : 2.165 (57.53%)

Information in  $(0, 4)$ : 0.768 (20.4%)

# Pasando de scores observados a scores latentes

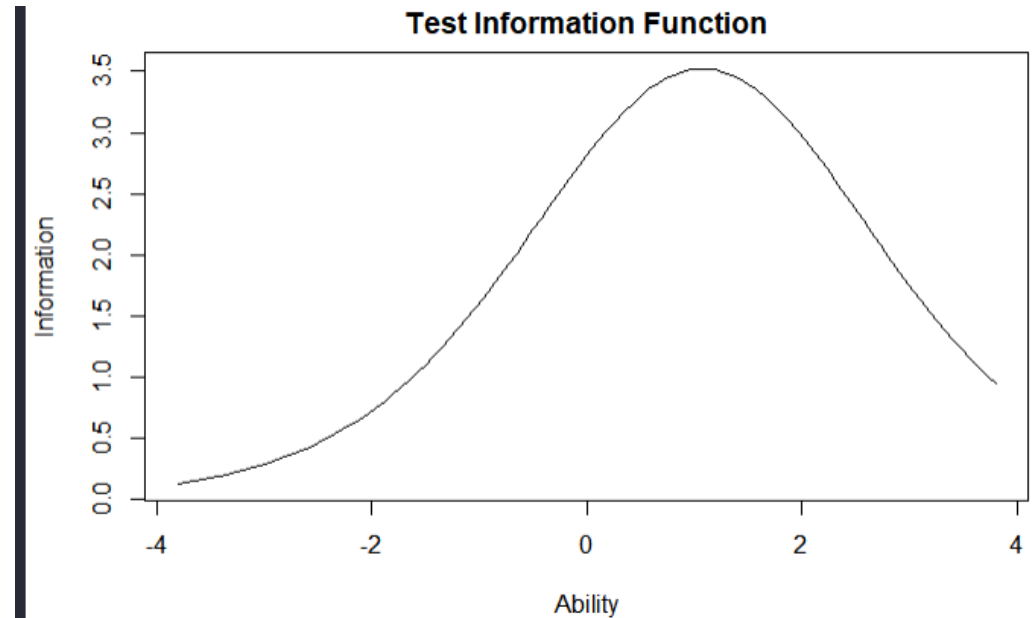
##	Lavadora	Televisión	Agua entubada	Refrigerador	GasNatural	Drenaje
## 1	0	0	0	0	0	0
## 2	0	0	0	0	0	0
## 3	0	0	0	0	0	0
## 4	0	0	0	0	0	0
## 5	0	0	0	0	0	0
## 6	0	0	0	0	0	0

##	Cisterna	Celular	Computadora	Netflix	Panel Solar	Jardin	Estufa
## 1	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	1
## 6	0	0	0	0	0	0	1

##	AireAcondicionado	Bicicleta	Obs	Exp	z1	se.z1
## 1	0	0	260	308.583613	-1.3723183	0.6710940
## 2	0	1	78	26.815590	-0.9565429	0.6137046
## 3	1	0	80	25.740288	-0.9565429	0.6137046
## 4	1	1	31	3.339630	-0.6034440	0.5701628
## 5	0	0	100	27.592368	-0.9565429	0.6137046
## 6	0	1	32	3.579925	-0.6034440	0.5701628

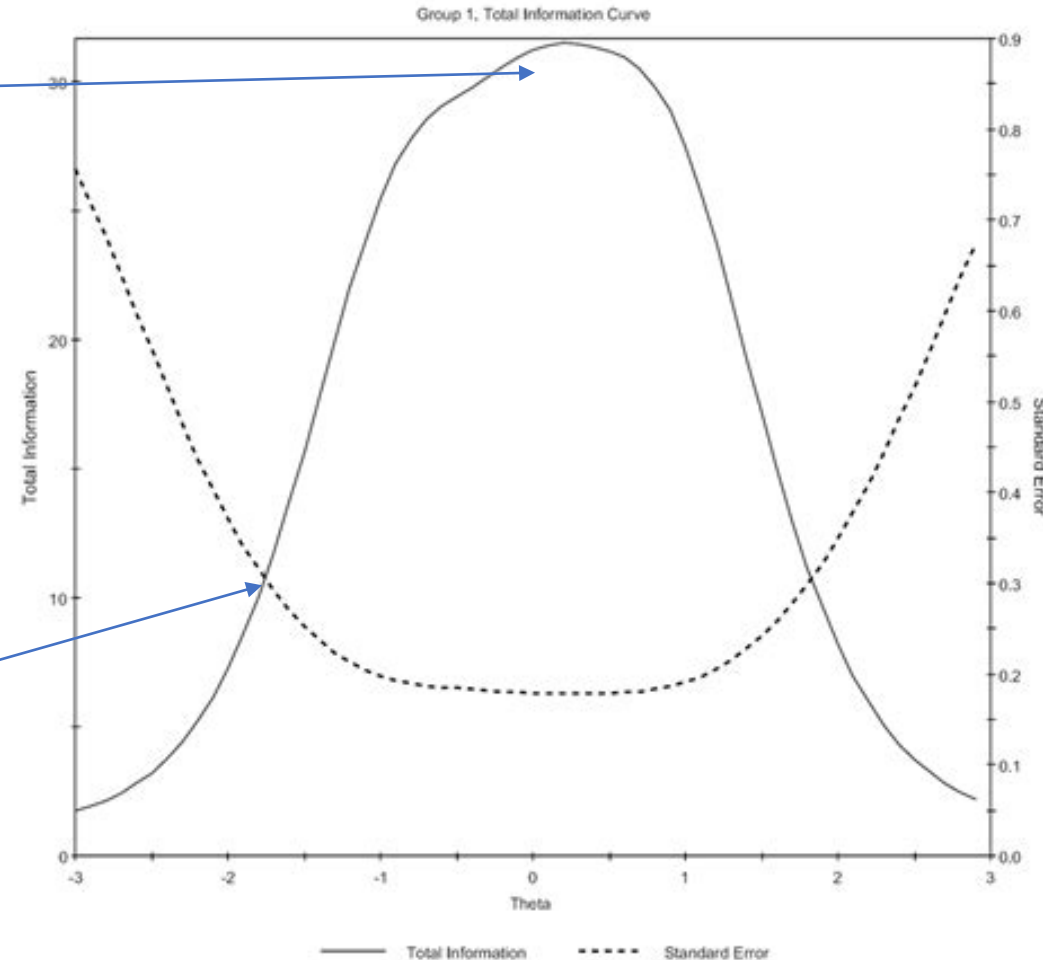


¿Cómo se obtienen?

