Informe técnico metodológico de DocenteAcreditado 2025

docente acreditado
Prueba de Reconocimiento Universitario de la Formación Docente

INEEd
Instituto Nacional de Evaluación Educativa

El autor de este documento es Mario Luzardo

Montevideo, 2025
ISBN: 978-9915-9820-3-8

Edificio Los Naranjos, planta alta, Parque Tecnológico del LATU
Av. Italia 6201, Montevideo, Uruguay
(+598) 2604 4649 – 2604 8590
ineed@ineed.edu.uy
www.ineed.edu.uy

Cómo citar: INEEd (2025). *Informe técnico metodológico de DocenteAcreditado 2025*. Recuperado de https://www.ineed.edu.uy/images/publicaciones/docenteacreditado/informe-tecnico-metodologico-docenteacreditado-2025.pdf

Índice

Introducción

Este informe presenta los métodos utilizado durante el análisis psicométrico de las diferentes pruebas de DocenteAcreditado y los métodos de equiparación para establecer los puntos de corte.

Identificación y eliminación de ítems de mala performance

La primera fase del estudio psicométrico se dedicó a identificar ítems que no tuvieron un comportamiento adecuado en las pruebas y a definir si se eliminaban o mantenían en las pruebas.

Identificación mediante medidas asociadas a la dificultad y a la correlación

Primeramente se identificaron ítems con niveles de dificultad extremos (muy fáciles o muy difíciles). Se calculó la dificultad por TCT. Si bien en general se considera que los valores aceptables de dificultad del ítem están entre 0,2 y 0,8, como en 2023 se mantuvieron ítems con dificultades fuera de ese rango, se fue más flexible y se consideraron aceptables ítems con dificultad TCT 0,1 y 0,9.

Tabla 1
Valores de referencia para la dificultad de un ítem

Valor de dificultad	Interpretación
0,91 – 1,00	Muy fácil
0,61 – 0,90	Fácil
0,41 – 0,60	Moderada
0,11 – 0,40	Difícil
0,00 – 0,10	Muy difícil

También se determinaron los ítems que contribuyen eficazmente a la diferenciación de los niveles de habilidad. La discriminación de un ítem mide su capacidad para diferenciar entre participantes con niveles altos y bajos de habilidad en la prueba. En esta etapa se utilizaron los siguientes índices: índice D (con los percentiles 30 y 70), correlación biserial puntual (rbp), correlación ítem-test (r) (con el ítem excluido e incluido en el total del test) e índice Delta de Ebel (Δ). Las tablas 2, 3 y 4 muestran los valores aceptables de estos índices.

Tabla 2
Valores de referencia para el coeficiente de discriminación D

Valor de D	Interpretación
≥0,40	Muy buena discriminación
0,30 – 0,39	Buena discriminación
0,20 – 0,29	Aceptable (puede mejorarse)
0,10 – 0,19	Baja discriminación
<0,10	Muy baja o nula — revisar o eliminar el ítem
<0	Discriminación negativa — posible error de clave

Tabla 3

Valores de referencia para la correlación ítem test

Valor de r ítem test	Interpretación
≥0,40	Muy buena discriminación
0,30 – 0,39	Buena discriminación
0,20 – 0,29	Aceptable – mejora posible
<0,20	Poca discriminación – revisar el ítem
Negativo	Eliminar el ítem

Tabla 4

Valores de referencia para el Delta de Ebel

Valor de índice delta	Interpretación
0 – 20	Ítem muy difícil – baja discriminación
21 – 40	Ítem difícil – puede discriminar
41 – 60	Dificultad óptima – alta discriminación
61 – 80	Ítem fácil – puede discriminar
81 – 100	Ítem muy fácil – baja discriminación

Identificación mediante medidas asociadas a la fiabilidad

Se calcularon diversos índices de fiabilidad y en particular se detectaron ítems que afecten seriamente a la fiabilidad de la prueba. En esta instancia se calcularon: Alfa de Cronbach, Omega y Omega h de McDonald, Razón Señal/Ruido (S/R), WH Index (Wiener-Hayes) e Índices Lambda (λ).

Se analizó, además, el efecto que tendría la exclusión de cada ítem en la fiabilidad global del instrumento, con el objetivo de identificar aquellos que podrían estar debilitando la consistencia interna. Especialmente se analizó el impacto en el Alfa de Cronbach y del Omega de Mc Donald. Se estudió cómo cambian esos índices al eliminar un ítem específico. Un incremento significativo al excluir un ítem indicaría que este no contribuye adecuadamente a la homogeneidad del instrumento o está altamente influenciada por factores externos o aleatorios, los cuales podrían reducir la precisión general de la prueba. También se analizó la varianza explicada por el ítem en el puntaje total. Este análisis ayudó a optimizar el contenido del instrumento, eliminando ítems que no aportan valor o que distorsionan las mediciones (tablas 5 y 6).

Tabla 5

Valores de referencia para el coeficiente del Alfa de Cronbach

Alfa de Cronbach	Interpretación
≥0,90	Excelente (ideal en evaluación clínica o diagnóstica)
0,80 – 0,89	Buena (aceptable para decisiones individuales)
0,70 – 0,79	Aceptable (adecuada para investigación exploratoria)
0,60 – 0,69	Cuestionable (puede necesitar revisión)
0,50 – 0,59	Pobre (baja confiabilidad)
<0,50	Inaceptable

Tabla 6

Valores de referencia para omega total (ω), omega jerárquico (ωh) y WH Index

Índice	Valor	Interpretación
Omega total (ω)	≥0,90	Excelente fiabilidad total
Omega total (ω)	0,80 – 0,89	Buena fiabilidad total
Omega total (ω)	0,70 – 0,79	Aceptable en contextos exploratorios
Omega jerárquico (ωh)	≥0,80	Alta varianza atribuible al factor general (uso de puntaje total justificado)
Omega jerárquico (ωh)	0,65 – 0,79	Moderada influencia del factor general
Omega jerárquico (ωh)	<0,65	Influencia débil del factor general (considerar subescalas)
WH Index	≥0,90	Excelente estabilidad o equivalencia
WH Index	0,80 – 0,89	Buena estabilidad o equivalencia
WH Index	<0,80	Estabilidad o equivalencia insuficiente

Estudio de distractores y respuesta correcta

Se realizó también un análisis para estudiar el comportamiento de la respuesta correcta e identificar los distractores ineficaces.

Para estudiar la funcionalidad de la escala se contemplaron los siguientes aspectos:

a) Evaluación del Comportamiento de los Distractores. Se observó qué tan atractivos son los distractores para los participantes con bajos niveles de habilidad, asegurando que estos seleccionen los distractores en lugar de la respuesta correcta. Se consideran distractores efectivos aquellos que son seleccionados por participantes que no dominan el contenido evaluado, contribuyendo a la discriminación del ítem.
b) Distractores ineficaces. Opciones que rara vez o nunca son seleccionadas, lo que puede indicar un diseño pobre.
c) Distractores atractivos para niveles altos. Identificación de distractores que atraen a participantes con altos niveles de habilidad, lo que podría indicar ambigüedad en el ítem.

Para identificar distractores ineficaces, se realizó un análisis de frecuencias para identificar distractores seleccionados por pocos participantes (menos del 5% y entre 5% y 10%). También se estudió la distribución de selección por Grupo de Habilidad, comparando la frecuencia de selección de cada distractor entre grupos de habilidad (participantes en el tercio superior frente al tercio inferior).

Modelización mediante modelo de Rasch

Para las pruebas en las que la cantidad de personas que las tomaron es suficiente, se modelizaron los ítems mediante el modelo de Rasch. En un principio para detectar los ítems que presenten problemas de escala y luego para analizar el ajuste global del modelo a la escala. Se eligió este modelo para estar en concordancia con el estudio de 2023. Por lo tanto, se pudo aplicar el modelo de Rasch a las escalas de Habilidades tecnológicas y digitales, Humanidades, Matemática y estadística básicas, Habilidades para la educación inclusiva, Comprensión lectora y Pensamiento científico, pues son las que tienen tamaños muestrales para realizar este análisis.

Se analizaron la dificultad de los ítems (del modelo de Rasch) y tres estadísticos: el M2, el Root Mean Square Error of Approximation (RMSEA) y el Standardized Root Mean Square Residual (SRMSR). El estadístico M2 es una prueba de bondad de ajuste basada en la comparación entre la matriz observada y la esperada de covarianzas/residuos y bajo la hipótesis nula de buen ajuste tiene una distribución chi2. Cuando la muestra es grande o el modelo tiene muchos ítems, el M2 puede volverse demasiado sensible, detectando como significativas pequeñas discrepancias sin relevancia práctica. Esto lleva a $p<0,05$ aunque el modelo sea aceptablemente bueno en términos prácticos.

El RMSEA estima el error medio cuadrático de aproximación del modelo a la población y toma en cuenta los grados de libertad, penalizando los modelos muy complejos. A diferencia del M2, el RMSEA es menos sensible al tamaño de la muestra, por lo que podemos usarlo para evaluar el ajuste práctico del modelo, no solo su ajuste exacto. Es decir, incluso cuando el M2 rechaza el modelo si la muestra es grande y el RMSEA <0,05 puede no rechazarse el modelo. Si el RMSEA es <0,05 indica ajuste excelente; si es 0,05–0,08, ajuste razonable, y si es >0,10, indica un mal ajuste.

El SRMSR mide los residuos promedio entre lo observado y lo esperado. Es útil para identificar errores localizados: si hay muchos residuos pequeños, el SRMSR será bajo. Si el ajuste general es bueno pero hay algunos ítems con mal ajuste, el SRMSR puede ser algo alto. Los valores indicativos de ajuste son los mismos que los del RMSEA. También se calcularon las curvas características de los ítems (CCI) y la función de información.

Análisis específico para la prueba de Producción escrita y organización textual

En la prueba de Producción escrita y organización textual adicionalmente se estudiaron las correlaciones de los puntajes totales entre los jueces y las correlaciones entre los jueces por criterio.

Como la prueba 2023 estaba analizada por la puntuación inicial de dos jueces, se trató igualmente la prueba 2025. Por lo tanto, se estudió la fiabilidad interjueces mediante coeficientes de correlación intraclase. Este análisis implicó la determinación de varios coeficientes que detallamos a continuación.

Coeficiente de Correlación Intraclase (ICC)

El Coeficiente de Correlación Intraclase (ICC) es un índice de fiabilidad que estima la proporción de la varianza total atribuible a las diferencias verdaderas entre los sujetos evaluados. A diferencia del coeficiente de correlación de Pearson, que mide la asociación lineal entre dos variables distintas, el ICC cuantifica el grado en que las mediciones realizadas sobre los mismos individuos por diferentes jueces o en distintas condiciones son consistentes o intercambiables.
Matemáticamente, el ICC puede expresarse como:

$$ICC = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_R^2 + \sigma_E^2},$$

donde: - σ_S^2 es la varianza entre sujetos (variance between subjects), - σ_R^2 es la varianza entre jueces (variance due to raters) y - σ_E^2 es la varianza residual o error.

Dependiendo del modelo de efectos (fijos o aleatorios) y del diseño de medición, se distinguen varias formas del ICC, cada una con una interpretación específica.

ICC(1) o ICC(1,1): Single Raters Absolute

El coeficiente ICC(1) se basa en un modelo de una vía con efectos aleatorios, en el cual tanto los sujetos como los jueces son considerados muestras aleatorias de una población mayor. Su fórmula es:

$$ICC(1) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E},$$

donde: - MS_S = cuadrado medio entre sujetos (mean square between subjects), - MS_E = cuadrado medio del error y - k = número de jueces.

Evalúa el acuerdo absoluto entre las calificaciones de jueces individuales. Valores altos de ICC(1) indican que la mayor parte de la varianza proviene de diferencias verdaderas entre los sujetos, mientras que valores bajos sugieren desacuerdo entre jueces.

ICC(2) o ICC(2,1): Single Random Raters

El coeficiente ICC(2) se obtiene de un modelo de dos vías con efectos aleatorios, donde tanto los sujetos como los jueces son aleatorios. Su expresión es:

$$ICC(2) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E + \frac{k}{n}(MS_R - MS_E)},$$

donde MS_R es el cuadrado medio entre jueces (mean square between raters) y n es el número de sujetos.

Este coeficiente mide la consistencia relativa entre jueces, es decir, si los jueces mantienen un mismo ordenamiento o jerarquía en las puntuaciones, aunque difieran en severidad media. Se utiliza cuando los jueces se consideran una muestra aleatoria de una población mayor.

ICC(3) o ICC(3,1): Single Fixed Raters

El ICC(3) corresponde a un modelo de dos vías con efectos fijos para los jueces. Se utiliza cuando los jueces evaluados son los únicos de interés y no se pretende generalizar a otros posibles evaluadores. Su fórmula es:

$$ICC(3) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E}.$$

Aunque formalmente idéntica a la de ICC(1), su interpretación es distinta: mide la consistencia interna de las evaluaciones entre los jueces observados, eliminando las diferencias sistemáticas de nivel medio (por ejemplo, severidad general de un juez).

Versiones promedio: ICC(1k), ICC(2k) y ICC(3k)

Cuando se desea estimar la fiabilidad de la media de las calificaciones de varios jueces en lugar de una sola, se utilizan las versiones promedio o agregadas:

$$ICC(1k) = \frac{MS_S - MS_E}{MS_S},$$
$$ICC(2k) = \frac{MS_S - MS_E}{MS_S + \frac{MS_R - MS_E}{n}},$$
$$ICC(3k) = \frac{MS_S - MS_E}{MS_S}.$$

Estas fórmulas derivan directamente de las versiones individuales, pero ajustadas al promedio de k observaciones, lo que reduce el error aleatorio y eleva la fiabilidad. En

la práctica, los valores k son superiores a los individuales porque el promedio atenúa la variabilidad entre jueces.

Interpretación de los valores del ICC

Los coeficientes ICC(1) e ICC(2) estiman la fiabilidad de una puntuación individual, mientras que ICC(1k) e ICC(2k) cuantifican la fiabilidad de la media de varios jueces. Por su parte, las variantes ICC(3) e ICC(3k) reflejan la consistencia interna cuando los jueces son considerados fijos.

Los rangos interpretativos propuestos por Landis y Koch (1977) son presentados en la tabla 7.

En contextos de evaluación del desempeño, rúbricas o exámenes de escritura, valores superiores a 0,80 en ICC(2k) o ICC(3k) suelen considerarse adecuados para decisiones de alta consecuencia, como certificaciones o evaluaciones sumativas de gran impacto.

Tabla 7

Interpretación orientativa de ICC

Rango	Interpretación
<0,40	Baja
0;40–0,59	Moderada
0,60–0,74	Buena
≥0,75	Excelente

Fiabilidad interjueces con puntuación total

Encontramos que los coeficientes ICC(2,k) e ICC(3,k) son mucho mayores a 0,8, por lo tanto, la fiabilidad media es adecuada en el contexto de la prueba de Producción escrita y organización textual. Las fiabilidades individuales son moderadas.

Tabla 8

Fiabilidad interjueces sobre Score total (ICC)

Medida	Valor	LI 95%	LS 95%
ICC(2,1) acuerdo (two-way random)	0,5159918	0,4877542	0,5436126
ICC(3,1) consistencia (two-way fixed)	0,5453846	0,5276829	0,5634249
ICC(2,k) acuerdo promedio	0,9372062	0,9302194	0,9434251
ICC(3,k) consistencia promedio	0,9438053	0,9399078	0,9475555

ICC(2,1): jueces aleatorios (two-way random, single); ICC(3,1): jueces fijos (two-way fixed, single).
ICC(2,k)/(3,k): fiabilidad del promedio de k jueces; IC 95% entre corchetes.

ANOVA de dos vías con juez y criterio como factores

También se realizó un análisis de varianza de dos vías aplicado a las puntuaciones otorgadas en la prueba de Producción escrita y organización textual, considerando como factores a los jueces y a las preguntas. El modelo permite identificar la magnitud y la significación de las diferencias asociadas a la severidad de los evaluadores, la dificultad relativa de los ítems y la posible interacción entre ambas fuentes de variación.

Resultados

A partir de los resultados se encontraron 65 ítems con problema y finalmente se decidió eliminar 31.

Tabla 9

Resumen de ítems con anomalías detectadas por escala

Escala	Ítem	Justificación	Decisión
Pensamiento científico	DACIE20	Dificultad muy baja (-3,102)	SE ELIMINA
Humanidades	DAHUM25	Baja correlación ítem-test	SE ELIMINA
Habilidades para la educación inclusiva	DAINC23	Muy fácil (p = 0,98)	SE ELIMINA
Habilidades para la educación inclusiva	I_1991248	Muy fácil (p = 0,97)	SE ELIMINA
Habilidades para la educación inclusiva	I_1991292	Muy fácil (p = 0,98)	SE ELIMINA
Comprensión lectora	DALEC3	Baja correlación ítem-test	SE ELIMINA
Comprensión lectora	I_1693234	Muy fácil (p = 0,97)	SE ELIMINA
Comprensión lectora	I_1837819	Dificultad muy baja (-4,117)	SE ELIMINA
Español	DAESP17	Correlación negativa	SE ELIMINA
Español	DAESP18	Correlación negativa	SE ELIMINA
Español	DAESP24	Correlación negativa	SE ELIMINA
Español	DAESP26	Correlación negativa	SE ELIMINA
Español	DAESP28	Baja correlación y mejora fiabilidad si se elimina	SE ELIMINA
Español	I_1994499	Correlación negativa	SE ELIMINA
Español	I_1994590	Correlación negativa	SE ELIMINA
Inglés	DAING23	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Inglés	DAING6	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Inglés	DAING9	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Inglés	I_1989649	Correlación negativa	SE ELIMINA
Portugués	DAPOR10	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Portugués	DAPOR11	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA

Escala	Ítem	Justificación	Decisión
Portugués	DAPOR3	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Portugués	I_2000194	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Portugués	I_2000404	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Artísticas	I_1990594	No hay correlación	SE ELIMINA
Artísticas	I_1990623	Muy fácil (p = 0,98)	SE ELIMINA
Italiano	I_1995859	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Italiano	I_1999270	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Italiano	I_1999488	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Italiano	I_1999508	Correlación negativa y mejora fiabilidad si se elimina	SE ELIMINA
Italiano	I_2001178	Correlación negativa	SE ELIMINA

Estudio psicométrico de las escalas

Mediante TCT

Una vez eliminados los ítems de mala performance, se realizó un estudio psicométrico de las escalas. Se repitieron los análisis descriptos en la sección anterior a modo de verificación y se agregaron nuevos estudios de mayor profundidad. Se sumó la correlación biserial puntual, la tabla 10 muestra los valores de referencia.

