Capítulo 20

EQUIPARACIÓN EN LAS PRUEBAS DE DESEMPEÑO: ALGUNAS TÉCNICAS MÁS COMUNES EN LA EVALUACIÓN EDUCATIVA

Rafael Enrico Sánchez Mayorga, Laura Delgado Maldonado

"La calidad no es un acto, sino un hábito."
Aristóteles

INTRODUCCIÓN

Las universidades y las distintas instancias de evaluación educativa cuentan con programas de evaluación que implican una variedad de recursos sea para mejorar el aprendizaje de los estudiantes o para medir su desempeño académico, haciendo inferencias válidas a partir de los resultados de las evaluaciones con las que cuentan; por ejemplo, para revisar el plan de estudios; como un diagnóstico para informar a los estudiantes de su progreso; informar al público sobre el desempeño escolar; para ser utilizado como guía en la toma de decisiones sobre estudiantes, maestros o escuelas; para el ingreso, egreso o certificación de un nivel académico proporcionando información útil que apoye las inferencias de los resultados.

Por lo que, cada vez es necesario establecer y mantener la calidad técnica de un programa de pruebas, ya que esta impacta directamente en la validez de las evaluaciones y el grado en que la evidencia y la teoría apoyan las interpretaciones de los puntajes de las pruebas (AERA, APA y NCME, 2014). Una de las técnicas que apoya a alcanzar este objetivo es la equiparación (*Equating),* la cual es un método que se lleva a cabo para establecer puntuaciones comparables entre las diferentes formas de una prueba, lo cual permite que se usen indistintamente en un periodo de tiempo determinado.

En el diseño de una prueba nos preguntamos, porqué es importante contar con más de una forma para su aplicación o porqué es necesario llevar a cabo la equiparación entre las distintas formas que se aplican de un mismo instrumento a través del tiempo; la respuesta puede estar en que los responsables de las pruebas deben establecer desde su diseño estrategias que respondan a las necesidades de la evaluación como pueden ser: establecer la validez entre formas y años; garantizar la justicia y equidad de los resultados para los aspirantes

que apliquen cualquiera de las formas de la prueba; para la seguridad de administración de las pruebas en donde se aplique más de una ocasión a más de un grupo de sustentantes, lo que puede conducir a una sobreexposición de los reactivos por el número de observaciones, amenazando la seguridad de esta y, cada vez más, en los programas de evaluación se vislumbra la necesidad de la elaboración constante de reactivos como un proceso iterativo en el mantenimiento de la prueba.

Por lo que, el objetivo de este apartado es proporcionar una perspectiva de los métodos de equiparación y brindar un recurso útil para informar sobre las decisiones que se necesitan considerar al equiparar las formas de una prueba. Como tal, puede ser una guía para formular soluciones que sean prácticas, factibles y técnicamente sólidas cuando se toman decisiones que involucran distintas formas de prueba a partir de un mismo constructo.

El primer punto importante que debemos definir es, qué es una forma de la prueba.

Una forma está definida como una de las posibles múltiples pruebas de un mismo instrumento que mide el mismo constructo, frecuentemente estas formas suelen denominarse formas alternativas. Cuando dos formas de una prueba se han equiparado con éxito, se puede interpretar válidamente que una forma de la prueba funciona como si tuviera el mismo significado sustantivo en comparación con la puntuación equiparada de la otra forma de prueba.

Es importante considerar que, en el campo de la evaluación educativa, tanto en pruebas de alto impacto como en las pruebas a gran escala, los sustentantes responden a la selección de un conjunto de reactivos destinados a representar la totalidad del dominio de los conocimientos, habilidades o tareas posibles, con el propósito de brindar a estos una descripción precisa de lo que ellos saben y pueden hacer. Por lo que, una parte importante radica en que las formas estén construidas de acuerdo con los mismos contenidos especificados en su diseño y sean desarrollas de forma equivalente en la dificultad a lo largo de toda la prueba.

Los estándares *educativos y pruebas psicológicas* representan un marco de referencia internacional que establecen criterios para el desarrollo y la evaluación de pruebas y prácticas de desarrollo de pruebas y enmarcan las pautas para determinar la validez de las interpretaciones de los puntajes de las pruebas para los usos previstos de mismas. (AERA, APA, y NCME, 2018, p.1). Particularmente es importante en el Estándar 5.6, refiere que en los programas de evaluación las pruebas demuestren la estabilidad de sus escalas en la que reportan sus puntajes a lo largo del tiempo. Asimismo, los que refieren a la Vinculación de puntajes en los estándares 5.12 al 5.20 que requieren respaldos de evidencia de la comparabilidad de puntajes de una forma a otra y de un año a otro, así como de reunir datos adecuados y aplicar la metodología estadística apropiada para realizar la equiparación de puntajes en las formas alternativas de la prueba.

Un primer paso para la introducción a la equiparación es distinguir entre los conceptos que se han utilizado en la literatura; Holland y Dorans (2006) hicieron distinciones entre diferentes tipos de alineamiento (*linking*) y enfatizaron que estas distinciones están relacionadas con, cómo se utilizan e interpretan las puntuaciones vinculadas, dividiendo los métodos de alineamiento en tres categorías básicas: predicción, alineación de escala y

equiparación, considerando esencial comprender por qué estas categorías difieren y por qué pueden confundirse en la práctica.

El método de predicción para Dorans, Moses y Eignor (2010) tiene como objetivo predecir el puntaje de un sustentante en una prueba basándose en otra información sobre ese mismo sustentante. Por ejemplo, las puntuaciones de otras pruebas, calificaciones en cursos, el promedio de calificación, entre otros. Las relaciones establecidas mediante la predicción no son simétricas, por lo que no puede utilizarse para equiparar puntuaciones o producir puntuaciones con propiedades de puntuación comparables.

El método de alineación de escalas se refiere a las formas de pruebas que no se han diseñado de acuerdo con las mismas especificaciones; es decir, formas que pueden diferir en longitud o contenido; en este caso, las escalas vinculadas se consideran similares, pero no intercambiables. Esta relación se lleva a cabo a través de una función de *linking*. De acuerdo con Dorans et al. (2010) la alineación de escalas *(linking)* y la equiparación de puntuaciones a menudo se confunden porque los procedimientos estadísticos utilizados para la alineación de escalas también se pueden utilizar para equiparar pruebas; por lo que la alineación de escalas no puede ser considerada para equiparar puntuaciones.

En cambio, el método de equiparación se define en una relación estadística funcional entre múltiples distribuciones de puntaje de formas de una prueba y, por lo tanto, entre múltiples escalas de puntaje. Cuando las formas de las pruebas se han creado de acuerdo con las mismas especificaciones y son similares en características estadísticas esta relación funcional se denomina función de equiparación y sirve para convertir las puntuaciones de una escala directamente a sus valores equivalentes en otra.

A diferencia de *linkin*g, la equiparación apoya la afirmación de que un sustentante que aplica la prueba y obtiene una puntuación determinada en un año sabe y se comporta de manera similar a otro sustentante que obtiene el puntaje equivalente en el año siguiente. Esto es especialmente importante para mantener un significado equiparable de puntos de corte y niveles de rendimiento de uno año a otro, garantizando la equidad de los resultados entre las diferentes poblaciones que aplican cualquiera de las formas de la prueba.