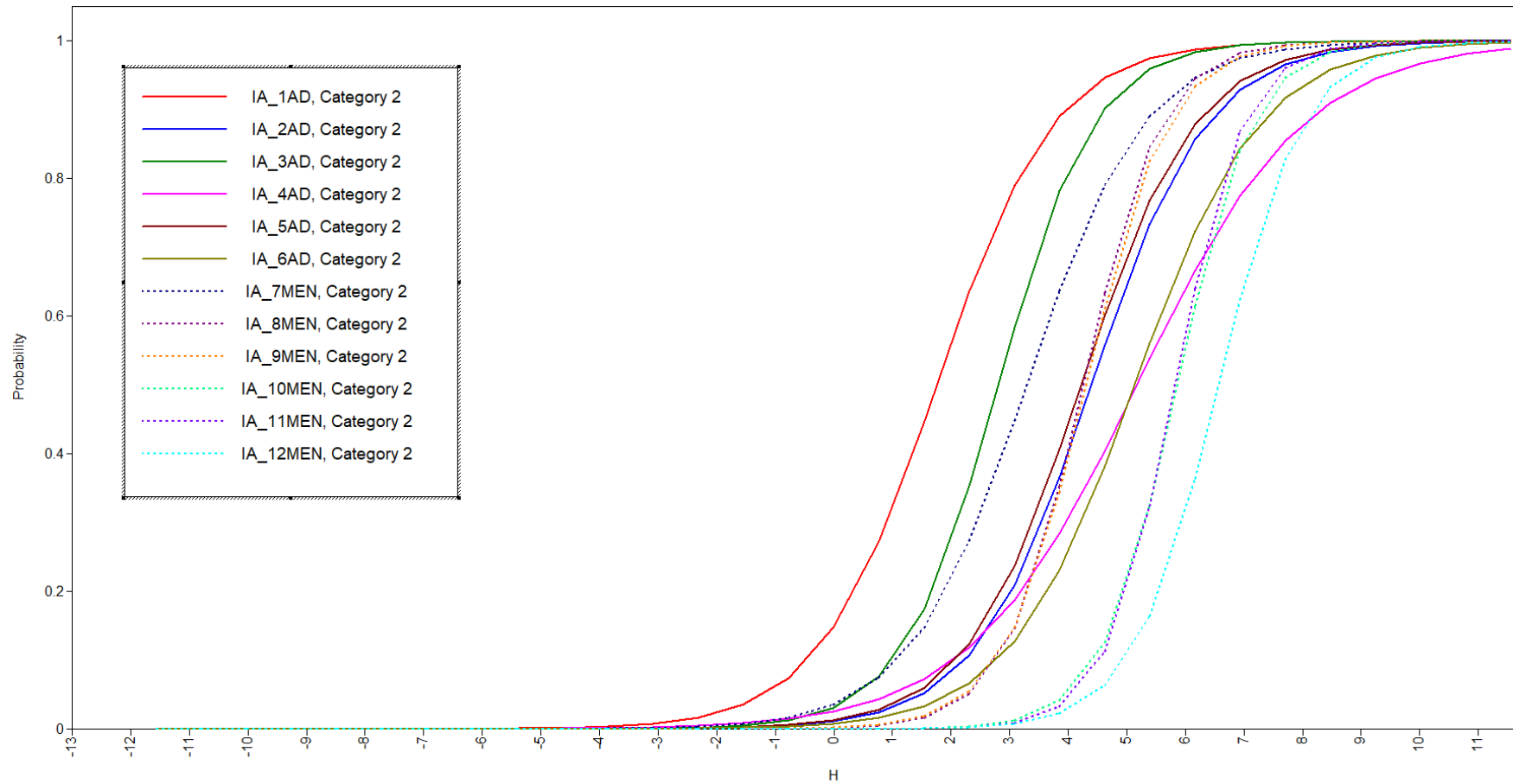
# Curva de información

Mucha  
información:  
Podemos hacer  
inferencia

Poca  
información  
: No  
podemos  
hacer  
buenas  
inferencias

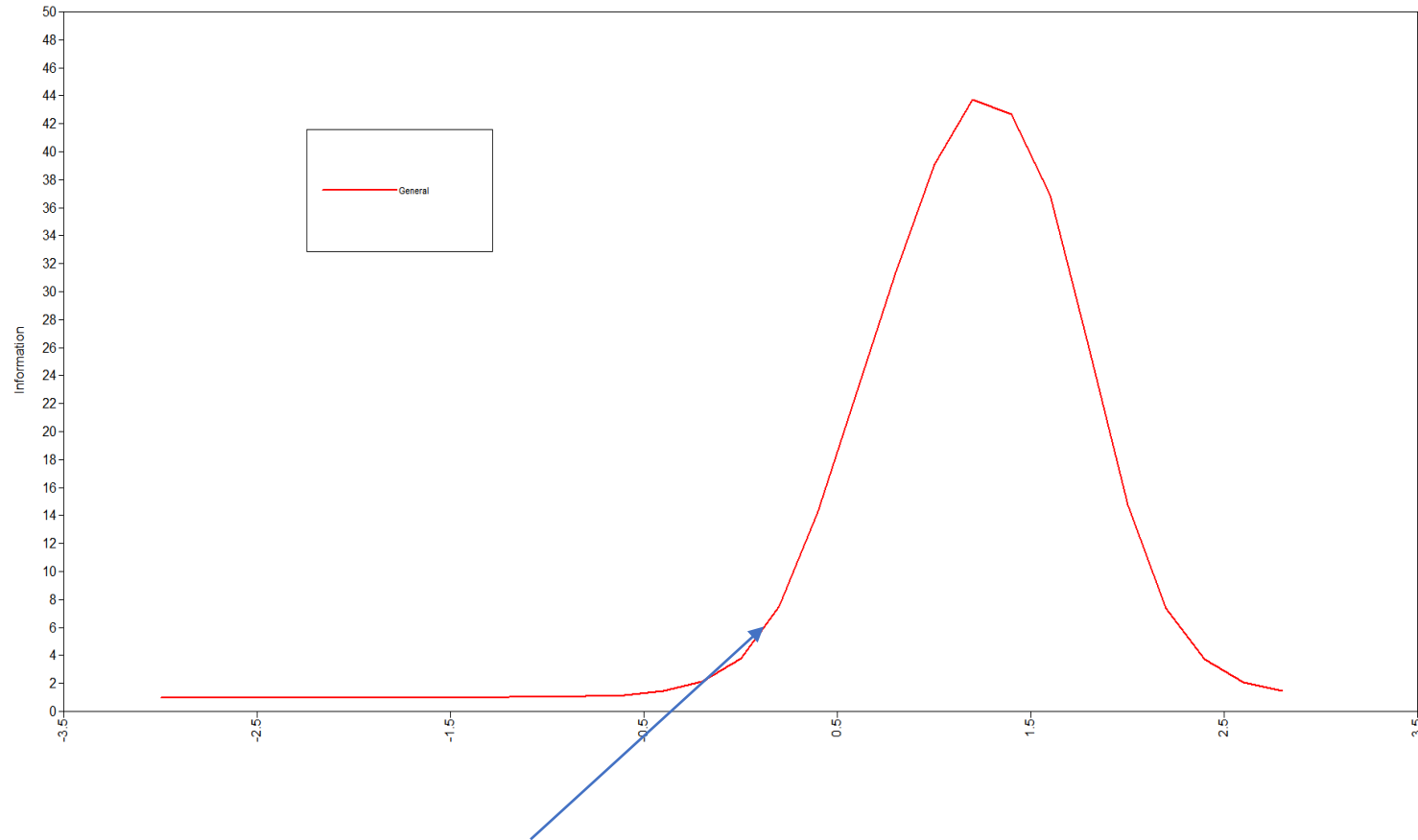


# Ejemplo: Escala Alimentaria



##	param	est
## 13	IA_1AD	0.455
## 15	IA_3AD	0.729
## 19	IA_7MEN	0.854
## 17	IA_5AD	1.096
## 20	IA_8MEN	1.105
## 21	IA_9MEN	1.118
## 14	IA_2AD	1.141
## 18	IA_6AD	1.333
## 16	IA_4AD	1.343
## 