Tabla 10

Valores de referencia para el coeficiente de correlación biserial puntual

Valor de correlación biserial puntual	Interpretación
≥0,40	Muy buena discriminación. El ítem diferencia claramente entre sujetos de bajo y alto desempeño.
0,30 – 0,39	Buena discriminación. El ítem contribuye de manera adecuada a la medición.
0,20 – 0,29	Discriminación aceptable, aunque moderada. El ítem aporta, pero podría mejorarse.
0,10 – 0,19	Discriminación baja. El ítem ofrece información limitada y debe revisarse.
<0,10	Discriminación deficiente. El ítem prácticamente no diferencia entre niveles de habilidad.
<0	Discriminación inversa (problemática). Los sujetos de alto desempeño tienden a fallar más que los de bajo desempeño; el ítem es inconsistente con el constructo medido.

Unidimensionalidad

La evaluación de la dimensionalidad constituye un paso central en el análisis psicométrico, ya que permite establecer si los ítems de una prueba se organizan en torno a un único constructo (unidimensionalidad) o si, por el contrario, reflejan múltiples factores. Este aspecto es crítico para justificar la interpretación de las puntuaciones y el cálculo de índices de fiabilidad y validez.

Dado que la estabilidad de las estimaciones depende fundamentalmente del tamaño muestral, se adoptaron diferentes estrategias metodológicas según la cantidad de casos disponibles en cada prueba (todas cuentan con un número suficiente de ítems). La aproximación se organiza en tres escenarios:

1. Con muestras muy reducidas (N<50) los procedimientos factoriales y de TRI no son confiables. En este caso, la evidencia empírica debe considerarse exploratoria y complementarse con el marco teórico de la prueba. Se utilizan:
   - **Alfa de Cronbach** y análisis de impacto de eliminar ítems.
   - **Correlaciones ítem-total corregidas**.

- **Valores propios** del análisis de componentes principales (ACP).
- **Razón λ1/λ2** (primer valor propio dividido por el segundo).

2. Con tamaños moderados (50 < N < 200) es posible realizar análisis exploratorios, aunque con cautela respecto a la generalización. Se incorporan:
- **Alfa de Cronbach** y **omega total** como índices de consistencia interna.
- **Correlaciones tetracóricas** entre ítems dicotómicos.
- **Valores propios** y **razón λ1/λ2**.
- Interpretación cualitativa a partir de los contenidos de la prueba.

3. Con un número suficiente de participantes (N>200) se dispone de evidencia empírica robusta y es posible aplicar procedimientos factoriales y de TRI para la detección de dimensionalidad. Se consideran los siguientes indicadores:
- **Análisis paralelo**

El análisis paralelo es un procedimiento estadístico diseñado para determinar el número óptimo de factores o componentes a retener en un análisis factorial exploratorio (AFE) o en un análisis de componentes principales (ACP). Fue propuesto originalmente por Horn (1965) como una alternativa más rigurosa a la tradicional regla de Kaiser (autovalores mayores a 1).

La idea central del método es comparar los autovalores obtenidos de la matriz de correlaciones de los datos observados con los autovalores obtenidos a partir de datos aleatorios simulados de igual tamaño (mismo número de sujetos y de variables). Si el autovalor de un factor en los datos reales es mayor que el correspondiente autovalor promedio de los datos aleatorios, se interpreta que ese factor explica más varianza que lo esperado por azar y, por tanto, debe retenerse.

Sea $\mathbf{R}_{\text{obs}}$ la matriz de correlaciones de los datos reales, de dimensión $p \times p$ con N sujetos.
Calculamos sus autovalores:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p.$$

Se generan M réplicas de matrices aleatorias de tamaño $N \times p$, cuyos elementos son simulados típicamente a partir de una distribución normal estándar.

De cada réplica m, se obtiene la matriz de correlaciones $\mathbf{R}_{\text{sim}}^{(m)}$ y sus autovalores $\lambda_1^{(m)}, \ldots, \lambda_p^{(m)}$.
Para cada posición j se calcula el promedio de los autovalores simulados:

$$\bar{\lambda}_j = \frac{1}{M} \sum_{m=1}^{M} \lambda_j^{(m)}.$$

Se compara λ_j (autovalor real) con $\bar{\lambda}_j$ (autovalor simulado).

El número de factores a retener es:

$$q = \max\{j : \lambda_j > \bar{\lambda}_j\}.$$

Si un autovalor real es mayor que el correspondiente autovalor simulado, el factor asociado se considera no atribuible al azar y se retiene. Si el autovalor real es menor o igual, se interpreta que no hay evidencia suficiente para retener ese factor.

El análisis paralelo suele dar soluciones más parsimoniosas y realistas que la regla de Kaiser, evitando la retención de factores espurios.

El método requiere definir el número de simulaciones M (usualmente entre 100 y 1.000, según la precisión deseada) y puede aplicarse tanto en matrices de correlaciones de Pearson como en matrices de correlaciones tetracóricas (para ítems dicotómicos).

- **Test MAP de Velicer** (*VSS::map*) → número de factores óptimo.

El test Minimum Average Partial (MAP), propuesto por Velicer (1976), es un procedimiento estadístico para determinar el número adecuado de factores a retener en un análisis factorial exploratorio. A diferencia de métodos basados en la magnitud de los autovalores (como la regla de Kaiser o el análisis paralelo), el MAP se centra en evaluar la cantidad de varianza común y residual explicada por los factores extraídos.

El MAP test busca identificar el número de factores que minimiza la varianza residual entre las variables, es decir, el punto en el que la inclusión de factores adicionales no aporta mejora sustantiva en la explicación de las correlaciones.

A partir de la matriz de correlaciones observada $\mathbf{R}$ de p variables, se van extrayendo factores secuencialmente (mediante componentes principales o un método factorial).

Para cada número de factores m, se calcula la matriz de correlaciones parciales residuales después de extraer esos m factores.

Se obtiene el promedio de los cuadrados de las correlaciones parciales:

$$\text{MAP}(m) = \frac{1}{p(p-1)} \sum_{i \neq j} \left(r_{ij}^{(m)}\right)^2,$$

donde $r_{ij}^{(m)}$ es la correlación parcial residual entre las variables i y j tras extraer m factores.

El número óptimo de factores se determina como aquel que minimiza este índice:

$$q = \arg\min_m \text{MAP}(m).$$

El valor de m que produce el mínimo promedio cuadrático de las correlaciones parciales indica el número de factores a retener. Valores decrecientes de MAP reflejan que la extracción de factores reduce las correlaciones residuales; cuando el valor comienza a aumentar, significa que se están introduciendo factores espurios que capturan solo ruido.

Es un método complementario al análisis paralelo: mientras este último compara con datos aleatorios, el MAP evalúa la estructura residual de los datos reales.

Valores propios y ratio λ_1/λ_2

En el análisis factorial exploratorio (AFE), los valores propios (λ) de la matriz de correlaciones de los ítems constituyen un insumo fundamental para evaluar la dimensionalidad de una prueba.

Si $\mathbf{R}$ es la matriz de correlaciones ($p \times p$) entre ítems, los valores propios se obtienen resolviendo:

$$\det(\mathbf{R} - \lambda\mathbf{I}) = 0,$$

donde $\mathbf{I}$ es la matriz identidad. El conjunto de soluciones $\{\lambda_1, \lambda_2, \dots, \lambda_p\}$ corresponde a los valores propios, ordenados de mayor a menor:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

λ_1 indica la proporción de varianza común explicada por el primer factor y λ_2 y los siguientes reflejan la magnitud de dimensiones adicionales.

Un criterio empírico ampliamente utilizado es el ratio entre el primer y segundo valor propio:

$$\text{Ratio} = \frac{\lambda_1}{\lambda_2}.$$

Este cociente permite evaluar la dominancia de un factor general frente a otros posibles factores.

Ratios elevados (≥ 3) sugieren una estructura fuertemente unidimensional. Ratios intermedios (≈ 2) pueden considerarse evidencia ambigua y ratios bajos (< 2) indican que existen múltiples dimensiones relevantes en los datos.

Mediante TRI

En el contexto de la Teoría de Respuesta al Ítem (TRI), la decisión entre un modelo unidimensional (1 factor) y un modelo multidimensional (2 factores) puede realizarse comparando la bondad de ajuste y la parsimonia de los modelos estimados. Para ello se utilizan criterios de información y medidas de ajuste global.

- **Criterios de información: AIC y BIC**

El Criterio de Información de Akaike (AIC) y el Criterio Bayesiano de Información(BIC) se definen como:

$$\text{AIC} = -2\ell + 2k,$$
$$\text{BIC} = -2\ell + k\ln(N),$$

donde:
- ℓ es el logaritmo de la verosimilitud maximizada del modelo,
- k es el número de parámetros estimados,
- N es el tamaño muestral.

Ambos penalizan la complejidad del modelo, aunque el BIC impone una penalización mayor para muestras grandes. Valores más bajos de AIC y BIC indican un mejor equilibrio entre ajuste y parsimonia.

- **Diferencia de BIC**

Una medida útil en la comparación de modelos es:

$$\Delta\text{BIC} = \text{BIC}_{2f} - \text{BIC}_{1f},$$

donde BIC_{2f} corresponde al modelo de dos factores y BIC_{1f} al de un factor.

Si $\Delta\text{BIC} < -10$, existe evidencia fuerte a favor del modelo de 2 factores. Si $-10 \leq \Delta\text{BIC} \leq -6$, la evidencia es moderada. Si $\Delta\text{BIC} > 0$, el modelo unidimensional resulta más parsimonioso.

- **Estadístico M2**

El estadístico M2 (Maydeu-Olivares & Joe, 2006) evalúa el ajuste global de los modelos de TRI a partir de la matriz de covarianzas y correlaciones de los ítems. Se interpreta de manera similar a una prueba de bondad de ajuste chi-cuadrado: valores elevados con $p < ,05$ sugieren mal ajuste.

Sea s el vector de momentos observados (proporciones de respuesta y correlaciones ítem-ítem), $\hat{\mu}$ el vector de momentos esperados por el modelo estimado y $\widehat{W}$ una matriz de ponderación consistente (típicamente la matriz de covarianzas asintóticas de s). Entonces:

$$M_2 \;=\; (s - \hat{\mu})^{\top}\, \widehat{W}^{-1}\, (s - \hat{\mu}).$$

Bajo el modelo correcto y para N grande:

$$M_2 \;\overset{a}{\sim}\; \chi^2_{\text{df}},$$

donde df es el número de momentos utilizados menos el número de parámetros libres del modelo (ajustado por restricciones).

- **RMSEA**

El RMSEA se utiliza como índice de ajuste aproximado:

$$\text{RMSEA} = \sqrt{\frac{\max(\chi^2 - df,\, 0)}{df \cdot (N-1)}},$$

donde χ^2 es el estadístico de ajuste del modelo y df los grados de libertad.

- RMSEA ≤ ,05 indica **buen ajuste**,
- entre ,05 y ,08 ajuste **aceptable**,
- entre ,08 y ,10 ajuste **mediocre**,
- ,10 ajuste **deficiente**.

La comparación de modelos Rasch (1PL) y 2PL puede realizarse con los mismos criterios, teniendo en cuenta que el 2PL introduce un parámetro de discriminación por ítem (a_i), lo que suele mejorar el ajuste pero a costa de mayor complejidad. El uso combinado de AIC, BIC, ΔBIC, M2 y RMSEA permite evaluar tanto la parsimonia como el ajuste global, brindando una base sólida para decidir entre modelos unidimensionales y multidimensionales.

Método DETECT

También se utilizó el método denominado Dimensionality Evaluation To Enumerate Contributing Traits (DETECT), que constituye un enfoque diseñado para evaluar la dimensionalidad de un conjunto de ítems bajo el marco de la TRI. Se basa en el análisis de la dependencia local condicional entre ítems, es decir, la correlación residual que permanece tras controlar por un rasgo latente común.

Los estadísticos principales de esta técnica son:

1. **DETECT (D):**

   Es el índice central del procedimiento. Mide el grado de dependencia local condicional promedio entre ítems agrupados en distintas dimensiones.
   - Valores bajos ($< 0{,}20$) sugieren unidimensionalidad.
   - Valores intermedios ($0{,}20 \leq D \leq 0{,}40$) sugieren una estructura ambigua o débilmente multidimensional.
   - Valores altos ($> 0{,}40$) sugieren multidimensionalidad.

   Formalmente, si Q denota la partición de ítems en clusters y $\hat{r}_{ij}$ la correlación condicional entre ítems i y j, entonces:

$$D(Q) = \frac{1}{|P|} \sum_{(i,j)\in P} \hat{r}_{ij},$$

   donde P es el conjunto de pares de ítems asignados a diferentes clusters.

El índice DETECT se define como:

$$\text{DETECT} = \max_{Q} D(Q),$$

es decir, el valor máximo de $D(Q)$ a través de todas las posibles particiones Q.

2. **ASSI (Approximate Simple Structure Index):**

   Evalúa el grado en que la estructura encontrada se aproxima a una estructura simple (cada ítem carga fuertemente en una sola dimensión).
   - Rango: $0 \leq \text{ASSI} \leq 1$.
   - Valores cercanos a 1 indican una estructura bien definida.
   - Valores cercanos a 0 indican ausencia de estructura simple.

3. **RATIO:**

   Es un ajuste de DETECT que incorpora la varianza explicada por la estructura dimensional encontrada. Proporciona una medida relativa más robusta, especialmente cuando el número de ítems o la varianza común es limitada. Aunque no posee puntos de corte universalmente estandarizados, un RATIO elevado combinado con un DETECT alto refuerza la conclusión de multidimensionalidad.

DETECT no requiere especificar a priori el número de dimensiones; en cambio, busca la partición que maximiza la dependencia local condicional entre *clusters*. Siempre se recomienda complementar DETECT con otras evidencias de dimensionalidad (análisis paralelo, MAP de Velicer, ratios de valores propios, ajuste de modelos confirmatorios).

La decisión final se establece integrando todos estos indicadores. Como criterio operativo, se considera evidencia suficiente de unidimensionalidad cuando:
- Parallel analysis y MAP sugieren un solo factor.
- La razón $\lambda 1/\lambda 2$ es elevada (≥ 3).
- El omega jerárquico es $\geq 0{,}50$.
- El modelo unidimensional no es claramente superado por el modelo de 2 factores (ΔBIC > −10 y RMSEA aceptable).
- El índice DETECT es bajo (<0,20) y no se observa estructura simple en ASSI o RATIO.

Tabla 11

Resumen de indicadores por escenario muestral

Escenario	N	Indicadores aplicables	Alcance de la evidencia
1. Muestras pequeñas	N< 50	Alfa de Cronbach, correlaciones ítem-total, eigenvalores, ratio λ1/λ2	Evidencia exploratoria; se complementa con el marco teórico
2. Muestras intermedias	50≤ N< 200	Alfa de Cronbach, omega total, correlaciones tetracóricas, eigenvalores, ratio λ1/λ2	Evidencia preliminar; análisis descriptivo y de consistencia
3. Muestras adecuadas	N≥ 200	Parallel analysis, MAP, eigenvalores, ratio λ1/λ2, omega jerárquico (ωh), comparación de modelos mirt (AIC, BIC, ΔBIC, M2, RMSEA), método DETECT (D, ASSI, RATIO)	Evidencia robusta para decidir unidimensionalidad o multidimensionalidad

Esta metodología se adapta a la disponibilidad de casos en cada una de las pruebas. Con muestras reducidas, se enfatizan los índices de consistencia interna y la fundamentación teórica; con muestras intermedias, se aplican análisis exploratorios, y con muestras adecuadas, se implementa la batería completa de procedimientos, incluyendo el método DETECT, para obtener una evaluación integral de la dimensionalidad.

Funcionamiento diferencial de los ítems

Una aproximación clásica al estudio del Funcionamiento Diferencial de Ítems (DIF) consiste en comparar las proporciones de aciertos entre dos grupos previamente definidos (grupo de referencia y grupo focal).

Sea el ítem i y dos grupos con tamaños n_{ref} y n_{foc}. Definimos:

$$p_{\text{ref},i} = \frac{1}{n_{\text{ref}}} \sum_{j=1}^{n_{\text{ref}}} X_{ij}^{(\text{ref})}, \quad p_{\text{foc},i} = \frac{1}{n_{\text{foc}}} \sum_{j=1}^{n_{\text{foc}}} X_{ij}^{(\text{foc})},$$

donde $X_{ij} = 1$ si el sujeto j respondió correctamente el ítem i, y $X_{ij} = 0$ en caso contrario.

El índice de diferencia de proporciones se define como:

$$\Delta p_i = p_{\text{foc},i} - p_{\text{ref},i}.$$

- Valores de $\Delta p_i > 0$ indican que el ítem resulta relativamente más fácil para el grupo focal.
- Valores de $\Delta p_i < 0$ indican que el ítem resulta relativamente más fácil para el grupo de referencia.
- Valores cercanos a 0 sugieren ausencia de DIF.

Para evaluar la significancia estadística de esta diferencia, se utiliza una prueba de contraste basada en el estadístico z, definido como:

$$z_i = \frac{p_{\text{foc},i} - p_{\text{ref},i}}{\sqrt{\frac{p_{\text{ref},i}(1 - p_{\text{ref},i})}{n_{\text{ref}}} + \frac{p_{\text{foc},i}(1 - p_{\text{foc},i})}{n_{\text{foc}}}}}.$$

Bajo la hipótesis nula de igualdad de proporciones ($H_0: p_{\text{ref},i} = p_{\text{foc},i}$), el estadístico z_i se distribuye aproximadamente como una normal estándar, lo que permite obtener un valor-p bilateral:

$$p\text{-valor}_i = 2 \cdot \Phi(-|z_i|),$$

donde $\Phi(\cdot)$ es la función de distribución acumulada de la normal estándar.