PROPIEDADES QUE DEBE CUMPLIR PARA LA EQUIPARACIÓN

En el diseño se puede comprender que, en caso de ser requerido, se utilice la metodología de equiparación para corregir las diferencias en dificultad de las formas de la prueba. Lord (1980) establecieron requisitos específicos para equiparar puntajes entre formas y así definir las diferencias entre las distintas metodologías de alineamiento estas son: igualdad de constructos, confiabilidad similar, invarianza poblacional, simetría y equidad. De este modo, es necesario considerar (Dorans y Holland, 2000):

1) Igualdad de construcción: las distintas formas de la prueba deben ser medidas a partir del mismo constructo (aptitud, rasgo latente, habilidad).
2) Confiabilidad similar: las formas deben tener un nivel similar de confiabilidad.

3) Simetría: la transformación de equiparación para mapear las puntuaciones de B a los de A deberá ser el inverso de la transformación de equiparación para mapear las puntuaciones de A a las de B.
4) Equidad: deberá ser una cuestión de indiferencia para un examinado en cuanto a cuál de las formas de la prueba realiza realmente el examinado.
5) Invarianza de la población: la función de igualación utilizada para vincular las puntuaciones de A y B debe ser la misma, independientemente de la elección de la subpoblación de la que se deriva.

En la práctica, es difícil que se cumpla el conjunto de condiciones establecidas y no existe un acuerdo sobre cuáles son realmente las que se deben exigir en el proceso de equiparación. Sin embargo, con respecto a las mejores prácticas de la evaluación Dorans et al. (2010), consideran que los requisitos de igualdad de construcción y confiabilidad significan que las distintas formas de las pruebas deben construirse a partir de las mismas especificaciones, mientras que la simetría excluye los métodos de regresión como una forma de equiparación de prueba. Asimismo, en lo que respecta a la Equidad estos comprenden los dos primeros requisitos. Sin embargo, este requisito es difícil de evaluar empíricamente y su uso es principalmente teórico (Lord, 1980). Por último, Holland y Dorans (2006), consideraron que se pueden utilizar los dos primeros requisitos para explicar, si las dos pruebas miden cosas diferentes o no son igualmente confiables, por lo que, los métodos de vinculación estándar no producirán resultados que sean invariables en ciertas subpoblaciones de examinados.

DISEÑO PARA EL LEVANTAMIENTO DE DATOS

Para realizar la equiparación es preciso vincular la información de las formas, lo que implica recoger datos en muestras de sujetos. Por lo que en la literatura se han propuesto diversas formas de recolección, conocidas como diseños de equiparación (Kolen y Brennan, 2010; Holland y Dorans, 2006; y von Davier et al., 2004). Estos autores consideran que los diseños más utilizados en la práctica son: Diseño de un solo grupo, Diseño de grupo único contrabalanceado, Diseño de grupos equivalentes o grupos aleatorios y Diseño de grupos no equivalentes con reactivos comunes.

Diseño de un solo grupo

En los diseños de un solo grupo se administra en el mismo grupo de sujetos las dos formas de la prueba que se desean equiparar. Ambas formas deben medir las mismas especificaciones y presentar el mismo grado de dificultad. El diseño de un solo grupo puede proporcionar resultados de equiparación precisos con tamaños de muestra relativamente pequeños.

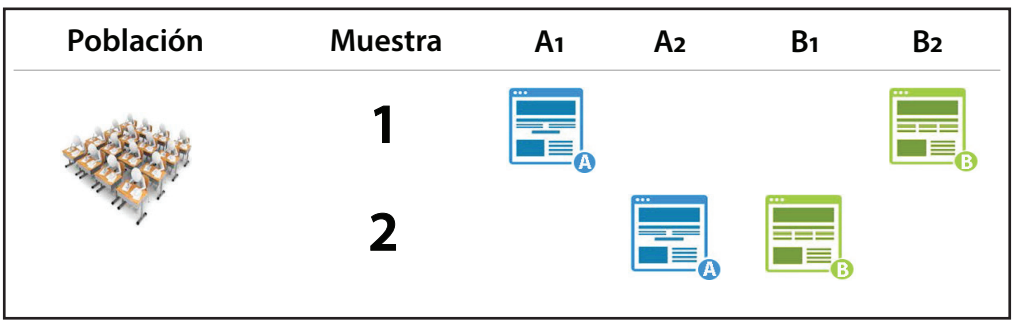
Figura 1. Diseño de un solo grupo

Este diseño se debe tener en cuenta que derivado del orden de administración de las formas de la prueba y los efectos de fatiga o aprendizaje que este puede producir, se debe asumir que el valor de las puntuaciones obtenidas por los sujetos en la segunda forma de la prueba, no están afectadas por habérseles administrado en una primera forma.

Diseño de un solo grupo contrabalanceado

Para asegurar la inexistencia de estos efectos, es recomendable utilizar una variante, el diseño de un solo grupo contrabalanceado, es una manera de poder evitar los posibles efectos del orden de administración de las dos formas de la prueba. En este caso, subdivide a los sujetos en dos grupos incluyendo en cada uno 50% de la muestra. A continuación, se administra a ambos subgrupos las dos formas de la prueba en orden inverso. De esta forma, se puede asegurar que ambas formas se verán afectadas por igual, por los efectos del orden de aplicación. En la figura 2 se observa que A_1 es cuando se toma primero y A_2 cuando se toma en segundo lugar, y de manera similar para B_1 y B_2, se describe el diseño de los grupos contrabalanceados.

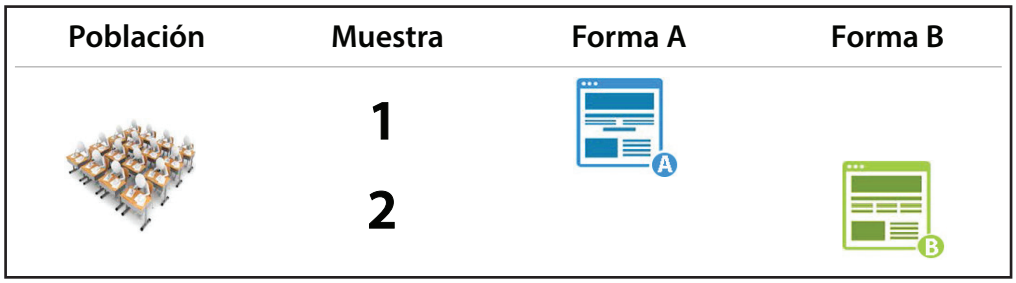
Figura 2. Diseño de un solo grupo con contrabalanceo



Diseño de grupos equivalentes (grupos aleatorios)

En este diseño se extraen de la población de forma aleatoria dos muestras de sujetos, y a cada muestra se le aplica una forma de la prueba. Por lo tanto, cada sujeto responde solamente a una de las formas.