23	IA_11MEN	1.512
## 22	IA_10MEN	1.522
## 24	IA_12MEN	1.705

# Total Information Curve

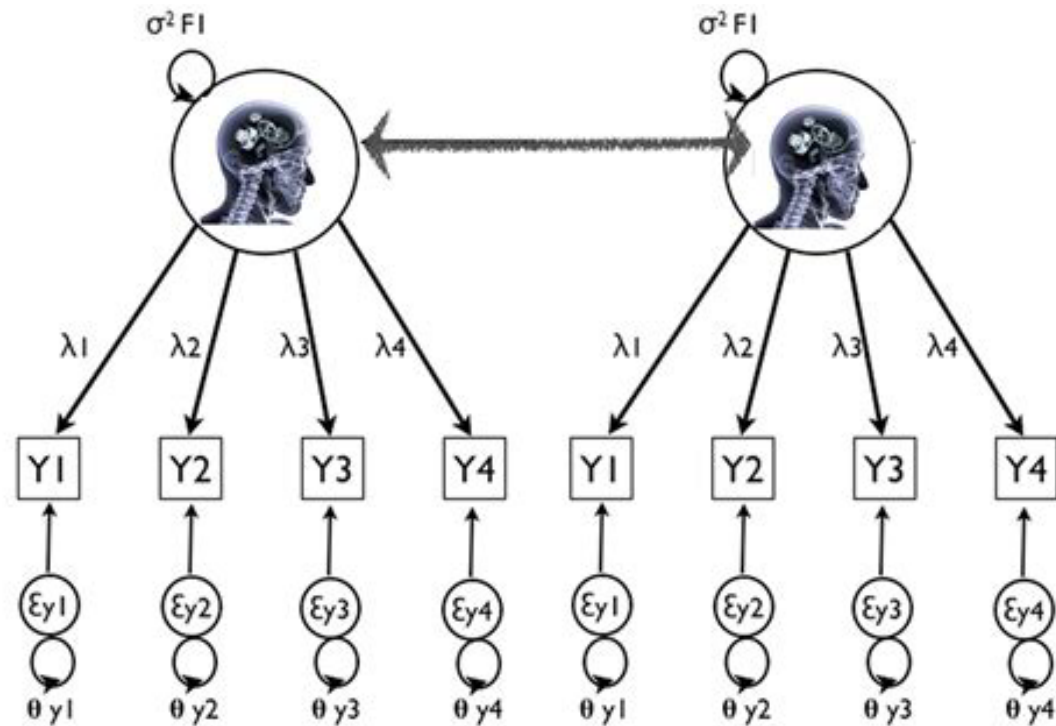


##	TIC	SEE	REL
## 1	1	1.00	0.00
## 2	2	0.71	0.50
## 3	3	0.58	0.67
## 4	4	0.50	0.75
## 5	5	0.45	0.80
## 6	6	0.41	0.83
## 7	7	0.38	0.86
## 8	8	0.35	0.88
## 9	9	0.33	0.89

Poca  
informació

# Usos de TRI

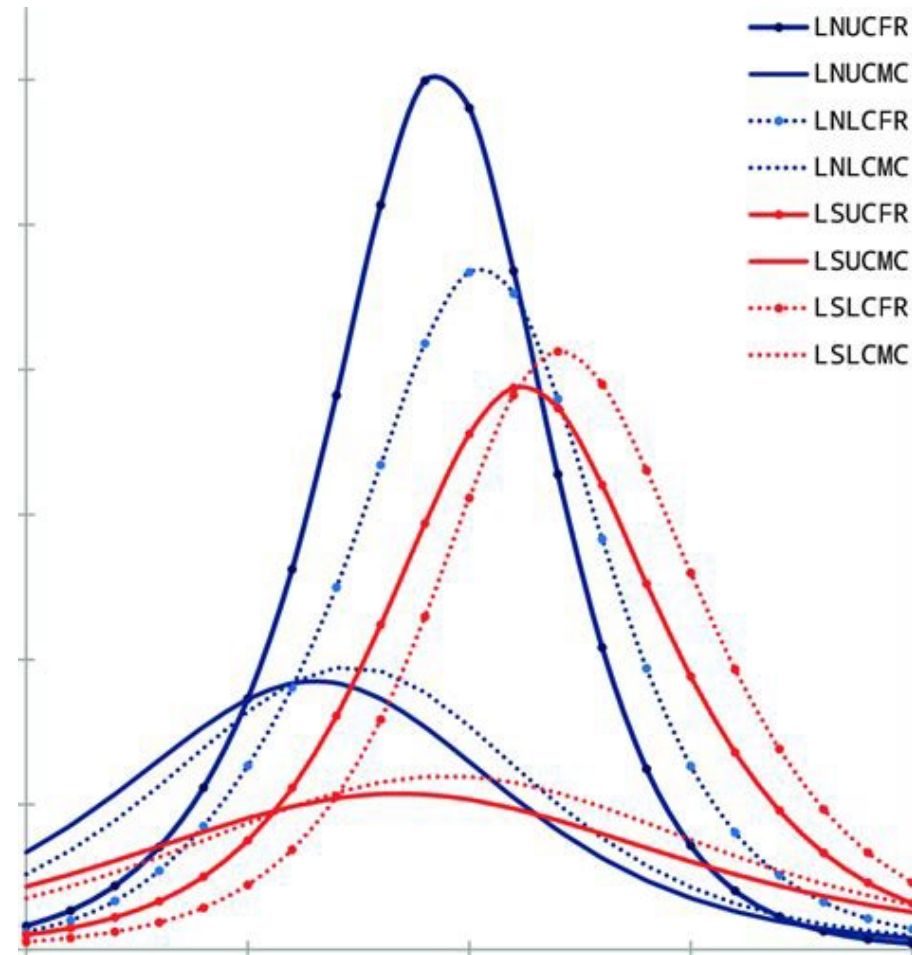
- Comparación de los puntajes de grupos a partir de variables categóricas