Este procedimiento provee tanto una medida del tamaño del efecto (Δp_i) como una evaluación de su significancia estadística (z_i), constituyendo un enfoque descriptivo inicial para la detección de posibles ítems con DIF

Otro procedimiento es el de Mantel–Haenszel (MH), que constituye uno de los enfoques clásicos y más difundidos para la detección del Funcionamiento Diferencial de Ítems (DIF), especialmente en contextos de ítems dicotómicos bajo la Teoría Clásica de los Tests (TCT).

Construcción de las tablas 2×2 por estrato

Para cada ítem i, los sujetos se dividen en grupos de referencia y focal, y se estratifican según su puntaje total en la prueba (o en un conjunto de ítems de anclaje). En cada estrato $s = 1, \ldots, S$ se construye una tabla de contingencia 2×2:

	Correcto	**Incorrecto**
Referencia	A_s	B_s
Focal	C_s	D_s

donde $A_s + B_s + C_s + D_s = n_s$ es el tamaño del estrato.

El estadístico de Mantel–Haenszel estima una razón de odds común a través de los estratos:

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S} \frac{A_s D_s}{n_s}}{\sum_{s=1}^{S} \frac{B_s C_s}{n_s}}.$$

- $\hat{\alpha}_{MH} = 1$ sugiere ausencia de DIF.
- $\hat{\alpha}_{MH} > 1$ indica ventaja del grupo focal.
- $\hat{\alpha}_{MH} < 1$ indica ventaja del grupo de referencia.

El contraste de hipótesis se formula como $H_0: \alpha_{MH} = 1$.

El estadístico de chi-cuadrado corregido de Mantel–Haenszel es:

$$\chi^2_{MH} = \frac{(|\sum_{s=1}^{S}(A_s - E[A_s])| - 0.5)^2}{\sum_{s=1}^{S} \text{Var}\,(A_s)},$$

donde $E[A_s]$ y $\text{Var}(A_s)$ son la esperanza y varianza de A_s bajo la hipótesis nula de independencia condicional.

Bajo H_0, χ^2_{MH} se distribuye aproximadamente como χ^2 con 1 grado de libertad, permitiendo obtener un valor-p.

Para facilitar la interpretación del tamaño del efecto, la razón de odds se transforma a la escala Delta (ETS):

$$\Delta_{MH} = -2.35 \ln(\hat{\alpha}_{MH}),$$

con la siguiente clasificación propuesta por ETS:

- $|\Delta_{MH}| < 1$: DIF despreciable (Categoría A).
- $1 \leq |\Delta_{MH}| < 1.5$: DIF moderado (Categoría B).
- $|\Delta_{MH}| \geq 1.5$: DIF grande (Categoría C).

Este método es robusto y ampliamente usado en aplicaciones operativas. Controla por habilidad general a través de estratos, reduciendo sesgos.

MH proporciona simultáneamente un estadístico para pruebas de significación (χ^2_{MH}) y un indicador de tamaño del efecto (Δ_{MH}).

El método de regresión logística constituye una extensión natural de los procedimientos de comparación de proporciones, ya que modela directamente la probabilidad de éxito en un ítem en función de la habilidad general (aproximada por el puntaje total) y del grupo de pertenencia.

Sea Y_{ij} la respuesta al ítem i del sujeto j, codificada como 1 si es correcta y 0 en caso contrario.

Se modela:

$$\text{logit}[P(Y_{ij} = 1)] = \beta_0 + \beta_1 \cdot \text{Total}_j + \beta_2 \cdot G_j,$$

donde:

- Total_j es el puntaje total del sujeto j en la prueba (variable de matching).
- G_j es un indicador binario del grupo (0 = referencia, 1 = focal).

En este modelo, β_2 captura el efecto uniforme de DIF: una diferencia sistemática en la probabilidad de acierto entre grupos, independiente del nivel de habilidad.

Para detectar posibles interacciones entre habilidad y grupo, se ajusta un modelo extendido:

$$\text{logit}[P(Y_{ij} = 1)] = \beta_0 + \beta_1 \cdot \text{Total}_j + \beta_2 \cdot G_j + \beta_3 \cdot (\text{Total}_j \times G_j).$$

En este caso:

- β_2: DIF uniforme.
- β_3: DIF no uniforme (el efecto del grupo varía según el nivel de habilidad).

Se realizan dos pruebas principales:

4. **DIF uniforme:** contraste $H_0: \beta_2 = 0$.
5. **DIF no uniforme:** contraste $H_0: \beta_3 = 0$.

Ambos se evalúan mediante pruebas de Wald (z) o de razón de verosimilitud (χ^2), obteniendo valores-p para determinar significancia.

- Un valor-p pequeño para β_2 indica presencia de DIF uniforme, es decir, un grupo tiene mayor probabilidad de responder correctamente en todos los niveles de habilidad.
- Un valor-p pequeño para β_3 indica DIF no uniforme, es decir, la ventaja relativa de un grupo depende del nivel de habilidad.
- Si ambos parámetros son no significativos, se concluye ausencia de DIF.

El análisis factorial confirmatorio (AFC) permite contrastar explícitamente la hipótesis de unidimensionalidad, especificando un modelo en el que todos los ítems cargan en un único factor latente.

Modelo básico

Sea Y_{ij} la respuesta observada del sujeto j al ítem i. El modelo de un factor se expresa como:

$$Y_{ij} = \lambda_i F_j + \varepsilon_{ij},$$

donde:

- λ_i es la carga factorial del ítem i.
- F_j es el puntaje latente del factor común.
- ε_{ij} es el error específico del ítem i, con varianza θ_i.

En notación matricial:

$$\mathbf{y}_j = \mathbf{\Lambda} F_j + \boldsymbol{\varepsilon}_j,$$

donde $\mathbf{\Lambda}$ es el vector de cargas factoriales.

Una vez estimado el modelo mediante métodos como WLSMV (Weighted Least Squares Mean and Variance adjusted) para ítems categóricos, se calculan los siguientes índices:

- **CFI (Comparative Fit Index)**

$$CFI = 1 - \frac{\max(\chi^2_{\text{modelo}} - gl_{\text{modelo}},\, 0)}{\max\left(\chi^2_{\text{independencia}} - gl_{\text{independencia}},\, 0\right)}$$

  Valores ≥0.95 indican muy buen ajuste; ≥0,90 ajuste aceptable.

- **TLI (Tucker–Lewis Index)**

$$TLI = \frac{(\chi^2_{\text{independencia}}/gl_{\text{independencia}}) - (\chi^2_{\text{modelo}}/gl_{\text{modelo}})}{\left(\chi^2_{\text{independencia}}/gl_{\text{independencia}}\right) - 1}$$

  Valores ≥0,95 indican buen ajuste; ≥0,90 aceptable.

- **RMSEA (Root Mean Square Error of Approximation)**

$$RMSEA = \sqrt{\frac{\chi^2_{\text{modelo}} - gl_{\text{modelo}}}{gl_{\text{modelo}}(N-1)}}$$

Valores ≤0,05 reflejan muy buen ajuste; 0,05–0,08 aceptable; >0,10 pobre ajuste. Se acompaña con un intervalo de confianza al 90%.

- **SRMR (Standardized Root Mean Square Residual)**

$$SRMR = \sqrt{\frac{2}{p(p+1)}\sum_{i<j}\left(r_{ij} - \hat{r}_{ij}\right)^2}$$

donde r_{ij} son correlaciones observadas y $\hat{r}_{ij}$ las estimadas. Valores ≤0,05 son excelentes; ≤0,08 aceptables.

- **χ^2/gl (chi-cuadrado sobre grados de libertad)**

$$\frac{\chi^2}{gl}$$

Valores ≤2 sugieren buen ajuste, ≤3 aceptable.

Un modelo unidimensional será sostenido empíricamente si la mayoría de los índices de ajuste se ubican dentro de los rangos adecuados (CFI/TLI ≥,90, RMSEA y SRMR bajos, χ^2/gl reducido). Si varios índices se encuentran fuera de rango, se considera evidencia de multidimensionalidad o de especificación inadecuada del modelo.

El método de Wald para la detección de Funcionamiento Diferencial del Ítem (DIF) en modelos de la TRI se basa en contrastar la igualdad de parámetros del ítem entre dos o más grupos previamente definidos. Bajo el supuesto nulo de no-DIF, se espera que los parámetros estimados (por ejemplo, dificultad b_i y discriminación a_i) sean equivalentes entre los grupos.

Sea $\widehat{\boldsymbol{\theta}}_i$ el vector de parámetros estimados para el ítem i en el grupo focal, y $\widehat{\boldsymbol{\theta}}_{i0}$ los parámetros en el grupo de referencia (o bajo restricción de igualdad). El contraste de Wald se formula como:

$$W = \left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i0}\right)^{\top}\left[\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i0}\right)\right]^{-1}\left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i0}\right),$$

donde $\mathrm{Var}(\cdot)$ es la matriz de covarianza asintótica de las estimaciones de parámetros.

Bajo la hipótesis nula de invarianza de parámetros (ausencia de DIF), el estadístico W se distribuye aproximadamente como una chi–cuadrado:

$$W \sim \chi^2_{df},$$

con grados de libertad igual al número de parámetros contrastados (por ejemplo, $df = 1$ en el modelo de Rasch, donde solo se contrasta la dificultad b_i).

Si $p > 0{,}05$: no se rechaza la hipótesis nula, no hay evidencia estadística de DIF en el ítem.
Si $p \leq 0{,}05$: se rechaza la hipótesis nula, indicando que el ítem presenta DIF significativo, es decir, el parámetro difiere entre grupos más allá del error muestral.

El método de Wald es sensible tanto a DIF uniforme (diferencias sistemáticas en dificultad) como a DIF no uniforme (diferencias en parámetros de discriminación o interacción con el nivel de habilidad, según el modelo especificado).

Este método utiliza directamente las estimaciones de parámetros y su varianza, evitando la necesidad de comparar modelos anidados mediante verosimilitud. Es computacionalmente eficiente y ampliamente implementado en *software* psicométrico. Depende fuertemente de la calidad de la estimación de los parámetros. Si el tamaño muestral es pequeño o los ítems presentan categorías con frecuencias muy bajas, los resultados pueden ser inestables.

En la práctica, el método de Wald se complementa con otros enfoques (como el LRT o los contrastes de regresión logística) para obtener un panorama más robusto de la presencia de DIF.

Método de Lord

El método de Lord (1980) es un procedimiento clásico para DIF. Se basa en contrastar los parámetros de dificultad de los ítems entre grupos, ajustando por posibles sesgos de escala a través de un conjunto de ítems considerados de anclaje (sin DIF).

En el modelo logístico de 1 parámetro (modelo de Rasch), la probabilidad de respuesta correcta al ítem i por un individuo con habilidad θ en el grupo g se expresa como:

$$P(X_{ij} = 1 \mid \theta, b_{ig}) = \frac{1}{1 + \exp[-(\theta - b_{ig})]},$$

donde b_{ig} es el parámetro de **dificultad** del ítem en el grupo g.

El método de Lord compara los parámetros de dificultad entre el grupo de referencia y el grupo focal. Sea:

$$\Delta b_i = \hat{b}_{iF} - \hat{b}_{iR},$$

la diferencia en dificultad del ítem i entre el grupo focal (F) y el de referencia (R).

El estadístico de Lord se calcula como:

$$\chi_i^2 = \frac{(\hat{b}_{iF} - \hat{b}_{iR})^2}{\mathrm{Var}(\hat{b}_{iF}) + \mathrm{Var}(\hat{b}_{iR})},$$

que, bajo la hipótesis nula de ausencia de DIF, sigue aproximadamente una distribución χ^2 con 1 grado de libertad.

Además de la significancia estadística, el método de Lord utiliza un índice de tamaño del efecto expresado en la escala ETS Delta:

$$\Delta_i^{Lord} = -2.35 \cdot \Delta b_i,$$

donde el factor 2,35 transforma la diferencia de parámetros de dificultad a la métrica de la escala Delta de ETS (Educational Testing Service).

- Valores próximos a 0 sugieren ausencia de DIF.
- Valores positivos/negativos reflejan ventaja para uno u otro grupo.

ETS clasifica la magnitud del DIF en categorías:
- **A**: efecto despreciable,
- **B**: efecto moderado,
- **C**: efecto grande.

Método de Raju

El método de Raju (1988, 1990) se basa en la comparación de las curvas características del ítem (CCI) entre el grupo de referencia y el grupo focal, evaluando el área entre ambas como indicador de DIF.

Sea $P_{iR}(\theta)$ y $P_{iF}(\theta)$ las probabilidades de respuesta correcta al ítem i en el grupo de referencia (R) y en el grupo focal (F), respectivamente, de acuerdo con el modelo TRI considerado (por ejemplo, 2PL o Rasch).

El estadístico de Raju cuantifica el área entre las dos CCI en el rango de habilidades θ:

$$\Delta_i = \int_{-\infty}^{+\infty} [P_{iF}(\theta) - P_{iR}(\theta)]\, f(\theta)\, d\theta,$$

donde $f(\theta)$ es la distribución de la habilidad en la población (habitualmente $N(0,1)$).

Este índice refleja el desplazamiento promedio de las probabilidades de respuesta** entre los dos grupos: - Si $\Delta_i \approx 0$, el ítem no muestra DIF.

- Valores positivos o negativos de Δ_i indican sesgo a favor del grupo focal o del grupo de referencia, respectivamente.

Raju propone una prueba de significación para evaluar si Δ_i difiere significativamente de cero, con base en un estadístico normalizado:

$$Z_i = \frac{\hat{\Delta}_i}{\mathrm{SE}(\hat{\Delta}_i)},$$

donde $\mathrm{SE}(\hat{\Delta}_i)$ es el error estándar de la estimación del área.

Bajo la hipótesis nula de ausencia de DIF ($\Delta_i = 0$), el estadístico Z_i se distribuye aproximadamente como una normal estándar.

- Si $|Z_i|$ excede el valor crítico (por ejemplo, 1,96 para $\alpha = 0{,}05$), se concluye que el ítem presenta DIF estadísticamente significativo.
- El signo de Δ_i indica la dirección del sesgo (a favor del grupo focal si es positivo o del de referencia si es negativo).
- La magnitud de Δ_i puede además clasificarse mediante umbrales establecidos en la práctica (ETS A/B/C o criterios alternativos).

El método de Raju es particularmente atractivo porque no se limita a comparar parámetros de los modelos, sino que considera la discrepancia en las *curvas de probabilidad de respuesta, ofreciendo una medida integral y más directamente interpretable del DIF. Además, es aplicable tanto al modelo Rasch como a modelos logísticos de 2 y 3 parámetros.

Regresión logística sobre θ

El método de regresión logística sobre la habilidad estimada (θ) es un enfoque flexible para detectar DIF, ya sea uniforme o no uniforme. Este procedimiento se basa en modelar la probabilidad de responder correctamente a un ítem como función de la habilidad latente y del grupo de pertenencia.

Sea Y_{ij} la respuesta al ítem i del sujeto j ($Y_{ij} \in \{0,1\}$), θ_j la habilidad estimada del sujeto mediante un modelo TRI y G_j una variable indicadora de pertenencia al grupo (0 = referencia, 1 = focal). El modelo de regresión logística es:

$$\log\left(\frac{\Pr(Y_{ij}=1)}{1-\Pr(Y_{ij}=1)}\right) = \beta_0 + \beta_1\theta_j + \beta_2 G_j + \beta_3(\theta_j \cdot G_j).$$

- β_0: intercepto (nivel base del grupo de referencia).
- β_1: efecto de la habilidad θ sobre la probabilidad de acierto.
- β_2: efecto del grupo, **DIF uniforme** (desplazamiento paralelo de la CCI entre grupos).
- β_3: interacción habilidad × grupo, **DIF no uniforme** (diferencias en pendiente de la CCI).

Se evalúan hipótesis anidadas comparando modelos:

6. **Modelo nulo (sin DIF):**

$$\log\frac{p}{1-p} = \beta_0 + \beta_1\theta_j.$$

7. **Modelo con DIF uniforme:**

$$\log\frac{p}{1-p} = \beta_0 + \beta_1\theta_j + \beta_2 G_j.$$

8. **Modelo con DIF no uniforme:**

$$\log\frac{p}{1-p} = \beta_0 + \beta_1\theta_j + \beta_2 G_j + \beta_3(\theta_j \cdot G_j).$$

- **Prueba de** β_2: DIF uniforme.
- **Prueba de** β_3: DIF no uniforme.
- Contrastes mediante **Wald** o **Likelihood Ratio Test (LRT)**.

Un β_2 significativo indica que, a igual nivel de habilidad, un grupo tiene ventaja sistemática sobre el otro (DIF uniforme).

Un β_3 significativo indica que la relación entre habilidad y probabilidad de éxito difiere entre grupos (DIF no uniforme).

Este enfoque permite separar claramente los dos tipos de DIF, además de ofrecer medidas de magnitud a través de los coeficientes.

Método SIBTEST

El Simultaneous Item Bias Test (SIBTEST), desarrollado por Shealy y Stout (1993), es un procedimiento no paramétrico diseñado para detectar DIF en el contexto de modelos unidimensionales. Se basa en la idea de comparar el desempeño de grupos de referencia y focal, emparejados en función de una medida del rasgo latente.

El método utiliza una partición de los ítems en dos conjuntos:

- **Pestaña de anclaje (valid subtest):** conjunto de ítems libres de DIF, que se usan para estimar la habilidad latente de los sujetos.
- **Subtest sospechoso (suspect subtest):** ítems que se desea evaluar respecto a posible presencia de DIF.

La lógica de SIBTEST es que, si el subtest sospechoso no contiene DIF, las diferencias de puntuación entre grupos (emparejados en habilidad mediante el subtest válido) deberían ser nulas. Diferencias sistemáticas se interpretan como evidencia de DIF.

El estadístico principal del método es:

$$\hat{\beta} = \mathbb{E}[X_S \mid \theta]_F - \mathbb{E}[X_S \mid \theta]_R,$$

donde: - X_S es la puntuación en el subtest sospechoso.
- Los subíndices F y R indican grupos focal y de referencia, respectivamente.