Figura 3. Diseño de grupos equivalentes

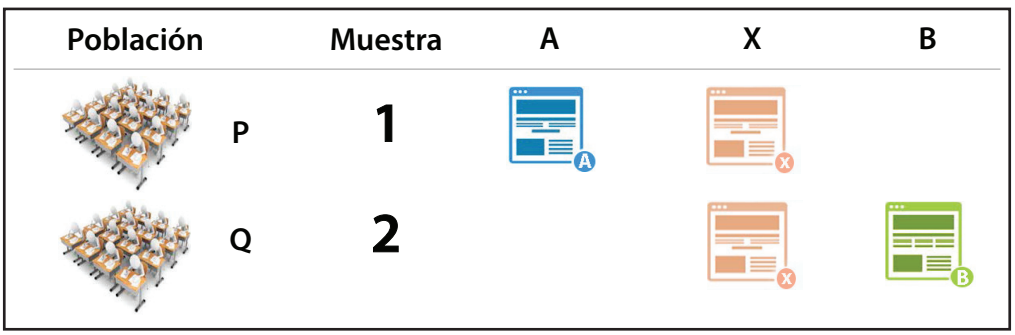


En la mayoría de las aplicaciones para obtener muestras aleatorias y equivalentes alternar las formas en cada grupo, de tal manera, que al primer sujeto se le entregue la forma A, al segundo la forma B, y así sucesivamente. Este tipo de muestreo se le denomina muestra en espiral. El diseño es bastante conveniente de administrar, no requiere que las dos pruebas tengan reactivos en común, sin embargo, este diseño se puede utilizar incluso cuando tienen reactivos en común. Una limitación es que requiere grandes tamaños de muestra para tener resultados de equiparación precisos.

Diseño de grupos no equivalentes con reactivos comunes

En los diseños de prueba de anclaje hay dos poblaciones, P y Q, con una muestra de sustentantes de P, a quienes se les administra la forma de la prueba A, y una muestra de Q quien se les administra la prueba B. La diferencia con el diseño de grupos equivalentes consiste en que ambas muestras no tienen por qué ser equivalentes entre sí y, además, ambas muestras se les aplica un subconjunto de reactivos de prueba anclajes X en cada una de las formas de prueba deben equipararse. Por lo tanto, cada sujeto contesta un subconjunto de reactivos de la prueba diferente y uno común. von Davier et al. (2004). Kolen y Brennan (2010).

Figura 4. Diseño de grupos no equivalentes con reactivos comunes



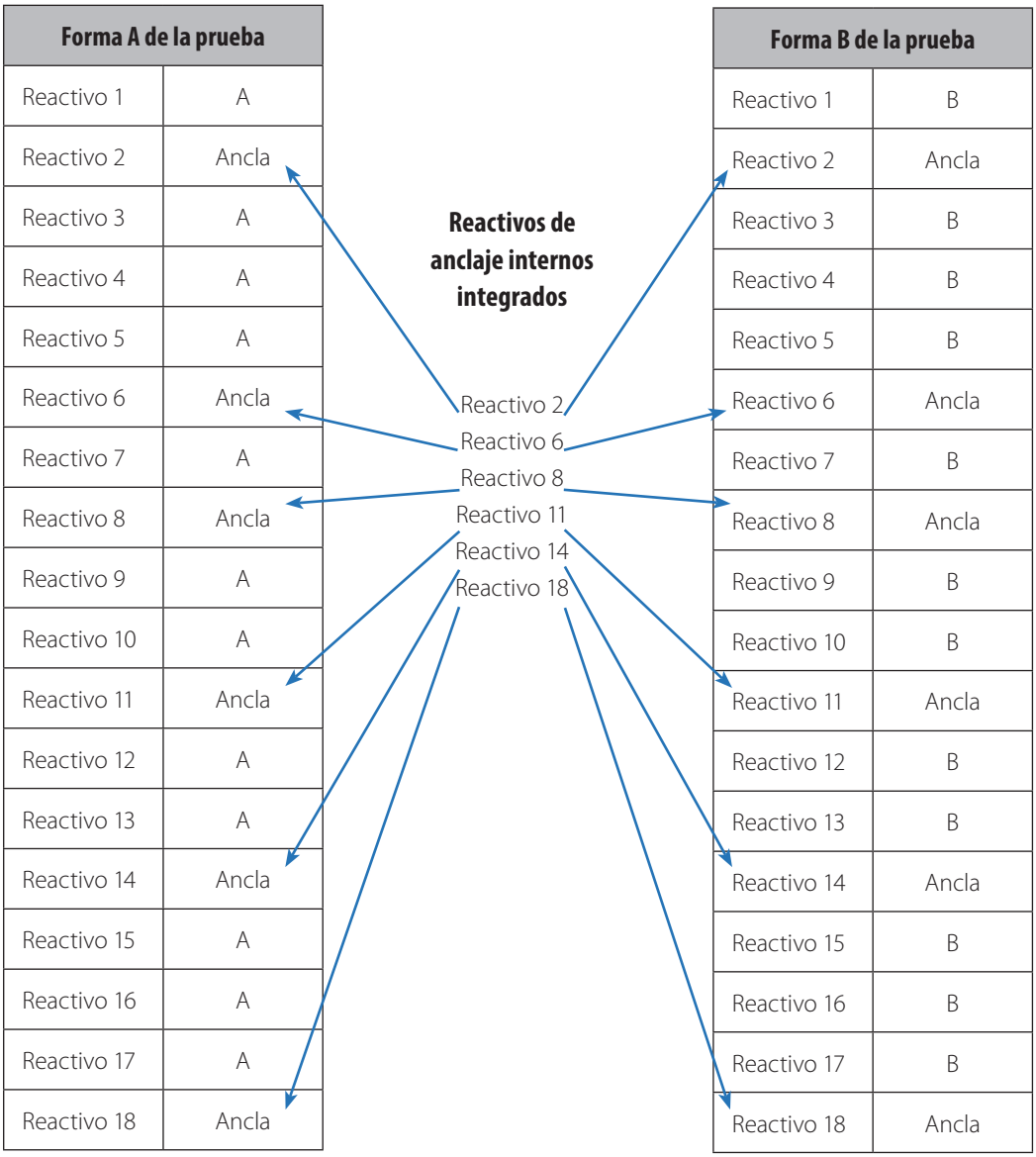
La selección de reactivos destinados a cumplir la función de anclas es particularmente importante. La representación del contenido proporcional de los reactivos en el conjunto de ancla debe ser similar a la representación de contenido proporcional de toda la

forma de prueba, incluso hasta el punto de considerar que el conjunto de ancla es una "pequeña-versión" de la forma de prueba completa (Kolen y Brennan, 2010).

Este diseño presenta dos posibles modalidades: la prueba de anclaje interno y la prueba de anclaje externo.