# Usos TRI: Comparación de escalas

Comparar escalas a partir de su información

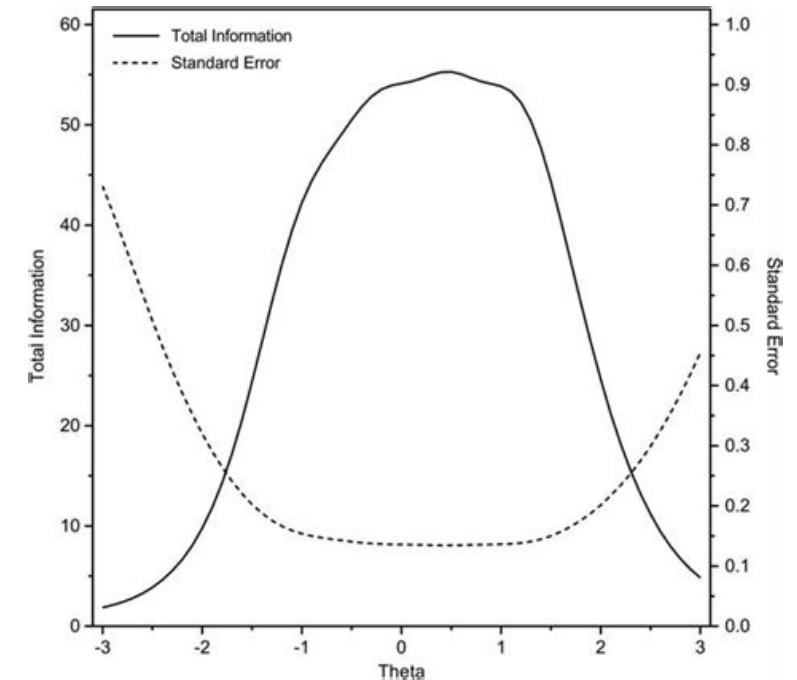
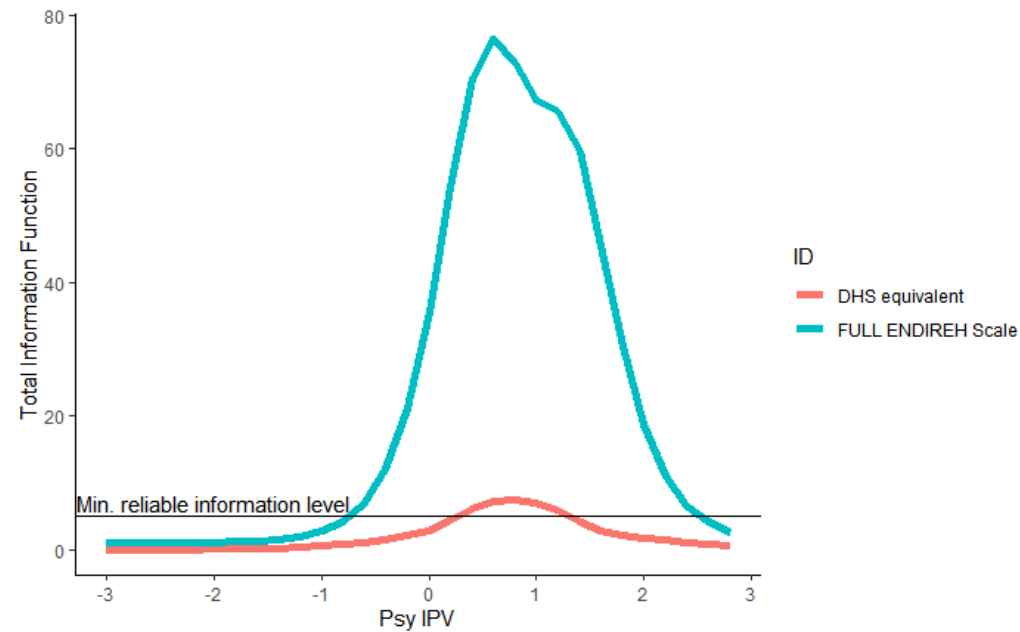
Hay cierta relación entre información y confiabilidad



# Ejemplo: Escala y subescala

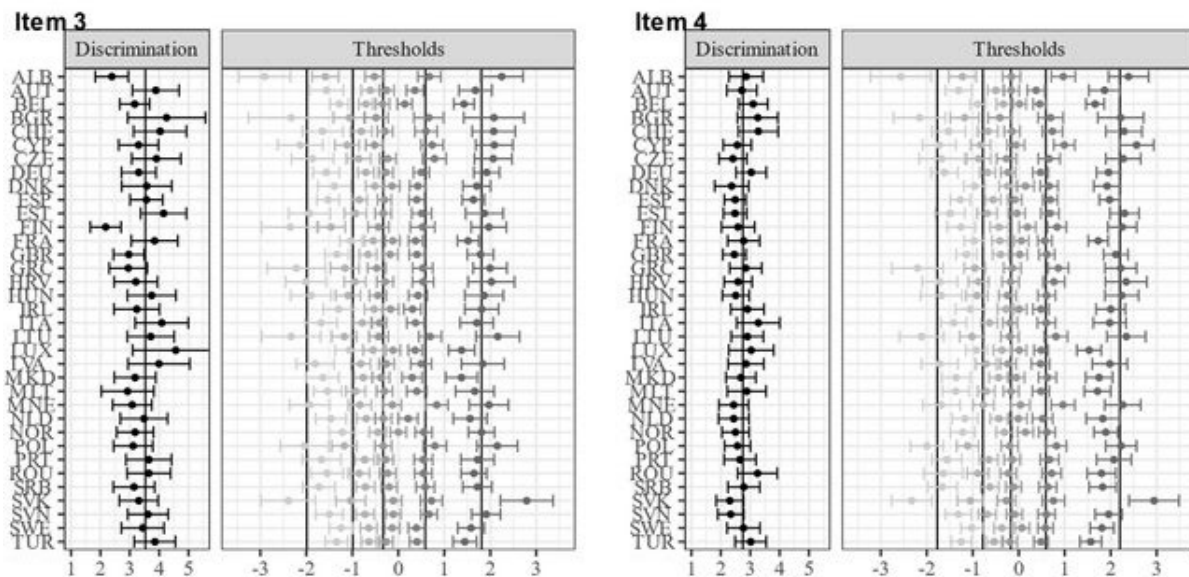
Esta es la información total de una escala - suma de las curvas de información-.

¿Mínimo de información?