El estimador $\hat{\beta}$ mide la diferencia media condicional entre grupos en el subtest sospechoso.

El estadístico de contraste es:

$$T_{\text{SIB}} = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})},$$

que bajo la hipótesis nula de ausencia de DIF se distribuye aproximadamente como una normal estándar.

- Si $\hat{\beta} > 0$, el ítem o subtest favorece al grupo focal.
- Si $\hat{\beta} < 0$, el ítem o subtest favorece al grupo de referencia.
- Magnitudes grandes de $|T_{\text{SIB}}|$ sugieren la presencia de DIF estadísticamente significativo.

Existen variantes del método: - **PUF-SIBTEST (Purification SIBTEST):** depura iterativamente el conjunto de anclaje eliminando ítems con evidencia de DIF.

- **NCDIF (Non-Compensatory DIF Index):** generalización de SIBTEST que cuantifica la magnitud del DIF en escalas continuas.

SIBTEST es especialmente útil en aplicaciones con ítems dicotómicos y donde no se desea imponer un modelo paramétrico estricto como Rasch o 2PL. Se considera un método robusto y apropiado para estudios exploratorios de DIF, aunque requiere tamaños muestrales moderados a grandes (al menos 200 sujetos por grupo para buena estabilidad de los estimadores).

Estudio específico para la prueba de Producción escrita y organización textual

En la prueba de Producción escrita y organización textual se formuló un Modelo de Rasch de Múltiples Facetas (Many-Facet Rasch Model, MFRM) con cuatro facetas: personas (n), criterios/ítems (i), jueces (j) y formas de prueba (f), especificado bajo dos esquemaa Rating Scale Model (RSM) y Partial Credit Model (PCM).

Rating Scale Model

Consideremos $X_{nijf} \in \{0,1,\dots,m\}$ la puntuación observada para la persona n, en el criterio (o ítem) i, calificada por el juez j, en la forma de prueba f, donde m es el

número máximo de categoría (común a todos los criterios en RSM). Se definen las siguientes cantidades latentes en la métrica logit (escala de Rasch):

- β_n: habilidad de la persona n.
- δ_i: dificultad (o severidad del criterio) del ítem/criterio i.
- γ_j: severidad del juez j.
- φ_f: efecto de la forma de prueba f.
- τ_k: umbral de categoría común (RSM) para el paso k ($k = 1, \dots, m$), con $\tau_0 \equiv 0$ por convenio.

En el esquema RSM todos los criterios comparten los mismos umbrales $\{\tau_k\}_{k=1}^{m}$, a diferencia del PCM donde los umbrales pueden variar por criterio.

Bajo el MFRM–RSM, la probabilidad de observar la categoría k ($k = 0, \dots, m$) viene dada por

$$P(X_{nijf} = k) = \frac{\exp(\sum_{s=1}^{k}[\beta_n - \delta_i - \gamma_j - \varphi_f - \tau_s])}{\sum_{h=0}^{m}\exp(\sum_{s=1}^{h}[\beta_n - \delta_i - \gamma_j - \varphi_f - \tau_s])}, \quad (1)$$

entendiendo que la suma vacía para $k = 0$ es cero. Equivalentemente, los logits adyacentes toman la forma

$$\log\frac{P(X_{nijf}=k)}{P(X_{nijf}=k-1)} = (\beta_n - \delta_i - \gamma_j - \varphi_f) - \tau_k, \qquad k = 1, \dots, m. \quad (2)$$

La expresión (2) muestra que el contraste entre categorías consecutivas depende de: (i) la localización neta del desempeño β_n frente a la severidad/dificultad agregada $\delta_i + \gamma_j + \varphi_f$, y (ii) el umbral común de paso τ_k.

El modelo está sobredeterminado en su localización absoluta; por lo tanto, se requieren restricciones de identificabilidad. Dos estrategias comunes son:

1. **Centraje por suma-cero** en cada faceta:
   $\sum_n \beta_n = 0, \quad \sum_i \delta_i = 0, \quad \sum_j \gamma_j = 0, \quad \sum_f \varphi_f = 0, \quad \sum_{k=1}^{m} \tau_k = 0.$

2. **Anclaje respecto de una referencia** en cada faceta (por ejemplo, fijar a cero un parámetro por faceta):
   $\beta_{n^\star} = 0, \quad \delta_{i^\star} = 0, \quad \gamma_{j^\star} = 0, \quad \varphi_{f^\star} = 0, \quad$ y $\sum_{k=1}^{m} \tau_k = 0$ (o fijar uno de los τ_k).

En RSM, es habitual imponer $\sum_{k=1}^{m} \tau_k = 0$ o fijar $\tau_1 = 0$ para definir la escala de los umbrales.

Bajo independencia condicional dada la estructura de facetas, la verosimilitud para un conjunto de observaciones $\mathcal{D} = \{x_{nijf}\}$ es

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_{n,i,j,f}\prod_{k=0}^{m}\left[P(X_{nijf} = k)\right]^{\mathbb{I}(x_{nijf}=k)}, \quad (3)$$

donde $\boldsymbol{\theta} = \{\beta_n\} \cup \{\delta_i\} \cup \{\gamma_j\} \cup \{\varphi_f\} \cup \{\tau_k\}$ y $P(X_{nijf} = k)$ se define en (1). La log-verosimilitud correspondiente es

$$\ell(\boldsymbol{\theta}) = \sum_{n,i,j,f}\sum_{k=0}^{m}\mathbb{I}(x_{nijf} = k)\log P(X_{nijf} = k). \quad (4)$$

En la métrica logit, incrementos positivos de β_n aumentan la probabilidad de categorías superiores; incrementos de δ_i, γ_j o φ_f disminuyen tal probabilidad (mayor dificultad/severidad/efecto de forma). Los umbrales τ_k ordenan el progreso entre

categorías, compartidos por todos los criterios (RSM), lo que favorece la comparabilidad transversal entre criterios y formas.

Modelo de Crédito Parcial

Si usamos el esquema PCM (Masters, 1982) sea $X_{nijf} \in \{0,1,\dots,m_i\}$ la puntuación observada para la persona n, en el criterio (o ítem) i, calificada por el juez j, en la forma de prueba f, donde m_i es el número máximo de categoría del criterio i. Se definen las siguientes cantidades latentes en la métrica logit (escala de Rasch):

- β_n: habilidad de la persona n.
- δ_i: dificultad (o severidad del criterio) del ítem/criterio i.
- γ_j: severidad del juez j.
- φ_f: efecto de la forma de prueba f.
- τ_{ik}: umbral de categoría del criterio i para el paso k ($k = 1,\dots,m_i$), con $\tau_{i0} \equiv 0$ por convenio.

En el PCM, los umbrales $\{\tau_{ik}\}$ pueden variar por criterio, permitiendo distintas estructuras de categorías entre criterios.

Bajo el MFRM–PCM, la probabilidad de observar la categoría k ($k = 0,\dots,m_i$) viene dada por

$$P(X_{nijf} = k) = \frac{\exp(\sum_{s=1}^{k}[\beta_n - \delta_i - \gamma_j - \varphi_f - \tau_{is}])}{\sum_{h=0}^{m_i}\exp(\sum_{s=1}^{h}[\beta_n - \delta_i - \gamma_j - \varphi_f - \tau_{is}])}, \quad (5)$$

con la convención de que la suma vacía para $k = 0$ es cero. Equivalentemente, los logits adyacentes toman la forma

$$\log\frac{P(X_{nijf}=k)}{P(X_{nijf}=k-1)} = (\beta_n - \delta_i - \gamma_j - \varphi_f) - \tau_{ik}, \quad k = 1,\dots,m_i. \quad (6)$$

La expresión (6) muestra que el contraste entre categorías consecutivas depende de: (i) la localización neta del desempeño β_n frente a la severidad/dificultad agregada $\delta_i + \gamma_j + \varphi_f$ y (ii) el umbral específico del criterio τ_{ik}.

El modelo requiere restricciones de identificabilidad para fijar el origen de la escala. Dos estrategias comunes son:

1. **Centraje por suma-cero** en cada faceta:

$$\sum_n \beta_n = 0, \quad \sum_i \delta_i = 0, \quad \sum_j \gamma_j = 0, \quad \sum_f \varphi_f = 0, \quad \sum_{k=1}^{m_i} \tau_{ik} = 0 \quad \forall i.$$

2. **Anclaje respecto de una referencia** en cada faceta (por ejemplo, fijar a cero un parámetro por faceta):

$$\beta_{n^\star} = 0, \quad \delta_{i^\star} = 0, \quad \gamma_{j^\star} = 0, \quad \varphi_{f^\star} = 0, \quad \text{yparacada} i: \sum_{k=1}^{m_i} \tau_{ik} = 0 \text{ ofijarunodelos} \tau_{ik}.$$

En PCM, lo habitual es imponer $\sum_{k=1}^{m_i} \tau_{ik} = 0$ por cada criterio i o fijar $\tau_{i1} = 0$.

Bajo independencia condicional dada la estructura de facetas, la verosimilitud para un conjunto de observaciones $\mathcal{D} = \{x_{nijf}\}$ es

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) \;=\; \prod_{n,i,j,f} \prod_{k=0}^{m_i} \left[P\big(X_{nijf} = k\big)\right]^{\mathbb{I}(x_{nijf}=k)}, \qquad (7)$$

donde $\boldsymbol{\theta} = \{\beta_n\} \cup \{\delta_i\} \cup \{\gamma_j\} \cup \{\varphi_f\} \cup \{\tau_{ik}\}$ y $P(X_{nijf} = k)$ se define en (1). La log-verosimilitud correspondiente es

$$\ell(\boldsymbol{\theta}) \;=\; \sum_{n,i,j,f} \sum_{k=0}^{m_i} \mathbb{I}(x_{nijf} = k) \log P\big(X_{nijf} = k\big). \qquad (8)$$

En la métrica logit, aumentos en β_n incrementan la probabilidad de categorías superiores; aumentos en δ_i, γ_j o φ_f la reducen (mayor dificultad/severidad/efecto de forma). Los umbrales τ_{ik} ordenan el progreso entre categorías *dentro* de cada criterio, permitiendo estructuras de respuesta heterogéneas entre criterios.

El MFRM conserva la estructura tipo Rasch en el sentido de que los totales por faceta actúan como estadísticas suficientes para los parámetros correspondientes bajo JMLE; por ejemplo, el total por persona $r_n = \sum_{i,j,f} x_{nijf}$ informa sobre β_n una vez controladas las restantes facetas. No obstante, a diferencia del Rasch simple, la condicionalización exacta para eliminar todas las facetas no es trivial; en la práctica se emplean JMLE, MMLE o enfoques bayesianos.

Para la estimación de los parámetros son habituales las siguientes estrategias:

- **JMLE** (Joint MLE): maximización conjunta de (4) con restricciones de identificabilidad. Se implementa típicamente vía Newton–Raphson, Fisher scoring o algoritmos alternantes por bloques.
- **MMLE/EM**: máxima verosimilitud marginal tratando $\{\beta_n\}$ como efectos aleatorios con distribución a priori (por ejmplo, normal), estimando el resto de parámetros por EM/Laplace.
- **Bayesiano**: estimación por MCMC (por ejemplo, Gibbs/Metropolis) con *priors* adecuados para cada faceta.

En todos los casos, la matriz de información observada (o su aproximación) proporciona errores estándar y permite pruebas de ajuste y constrastes entre facetas (por ejemplo, diferencias de severidad entre jueces).

Si en lugar de RSM se especifica PCM, la ecuación (2) reemplaza τ_k por umbrales específicos por criterio τ_{ik}, perdiéndose la homogeneidad entre criterios pero ganándose flexibilidad:

$$\log \frac{P(X_{nijf}=k)}{P(X_{nijf}=k-1)} \;=\; \big(\beta_n - \delta_i - \gamma_j - \varphi_f\big) - \tau_{ik}.$$

Si se impusiera que los umbrales son comunes a todos los criterios ($\tau_{ik} \equiv \tau_k$), el PCM se reduce al RSM, recuperando una estructura homogénea de categorías entre criterios y mayor parquedad paramétrica.

El Facets otorga para cada faceta información relevante da una tabla para cada faceta (persona, juez, ítem o criterio). Los primeros indicadores son:

- **Total Score**: suma de puntuaciones observadas $S = \sum_t x_t$.
- **Total Count**: número de observaciones consideradas N.
- **Observed Average**: promedio empírico $\bar{x} = S/N$.
- **Fair Average (M)**: promedio ajustado al modelo Rasch, corrigiendo por las restantes facetas.

También se obtiene la estimacion de la medida y su error estandar. El parámetro fundamental estimado es la **medida de Rasch** el logit de la probabilidad ajustada $\hat{\theta}$ que representa habilidad (personas), severidad (jueces), dificultad (criterios) o efecto de forma (formas de prueba). El **error estándar** se obtiene a partir de la matriz de información observada.

Los índices de ajuste *fit* evalúan la consistencia de los datos respecto al modelo Rasch. $\text{InfitMnSq} = \frac{\sum w_i (O_i - E_i)^2}{\sum w_i}$, y $\text{OutfitMnSq} = \frac{\sum (O_i - E_i)^2}{N}$,
donde O_i es la respuesta observada, E_i la probabilidad esperada, w_i un peso proporcional a la varianza esperada y N el número de observaciones.

Los valores orientativos ideales del Infit y el Outfit son ≈ 1 con un intervalo aceptable: 0,7 a 1,3

La versión tipificada $ZStd$ se define como $ZStd = \frac{\text{MnSq}-1}{\text{ErrorEst.}}$, indicando la magnitud de desviación respecto a lo esperado (valores $|ZStd| > 2$ sugieren desajuste significativo).

Algunos modelos permiten estimar un parámetro de discriminación α:

$$P(X = 1|\theta) = \frac{\exp(\alpha(\theta - \delta))}{1 + \exp(\alpha(\theta - \delta))},$$

donde $\alpha = 1$ corresponde al Rasch estricto; valores distintos reflejan variaciones en la pendiente de la curva característica del ítem o faceta.

La correlación punto-medida empírica se define como

$$r_{pt-mea} = \frac{\sum (x_i - \bar{x})(\hat{\theta}_i - \bar{\theta})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (\hat{\theta}_i - \bar{\theta})^2}},$$

comparándose con el valor esperado bajo el modelo r_{pt-exp}. Correlaciones bajas o negativas sugieren mal funcionamiento del ítem o juez.

La **Exact Agreement** evalúa el porcentaje de coincidencia entre las respuestas observadas y las categorías predichas como más probables por el modelo:

$$\text{ExactAgreement} = \frac{\text{Númeroderespuestascoincidentes}}{\text{Númerototalderespuestas}} \times 100\%.$$

Sea $\hat{\theta}_p$ la medida estimada (en logits) para la entidad p de una faceta (persona, juez, etc.), con error estándar modelo SE_p. Defínanse:

$$\text{RMSE} \equiv \sqrt{\frac{1}{P}\sum_{p=1}^{P} SE_p^2},$$

$$\text{SD}^2 \equiv \frac{1}{P-1}\sum_{p=1}^{P} \left(\hat{\theta}_p - \bar{\theta}\right)^2, \qquad \bar{\theta} = \frac{1}{P}\sum_{p=1}^{P} \hat{\theta}_p,$$

$$\text{Adj(True)S. D.} \equiv \sqrt{\max(0,\ \text{SD}^2 - \text{RMSE}^2)}.$$

La **separación** y el número de estratos se definen por

$$\text{Separation} \equiv \frac{\text{Adj(True)S.D.}}{\text{RMSE}},$$
$$\text{Strata} \equiv \frac{4\,\text{Separation}+1}{3}.$$

La **fiabilidad** (*Reliability*) es la proporción de varianza verdadera en la varianza observada; es algebraicamente equivalente a

$$\text{Reliability} \equiv \frac{\text{Adj(True)S.D.}^2}{\text{Adj(True)S.D.}^2+\text{RMSE}^2} = \frac{\text{Separation}^2}{1+\text{Separation}^2}.$$

Tabla 12.
Valores de referencia

Métrica	**Rango de referencia esperado**	**Interpretación**
RMSE (Root Mean Square Error)	Lo más bajo posible (≈ 0,1–0,3 en aplicaciones típicas)	Error medio cuadrático de las estimaciones. Indica la precisión promedio de las medidas; valores bajos = mayor precisión.
Adj(True) S.D.	Mayor que el RMSE y representativo de la varianza real (≈ 0,3–1,0)	Desviación estándar "verdadera", ajustada por error de medición. Refleja la varianza real entre entidades.
Separation	>2,0 (aceptable), >3,0 (bueno), >4,0 (muy bueno)	Índice de separación: cuántas veces la dispersión verdadera excede al error. Valores altos indican mayor capacidad discriminativa.
Strata	>3 (adecuado), >4 (bueno), >5 (muy bueno)	Número de grupos distinguibles en la escala. Aproxima cuántos niveles distintos pueden diferenciarse confiablemente.
Reliability	>0,80 (aceptable), >0,90 (bueno), >0,95 (excelente)	Consistencia de la medida: proporción de varianza verdadera en la varianza observada. Equivalente al coeficiente de fiabilidad.

Equiparación

Se equipararon las pruebas 2023 y 2025 y se determinaron los puntos de corte equivalentes de las pruebas utilizando según el tamaño muestras TCT o TRI.

Procedimientos basados en TCT

En el diseño NEAT (Nonequivalent groups with Anchor Test) se busca colocar en una misma escala los puntajes de dos formas de prueba administradas a grupos no equivalentes, utilizando un test ancla como puente estadístico. La idea central es que los puntajes observados en las dos aplicaciones provienen de poblaciones distintas y, por tanto, no es apropiado suponer equivalencia muestral. La literatura propone modelar una población sintética como combinación ponderada de las poblaciones que rinden cada forma, con pesos que suman uno. Sobre esa población sintética se definen los parámetros de interés y se construye la función de equiparación. En términos de notación, si X es la forma del primer año, Y la del segundo, V el ancla y w_1, w_2 los pesos sintéticos, entonces las medias y varianzas de X e Y en la población sintética satisfacen

$$\mu(X_s) = w_1\mu(X_1) + w_2\mu(X_2), \qquad \sigma^2(X_s) = w_1\sigma^2(X_1) + w_2\sigma^2(X_2) + w_1w_2[\mu(X_1) - \mu(X_2)]^2,$$

y análogamente para Y. Dado que en NEAT X no se observa en el segundo grupo ni Y en el primero, los términos faltantes se imputan mediante el ancla y supuestos de relación entre total y ancla, produciendo versiones ajustadas de $\mu(\cdot)$ y $\sigma^2(\cdot)$ que dependen del método elegido.