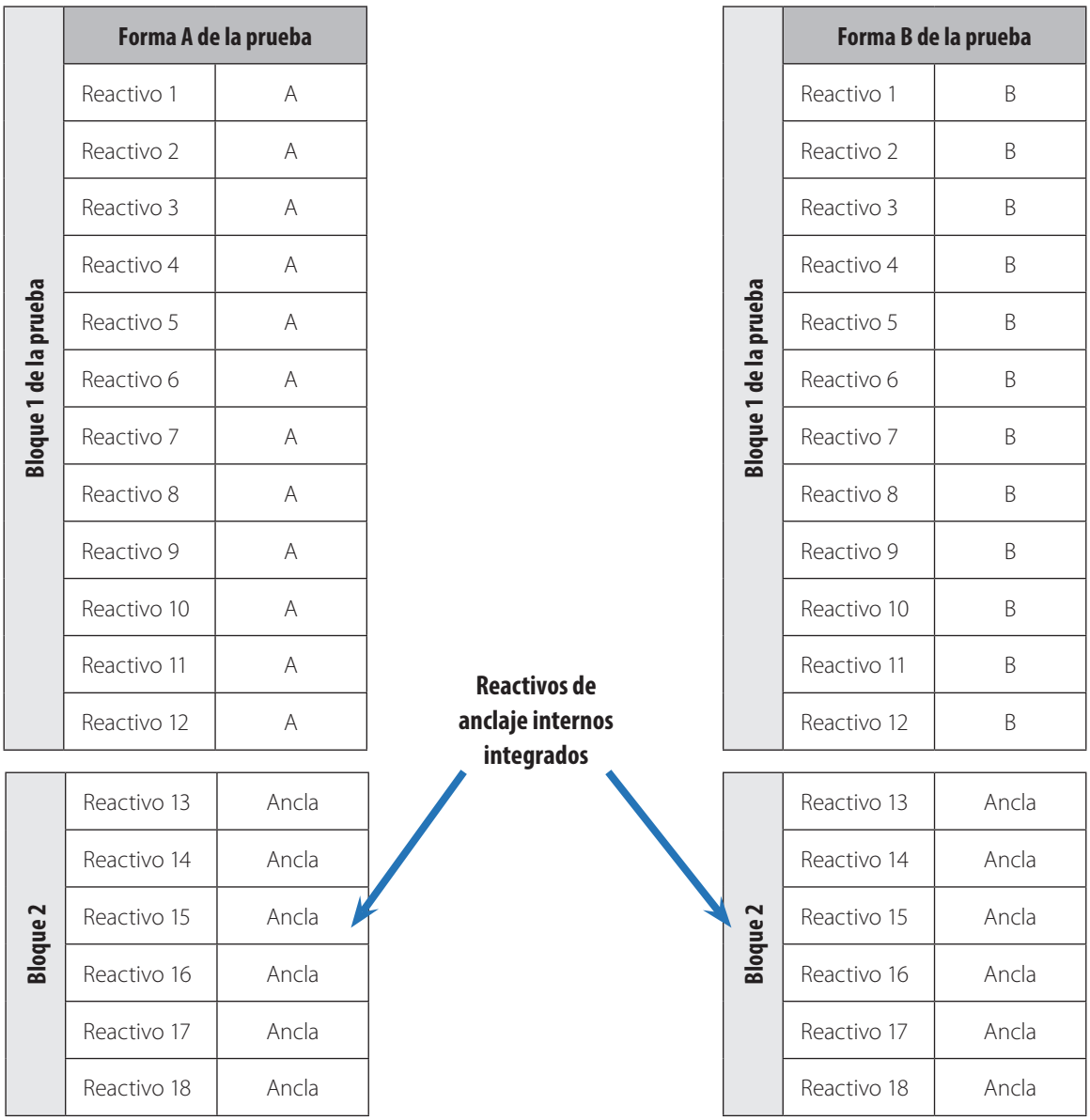
La prueba de anclaje interno se utiliza un conjunto de reactivos comunes en ambas formas de la prueba y estos aparecen intercalados con el resto de los reactivos que son propios de las dos formas A y B, cuyas puntuaciones se quieren equiparar. Las puntuaciones obtenidas en los reactivos comunes se incluyen en la puntuación total de los sujetos en la prueba.

Figura 5. Reactivos de anclaje internos integrados



En la prueba de anclaje externo los ítems comunes aparecen formando una prueba independiente y las puntuaciones obtenidas por los sujetos en esa prueba no se utilizan en el cálculo de la puntuación total de los sujetos en las formas a equiparar.

Figura 6. Reactivos de anclaje externos integrados



En el primer caso se consideran reactivos de anclaje y en el segundo de prueba de anclaje. En ambos casos los ítems comunes deben ser lo más parecidos posible a los de las dos formas, aunque no sea condición imprescindible.

MÉTODOS DE EQUIPARACIÓN

Existen varias herramientas o procedimientos para la equiparación de las formas de la prueba, algunos asociados con la teoría clásica de los test (TCT) y otros con la teoría de respuesta al ítem (TRI). Todos estos procedimientos se pueden utilizar para equiparar y otros tipos de vinculación. Sin embargo, es esencial reconocer que cuando se usa para equiparación, estos procedimientos se aplican a las pruebas que han sido construidas para ser paralelos de modo que las puntuaciones en las formas tengan el mismo significado e interpretación.

Equiparación Identidad

La equiparación identidad es la más simple de todas, ya que las formas se consideran iguales y no es necesario equipararlas, considerando que un puntaje en la forma inicial (forma A) es equivalente a un puntaje idéntico en la forma nueva (forma B). La función matemática del método de identidad es:

$$B = ID\ B\ (A) = A$$

Donde, *A* se refiere a la puntuación bruta obtenida de la forma *A*, y *B* es la puntuación bruta equivalente a *A* para la forma B.

En la Gráfica 1, se puede observar la relación de identidad entre las dos formas en el puntaje 20. No es necesario un número mínimo de sustentantes, si el número de sustentantes es menor de 100 en alguna de las formas se recomienda utilizar el método de equiparación identidad, cuando las formas de prueba se consideran paralelas (Kolen y Brennan, 2010).

Equiparación de la media

Para este método de equiparación se considera que la forma A difiere de la B, esto debido a diferencias en dificultad. Por lo que, esta cantidad se considera constante a lo largo de la escala de puntuaciones, es decir, que, considerando un ejemplo, si A es más fácil en 4 puntos, estos 4 puntos son los mismos para los sujetos de alta habilidad que para los de baja. El supuesto básico es que las diferencias de los sujetos en las dos formas son iguales:

$$y - \mu\ (B) = x - \mu(A)$$

$y = m_B(x)$transformación a la forma B en función del puntaje de la forma A

$$m_B(x) = x - \mu(A) + \mu(B)$$

Donde:
x = un puntaje particular de la forma A
y = un puntaje particular de la forma B

$m_B(x)$ = un puntaje x en la forma A transformado a la escala de la forma B
$\mu(A)$ = media del puntaje de la forma A
$\mu\ (B)$ = media del puntaje de la forma B

Equiparación lineal

La equiparación lineal es una herramienta que se utiliza principalmente en TCT para determinar puntuaciones equivalentes entre dos formas de prueba paralelas que deben equipararse y se basa en la suposición de que las distribuciones de los puntajes de las formas A y la forma B son iguales, pero sus medias y desviaciones estándar son diferentes (Crocker y Algina, 1986).