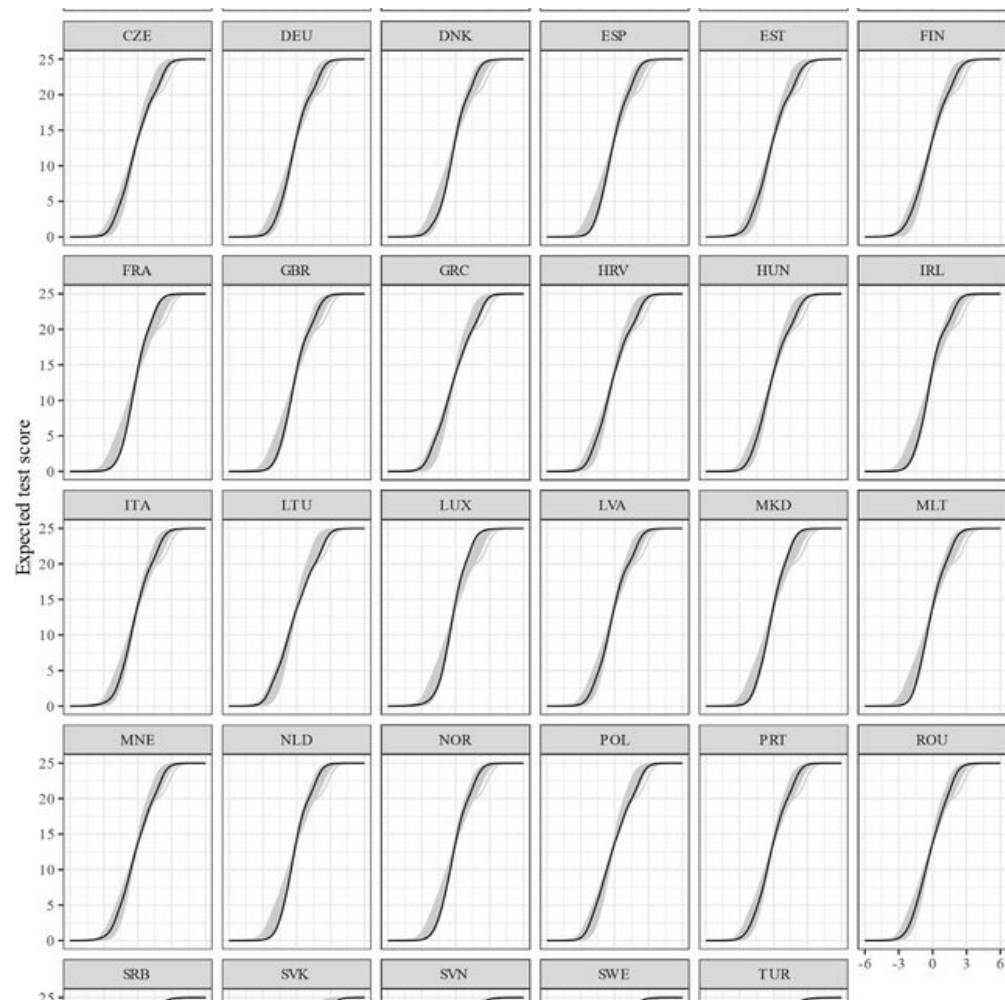




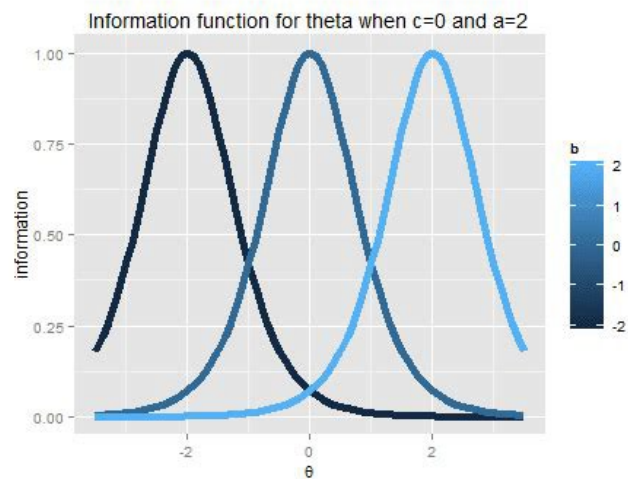
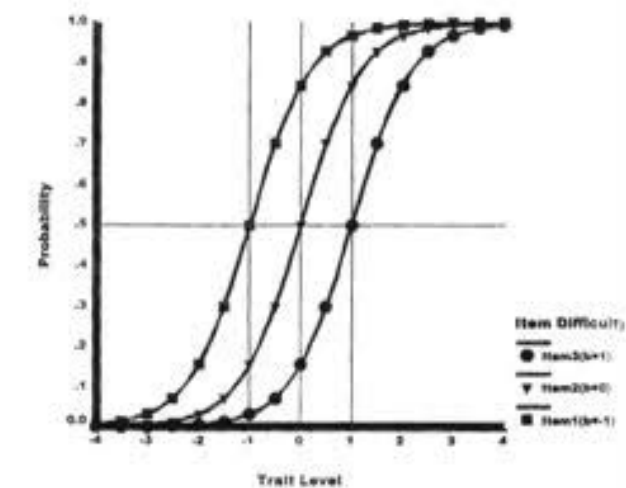
# Usos TRI



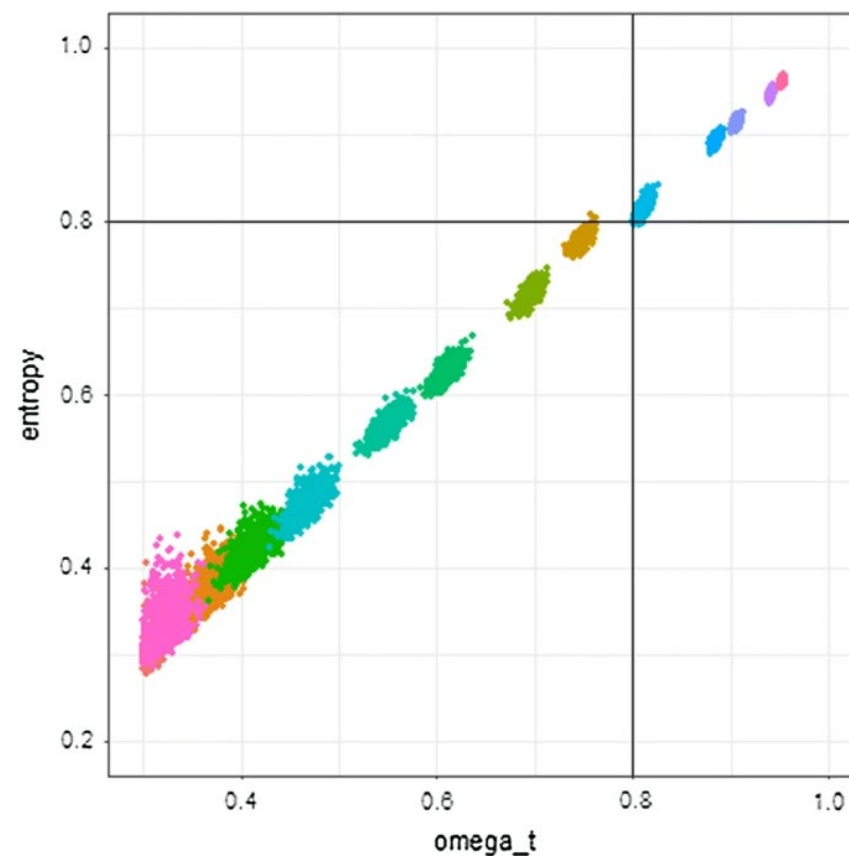
Medición invariante: Mido lo mismo y de la misma manera entre distintos grupos