Métodos lineales

Los métodos lineales comparten la estructura $e(x) = \alpha + \beta x$, donde $\beta = \sigma(Y_s)/\sigma(X_s)$ y $\alpha = \mu(Y_s) - \beta\,\mu(X_s)$. Se distinguen por cómo estiman los momentos sintéticos $\mu(\cdot)_s$ y $\sigma^2(\cdot)_s$.

El método de pesos nominales utiliza combinaciones ponderadas por tamaños muestrales sin introducir correcciones específicas por relación total–ancla. Su simplicidad lo hace estable en muestras pequeñas, aunque su capacidad de corregir no equivalencias entre grupos es limitada. Cuando las poblaciones difieren sustancialmente, el sesgo residual puede ser apreciable, en particular si el ancla es corto o poco representativo del constructo total.

El método de Tucker también se apoya en la población sintética, pero su tratamiento de los momentos no incorpora explícitamente la regresión entre total y ancla. En la práctica actúa como un ajuste de medias y varianzas que resulta adecuado cuando la no equivalencia entre grupos es moderada y el ancla tiene poder limitado. Tucker es una referencia clásica y constituye una línea de base útil para evaluar sensibilidad frente a alternativas más ricas.

El método de Levine de puntaje observado extiende a Tucker incorporando la relación entre total y ancla a través de coeficientes de regresión γ específicos por grupo. Para ancla interna, una parametrización frecuente escribe $\gamma_1 = \sigma^2(X_1)/\sigma(X_1, V_1)$ y $\gamma_2 =$

$\sigma^2(Y_2)/\sigma(Y_2, V_2)$; estas cantidades ajustan las medias y varianzas de X e Y cuando se proyectan a la población sintética. El efecto es compensar de manera explícita las diferencias entre poblaciones usando la información del ancla. En contextos de cohortes con perfiles heterogéneos, Levine suele ofrecer una corrección más adecuada que Tucker, siempre que el ancla tenga longitud y representatividad razonables respecto del total.

El método de Braun/Holland es una variante lineal menos difundida que modifica la construcción de momentos sintéticos, con especial atención a escenarios con múltiples ítems de anclaje. Puede otorgar estabilidad adicional cuando se cuenta con más de un conjunto de ítems comunes bien diseñados. No obstante, su menor presencia en informes operativos suele relegarlo a análisis técnicos o de sensibilidad.

Métodos no lineales

Los métodos no lineales abandonan la restricción afín y buscan preservar la estructura de las distribuciones. La estimación por frecuencias conduce a una equiparación equipercentil directa, en la que el puntaje transformado conserva su posición relativa en la distribución objetivo. Este enfoque captura diferencias de forma más allá de medias y varianzas, pero es sensible a la discreción y a la irregularidad de las frecuencias en muestras pequeñas, por lo que requiere presuavizado, típicamente log-lineal, para lograr curvas estables y evitar oscilaciones en colas.

El encadenado equipercentil (NEAT-CE) utiliza el ancla como puente y compone tres transformaciones: de la forma fuente al ancla en el primer grupo, del ancla del primer grupo al del segundo y del ancla del segundo grupo a la forma destino. El resultado es una función no lineal que preserva percentiles a lo largo del encadenamiento. Su ventaja radica en capturar no linealidades que los métodos lineales no modelan; su desventaja es la acumulación del error de cada eslabón y la necesidad de tamaños muestrales suficientes o, en su defecto, de presuavizado robusto.

La selección depende de objetivos, calidad y longitud del ancla, y tamaños muestrales. Cuando los grupos no son equivalentes, el marco sintético es indispensable; si, además, el ancla es informativo, Levine ofrece una corrección de primer orden al explotar la relación total–ancla. Tucker resulta apropiado como línea de base o cuando se busca una solución estable con ítems de anclaje débiles y no equivalencia moderada. Las alternativas no lineales, como equipercentil con estimación por frecuencias o encadenado, son preferibles cuando se espera que la relación entre las formas no sea bien representada por una transformación lineal, y cuando el tamaño muestral y el presuavizado permiten obtener funciones suaves y monotónicas. En presencia de múltiples ítems de anclaje bien construidas, Braun/Holland puede complementar el análisis lineal, aportando robustez adicional.

En aplicaciones operativas con decisiones de punto de corte definidas en puntaje observado, es frecuente adoptar Levine como método principal y reportar Tucker y, si el tamaño lo permite, NEAT-CE como análisis de sensibilidad. La consistencia de cortes transportados y la estabilidad en los extremos de la escala deben evaluarse

explícitamente, idealmente con intervalos de error estándar de equiparación obtenidos por delta o bootstrap.

Tabla 13

Comparación de métodos de equiparación en diseño NEAT

Método	Naturaleza	Uso ancla gamma	Fortaleza principal	Limitación principal	Escenarios recomendados
Pesos nominales	Lineal	No explícito	Simplicidad y estabilidad con N chico	Puede dejar sesgo con poblaciones distintas	Grupos casi equivalentes o ancla débil
Tucker	Lineal	No explícito	Baseline robusto en no equivalencia moderada	No corrige relación total–ancla; sesgo si no equivalencia alta	Referencia y estabilidad; ancla corta
Levine (puntaje observado)	Lineal	Sí, regresión total–ancla (γ)	Corrección explícita de no equivalencia con ancla informativa	Sensibilidad a calidad/longitud del ancla	Cohortes no equivalentes con ancla representativa
Braun/Holland	Lineal	Sí, construcción alternativa	Estabilidad con ítems de anclaje múltiples	Menor estandarización en reportes operativos	Múltiples ítems de anclaje bien diseñadas
Estimación por frecuencias (equipercentil)	No lineal	Sí, vía distribuciones	Captura diferencias de forma; preserva percentiles	Requiere suavizado y N suficiente	N grande y necesidad de forma completa de la distribución
Encadenado equipercentil (NEAT-CE)	No lineal	Sí, vía composición	Preserva percentiles a través del puente ancla	Acumula error en cada eslabón; requiere N o suavizado	Explorar no linealidad con ancla adecuado y N moderado

Método lineal de Levine para diseño NEAT

A continuación se presenta una descripción del método lineal de Levine para diseño NEAT aplicado.

En el diseño NEAT se desean equiparar dos formas X (2023) e Y (2025) administradas a grupos no equivalentes.

Cada grupo responde, además, un test de anclaje V: el grupo que toma la prueba$X$ tiene un puntaje V_1 en el test de anclaje y el que toma la prueba Y tiene un puntaje V_2 en el test de anclaje.

En términos del modelo clásico de test tenemos que $X = G + V$ e $Y = H + V$ donde G son ítems exclusivos de X (2023) y H son ítems exclusivos de Y (2025)

El objetivo del método es construir una transformación lineal $l(x) = \alpha + \beta x$ tal que los puntajes observados de X, transformados por $l(\cdot)$, tengan la misma media y desviación estándar que los de Y en una población sintética que pondera ambos grupos.

Usando la notación de Hanson et al. (1993), sea la población sintética una mezcla con pesos w_1 y w_2 $(w_1 + w_2 = 1)$, entonces el funcional de equiparación de Levine (observed-score) es

$$l(x) \;=\; \frac{\sigma(Y_s)}{\sigma(X_s)}\,[\,x - \mu(X_s)\,]\;+\;\mu(Y_s),$$

donde $\mu(\cdot)$ y $\sigma(\cdot)$ denotan media y desviación estándar en la población sintética "s".

Los momentos de la mezcla vienen dados por:

$$\mu(X_s) = w_1\mu(X_1) + w_2\mu(X_2), \qquad \mu(Y_s) = w_1\mu(Y_1) + w_2\mu(Y_2),$$
$$\sigma^2(X_s) = w_1\sigma^2(X_1) + w_2\sigma^2(X_2) + w_1w_2[\mu(X_1) - \mu(X_2)]^2,$$
$$\sigma^2(Y_s) = w_1\sigma^2(Y_1) + w_2\sigma^2(Y_2) + w_1w_2[\mu(Y_1) - \mu(Y_2)]^2$$

donde X_1 es el puntaje total de la forma X observado en el grupo 1, X_2 es el puntaje total que obtendría el grupo 2 en la forma X, Y_1 es el puntaje total que obtendría el grupo 1 en la forma Y e Y_2 es el puntaje total de la forma Y.

El problema práctico es que en el NEAT no se observan simultáneamente los componentes exclusivos G (ítems propios de X) y H (ítems propios de Y) en ambos grupos (pues G solo se encuentra en 2023 y H en 2025), de modo que $\mu(X_2)$, $\mu(Y_1)$, $\sigma^2(X_2)$ y $\sigma^2(Y_1)$ no son directamente estimables.

La solución clásica de Levine asume un modelo congénere clásico para G, H, V (Angoff/Feldt), en el cual, para $i = 1,2$,

$$G_i = \lambda_G T_i + \delta_G + (\lambda_G)^{1/2}E_{Gi}, \qquad H_i = \lambda_H T_i + \delta_H + (\lambda_H)^{1/2}E_{Hi},$$
$$V_i = \lambda_V T_i + \delta_V + (\lambda_V)^{1/2}E_{Vi},$$

con E_{Gi}, E_{Hi}, E_{Vi} incorrelados entre sí y con los verdaderos puntajes T_i. Bajo este supuesto, y usando el ancla para "proyectar" momentos entre grupos, Woodruff (1986) muestra que los momentos sintéticos de X e Y pueden reescribirse en términos de momentos observables de (X_1, V_1) y (Y_2, V_2) más dos razones γ_1, γ_2, que dependen de si el ancla es interna (sus ítems contribuyen al total) o externa (test separado). En

nuestro caso, tenemos un anclaje interno $(X_i = G_i + V_i,\ Y_i = H_i + V_i)$ quedando las fórmulas,

$$\mu(X_s) = \mu(X_1) - \gamma_1\,[\,\mu(V_1) - \mu(V_s)\,], \qquad \mu(Y_s) = \mu(Y_2) + \gamma_2\,[\,\mu(V_2) - \mu(V_s)\,],$$
$$\sigma^2(X_s) = \sigma^2(X_1) - \gamma_1^2\,[\,\sigma^2(V_1) - \sigma^2(V_s)\,], \qquad \sigma^2(Y_s) = \sigma^2(Y_2) + \gamma_2^2\,[\,\sigma^2(V_2) - \sigma^2(V_s)\,],$$

con

$$\gamma_1 = \frac{\sigma^2(X_1)}{\sigma(X_1, V_1)}, \qquad \gamma_2 = \frac{\sigma^2(Y_2)}{\sigma(Y_2, V_2)}.$$

Si el ancla es externa $(X_i = G_i,\ Y_i = H_i)$, las expresiones de γ_1 y γ_2 cambian a funciones de varianzas y covarianzas que ajustan la exclusión del ancla del total; Hanson et al. ofrecen las formas cerradas para ambos casos. Sustituyendo (14)–(17) en (3)–(6) se obtienen las expresiones finales de los momentos sintéticos usados en (2); por ejemplo,

$$\mu(X_s) = \mu(X_1) - w_2\,\gamma_1\,[\,\mu(V_1) - \mu(V_s)\,], \quad \mu(Y_s) = \mu(Y_2) + w_1\,\gamma_2\,[\,\mu(V_2) - \mu(V_s)\,],$$
$$\sigma^2(X_s) = \sigma^2(X_1) - w_2\,\gamma_1^2\,[\,\sigma^2(V_1) - \sigma^2(V_s)\,] + w_1 w_2\,\gamma_1^2\,[\,\mu(V_1) - \mu(V_s)\,]^2,$$
$$\sigma^2(Y_s) = \sigma^2(Y_2) + w_1\,\gamma_2^2\,[\,\sigma^2(V_2) - \sigma^2(V_s)\,] + w_1 w_2\,\gamma_2^2\,[\,\mu(V_2) - \mu(V_s)\,]^2.$$

Las medias y varianzas del ancla en la población sintética se obtienen por analogía con (3)–(6): $\mu(V_s) = w_1\mu(V_1) + w_2\mu(V_2)$ y $\sigma^2(V_s) = w_1\sigma^2(V_1) + w_2\sigma^2(V_2) + w_1 w_2[\,\mu(V_1) - \mu(V_2)\,]^2$. Con estos momentos, el par (α, β) de la función de Levine se escribe de manera explícita como

$$\beta = \frac{\sigma(Y_s)}{\sigma(X_s)}, \qquad \alpha = \mu(Y_s) - \beta\,\mu(X_s),$$

y el corte transportado desde el puntaje observado c_X de X se obtiene como

$$c_Y = l(c_X) = \alpha + \beta\,c_X.$$

El método de Levine (observed-score) es lineal y garantiza equidad de primer orden: tras la equiparación,

$$E\{\,l(X) \mid \theta\,\} = E\{\,Y \mid \theta\,\}$$

para todo θ en la población de interés, bajo el modelo congénere y la construcción de población sintética. La precisión de $l(x)$ puede cuantificarse con errores estándar de equiparación vía el método delta; sea $l(x)$ función de diez momentos muestrales (medias, varianzas y covarianzas de X_1, Y_2, V_1, V_2 y sus combinaciones en la mezcla), entonces

$$\mathrm{Var}\{\,\hat{l}(x)\,\} \;\approx\; \nabla l(\boldsymbol{\theta})^{\top}\,\mathrm{Var}\big(\widehat{\boldsymbol{\theta}}\big)\,\nabla l(\boldsymbol{\theta}),$$

donde $\boldsymbol{\theta}$ apila esos momentos y $\nabla l(\boldsymbol{\theta})$ recoge las derivadas parciales de l respecto de cada uno; las expresiones cerradas de las derivadas y de $\mathrm{Var}\big(\widehat{\boldsymbol{\theta}}\big)$ se encuentran tabuladas para el caso NEAT con ancla interna o externa.

En términos operativos, los parámetros $\mu(\cdot)$, $\sigma^2(\cdot)$, $\sigma(\cdot,\cdot)$ se estiman con sus contrapartes muestrales en cada grupo y luego se construyen $\mu(\cdot)_s$ y $\sigma^2(\cdot)_s$ por mezcla. El peso w_g suele fijarse a proporciones muestrales o a pesos poblacionales conocidos. Con ancla interna se emplean $\gamma_1 = \sigma^2(X_1)/\sigma(X_1, V_1)$ y $\gamma_2 = \sigma^2(Y_2)/\sigma(Y_2, V_2)$; con ancla externa se usan las fórmulas alternativas provistas en la literatura de NEAT.

El procedimiento está implementado en el paquete equate de R como “linear / method = ‘levine’”, que además ofrece presuavizado, encadenamiento y bootstrap para los errores estándar de equiparación; véase la documentación técnica del paquete para detalles de estimación y reporte.

Para situar el método en el contexto de la literatura, el "Levine observed-score" es uno de los tres métodos lineales clásicos en NEAT (junto con Tucker y el lineal encadenado). La formulación unificada elaborada en ETS muestra que todos son casos particulares de una misma familia lineal bajo diferentes funciones de método; diversas comparaciones analíticas y empíricas documentan su comportamiento e invariancia poblacional relativa.

Método lineal de Braun–Holland en diseño NEAT

El método lineal de Braun–Holland para el diseño NEAT formaliza la equiparación de puntajes observados construyendo una población sintética a partir de dos poblaciones no equivalentes y utilizando el test ancla como variable de post-estratificación. A diferencia de enfoques que corrigen los momentos mediante una relación global entre total y ancla, Braun–Holland define los momentos de la forma fuente y destino en la población sintética integrando, sobre la distribución del ancla en dicha población, las medias y las varianzas condicionales estimadas en los grupos donde cada forma fue administrada. Este planteo reduce la dependencia de supuestos paramétricos fuertes y hace explícito el rol del ancla como puente entre poblaciones.

Diseño NEAT y población sintética

Sea X la forma administrada al grupo P (año 2023), Y la forma administrada al grupo Q (año 2025) y V el puntaje del test ancla, observado en ambos grupos. Los grupos P y Q no son equivalentes y, por lo tanto, no comparten necesariamente la misma distribución de habilidad ni del ancla. Se define una población sintética T como mezcla convexa de P y Q, $T = w_P P + w_Q Q$ con $w_P, w_Q \geq 0$ y $w_P + w_Q = 1$. La distribución del ancla en T es $F_{V|T}(v) = w_P F_{V|P}(v) + w_Q F_{V|Q}(v)$, y su función de masa o densidad se denota $f_{V|T}(v)$. En el diseño NEAT únicamente se observan los pares (X, V) en P y (Y, V) en Q; los pares (X, V) en Q y (Y, V) en P son contrafactuales.

La idea central de Braun–Holland es definir los momentos de X e Y en la población sintética combinando, sobre la distribución del ancla en T, las medias y varianzas condicionales disponibles en los grupos donde cada forma se administró. Para el caso discreto, donde V toma valores $v \in \mathcal{V}$, se definen las funciones condicionales $m_X(v) = \mathbb{E}[X \mid V = v, P]$ y $s_X^2(v) = \mathrm{Var}(X \mid V = v, P)$, análogamente $m_Y(v) = \mathbb{E}[Y \mid V = v, Q]$ y $s_Y^2(v) = \mathrm{Var}(Y \mid V = v, Q]$. Los momentos sintéticos de primer y segundo orden se obtienen por las identidades de ley de la esperanza total y la varianza total evaluadas en T:

$$\mu_{X,T} = \mathbb{E}_T[X] = \sum_{v \in \mathcal{V}} m_X(v)\, f_{V|T}(v), \qquad \mu_{Y,T} = \mathbb{E}_T[Y] = \sum_{v \in \mathcal{V}} m_Y(v)\, f_{V|T}(v),$$

$$\sigma_{X,T}^2 = \mathrm{Var}_T(X) = \sum_{v \in \mathcal{V}} s_X^2(v)\, f_{V|T}(v) + \sum_{v \in \mathcal{V}} (m_X(v) - \mu_{X,T})^2 f_{V|T}(v),$$

$$\sigma_{Y,T}^2 = \mathrm{Var}_T(Y) = \sum_{v \in \mathcal{V}} s_Y^2(v)\, f_{V|T}(v) + \sum_{v \in \mathcal{V}} (m_Y(v) - \mu_{Y,T})^2 f_{V|T}(v).$$

Cuando V es continuo, las sumas se reemplazan por integrales con respecto a la densidad $f_{V|T}(v)$. Estas expresiones especifican que los momentos de X e Y en la población objetivo T resultan de post-estratificar por V utilizando en cada estrato del

ancla la información condicional del grupo donde el test correspondiente fue efectivamente administrado. La dependencia de w_P, w_Q entra únicamente a través de $f_{V|T}(v)$, de modo que cambiar los pesos sintéticos modifica la mezcla de estratos del ancla y, por ende, los momentos resultantes.