Angoff (1984) definió la equiparación lineal como puntajes que son equivalentes cuando los puntajes en dos formas de prueba corresponden a las mismas desviaciones de puntaje estándar. Los puntajes en las formas con una distancia igual con respecto a sus medias (en unidades de desviación estándar) son considerados equivalentes.

$$\frac{x - \mu(A)}{\sigma(A)} = \frac{y - \mu(B)}{\sigma(B)}$$

$y = l_y\ (x)$ transformación a la forma B en función del puntaje de la forma A

$$l_Y(x) = \frac{\sigma(B)}{\sigma(A)}\, x + \left[\mu(B) - \frac{\sigma(B)}{\sigma(A)}\, \mu(A)\right]$$

donde:

$l_y(x)$ = un puntaje x en la forma A convertido a la escala de la forma B
$\mu(A)$ = media de la forma A
$\mu(B)$ = media de la forma B
$\sigma(A)$ = desviación estándar de la forma A
$\sigma(B)$ = desviación estándar de la forma B

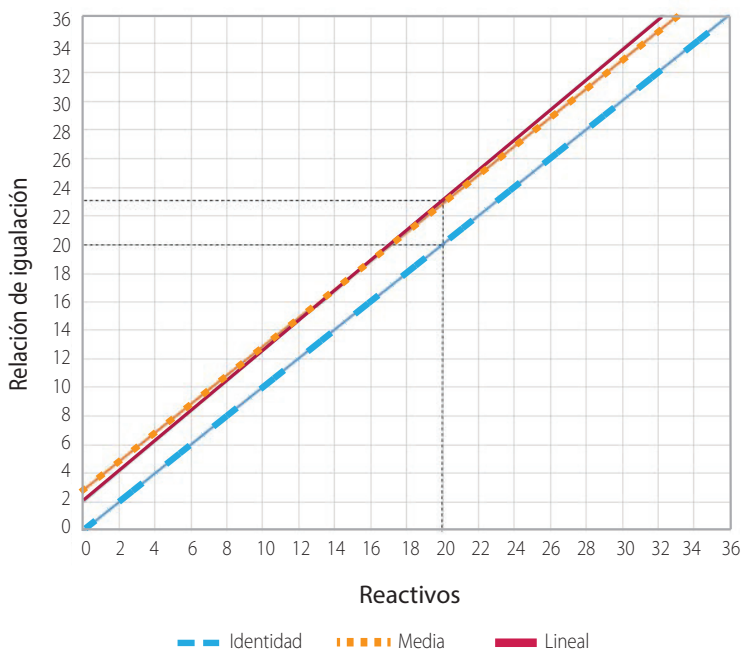
En la ecuación lineal toma su nombre del hecho de que la relación entre las puntuaciones de la forma de la prueba A y la forma de prueba B se puede mostrar como una línea recta en un gráfico. Por tanto, la línea representa la relación equivalente para todas las puntuaciones posibles. Como se muestra en la Gráfica 1, la forma de la prueba B en una prueba de 36 reactivos parece tener reactivos más difíciles que la Prueba A con 36 reactivos, basada en las puntuaciones medias (Prueba A media = 18.5 y Prueba B media = 17).

La Gráfica 1 muestra algunas de las limitaciones de este método, teniendo en cuenta que las puntuaciones de la forma B (media 20) son discretas: no tienen valores decimales. Pero no se transforman en puntajes brutos discretos en la forma A. Por ejemplo, un puntaje de 20 en la forma B es el equivalente ajustado del puntaje 18.5 en la forma A. Sin

embargo, los estudiantes que toman la forma A no pueden recibir un puntaje bruto de 18.5. Para abordar el problema de las puntuaciones equivalentes no discretas, se han utilizado varios enfoques para redondear las puntuaciones equivalentes de manera que les permitan informar puntuaciones discretas. Sin embargo, el redondeo introduce su propio tipo de error de equiparación.

Asimismo, se muestra una relación lineal en la que una puntuación muy alta en la forma B da como resultado un puntaje que está fuera del rango de puntajes posibles para la forma A. El gráfico parece sugerir que un puntaje de 36 en la forma B equivale a una puntuación de 39.9 o más en la forma A, lo cual no es posible. Sin embargo, Esto no es un error en la forma en que se traza el gráfico; si no, es parte de la naturaleza de la equiparación lineal.

Gráfica 1. Relación entre equiparación de identidad, de media y de igualación lineal



Método equipercentil

El método equipercentil proporciona precisión en la equiparación de resultados a lo largo de toda la escala de puntuación. También permite una mayor precisión que la ecuación lineal cuando las formas de prueba difieren en el nivel de dificultad general (Kolen y Brennan, 2010).

Para definir los rangos percentilares:

K_x = Número de reactivos en la forma A
$f(x)$ = Proporción de sustentantes que tuvieron el puntaje x
$f(x) \geq$ para los puntajes enteros $x = 0,1, K_x$;
$f(x) = 0$ para otro puntaje

$$\sum f(x) = 1$$

$f(x)$ = Proporción acumulada igual o debajo del puntaje

$$f(x) = 0 \text{ para } x < 0; y$$

$$f(x) = 1 \text{ para } x > K_x$$

$0 \leq F(x) \leq 1$ para x = 0,1, K_x;

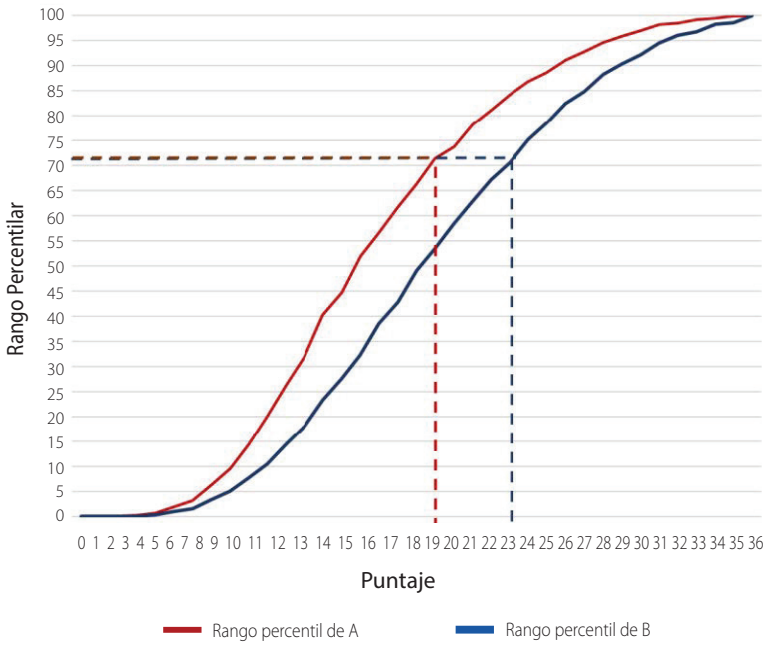
Si tenemos un valor de que no es entero, entonces x^* es el entero más cercano tal que $x^* - .5 \leq x < x^* + .5$

La función para calcular rango percentilar de la forma A es:

$$P(x) = 100\{F(x^* - 1) + [x - (x^* - .5)][F(x^*) - F(x^* - 1)]\},$$
$$-.5 \leq x < K_x + .5$$
$$= 0, x < -.5$$
$$= 100, x \geq K_x + .5$$