# TRI y clasificación



## Entropía relativa



# IRT: Avanzado

SEM es un marco potente y flexible

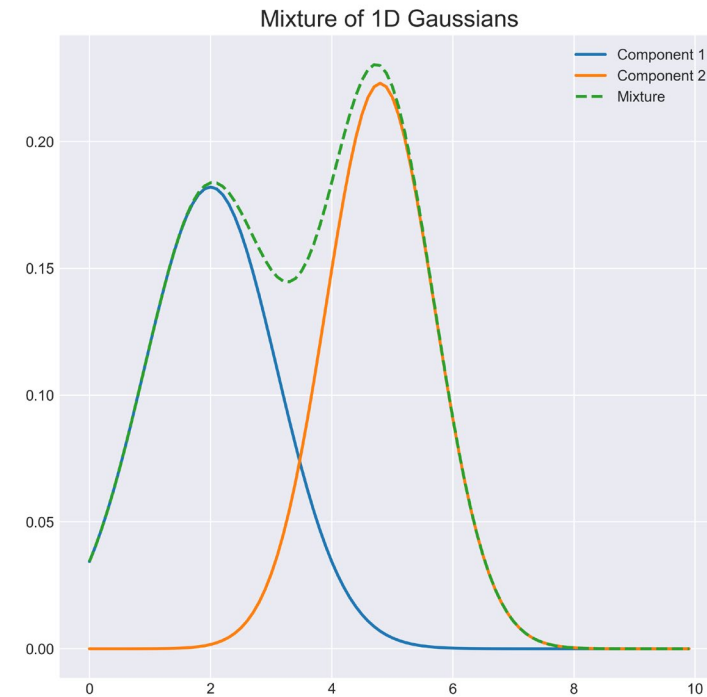
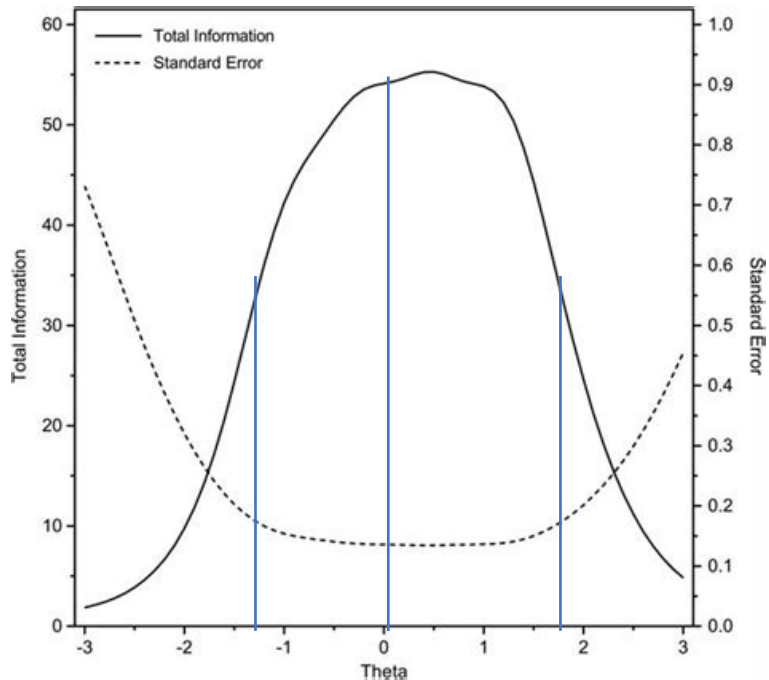
IRT es un caso especial de SEM que puede extenderse para abordar distintos problemas de investigación:

- Problemas de clasificación
- Problemas de muchos ceros



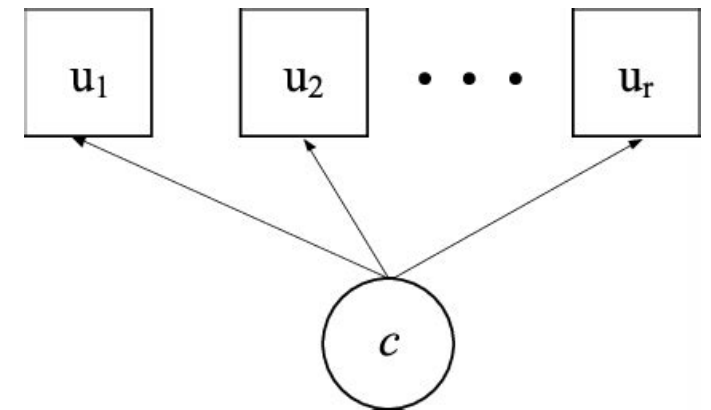
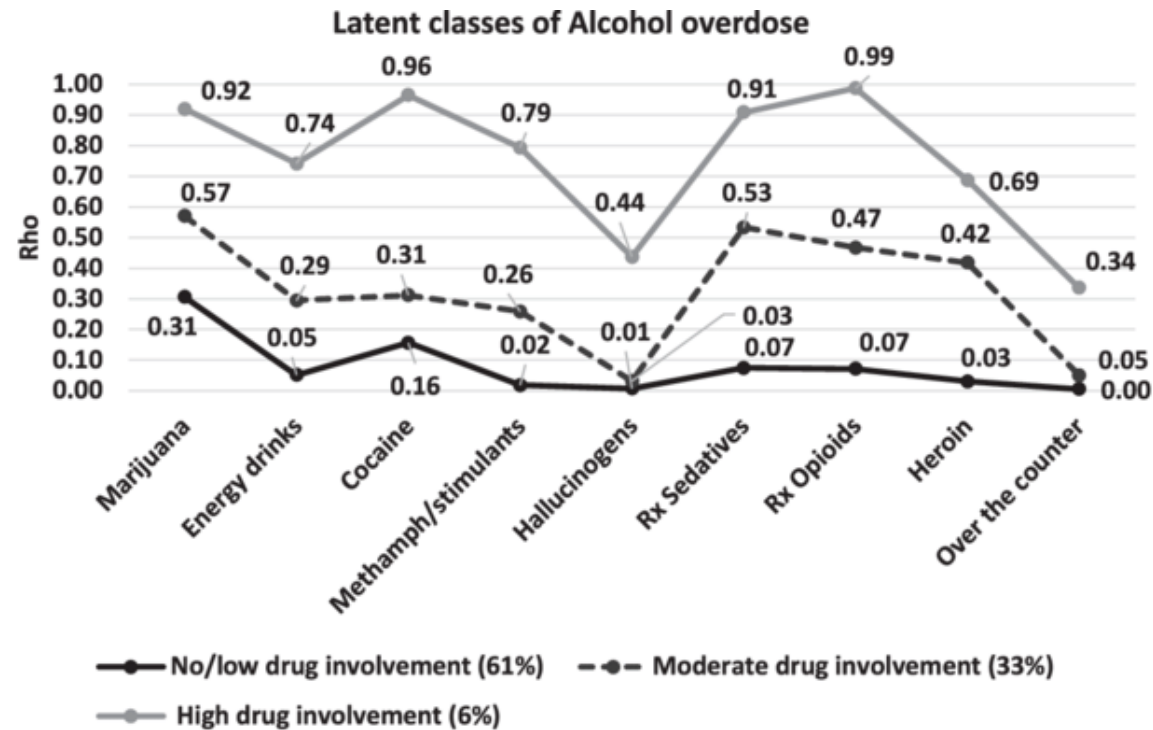
# Clasificación IRT