Función lineal de equiparación

Con $\mu_{X,T}, \mu_{Y,T}, \sigma_{X,T}, \sigma_{Y,T}$ definidos como arriba, la equiparación lineal en la población sintética se expresa en su forma de momentos como

$$e_{X\to Y}(x) \;=\; \mu_{Y,T} + \frac{\sigma_{Y,T}}{\sigma_{X,T}}\big(x - \mu_{X,T}\big) \;=\; \alpha + \beta x,$$

con

$$\beta = \frac{\sigma_{Y,T}}{\sigma_{X,T}}, \qquad \alpha = \mu_{Y,T} - \beta\,\mu_{X,T}.$$

Esta transformación garantiza coincidencia de medias y desviaciones estándar de X transformado y Y en la población sintética T. En aplicaciones con puntos de corte definidos en la forma X, el corte equiparado en Y se obtiene como $c_Y = \alpha + \beta c_X$.

En la práctica, $m_X(v)$ y $s_X^2(v)$ se estiman a partir de las observaciones de (X, V) en P y $m_Y(v), s_Y^2(v)$ a partir de (Y, V) en Q. Para V discreto se utilizan medias y varianzas condicionales por puntaje del ancla; cuando V es continuo o el soporte discreto es extenso, se emplean discretizaciones o suavizados (por ejemplo, agrupando en bandas de V o ajustando modelos suavizados de $m_X(v)$ y $m_Y(v)$). La distribución del ancla en la población sintética se estima como mezcla empírica $\hat{f}_{V|T}(v) = w_P \hat{f}_{V|P}(v) + w_Q \hat{f}_{V|Q}(v)$, con $\hat{f}_{V|P}, \hat{f}_{V|Q}$ obtenidas de las frecuencias relativas del ancla en cada grupo. Sustituyendo estos estimadores en las fórmulas anteriores se obtienen $\hat{\mu}_{X,T}, \hat{\mu}_{Y,T}, \hat{\sigma}^2_{X,T}, \hat{\sigma}^2_{Y,T}$, y por ende $\hat{\alpha}, \hat{\beta}$.

El enfoque de Braun–Holland es lineal en el sentido de la transformación $e_{X\to Y}$, pero su construcción de momentos se apoya en una descomposición condicional en función del ancla. Esto lo diferencia de métodos lineales que corrigen momentos mediante relaciones globales total–ancla, ya que aquí la corrección se realiza estrato por estrato a través de V y luego se integra con la mezcla sintética. Si el ancla es representativo del constructo total y discrimina homogéneamente en el rango de la escala, la post-estratificación por V captura de forma transparente las diferencias de composición entre P y Q. En muestras pequeñas, la estabilidad de las estimaciones condicionales puede requerir suavizado o agregación de estratos; en presencia de múltiples ítems de anclaje, la estimación conjunta sobre $V = (V_1, \ldots, V_k)$ puede mejorar la precisión a costa de una mayor complejidad en la tabulación.

Los errores estándar de equiparación pueden obtenerse por método delta propagando la varianza de las estadísticas condicionales e incondicionales a $(\hat{\alpha}, \hat{\beta})$ o mediante remuestreo bootstrap replicando la construcción de tablas y momentos. En el reporte técnico se deben explicitar los pesos sintéticos w_P, w_Q, la definición y el tamaño del ancla, el esquema de estratificación o suavizado aplicado a V, y la consistencia de los resultados en análisis de sensibilidad frente a variaciones razonables de w_P, w_Q y de la granularidad de V.

Método de equiparación encadenado equipercentil (NEAT-CE)

El método de equiparación encadenado equipercentil (NEAT-CE, Nonequivalent groups with Anchor Test–Chain Equipercentile) es una extensión del diseño NEAT que busca construir una transformación no lineal entre las formas X y Y, utilizando un test de anclaje V como puente. A diferencia del método lineal de Levine, el encadenado equipercentil no supone linealidad ni varianzas constantes, sino que preserva las funciones de distribución acumulada (FDA) de los puntajes observados en cada forma.

En el diseño NEAT, tenemos dos grupos no equivalentes:
Grupo 1: toma la forma $X = G + V$, de donde se observan X_1 y V_1.
Grupo 2: toma la forma $Y = H + V$, de donde se observan Y_2 y V_2.

Como en todo NEAT, los componentes exclusivos G y H nunca se observan simultáneamente, y se utiliza el ancla V para conectar ambos grupos.

La equiparación equipercentil entre dos distribuciones A y B se define como la transformación $e_{A\to B}(\cdot)$ tal que los percentiles acumulados se conservan:

$$e_{A\to B}(x) = F_B^{-1}\big(F_A(x)\big),$$

donde:

- $F_A(x) = P(A \le x)$ es la función de distribución acumulada de A.
- $F_B^{-1}(p) = \inf\{y: F_B(y) \ge p\}$ es la cuantil correspondiente en B.

En palabras: se busca el puntaje en B cuya posición relativa en la distribución sea la misma que la del puntaje x en A.

Dado que los grupos son no equivalentes, no podemos aplicar equipercentil directamente entre X_1 y Y_2. En su lugar, se utilizan las distribuciones del ancla V como puente:

- **Primer paso:** equiparar la forma X con el ancla en el grupo 1:
$$e_{X\to V_1}(x) = F_{V_1}^{-1}\Big(F_{X_1}(x)\Big).$$
- **Segundo paso:** equiparar el ancla entre grupos:
$$e_{V_1\to V_2}(v) = F_{V_2}^{-1}\Big(F_{V_1}(v)\Big).$$
- **Tercer paso:** equiparar el ancla con la forma Y en el grupo 2:
$$e_{V_2\to Y}(v) = F_{Y_2}^{-1}\Big(F_{V_2}(v)\Big).$$

Finalmente, la equiparación encadenada NEAT-CE se obtiene como la composición de estas tres transformaciones:

$$e_{X\to Y}^{\text{NEAT-CE}}(x) = e_{V_2\to Y}\Big(e_{V_1\to V_2}\Big(e_{X\to V_1}(x)\Big)\Big).$$

A diferencia de Levine, esta transformación es no lineal y sigue la forma de las distribuciones empíricas. También garantiza que un examinando en el percentil p de X quede en el percentil p de Y, tras la equiparación.

El método tiene dependencia del test de anclaje dado que la calidad del ancla (su representatividad y discriminación) es crucial para la estabilidad de la equiparación. Con muestras pequeñas, las FDA empíricas pueden producir curvas con saltos o irregularidades; por eso se recomienda realizar un suavizado (por ejemplo, modelos log-lineales o kernel).

Si el punto de corte en la forma X es c_X, el punto equivalente en Y bajo NEAT-CE es:

$$c_Y = e_{X\to Y}^{\text{NEAT-CE}}(c_X).$$

En aplicaciones reales:

- Se construyen las distribuciones de frecuencias de X_1, Y_2, V_1, V_2.
- Se suavizan (opcional, por ejemplo, ajuste log-lineal).
- Se aplican las tres equiparaciones parciales.
- Se compone la función encadenada.
- Se estima el corte c_Y aplicando la transformación al corte c_X.

Método de Tucker

El método de Tucker es una variante lineal del NEAT, cuyo propósito es estimar una función de equiparación entre dos formularios de prueba, denotados como X y Y. A diferencia de Levine, que incorpora explícitamente la covarianza con la prueba de anclaje, Tucker utiliza un modelo lineal de ajuste en el cual los parámetros de la distribución de los puntajes en la población sintética se obtienen como combinaciones ponderadas de las distribuciones observadas en los grupos de referencia. El supuesto esencial es que las poblaciones de los dos grupos de examinados pueden considerarse mezclas de una población sintética común.

Sea X el puntaje total en la forma X aplicado al grupo P, y Y el puntaje total en la forma Y aplicado al grupo Q. Sea V la puntuación en la prueba de anclaje común a ambos grupos. Se definen como μ_{XP}, σ_{XP}^2 la media y varianza de X en el grupo P; μ_{YQ}, σ_{YQ}^2 las de Y en Q; y $\mu_{VP}, \mu_{VQ}, \sigma_{VP}^2, \sigma_{VQ}^2$ los momentos de la prueba de anclaje en cada grupo. El peso w_P representa la proporción del grupo P en la población sintética, con $w_Q = 1 - w_P$.

Los parámetros sintéticos en el método Tucker se calculan como combinaciones ponderadas de las medias y varianzas ajustadas, en la forma:

$$\mu_X = w_P\mu_{XP} + w_Q\left(\mu_{XP} + \rho_{XV,P}\,\sigma_{XP}/\sigma_{VP}\,(\mu_{VQ} - \mu_{VP})\right),$$
$$\mu_Y = w_Q\mu_{YQ} + w_P\left(\mu_{YQ} + \rho_{YV,Q}\,\sigma_{YQ}/\sigma_{VQ}\,(\mu_{VP} - \mu_{VQ})\right).$$

Aquí, $\rho_{XV,P}$ es la correlación entre X y V en el grupo P, y $\rho_{YV,Q}$ es la correlación entre Y y V en el grupo Q. Estas expresiones ajustan las medias de X y Y al desplazar las diferencias en el anclaje entre grupos.

Las varianzas sintéticas se obtienen como:

$$\sigma_X^2 = w_P\sigma_{XP}^2 + w_Q\sigma_{XP}^2\left(1 - \rho_{XV,P}^2\right) + w_Pw_Q\left(\rho_{XV,P}\,\sigma_{XP}/\sigma_{VP}\,(\mu_{VQ} - \mu_{VP})\right)^2,$$
$$\sigma_Y^2 = w_Q\sigma_{YQ}^2 + w_P\sigma_{YQ}^2\left(1 - \rho_{YV,Q}^2\right) + w_Pw_Q\left(\rho_{YV,Q}\,\sigma_{YQ}/\sigma_{VQ}\,(\mu_{VP} - \mu_{VQ})\right)^2.$$

La función de equiparación de Tucker es lineal y se expresa como:

$$e(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\,(x - \mu_X).$$

Este modelo implica que los puntajes en la forma X se transforman en puntajes equivalentes en la forma Y mediante un ajuste lineal basado en medias y varianzas de la población sintética, corregidas por las diferencias en la prueba de anclaje. En comparación con el método de Levine, Tucker asume un ajuste menos dependiente de la covarianza directa con el anclaje, y tiende a ser más estable cuando el tamaño muestral es limitado y las correlaciones con la prueba de anclaje no son elevadas.

Punto de corte en θ bajo TRI con muestra pequeña

En las pruebas en las que se operacionalizan los puntos de corte mediante el puntaje total observado y las muestras de aplicación son pequeña no usaremos la TRI. Sin embargo, podemos, cuando la muestra lo permita, disponer también de un punto de corte en la escala de habilidad latente (θ), en nuestro caso, para informes internos y validación. No se usará esta técnica para calcular el punto de corte, solo para comparar con los propuestos en 2023 en puntaje logit y validar lo encontrado.

Lo adecuado en este caso es no utilizar una calibración libre completa en TRI, ya que con muestras reducidas la estimación de parámetros ítem puede ser inestable y producir resultados poco fiables. La alternativa práctica es una variante híbrida que aprovecha los parámetros ya estimados en una aplicación grande (2023) y limita la estimación en la aplicación pequeña (2025).

Se supone que existe un banco de ítems calibrados en un año de muestra grande (2023), del cual se dispone de parámetros (b_j) de modelo de Rasch en nuestro caso.

En el año de muestra pequeña (2025) se administra una forma que contiene: ítems de anclaje (compartidos con 2023, ya calibrados) e ítems nuevos (no calibrados).

El objetivo es transportar un punto de corte definido originalmente en puntaje observado 2023 (c_X) hacia la escala de habilidad latente, para verificar la correspondencia con el puntaje de 2025.

Utilizaremos como estrategia que los parámetros de los ítems comunes se mantienen fijos con los parámetros obtenidos en la calibración 2023. Solo se estiman los parámetros de los ítems exclusivos de 2025.

La escala queda así anclada en la métrica de 2023, y se pueden estimar habilidades θ para los examinados de 2025.

Si el número de ítems nuevos es reducido, o si la prioridad es solo un chequeo de consistencia, puede optarse por no calibrar nada en 2025.

Se aplica la forma 2025, pero los puntajes se transforman a θ usando directamente los parámetros de 2023 para los ítems comunes.

Los ítems nuevos se pueden excluir o tratarlos como adicionales para la clasificación, pero no como base de la equiparación.

Procedimiento para obtener el punto de corte en θ

Sea Y la forma objetivo (por ejemplo, año 2025) cuyo corte operativo en puntaje observado es c_Y. Sean b_i los parámetros de dificultad del ítem i de Y expresados en una única escala, en la de referencia 2023. La probabilidad de acierto en el ítem i para un examinado con habilidad θ se define como

$$P_i(\theta) = \frac{1}{1 + exp^{-(\theta - b_i)}}$$

La curva característica del test (TCC) es la suma de probabilidades (o puntuación esperada) a través de los ítems del formulario:

$$\mathrm{TCC}_Y(\theta) = \sum_{i \in Y} P_i(\theta).$$

El punto de corte en θ se define como la solución θ^* de la ecuación

$$\mathrm{TCC}_Y(\theta^*) = c_Y.$$

Esta ecuación se resuelve numéricamente porque TCC_Y no tiene generalmente inversa analítica.

Primeramente se deben expresar los parámetros de los ítems de Y en la misma escala que los de referencia (se ancla a 2023). Para los ítems de 2025 que no son de anclaje se estiman con los parámetros fijos de los ítems de anclaje. Si bien tenemos una muestra pequeña, como hay numerosos ítems de anclaje, lo haremos así. Evitaremos usar los parámetros de diseño pues darán mayor error.

Se usará el como punto de corte operativo el obtenido mediante equiparación de Levine c_Y . Se resuelve $\mathrm{TCC}_Y(\theta) = c_Y$ con métodos numéricos obteniendo θ^*, el punto de corte en habilidad latente en escala 2023 coherente con el corte operativo en 2025 en puntaje total.

Este método aprovecha la estabilidad de los parámetros estimados en la aplicación grande (2023), evitando calibraciones inestables en 2025.

El punto de corte en θ no se usará para toma de decisiones oficiales, pues la muestra 2025 es pequeña, sino solo como validación convergente o para documentos internos.

El punto de corte oficial sigue siendo el observado (*score*) obtenido mediante los métodos robustos en muestras pequeñas.

Métodos basados en TRI

Calibración separada y enlace posterior en TRI

La equiparación de formas de prueba bajo la TRI puede abordarse mediante transformaciones lineales de escala que alinean momentos de distribuciones relevantes. El objetivo es vincular dos métricas latentes, θ^X y θ^Y, a través de una transformación afín $\theta^X = A\,\theta^Y + B$, con $A > 0$ y $B \in \mathbb{R}$. Esta relación induce transformaciones coherentes sobre los parámetros de ítem y permite transportar estándares de desempeño entre formas.

Considérense dos formas de prueba, X e Y, calibradas por separado bajo un modelo logístico. Sea θ^X la habilidad en la métrica de X y θ^Y la habilidad en la métrica de Y. El enlace busca constantes (A, B), con $A > 0$, tales que

$$\theta^X = A\,\theta^Y + B.$$

En modelos dicotómicos 1PL/2PL/3PL, la transformación inducida sobre los parámetros de Y para llevarlos a la métrica de X es

$$a_i^{Y\to X} = \frac{a_i^Y}{A}, \qquad b_i^{Y\to X} = A\,b_i^Y + B, \qquad c_i^{Y\to X} = c_i^Y.$$

En modelos politómicos (por ejemplo, GRM, PCM/GPCM), las discriminaciones se dividen por A y los umbrales o pasos se trasladan como $A(\cdot) + B$.

Los métodos de Media–Media y Media–Sigma (también denominado Media–Varianza cuando se trabaja con varianzas) son procedimientos de estimación de A y B que se basan en igualar primer y segundo momento de una distribución de referencia, ya sea de habilidades de personas comunes o de dificultades de ítems de anclaje.

La estimación por momentos es transparente y poco exigente computacionalmente, pero su precisión depende de la calidad y representatividad del conjunto utilizado para los momentos. Cuando se usa el conjunto de ítems de anclaje, conviene que cubra el rango del rasgo y el dominio de contenidos; cuando se usan personas comunes, deben controlarse efectos de orden, práctica y fatiga, así como la representatividad del rango de habilidades. Versiones ponderadas que asignan pesos inversamente proporcionales a errores estándar de b_i o θ_j pueden reducir la influencia de estimaciones imprecisas. En presencia de discriminaciones heterogéneas o formatos politómicos complejos, los métodos por momentos suelen ser útiles como punto de partida o como chequeo de sensibilidad frente a métodos de enlace basados en curvas (Stocking–Lord, Haebara). En MIRT, la generalización es una transformación afín $\boldsymbol{\theta}' = \mathbf{T}\,\boldsymbol{\theta} + \mathbf{c}$, donde $\mathbf{T}$ puede estimarse haciendo coincidir medias y covarianzas de $\boldsymbol{\theta}$ o de parámetros estructurales seleccionados; la transformación de cortes definidos sobre combinaciones lineales de dimensiones sigue de forma directa.