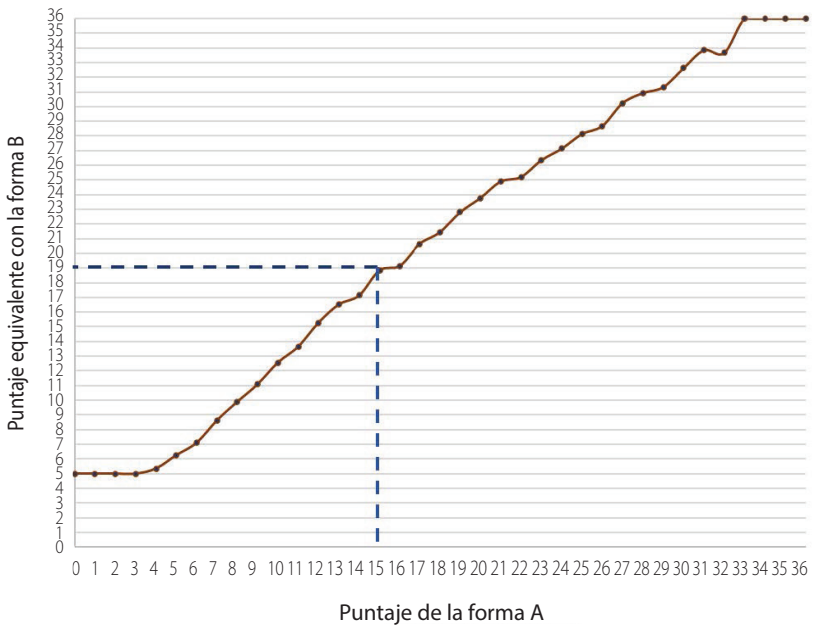
Se puede representar gráficamente las dos distribuciones de percentiles. Para ello, en el eje de abscisas ponemos las puntuaciones de la prueba A y de la forma B. En el eje de ordenadas los rangos percentiles. A continuación, trazan la curva correspondiente a cada prueba. Por lo que se obtiene las puntuaciones equivalentes en las dos formas A y B. En la Gráfica 2, se observa que el puntaje por debajo de 19 que representa 71% de los sustentantes para la forma A y por debajo del puntaje de 23 representa el mismo 71% en la forma B.

Gráfica 2. Comparación entre los rangos de las dos formas A y B



En la Gráfica 3 de equivalencias muestra las relaciones equipercentilares entre la forma A y la forma B de los datos, se observa que el puntaje 15 de la forma A es equivalente a 19 con la forma B.

Gráfica 3. Comparación entre los rangos de las dos formas A y B



DISEÑO DE GRUPOS NO EQUIVALENTES CON REACTIVOS COMUNES

Para estos métodos de equiparación, al ser bajo el diseño de grupos no equivalentes, consideran dos poblaciones diferentes: la población que contestó la forma A y la población que contestó la forma B; sin embargo, la función de equiparación está definida para una sola población.

En la literatura es uno de los diseños que más se utilizan. En cada muestra de sujetos se administra solamente una forma de la prueba, con la particularidad de que en ambas muestras se administra un conjunto de reactivos en común, que permite establecer la equivalencia entre las formas a equiparar.

Población sintética

El método de equiparación involucra dos poblaciones diferentes. Sin embargo, una función de equiparación de puntajes se considera como definida sobre una población única. Por lo tanto, las poblaciones 1 y 2 que corresponden a las poblaciones donde se aplicó la forma A (inicial) y B (nueva), deben ser combinadas para obtener una población única con el fin de definir una relación de equiparación.

Esta única población se conoce como población sintética Braun y Holland (1982), en la cual se le asignan pesos w_1 y w_2 a las poblaciones 1 y 2, respectivamente, esto es, $w_1 + w_2 = 1$ y $w_1, w_2 \geq 0$. Lo recomendable es utilizar

$$w_1 = \frac{N_1}{N_1 + N_2}$$

y

$$w_2 = \frac{N_2}{N_1 + N_2}$$

Donde N_1 corresponde al tamaño de la población 1 y N_2 corresponde al tamaño de la población 2.
Los puntajes de la forma nueva, aplicada a la población 1, serán denotados por X; Los puntajes de la forma antigua, aplicada a la población 2, serán denotados por .

Anclaje interno

Los puntajes comunes serán denotados por V y expresamos que los reactivos comunes corresponden a un anclaje interno cuando V se utiliza para calcular los puntajes totales de ambas poblaciones.

Métodos Lineales

Usando el concepto de población sintética, la relación lineal de equiparación de puntajes para el diseño de grupos no equivalentes con reactivos comunes se escribe de la siguiente forma:

$$l_{Y_s}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y)$$

Donde denota la población sintética y

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)]$$

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)]$$

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_1^2[\mu_1(V) - \mu_2(V)]^2$$

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_2^2[\mu_1(V) - \mu_2(V)]^2$$

Donde los subíndices 1 y 2 se refieren a las poblaciones 1 y 2 respectivamente.

$$\gamma_1 = \frac{\sigma_1(X,V)}{\sigma_1^2(V)}$$

y

$$\gamma_2 = \frac{\sigma_2(X,V)}{\sigma_2^2(V)}$$

En la variante del método que se va a utilizar, las *y' s* se pueden expresar de la siguiente manera:

$$\gamma_1 = \frac{\sigma_1^2(X)}{\sigma_1(X,V)}$$

$$\gamma_2 = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)}$$

En la aplicación de este método basta con reemplazar estos coeficientes en las ecuaciones lineales antes descritas. Kolen y Brennan en 2010, proporcionan justificaciones para usar esta aproximación.

MÉTODOS DE EQUIPARACIÓN BASADOS EN EL MODELO DE RESPUESTA AL ÍTEM (TRI)

Existen una variedad de enfoques para equiparar en TRI, entre los que se encuentran los más utilizados que son la equiparación a través de reactivos comunes. Se calibran cada una de las formas de la prueba usando una métrica común, completamente por separado. Posteriormente se evalúa la relación entre los parámetros TRI en cada forma y se usa para estimar la relación para convertir las puntuaciones de los sustentantes. Lo que se hace es alinear los parámetros TRI de los reactivos comunes (ancla) se pueda aplicar esa transformación lineal a las puntuaciones. Los métodos de calibración se pueden analizar bajo los métodos de momento (media-media, media-sigma) y métodos de curva característica (Stocking Lord y Haebara). Las dos escalas del mismo constructo difieren solo en una transformación lineal simple, por lo que requiere encontrar la pendiente y la intersección de esa transformación.

Métodos de transformación Media-Sigma y Media-Media

Considerando que *I* y *J* son dos escalas construidas como escalas de TRI de 3 parámetros, que toma valores de θ para las dos escalas se relacionan de la siguiente forma:

$$\theta_{Ji} = A\theta_{Ii} + B$$

donde A y B son constantes en la ecuación lineal y θ_{ji} y θ_{Ii} son los valores de θ para el individuo *i* en las escalas *J* e *I*. Los parámetros de los reactivos en las dos escalas están dados por:

$$a_{Jj} = \frac{a_{Ij}}{A},$$
$$b_{Jj} = Ab_{Ij} + B,$$
$$c_{Jj} = c_{Ij}$$

donde y $a_{Jj,}$ b_{Jj} y c_{Jj} son los parámetros del ítem *j* de la escala J, a_{Ij}, b_{Ij} y c_{Ij} son los parámetros del ítem *j* de la escala I.

Para dos individuos *i* e *i** y dos ítems j y *j**, A y B pueden ser expresados como:

$$A = \frac{\theta_{Ji} - \theta_{Ji^*}}{\theta_{Ii} - \theta_{Ii^*}} = \frac{b_{Jj} - b_{Jj^*}}{b_{Ij} - b_{Ij^*}} = \frac{a_{Ij}}{a_{Jj}}$$
$$B = b_{Jj} - Ab_{Ij} = \theta_{Ji} - A\theta_{Ii}$$

De las anteriores ecuaciones se deriva que A y B pueden ser expresados como:

$$A = \frac{\sigma(b_J)}{\sigma(b_J)}$$
$$= \frac{\mu(a_I)}{\mu(a_J)}$$
$$= \frac{\sigma(\theta_J)}{\sigma(\theta_I)},$$
$$B = \mu(b_J) - A\mu(b_I)$$
$$= \mu(\theta_J) - A\mu(\theta_I)$$

donde las medias $\mu(b_J)$, $\mu(b_I)$, $\mu(a_J)$, $\mu(a_I)$ y las desviaciones estándar $\sigma(b_J)$ y $\sigma(b_I)$ son definidas sobre los parámetros de los ítems en cada escala *I* y *J*. Las medias $\mu(\theta_J)$ y $\mu(\theta_I)$ se definen sobre los sustentantes, de acuerdo con los parámetros relacionados con la habilidad en cada escala.