Dado un patrón de respuestas, nos gustaría clasificar a la población en grupos



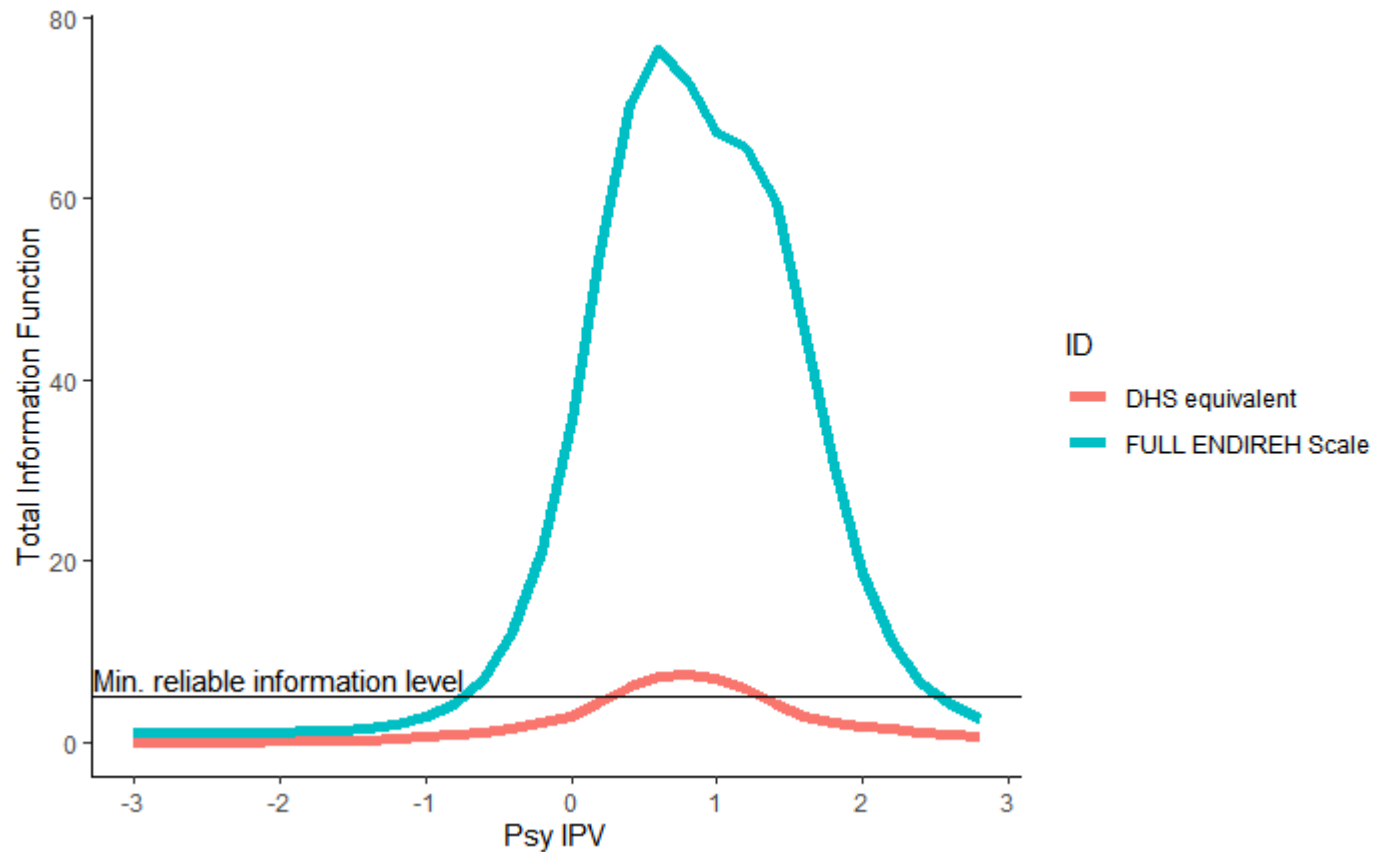
# Clasificación IRT

- Análisis de clases latentes (Mixture modelling)



La clasificación será tan buena como la confiabilidad de los scores

# IRT: Muchos ceros



Hay mediciones que requieren muchas preguntas para poder estimar con confiabilidad scores para ciertas áreas de la distribución

Cuando no se tienen esas escalas, uno tiene altos falsos positivos y negativos

# IRT + Clases latentes

- ¿Qué tal si combinamos un modelo factorial (IRT) con un modelo de clasificación?