El enlace por ítems comunes en el marco del modelo de Rasch ofrece una vía parsimoniosa y estable para colocar múltiples formas en una métrica común siempre que los ítems de anclaje conserven su comportamiento psicométrico entre aplicaciones. Los métodos por momentos y de tipo Stocking–Lord conducen a

transformaciones lineales que, aplicadas a los parámetros de dificultad y a los valores de habilidad, permiten transportar criterios de decisión como puntos de corte con garantías de coherencia en la interpretación. En aplicaciones de alto impacto conviene acompañar el resultado con estudios de sensibilidad frente a la selección de ítems de anclaje, diagnósticos de estabilidad de parámetros y estimaciones de incertidumbre mediante procedimientos de remuestreo o simulación paramétrica.

Considérense dos formas de prueba, X (origen en 2023) y Y (destino, 2025), con un conjunto de ítems de anclaje administrados en ambas que siguen un modelo de Rasch. Denótese por $\{b_i^X\}$ y $\{b_i^Y\}$ los parámetros de dificultad de los ítems de anclaje estimados separadamente en cada forma. Si los ítems de anclaje son estables (sin deriva ni funcionamiento diferencial relevante), cualquier diferencia sistemática entre $\{b_i^X\}$ y $\{b_i^Y\}$ se explica por un cambio de escala latente que puede representarse mediante una transformación lineal

$$\theta^X = A\,\theta^Y + B, \qquad A > 0,\ B \in \mathbb{R}.$$

La transformación inducida sobre los parámetros de ítem satisface

$$b_i^{Y\to X} = A\,b_i^Y + B\,.$$

El problema de enlace consiste en estimar (A, B) a partir de los ítems de anclaje. En Rasch, al existir un único parámetro por ítem, el enlace por parámetros comunes resulta particularmente transparente y puede fijarse por criterios de momentos o por criterios basados en la proximidad de las curvas características.

Las constantes (A, B) se eligen minimizando una discrepancia entre calibraciones separadas, ya sea al nivel de la curva característica de la prueba (Stocking–Lord) o al nivel de curvas/características de ítem (Haebara).

Estimación por momentos basada en ítems de anclaje

Cuando dos formas comparten un conjunto de ítems de anclaje I_A calibrados separadamente, las diferencias entre sus dificultades estimadas $\{b_i^X\}_{i\in I_A}$ y $\{b_i^Y\}_{i\in I_A}$ pueden atribuirse, bajo invariancia, a un cambio lineal de escala. Denótese por $\bar{b}^X$ y s_b^X la media y la desviación estándar muestral de las dificultades ancla en la forma X, y análogamente $\bar{b}^Y$ y s_b^Y en Y. El método Media–Sigma determina A y B imponiendo

$$\mathbb{E}(b^{Y\to X}) = \bar{b}^X, \qquad \mathrm{SD}(b^{Y\to X}) = s_b^X,$$

lo que conduce a

$$A = \frac{s_b^X}{s_b^Y}, \qquad B = \bar{b}^X - A\,\bar{b}^Y.$$

Cuando solamente se desea alinear niveles, el método Media–Media fija $A = 1$ y establece

$$B = \bar{b}^X - \bar{b}^Y.$$

Estas expresiones son válidas en Rasch, donde solo hay dificultades, y se extienden a 2PL/3PL si la comparación se restringe a las dificultades de los ítems de anclaje; en presencia de discriminaciones heterogéneas, el criterio Media–Sigma ofrece una

aproximación simple, pero puede ser subóptimo respecto de métodos que alinean curvas características.

En un ajuste multi-grupo, si se fija un grupo de referencia con $\theta \sim \mathcal{N}(0,1)$ y se estiman (μ_g, σ_g) para otros grupos, la relación $\theta^{(\text{ref})} = (\theta^{(g)} - \mu_g)/\sigma_g$ es formalmente equivalente a una transformación lineal con $A = 1/\sigma_g$ y $B = -\mu_g/\sigma_g$. Por tanto, los enlaces Media–Sigma y Media–Media pueden interpretarse como casos particulares de una identificación de escala basada en momentos cuando la calibración es concurrente.

Transformación de puntos de corte

Sea θ_c un punto de corte definido en la escala latente de la forma X. Tras estimar A y B por Media–Sigma o Media–Media, el corte en la métrica de Y se obtiene como $\theta'_c = A\,\theta_c + B$. Si el estándar está definido como una puntuación observada s_c en X, se determina primero θ_c resolviendo $S_X(\theta_c) = s_c$, donde la curva característica de la prueba es $S_X(\theta) = \sum_i \sum_k w_{ik} P_{ik}(\theta)$. El punto de corte equivalente en la forma Y se calcula evaluando su TCC en θ'_c: $s_c' = S_Y(\theta'_c)$. La monotonicidad de S en modelos bien comportados garantiza unicidad de la solución para θ_c.

Método de Stocking–Lord

El criterio de Stocking–Lord busca que, tras transformar la escala de Y a X, las puntuaciones verdaderas esperadas por θ sean, en promedio y a lo largo de la escala, las más cercanas posibles. En el caso de Rasch se plantea el problema de optimización

$$(A^*, B^*) \;=\; \arg\min_{A>0,B} \int \{S_X(\theta) - S_Y(A\theta + B)\}^2\, w(\theta)\, d\theta,$$

donde S_X y S_Y son las curvas características de las pruebas X e Y respectivamente, y $w(\theta)$ es una función de ponderación que suele elegirse como la densidad normal estándar. En la práctica la integral se aproxima por cuadratura o por una suma discreta sobre una rejilla $\{\theta_t\}$ con pesos $\{w_t\}$. En Rasch, al coincidir las discriminaciones, la solución típicamente resulta muy próxima a la solución por momentos cuando los ítems de anclaje cubren adecuadamente el rango de dificultad, aunque el criterio de Stocking–Lord es preferible cuando interesa alinear la función de puntuación verdadera en toda la escala.

Método de Haebara

El método de Haebara alinea, en promedio, las probabilidades de respuesta (o puntuaciones esperadas por ítem) entre calibraciones. En el caso dicotómico, con $P_{Xi}(\theta)$ y $P_{Yi}(\theta)$ las probabilidades de acierto del ítem i en cada métrica, las constantes se obtienen de

$$(A^*, B^*) = \arg\min_{A>0,B} \sum_{i \in I_A} \int\, [P_{Xi}(\theta) - P_{Yi}(A\theta + B)]^2\, w(\theta)\, d\theta,$$

donde I_A es el conjunto de ítems de anclaje utilizados para el enlace y $w(\theta)$ es, nuevamente, una densidad de referencia. En modelos politómicos se reemplaza $P_{Xi}(\theta)$ por la puntuación esperada por ítem o por el vector de probabilidades por categoría con una métrica de discrepancia adecuada.

Elección de la población de referencia y ponderaciones

La función $w(\theta)$ determina la población de referencia del enlace. Elegir w proporcional a la densidad de θ en el grupo de referencia concentra el ajuste donde hay mayor densidad empírica; elegir w normal estándar estabiliza cuando se busca una referencia abstrayéndose de una muestra concreta. Cualquiera sea la elección, debe reportarse explícitamente, ya que afecta los valores de (A, B) y, por ende, la proyección de puntos de corte.

Transformación del punto de corte entre pruebas

Sea θ_c un punto de corte definido en la escala latente de la prueba origen X. Una vez estimados A y B, el punto de corte en la escala latente de la prueba destino se obtiene como $\theta_c' = A\,\theta_c + B$. Si el punto de corte está fijado originalmente en unidades de puntuación observada s_c en X, se determina primero el valor latente que lo induce resolviendo $S_X(\theta_c) = s_c$. Dado que S_X es estrictamente creciente, el valor θ_c se obtiene por inversión numérica con métodos de búsqueda de raíces. El punto de corte equivalente en puntuación observada de la prueba destino resulta al evaluar la curva característica S_Y en θ_c', es decir $s_c' = S_Y(\theta_c')$. Cuando se requieren puntos de corte enteros sobre número-correcto, se aplica una regla explícita de redondeo o de corrección por continuidad coherente con la normativa de la prueba.

En el modelo de Rasch con parametrización logística $P_i(\theta) = \{1 + \exp[-D(\theta - b_i)]\}^{-1}$, con $D = 1{,}7$, la curva característica de la prueba es $S(\theta) = \sum_i P_i\,(\theta)$. En este caso, el método de Stocking–Lord se implementa minimizando la discrepancia entre $S_X(\theta)$ y $S_Y(A\theta + B)$, mientras que Haebara se formula a nivel de ítem mediante las probabilidades $P_{Xi}(\theta)$ y $P_{Yi}(A\theta + B)$.

Anclaje con ítems calibrados previamente en TRI

El anclaje con ítems calibrados previamente es un mecanismo de equiparación que fija desde el inicio la métrica latente de una nueva forma de prueba mediante un subconjunto de ítems cuyos parámetros han sido estimados con anterioridad y se asumen estables. La estrategia consiste en mantener constantes dichos parámetros de ancla durante la calibración de la nueva forma, de modo que tanto los parámetros de los ítems no ancla como las habilidades de los examinados queden expresados en la misma escala de referencia. Esta nota expone la notación general, la formulación de verosimilitud bajo la TRI, los casos particulares más usados (Rasch, 2PL/3PL y modelos politómicos), la variante bayesiana de anclaje blando y el procedimiento para trasladar puntos de corte entre formas usando la métrica anclada.

Modelo general

Considérese un conjunto de examinados $j = 1, \dots, N$ y un conjunto de ítems $i = 1, \dots, n$, con $A \subset \{1, \dots, n\}$ denotando los ítems de anclaje y $U = \{1, \dots, n\} \backslash A$ los ítems no ancla. Sea θ_j la habilidad latente del examinado j, con densidad a priori $f(\theta_j \mid \eta)$ dependiente de

hiperparámetros η que, salvo indicación, se toman como $\mathcal{N}(0,1)$. La probabilidad de respuesta se modela como $P(x_{ij} \mid \theta_j; \psi_i)$ donde ψ_i recoge los parámetros del ítem según el modelo elegido.

En el caso dicotómico 3PL con constante de escala D se tiene

$$P(X_{ij} = 1 \mid \theta_j; \psi_i) \;=\; c_i + (1 - c_i)\,\frac{1}{1 + \exp\{-D\,a_i(\theta_j - b_i)\}},$$

con $\psi_i = (a_i, b_i, c_i)$. En 2PL se fija $c_i = 0$ y en Rasch se fija además $a_i \equiv 1$. Para modelos politómicos, como el Modelo de Respuesta Graduada (GRM) o el Modelo de Crédito Parcial (PCM/GPCM), la verosimilitud se expresa en términos de probabilidades por categoría $P_{ik}(\theta_j)$ y puntuaciones w_{ik}, manteniendo la estructura general de estimación marginal sobre θ.

Verosimilitud con anclaje "duro"

El anclaje "duro" fija los parámetros de los ítems de anclaje en sus valores de referencia $\psi_i^\star$ y estima únicamente los parámetros $\{\psi_i\}_{i\in U}$ y, si corresponde, los hiperparámetros η. La verosimilitud marginal de los datos x_{ij} puede escribirse como

$$\mathcal{L}(\{\psi_i\}_{i\in U}, \eta) \;=\; \prod_{j=1}^{N} \int \left[\prod_{i\in A} P\left(x_{ij} \mid \theta_j; \psi_i^\star\right)\right] \left[\prod_{i\in U} P\left(x_{ij} \mid \theta_j; \psi_i\right)\right] f\left(\theta_j \mid \eta\right) d\theta_j.$$

Al maximizar $\mathcal{L}$ con respecto a $\{\psi_i\}_{i\in U}$ y η, la escala latente queda identificada por los $\psi_i^\star$, puesto que cualquier reparametrización lineal $\theta' = A\theta + B$ alteraría las probabilidades de los ítems de anclaje, lo que está prohibido por construcción. En consecuencia, los parámetros estimados para los ítems no ancla y las habilidades a posteriori de los examinados quedan automáticamente en la métrica de referencia.

En el esquema de estimación por máxima verosimilitud marginal tipo EM de Bock–Aitkin, el paso E utiliza cuadratura o aproximaciones de Gauss–Hermite para calcular las expectativas suficientes condicionadas a $\psi^\star$ y a los ψ actuales. El paso M actualiza exclusivamente los ψ_i de U y, si se modela, los hiperparámetros η, mientras que los términos correspondientes a $i \in A$ actúan como constantes. En Rasch con estimación condicional, el anclaje se implementa fijando las dificultades $b_i = b_i^\star$ en los sufijos del estimador; la identificación de escala proviene del conjunto ancla y no requiere restricciones adicionales tales como $\sum_i b_i = 0$.

Bajo anclaje, la relación de enlace entre una forma de referencia y la nueva forma es trivial. Si θ denota la habilidad en la escala de referencia, la nueva forma queda definida en la misma θ, sin necesidad de estimar constantes (A, B). En modelos dicotómicos con parámetros $\psi_i = (a_i, b_i, c_i)$, los ítems no ancla estimados satisfacen las mismas reglas de transformación que bajo un enlace lineal, pero con $A = 1$ y $B = 0$ por construcción; por tanto, la comparación entre formas es directa en la escala de referencia.

Anclaje “blando” bayesiano

Cuando se desea absorber pequeñas variaciones sin abandonar la métrica de referencia, se emplea un anclaje blando mediante *priors* informativos centrados en $\psi_i^\star$. Para cada ítem ancla se especifica

$$\psi_i \sim \mathcal{N}(\psi_i^\star, \Sigma_i), \qquad i \in A,$$

con Σ_i de varianza pequeña. La a posteriori resultante penaliza desviaciones respecto de la calibración de referencia sin impedirlas por completo, lo que puede mejorar el ajuste cuando existen microcambios no sustantivos en los ítems de anclaje. En Rasch esto equivale a $b_i \sim \mathcal{N}(b_i^\star, \sigma_i^2)$ con σ_i^2 pequeña. El límite $\Sigma_i \to \mathbf{0}$ recupera el anclaje duro.

La validez del anclaje depende de la estabilidad psicométrica de los ítems de anclaje. Antes de fijarlos, es recomendable verificar ausencia de deriva temporal y funcionamiento diferencial relevante entre poblaciones de referencia y de aplicación, por medio de análisis de invariancia y de DIF a nivel de ítem y, cuando corresponde, a nivel de categorías. Una vez calibrada la nueva forma, resulta útil inspeccionar el ajuste de los ítems de anclaje, la concordancia entre curvas características inducidas por los ítems de anclaje y los no ancla y la sensibilidad del resultado al retirar ítems de anclaje sospechosos. El reporte técnico debería documentar el conjunto de ítems de anclaje empleado, la evidencia de estabilidad, la estrategia de anclaje (duro o blando y, en este caso, las matrices Σ_i) y, cuando el uso es de alto impacto, intervalos de incertidumbre para los puntos de corte transformados, derivados mediante remuestreo paramétrico que reestima parámetros y repite el procedimiento de inversión y evaluación de TCC.

Transformación de puntos de corte entre formas

Si el estándar está definido en la escala latente de referencia como θ_c, una nueva forma calibrada con anclaje duro o blando suficientemente informativo queda automáticamente en la misma métrica, por lo que el punto de corte latente en la nueva forma es también θ_c. Cuando el estándar se expresa en puntuación observada de una forma de referencia Y como s_c, se determina primero θ_c resolviendo

$$S_Y(\theta_c) = s_c,$$

donde $S_Y(\theta) = \sum_{i \in Y} \sum_k w_{ik} P_{ik}(\theta)$ es la curva característica de la prueba de referencia, con w_{ik} las ponderaciones de categoría. Dada la monotonicidad de S_Y, la ecuación admite una única solución que puede hallarse por bisección o Newton–Raphson. El punto de corte equivalente en la nueva forma X se obtiene evaluando su TCC en el mismo valor latente,

$$s_c' = S_X(\theta_c) = \sum_{i \in X} \sum_k w_{ik} P_{ik}(\theta_c).$$

En pruebas dicotómicas de número-correcto, si la reglamentación exige valores enteros, se adopta una regla explícita de redondeo o una corrección por continuidad coherente con la política de decisión.

Estimación

Esta sección detalla dos procedimientos de estimación para el modelo de Rasch cuando se emplea anclaje con ítems calibrados previamente: máxima verosimilitud marginal

con algoritmo EM y máxima verosimilitud condicional. En ambos casos se asume el anclaje duro para los ítems del conjunto A, por lo que sus dificultades se toman como b_i en toda la estimación, mientras que las dificultades b_i de los ítems en U se estiman en la misma métrica.

EM marginal tipo Bock–Aitkin con cuadratura

El algoritmo se apoya en una representación discreta de la distribución a priori $f(\theta) = \mathcal{N}(0,1)$ mediante una cuadratura $\{(\theta_t, w_t)\}_{t=1}^{T}$. Sea $x_{ij} \in \{0,1\}$ la respuesta del examinado j al ítem i. Con $P_i(\theta) = \{1 + \exp[-D(\theta - b_i)]\}^{-1}$ y $D = 1.7$, el ciclo EM se describe de la siguiente manera. En el paso E se calcula, para cada examinado j, la distribución a posteriori discreta γ_{jt} en los nodos θ_t,

$$\tilde{\gamma}_{jt} \;=\; w_t \prod_{i \in A} P_i^{\star}(\theta_t)^{x_{ij}} [1 - P_i^{\star}(\theta_t)]^{1 - x_{ij}} \prod_{i \in U} P_i(\theta_t)^{x_{ij}} [1 - P_i(\theta_t)]^{1 - x_{ij}}, \qquad \gamma_{jt} = \frac{\tilde{\gamma}_{jt}}{\sum_{s=1}^{T} \tilde{\gamma}_{js}}.$$

Se obtienen las cuentas esperadas de aciertos por nodo para cada ítem no ancla, $\hat{y}_{i,t} = \sum_{j=1}^{N} \gamma_{jt}\, x_{ij}$, y las masas efectivas por nodo $n_t = \sum_{j=1}^{N} \gamma_{jt}$. En el paso M, cada dificultad b_i para $i \in U$ se actualiza resolviendo la ecuación de balance de aciertos esperados

$$\sum_{t=1}^{T} \hat{y}_{i,t} \;=\; \sum_{t=1}^{T} n_t \; P_i(\theta_t; b_i).$$

La actualización se implementa por Newton–Raphson sobre la función $g_i(b) = \sum_t n_t \; P_i(\theta_t; b) - \sum_t \hat{y}_{i,t}$ con derivada

$$g_i{}'(b) \;=\; D \sum_{t=1}^{T} n_t \; P_i(\theta_t; b)[1 - P_i(\theta_t; b)].$$

El ciclo EM se itera hasta convergencia en log-verosimilitud o en el máximo cambio $|\Delta b_i|$. Las cantidades asociadas a los ítems de anclaje se mantienen fijas en $b_i^{\star}$, de manera que la métrica queda determinada por estos valores.