Método de transformación Media-Sigma

Es un proceso de conversión de puntaje definido por Marco (1977). En este método, la desviación estándar se utiliza en la determinación de la curva de la ecuación; y las dificultades promedio de las pruebas se utilizan en la determinación de la constante de ecuación.

$$A = \frac{\sigma(b_J)}{\sigma(b_I)}$$

$$B = \mu(b_J) - A\,\mu(b_I)$$

$\mu(b_I)$: media de los parámetros de dificultad para la escala *I*
$\mu(b_J)$: media de los parámetros de dificultad para la escala *J*
$\sigma(b_I)$: desviación estándar de los parámetros de dificultad para la escala *I*
$\sigma(b_J)$: desviación estándar de los parámetros de dificultad para la escala *J*
A: pendiente
B: intercepto

Método de transformación media-media

El método media-media definido por Loyd y Hoover (1980) calcula los coeficientes A y B utilizando las medias de los parámetros de discriminación y de dificultad. La media de los parámetros de los reactivos comunes para despejar el valor de la constante A y la media de los parámetros de los reactivos comunes para despejar el valor de la constante B, esto es:

$$A = \frac{\mu(a_I)}{\mu(a_J)}$$

$$B = \mu(b_J) - A\mu(b_I)$$

$\mu(a_I)$: media de los parámetros de discriminación para la escala *I*
$A\mu(b_I)$: media de los parámetros de dificultad para la escala *I*
$\mu(a_J)$: media de los parámetros de discriminación para la escala *J*
$\mu(b_J)$: media de los parámetros de dificultad para la escala *J*
A: Pendiente
B: intercepto

Método Haebara

En el método Haebara, la curva de la ecuación y la constante de la ecuación se obtienen utilizando la diferencia entre las curvas características del reactivo. Este enfoque es desarrollado por Haebara (1980). Para los sustentantes que tienen un cierto nivel de habilidad, la diferencia entre las curvas características del reactivo es la suma de los cuadrados de las curvas características del reactivo que pertenecen a cada reactivo. Se intenta encontrar la constante de ecuación y la curva de ecuación que minimizan esta diferencia (Kolen y Brennan, 2004). Para un θ_i dado, la suma, sobre los reactivos, del cuadrado de las diferencias se puede ver como:

$$Hdiff\left(\theta_i\right) = \sum_{j:V}\left[p_{ij}\left(\theta_{Ji}, \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}\right) - p_{ij}\left(\theta_{Ji}, \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)\right]^2$$

La sumatoria es sobre los reactivos comunes (*j:V*). En esta ecuación, la diferencia entre la curva característica de cada reactivo en las dos escalas se eleva al cuadrado y se suma. *Hdiff* es entonces acumulada sobre los sustentantes. El procedimiento de estimación encuentra A y B que minimizan el siguiente criterio:

$$Hcrit = \sum_i Hdiff\left(\theta_i\right)$$

Método Stocking-Lord

En el método Stocking-Lord la curva de la ecuación y la constante de la ecuación se obtienen utilizando la diferencia entre las curvas características del ítem. A diferencia del enfoque de Haebara, en Stocking-Lord (1983), para los participantes que tienen un cierto nivel de habilidad, la diferencia entre las curvas características del ítem es el cuadrado de la suma de la diferencia entre las curvas características del ítem pertenecientes a cada ítem. Se intenta encontrar la constante de ecuación y la curva de ecuación que minimizan esta diferencia (Kolen y Brennan, 2004).

El enfoque de Stocking-Lord utiliza el cuadrado de la diferencia de las sumas sobre los reactivos, y la notación estadística se define como:

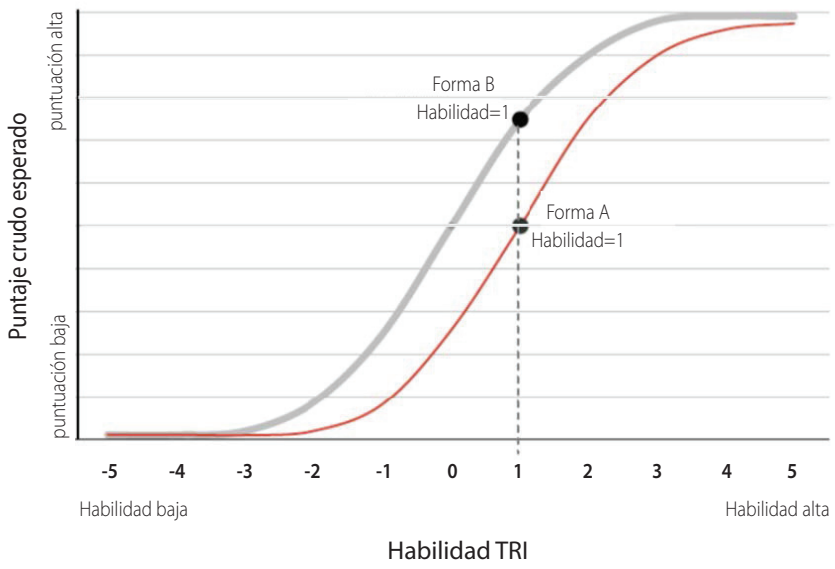
$$SLdiff\left(\theta_i\right) = \left[\sum_{j:V} p_{ij}\left(\theta_{Ji}, \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}\right) - \sum_{j:V} p_{ij}\left(\theta_{Ji}, \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)\right]^2$$

La expresión $SLdiff(\theta_i)$ es el cuadrado de las diferencias entre las curvas características de la prueba para una habilidad dada θ_i. El procedimiento de estimación es encontrar la combinación de A y B tal que se minimice el siguiente criterio:

$$SLcrit = \sum_i SLdiff\left(\theta_i\right)$$

A manera de ejemplo que se muestra en la gráfica 4, donde se observan las dos formas de prueba comparten reactivos comunes, pero cada forma también tiene su propio conjunto de reactivos únicos. En la forma de prueba A, el examen de la derecha se muestra más difícil ya que la misma posición de habilidad de TRI en el eje horizontal se asigna a una puntuación bruta más baja que la puntuación bruta de la forma de prueba B que corresponde a la misma habilidad (Ryan, y Brockmann, 2022).

Gráfica 4. Curvas de características de la prueba



La equiparación de TRI es compleja, tanto conceptual como procedimentalmente. La definición de TRI en puntajes equiparados se basa en una abstracción, más que en estadísticas que realmente se pueden calcular. TRI se basa en suposiciones sólidas que, a menudo, no son una buena aproximación a la realidad de las pruebas (Livingston, 2014).