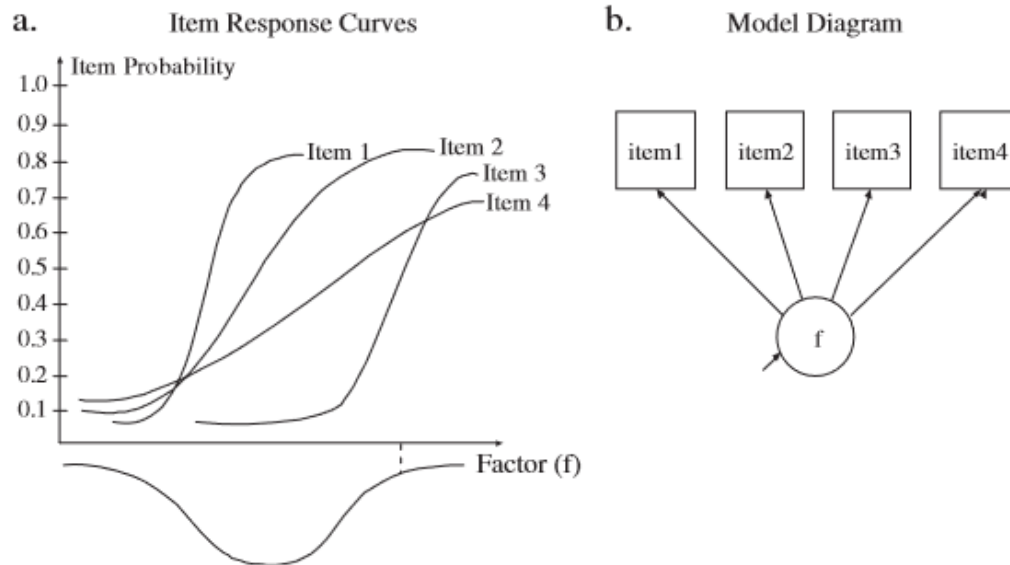
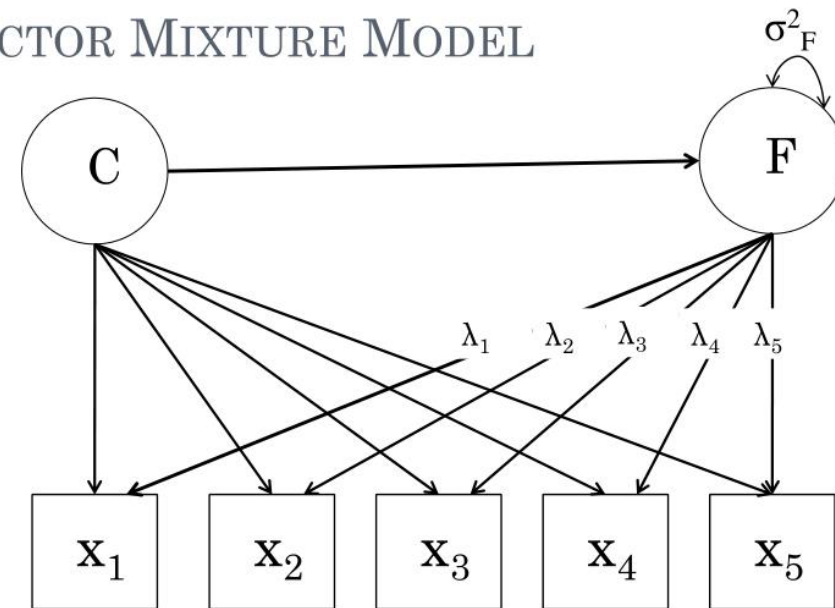


Fig. 1. Uni-dimensional factor analysis.

## FACTOR MIXTURE MODEL



Classes can be indicated by item thresholds (categorical)\ item means (continuous) or factor mean and variance



# IRT + Clases latentes

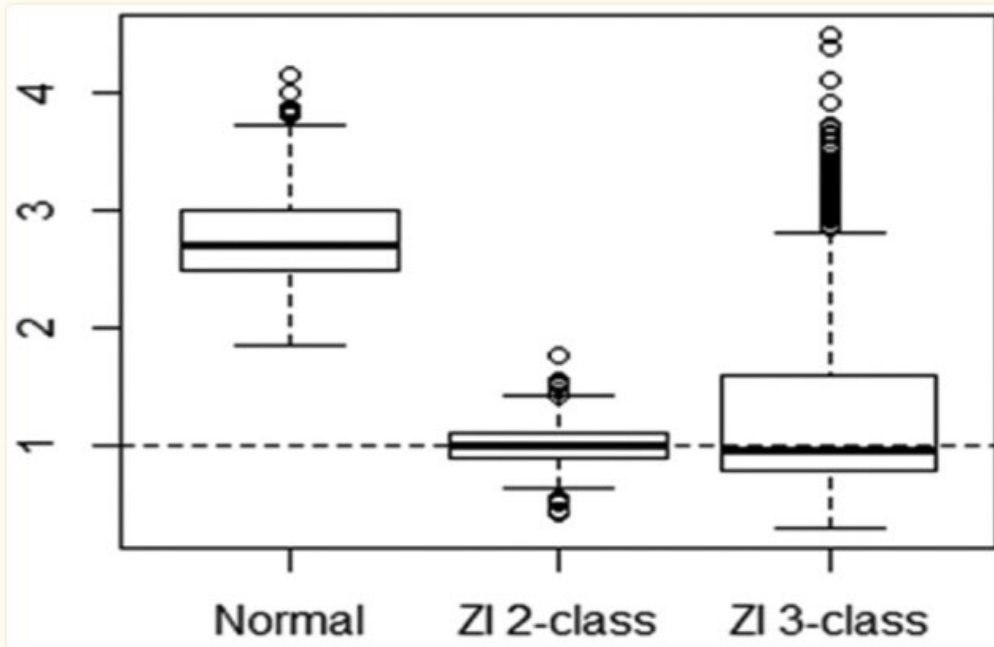


Figure 2.

Truth is ZI  $K = 2$  mixture with 75% non-pathological.

Note. Side-by-side boxplots of 100 estimates of all 10 of the discrimination parameter  $\beta_1 - \beta_{10}$  (true value = 1) for each of the three models. ZI = zero-inflated.

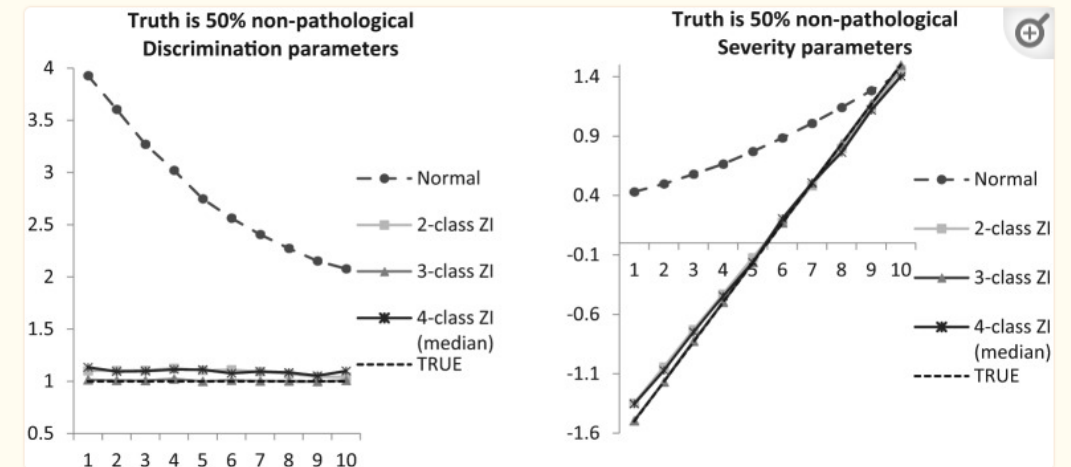


Figure 4.

Truth is ZI  $K = 3$  mixture with 50% of the population non-pathological.

Note. Empirical mean (and median for the four-class ZI mixture) item parameter estimates (discrimination on left, severity on right) across 100 simulated data sets estimated with the four different models. ZI = zero-inflated.



# Próxima clase

- Sesión práctica de TRI