Máxima verosimilitud condicional de Andersen con anclaje

La verosimilitud condicional del modelo de Rasch elimina los parámetros de persona condicionando en los puntajes brutos $r_j = \sum_{i=1}^{n} x_{ij}$. Para $b = (b_i)_{i \in U}$ y con $b_i = b_i^{\star}$ para $i \in A$, la log-verosimilitud condicional se expresa como

$$\ell_c(b) \;=\; \sum_{j=1}^{N} \left[\sum_{i=1}^{n} x_{ij}\,(-D b_i) \;-\; \log Z_j(b) \right], \qquad Z_j(b) \;=\; \sum_{\mathbf{x} \in \mathcal{X}(r_j)} \exp\left(-D \sum_{i=1}^{n} x_i \; b_i \right),$$

donde $\mathcal{X}(r_j)$ denota el conjunto de vectores de respuestas con suma r_j. Los términos $\log Z_j(b)$ se calculan por programación dinámica mediante funciones simétricas elementales. Las ecuaciones de verosimilitud, para $i \in U$, toman la forma

$$U_i \;=\; \sum_{j=1}^{N} \mathbb{E}_c \left[X_{ij} \mid r_j, b \right], \qquad U_i = \sum_{j=1}^{N} x_{ij},$$

y se resuelven por Newton–Raphson utilizando derivadas segundas obtenibles del mismo esquema dinámico. El anclaje se implementa manteniendo fijos los $b_i^{\star}$ en los términos correspondientes, por lo que la identificación de escala proviene exclusivamente de los ítems de anclaje. Este procedimiento evita especificar $f(\theta)$ y es

robusto a su mala especificación, a costa de un mayor costo computacional en el cálculo de $\log Z_j(b)$ para grandes n.

Versión bayesiana: anclaje duro y anclaje blando

En un planteo bayesiano para el Rasch dicotómico con $D = 1.7$, el modelo completo con anclaje duro se define por

$$X_{ij} \mid \theta_j, b \sim \text{Bernoulli}\left(\frac{1}{1+\exp\{-D(\theta_j - b_i)\}}\right), \quad \theta_j \sim \mathcal{N}(0,1), \quad b_i = b_i^\star \text{ si } i \in A, \quad b_i \sim \mathcal{N}(\mu_b, \sigma_b^2) \text{ si } i \in U.$$

La distribución a posteriori es proporcional a

$$p(\theta, b_U \mid X, b_A^\star) \propto \left[\prod_{j=1}^{N}\prod_{i\in A} \Pr\left(X_{ij} \mid \theta_j, b_i^\star\right)\right]\left[\prod_{j=1}^{N}\prod_{i\in U} \Pr\left(X_{ij} \mid \theta_j, b_i\right)\right]\left[\prod_{j=1}^{N} \phi\left(\theta_j\right)\right]\left[\prod_{i\in U} \phi_{\mu_b,\sigma_b}\left(b_i\right)\right],$$

donde ϕ indica densidad normal. El anclaje blando introduce *priors* informativos para los ítems de anclaje,

$$b_i \sim \mathcal{N}(b_i^\star, \tau_i^2), \qquad i \in A,$$

con τ_i^2 pequeño. El límite $\tau_i^2 \to 0$ recupera el anclaje duro. La inferencia puede implementarse por HMC/NUTS en Stan o por Gibbs aproximado mediante aumentación Pólya–Gamma, que convierte las verosimilitudes logísticas en gaussianas condicionales y permite muestreos alternados de θ y b a partir de normales multivariadas. En todos los casos, la métrica resultante coincide con la de referencia por fijación o fuerte anclaje de los b_i en A.

En este caso en aplicaciones de alto impacto se puede cuantificar la varianza del punto de corte simulando a partir de la distribución a posteriori. Sea s_c el corte en puntuación observada de una forma de referencia Y. En cada muestra m de la a posteriori, se obtiene $\theta_c^{(m)}$ resolviendo $S_Y^{(m)}(\theta) = s_c$ y luego se evalúa $s_c'^{(m)} = S_X^{(m)}(\theta_c^{(m)})$. El resumen de $s_c'^{(m)}$ aporta intervalos creíbles que reflejan la incertidumbre en parámetros y, por extensión, del procedimiento de transformación.

Calibración concurrente para equiparación en TRI

La calibración concurrente es un enfoque de equiparación que coloca dos o más formas de prueba en una misma métrica estimando, en una sola corrida, todos los parámetros del modelo de TRI. La equiparación es "por construcción": los ítems de anclaje compartidos entre formas se tratan como el mismo ítem con idénticos parámetros y la escala latente se fija mediante restricciones de identificación globales. El resultado es una métrica común en la que los parámetros de los ítems exclusivos de cada forma y las habilidades de los examinados quedan directamente comparables, sin necesidad de calcular posteriormente constantes de enlace lineal.

La validez del enfoque descansa en la invariancia de los ítems de anclaje entre formas y administraciones. Resulta necesario verificar ausencia de funcionamiento diferencial relevante y estabilidad temporal de parámetros, controlar efectos de orden, práctica o fatiga cuando hay personas comunes y realizar diagnósticos posteriores de

ajuste, como discrepancias entre curvas características de prueba e ítem, errores estándar de (μ_g, σ_g) y análisis de sensibilidad al retirar ítems de anclaje potencialmente inestables. El reporte técnico debería documentar el diseño de enlace, el conjunto de ítems de anclaje, las restricciones de identificación, los valores estimados de (μ_g, σ_g) y evidencia empírica de la coherencia de la métrica resultante.

Considérense dos formas de prueba, X e Y, que miden el mismo rasgo latente. Sea A el conjunto de ítems de anclaje administrados en ambas formas, U_X los ítems exclusivos de X y U_Y los exclusivos de Y. Denótese por G_X y G_Y los grupos de examinados que responden cada forma (diseño NEAT), admitiéndose además la presencia de personas comunes. En un modelo logístico dicotómico de tres parámetros, la probabilidad de acierto es

$$P(X_{ij} = 1 \mid \theta_j; \psi_i) = c_i + (1 - c_i)[1 + \exp\{-D\, a_i(\theta_j - b_i)\}]^{-1},$$

donde $\psi_i = (a_i, b_i, c_i)$ y $D > 0$ es la constante de escala (usualmente $D = 1.7$). En 2PL se fija $c_i = 0$ y en Rasch se fija además $a_i \equiv 1$. En calibración concurrente, cada ítem ancla $i \in A$ posee un único vector de parámetros ψ_i que se comparte en ambas formas, mientras que los ítems exclusivos tienen parámetros propios pero estimados en la misma métrica. La verosimilitud marginal conjunta para un diseño NEAT se escribe como

$$\mathcal{L} = \left[\prod_{j \in G_X} \int \prod_{i \in X} P\left(x_{ij} \mid \theta_j; \psi_i\right) f_X(\theta_j)\, d\theta_j\right]\left[\prod_{j \in G_Y} \int \prod_{i \in Y} P\left(x_{ij} \mid \theta_j; \psi_i\right) f_Y(\theta_j)\, d\theta_j\right],$$

donde $f_g(\theta)$ representa la distribución latente en el grupo g. La equiparación surge porque los ítems de anclaje imponen una métrica común a los parámetros de ítem, mientras que las diferencias de nivel o dispersión entre poblaciones se capturan en f_X y f_Y.

Para identificar la escala se fija un grupo de referencia, por ejemplo X, con $\theta \mid g = X \sim \mathcal{N}(0,1)$ o restricciones equivalentes. En un ajuste multi-grupo, las medias y varianzas latentes de los otros grupos se estiman como $\theta \mid g \sim \mathcal{N}(\mu_g, \sigma_g^2)$. Estas cantidades inducen una transformación lineal entre métricas

$$\theta^{(\text{ref})} = \frac{\theta^{(g)} - \mu_g}{\sigma_g},$$

que es formalmente equivalente a un enlace con constantes $A = 1/\sigma_g$ y $B = -\mu_g/\sigma_g$. En calibración concurrente no se calculan A, B a posteriori, ya que su efecto queda absorbido por (μ_g, σ_g) en el ajuste. Cuando se imponen $\mu_g = 0$ y $\sigma_g = 1$ para todos los grupos, todas las variables latentes quedan en la misma métrica por construcción.

Bajo calibración concurrente los parámetros de los ítems de anclaje son idénticos entre formas y los parámetros de los ítems exclusivos pertenecen a la misma escala por compartir la estructura de θ. Si un estándar está definido en la escala latente de referencia en θ_c, su valor es el mismo para cualquier forma calibrada concurrentemente. Cuando el estándar original está expresado como puntuación observada s_c en una forma particular, se obtiene el valor latente resolviendo $S(\theta_c) = s_c$, donde la curva característica de la prueba es

$$S(\theta) = \sum_i \sum_k w_{ik}\; P_{ik}(\theta).$$

El punto de corte equivalente en puntuación observada de otra forma se calcula como $s_c' = S_{\text{destino}}(\theta_c)$. En pruebas de número-correcto, la monotonicidad de S garantiza una solución única para θ_c; en la práctica se emplea bisección o Newton–Raphson para la inversión y cuadratura para la evaluación.

En el modelo de Rasch con parametrización logística $P_i(\theta) = \{1 + \exp[-D(\theta - b_i)]\}^{-1}$ y $D = 1{,}7$, la calibración concurrente se simplifica por la homogeneidad de las discriminaciones. Los ítems de anclaje comparten un único parámetro de dificultad b_i entre formas; los ítems exclusivos de cada forma tienen sus dificultades propias, pero todas quedan en la misma escala por coexistir en un único ajuste con θ común. En un planteo marginal multi-grupo se adopta

$$\theta \mid g \sim \mathcal{N}\left(\mu_g, \sigma_g^2\right), \qquad g \in \{X, Y\},$$

con $(\mu_X, \sigma_X) = (0{,}1)$ fijados para identificar la escala y (μ_Y, σ_Y) estimados por máxima verosimilitud junto con las dificultades $\{b_i\}$. En este marco, el mapeo entre métricas es $\theta^{(\text{ref})} = \left(\theta^{(Y)} - \mu_Y\right)/\sigma_Y$, por lo que un punto de corte θ_c definido en la referencia se proyecta a la métrica del grupo Y como $\theta_c^{(Y)} = \sigma_Y\,\theta_c + \mu_Y$ si se decide reportar en esa métrica, aunque no es necesario para la transformación de puntuaciones observadas, que se realiza evaluando las respectivas curvas características en la θ_c de referencia.

Una alternativa basada en verosimilitud condicional elude la especificación de $f_g(\theta)$. La verosimilitud condicional del Rasch condiciona en los puntajes brutos de cada examinado y permite estimar las dificultades b_i sin integrar sobre θ. En calibración concurrente condicional, los ítems de anclaje comparten un único b_i, la escala se fija por una restricción (por ejemplo, $\sum_i b_i = 0$ en un conjunto de referencia) y los datos de ambas formas se combinan en la misma función objetivo condicional. Este enfoque es robusto a la mala especificación de la distribución latente, aunque a costa de mayor complejidad computacional en los términos de partición condicional cuando el número de ítems es elevado.

En ambos enfoques, una vez estimados los parámetros en calibración concurrente, la transformación de un estándar entre formas se resuelve trabajando en θ común y proyectando a puntuación observada mediante la TCC de la forma destino. Cuando el estándar original es una puntuación cruda, se invierte la TCC de la forma de referencia para hallar θ_c y se evalúa la TCC de la forma destino en ese mismo valor latente. La coherencia métrica se garantiza por la igualdad de parámetros de los ítems de anclaje y por la identificación global de la escala.

Métodos basados en puntuaciones esperadas

La equiparación basada en puntuaciones esperadas bajo la TRI comprende dos enfoques estrechamente relacionados pero conceptualmente distintos. El primero es el True-Score Equating (TSE), que utiliza la curva característica de la prueba para igualar puntuaciones verdaderas esperadas a lo largo del continuo latente. El segundo es el Observed-Score Equating (OSE) bajo TRI, que opera sobre la distribución de puntuaciones observadas inducida por el modelo, integrando sobre la distribución de la habilidad en la población de referencia. Ambos enfoques permiten transformar estándares de desempeño entre formas de prueba cuando se dispone de un enlace métrico entre sus escalas latentes.

El TSE presenta ventajas cuando se pretende aislar la relación funcional entre pruebas independientemente de la distribución de habilidad de una muestra específica. El OSE, por su parte, preserva posiciones relativas en la distribución de puntajes observados y, por ello, es sensible a la elección de $f(\theta)$ que define la población de referencia. En ambos casos, la coherencia del método depende de la calidad del enlace métrico entre escalas; si las calibraciones se realizaron separadamente, el empleo de procedimientos de enlace como Stocking–Lord o Haebara es un requisito previo. Resulta recomendable reportar diagnósticos de ajuste, la población de referencia utilizada, análisis de sensibilidad y, cuando corresponda, intervalos de incertidumbre obtenidos por remuestreo paramétrico que reestime parámetros, repita la inversión de TCC y vuelva a calcular las distribuciones observadas modeladas.

Sea θ la habilidad latente y considérese, para la forma X, un conjunto de ítems con funciones de respuesta $P_{Xi}(\theta)$. En ítems dicotómicos con parametrización logística y constante $D > 0$, se tiene

$$P_{Xi}(\theta) \;=\; c_{Xi} + (1 - c_{Xi})[1 + \exp\{-D\, a_{Xi}(\theta - b_{Xi})\}]^{-1}.$$

La Curva Característica de la Prueba (TCC) es la puntuación verdadera esperada

$$S_X(\theta) \;=\; \sum_i \sum_k w_{ik}\; P_{Xi,k}(\theta),$$

que en el caso dicotómico reduce a $S_X(\theta) = \sum_i P_{Xi}\,(\theta)$. La TCC es estrictamente creciente bajo condiciones regulares, lo que habilita su inversión numérica para recuperar θ a partir de un nivel de puntuación verdadera.

True-Score Equating (TSE) bajo TRI

El TSE define la equiparación por la igualdad de puntuaciones verdaderas esperadas a un mismo nivel de habilidad. Si dos formas X y Y están en la misma métrica latente, el emparejamiento se expresa como

$$t \;=\; S_X(\theta) \quad\Leftrightarrow\quad t' \;=\; S_Y(\theta).$$

Cuando se requiere transformar un punto de corte definido en la escala latente, se toma θ_c y se evalúa $S_Y(\theta_c)$ para obtener el estándar en puntuación de Y. Si el estándar está dado como puntuación observada s_c de X, se determina primero θ_c resolviendo $S_X(\theta_c) = s_c$ y luego se calcula $s_c' = S_Y(\theta_c)$. Este procedimiento depende únicamente de las TCC modeladas y no de la distribución de θ en la población.

Observed-Score Equating (OSE) bajo TRI

El OSE bajo TRI trabaja con la distribución de la puntuación observada inducida por el modelo y una población de referencia. Para una puntuación cruda R_X en X, la distribución condicional dada θ se obtiene de la independencia local de ítems. En el caso dicotómico con número-correcto, R_X es la suma de Bernoulli independientes con parámetros $P_{Xi}(\theta)$. La distribución marginal en la población se define como

$$\Pr(R_X = r) \;=\; \int \Pr(R_X = r \mid \theta)\, f(\theta)\, d\theta, \qquad r = 0,1,\ldots,n_X,$$

donde $f(\theta)$ es la densidad de habilidad en la población de referencia. El emparejamiento equipercentil modelado selecciona r' en Y tal que

$$F_Y(r') \;=\; F_X(r), \qquad F_X(r) = \sum_{k \le r} \Pr\,(R_X = k),$$

con las probabilidades calculadas por integración numérica. El resultado preserva posiciones relativas en la distribución observada, pero la distribución es model-based y, por ende, depende de $f(\theta)$ y del ajuste TRI.

La densidad $f(\theta)$ puede fijarse como normal estándar, como la distribución latente estimada en el grupo de referencia o como una mezcla que represente una población objetivo. La elección de f influye en $\Pr(R_X = r)$ y, por tanto, en la función de equiparación observada. Es recomendable documentar explícitamente la opción adoptada y reportar análisis de sensibilidad.

En Rasch dicotómico con $P_i(\theta) = [1 + \exp\{-D(\theta - b_i)\}]^{-1}$ y $D = 1{,}7$, la TCC es $S(\theta) = \sum_i P_i\,(\theta)$. Para TSE, la transformación de un corte s_c de X a Y se implementa invirtiendo S_X para recuperar θ_c y evaluando luego S_Y en ese valor. Para OSE, la distribución condicional $\Pr(R = r \mid \theta)$ se calcula por convolución de Bernoulli con probabilidades $P_i(\theta)$ y la distribución marginal $\Pr(R = r)$ se obtiene por cuadratura sobre $f(\theta)$. El mapeo equipercentil se determina comparando las funciones de distribución acumulada modeladas.

Referencias bibliográficas

Hanson, B. A., Zeng, L. y Kolen, M. J. (1993). A comparison of item response theory observed-score equating methods. *Applied Psychological Measurement*, *17*(4), 345–360. https://doi.org/10.1177/014662169301700403

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Landis, J. R. y Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A. y Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional item response theory. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1294-4

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *142*(2), 197–207. https://doi.org/10.1177/014662169001400208

Shealy, R. y Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327. https://doi.org/10.1007/BF02293557

Woodruff, D. J. (1986). A distribution-free method for comparing means of two independent samples. *Communications in Statistics – Theory and Methods*, *15*(11), 3669–3692. https://doi.org/10.1080/03610928608829167