En la actualidad existe una variedad de programas tanto gratuitos como de licencia, entre ellos el desarrollado por Kollen y Brenan, el paquete se llama *CIPE,* sirve para calcular varios procedimientos de equiparación (Kolen y Brennan, 2010), Asimismo otros paquetes del software estadístico R; el paquete *equate* (Albano, 2016) que reproduce los resultados del *CIPE.* Asimismo, los programas de TRI como BILOG-MG, ICL, MULTILOG, PARSCALE, Xcalibre, IRTEQ, entre otros.

Kolen y Brennan, (2010) recomiendan que para que cualquiera de estos métodos se utilice adecuadamente las especificaciones de la prueba, los datos recopilados y los procedimientos de estandarización y control de calidad deben ser adecuados. Sin embargo, para decidir qué métodos estadísticos implementar para una equiparación particular dependerá de las características de las situaciones de equiparación para las cuales cada uno de los métodos puede ser apropiado. Por lo debe considerar la literatura de investigación sobre métodos de equiparación y la realización de investigaciones para el programa de prueba en cuestiones prácticas en la equiparación en el que se va a hacer la equiparación. En la Tabla 1 se presenta una lista de características de las situaciones de equiparación basada en (Kolen y Brennan, 2010; p.305-310)

Tabla 1. Métodos de equiparación

Diseño	Método						
	Características	Identidad	Media	Lineal	Equipercentil	Rasch	TRI 3 PL
Grupos aleatorios equivalentes y Grupo No equivalente con reactivos en común	Control de calidad o condiciones de estandarización	Deficiente	Adecuada	Adecuada	Adecuada	Adecuada	Adecuada
	Muestras	Muy pequeñas o ningún dato	Muy pequeñas	Pequeñas	Grandes	Pequeñas	Grandes
	Forma de prueba similares	Dificultades de forma de prueba similares	Dificultades de forma de prueba similares	Dificultades de forma de prueba similares	Las formas de prueba pueden diferir en el nivel de dificultad más que para un método lineal	Dificultades de forma de prueba similares	Las formas de prueba pueden diferir en el nivel de dificultad más que para un método lineal
	Entendimiento de tablas de conversión o ecuaciones, en la realización de análisis	Sí	Sí	Sí	No	No	No
	Facilidad para explicar el procedimiento a personas que no son psicómetras	Sí	Sí	Sí	No	No	No
	Precisión de los resultados	No (tolera resultados inexactos)	Sí (alrededor de la media)	Sí (alrededor de la media)	Sí (a lo largo de la escala de puntuación)	Sí (alrededor de la media).	Sí (a lo largo de la escala de puntuación)
	Cumple los supuestos del modelo IRT	No aplica	No aplica	No aplica	No aplica	Sí	Sí
Grupos No equivalentes con reactivos en común No	Conjunto de reactivos en común que sean representativos del constructo	No	Sí	Sí	Sí	Sí	Sí
	Grupos de sustentantes con nivel de logro similar	No	Sí	Sí	Sí	Sí	Sí

A MANERA DE CIERRE

La equiparación juega un papel importante en el sistema de responsabilidad escolar y, en general, para las instancias evaluadoras ya que permite comparar los puntajes de una forma de prueba a otra, de una aplicación a otra o de un año a otro. Por lo tanto, la calidad técnica de los métodos de equiparación utilizados y la documentación de los procesos de equiparación empleados, son relevantes para cualquier sistema de rendición de cuentas destinado a reflejar el crecimiento anual y la mejora continua en los procesos de evaluación.

Asimismo, como se considera importante que las decisiones que se tomen al momento de equiparar sean mejor informadas mediante la descripción de problemas comunes relacionados con la equiparación y garanticen la equidad de los resultados para los sustentantes que contestaron cualquiera de las formas de la prueba.

El primer paso para respaldar una inferencia de equivalencia comienza con la construcción de las formas de prueba. Para ser equiparadas, las formas de prueba deben construirse de acuerdo con los mismos contenidos y especificaciones. Los elaboradores de las pruebas deben solicitar el modelo y las especificaciones de esta y cualquier otra documentación que respalde la inferencia de que las formas son equivalentes. La evidencia de los procedimientos psicométricos debe incluir datos que verifiquen que todas las suposiciones hechas para respaldar la equiparación han sido examinadas e informadas. En la equiparación de reactivos comunes, se debe presentar la evidencia de la estabilidad de los reactivos de anclaje.

Por lo anterior, de acuerdo con los estándares, la equiparación apoya la interpretación de los resultados y el uso para el cual fue diseñado el instrumento, abonando a las evidencias de validez que requiere el diseño y mantenimiento de la prueba.

Error de Equiparación

En los procedimientos de equiparación existen diferencias en las características de las muestras seleccionadas al azar lo que puede reflejar un error de muestreo. A esto se suman los propios procedimientos de equiparación que conllevan cierta imprecisión o error en el proceso de equiparación, así como las formas de la prueba tienen un cierto grado de error de medición.

Para Ryan y Brockmann, (2022) el concepto de error es la diferencia entre el valor observado o esperado y el valor real de algo, encontrado en el proceso de resolución del problema. Por lo que se debe describir el error como el nivel de precisión o certeza que se sabe que producen las puntuaciones de las pruebas, los muestreos, las estimaciones de parámetros o los procedimientos de equiparación.

Los estándares de la AERA, APA, y NCME, (2018) establecen en su estándar 5.13 que se deben estimar y reportar siempre que sea posible los errores estándares de las funciones de equiparación. Asimismo, consideran que debe presentarse en unidades de la escala de puntajes reportada. Así como para los programas de evaluación con puntajes de corte, el error de equiparación cercano al punto de corte es de primordial importancia. En los programas proporcionan una aproximación al error de equiparación, de acuerdo con el tipo de métodos de equiparación con que se esté haciendo el procedimiento.

REFERENCIAS

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association, American Psychological Association, & National Council of Measurement in Education.

Albano, Anthony. (2016). equate: An R Package for Observed-Score Linking and Equating. *Journal of Statistical Software. 74*(8). https://doi.org/10.18637/jss.v074.i08

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 9–49). New York: Academic.

Brennan, R.J. (2006). Chained Linear Equating. CASMA Technical Note. Center for Advanced Studies in Measurement and Assessment, Universidad de Iowa, IA. Recuperado el 12 de Julio de 2010, de http://www.education.uiowa.edu/casma/documents/clinearreport3.pdf

Dorans, Neil & Moses, Tim & Eignor, Daniel. (2010). Principles and Practices of Test Score Equating. ETS Research Report Series. 2010. i-41. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research. 22*, pp. 144-149.

Holland P, Dorans N (2006). Linking and Equating. In R Brennan (ed.), Educational Measurement, 4th edition, pp. 187–220. Greenwood, Westport.

Kolen, M. J., & Brennan, R. L. (2010). Test Equating, Scaling, and Linking: Methods and Practices (Second ed.). (Springer, Ed.) New York

Marco, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. *Journal of Educational Measurement, 14*(2), 139-160.

Loyd, B. H. and Hoover, H. D. (1980). Vertical Equating Using the Rasch Model. *Journal of Educational Measurement, 17*(3), 179-193.

Livingston, S. (2004). Equating Test Scores (Withot IRT). Educational Testing Service, Princeton, NJ.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Ryan, J. & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on classical test and item response theory.* Washington, DC: Council of Chief State School Officer.

Stocking, M. & Lord, F. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement. 56*(4), pp. 570 – 584

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The kernel method of test equating. New York, NY: Springer.