UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"EQUIPARACIÓN DE PUNTUACIONES EN EL EXAMEN DE ADMISIÓN DE LA UNIVERSIDAD NACIONAL AGRARIA LA MOLINA UTILIZANDO LOS MÉTODOS LINEAL Y EQUIPERCENTIL"

Presentada por:

JOAO MANUEL RADO HUARINGA

TESIS PARA OPTAR EL GRADO DE MAESTRO

MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA

Lima-Perú

2019

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"EQUIPARACIÓN DE PUNTUACIONES EN EL EXAMEN DE ADMISIÓN DE LA UNIVERSIDAD NACIONAL AGRARIA LA MOLINA UTILIZANDO LOS MÉTODOS LINEAL Y EQUIPERCENTIL"

TESIS PARA OPTAR EL GRADO DE MAESTRO
MAGISTER SCIENTIAE

Presentada por:

JOAO MANUEL RADO HUARINGA

Sustentada y aprobada ante el siguiente jurado:

M. Sc. Grimaldo Febres Huamán	Mg. Sc. Jaime Porras Cerrón
PRESIDENTE	**ASESOR**
Mg. Raphael Valencia Chacón	M. A. Fernando Rosas Villena
MIEMBRO	**MIEMBRO**

DEDICATORIA

A Dios Todopoderoso, mi principal amor.

A mis padres Manuel Rado y Ana Huaringa,

mi hermana Yuliana Rado,

mis abuelos Carlos Tello e Irma Huaringa,

quienes han depositado su esperanza en mí,

me dieron los consejos y su apoyo constante,

y a mis sobrinos Mathías, María Julia, Valentina,

Manuel y Zoe, quienes son mi motivo para

seguir adelante a pesar de las adversidades.

AGRADECIMIENTOS

Quiero dar gracias a Dios Padre, Hijo y Espíritu Santo, que en su santa providencia puso en el desarrollo de esta tesis a tantas personas que me ayudaron a terminarla, en especial a las siguientes:

- Jaime Porras, mi asesor de tesis, por sus consejos, dedicación y revisión exhaustiva de esta investigación.
- Raphael Valencia, Fernando Rosas y Grimaldo Febres, miembros distinguidos del jurado, por sus observaciones que ayudaron a mejorar este trabajo.
- César Menacho, profesor del Departamento de Estadística e Informática de la UNALM, por el apoyo con los datos para realizar las aplicaciones de este trabajo.
- Celso Gonzales, profesor del Departamento de Estadística e Informática de la UNALM, por facilitarme un libro de psicometría que me ayudó a tener la idea de investigación.
- Jesús Salinas y Jorge Chue, profesores del Departamento de Estadística e Informática de la UNALM, por su aliento y motivación constante para terminar este trabajo.
- Diana Rebaza, Ana Vargas, Rolando Salazar y Yency Choque, colegas y amigos míos, por compartir su atención, tiempo y por ayudarme a preparar la exposición para esta tesis.
- Mis hermanos y hermanas del Movimiento de Retiros Parroquiales Juan XXIII, por su entendimiento y oración constante para culminar este trabajo.
- Mis hermanos y hermanas de la Orden de San Agustín, por sus oraciones y ayudarme a entender que con el desarrollo y finalización de este trabajo también se puede dar gloria a Dios.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

En esta investigación se realizó la aplicación de los métodos de equiparación lineal y equipercentil a los puntajes obtenidos de los postulantes a los exámenes de admisión 2016-I y 2016-II de la Universidad Nacional Agraria La Molina. El desarrollo se realizó en las seis áreas que se evalúan en el examen de admisión: Razonamiento Verbal, Razonamiento Matemático, Matemática, Física, Química y Biología. El indicador utilizado para comparar ambos métodos fue el error estándar de equiparación. Entre los resultados más importantes se encontró que el método de equiparación lineal tuvo un mejor ajuste que el método equipercentil. Respecto a la dificultad de los exámenes de admisión, se obtuvo que el examen 2016-II presentó una mayor dificultad que el examen 2016-I. Finalmente, en relación a las seis áreas evaluadas en los exámenes, fue Matemática la que presentó una mayor dificultad en el examen de admisión 2016-II que en el 2016-I.

Palabras claves: Teoría Clásica de los Test, Equiparación de puntuaciones, Equiparación Lineal, Equiparación Equipercentil, Análisis de ítems.

ABSTRACT

In this research, the use of the linear and equipercentile equating methods was applied to the scores obtained from the applicants for the admission exams 2016-I and 2016-II of the Universidad Nacional Agraria La Molina. The development was carried out in the six areas that are evaluated in the admission test: Verbal Reasoning, Mathematical Reasoning, Mathematics, Physics, Chemistry and Biology. The indicator used to compare both methods was the standard error of equating. Among the most important results, it was found that the linear equating method had a better fit than the equipercentile method. Regarding the difficulty of the admission exams, it was obtained that the 2016-II exam presented a greater difficulty than the 2016-I exam. Finally, in relation to the six areas evaluated in the exams, it was Mathematics that presented a greater difficulty in the 2016-II admission exam than in 2016-I.

Key words: Classical Test Theory, Equating of scores, Linear equating, Equipercentile equating, Item Analysis.

I. INTRODUCCIÓN

La psicometría es una disciplina cuyo objetivo es medir los aspectos psicológicos de una persona tales como: conocimientos, habilidades, rasgos de personalidad, capacidades mentales, etc. Sobre esta disciplina se han presentado varias técnicas estadísticas para la evaluación de pruebas aplicadas a individuos. Dentro de estas técnicas se encuentra la llamada *equating*.

La equiparación de puntuaciones o *equating* es una de las técnicas estadísticas que se usan en psicometría con el objetivo de homologar pruebas. Navas (2000), señala que la equiparación de puntuaciones garantiza una adecuada comparación de los puntajes en distintas pruebas que miden el mismo constructo o características.

Existen diversos métodos para poder equiparar puntuaciones, entre ellos se encuentran los basados en la teoría clásica de los *tests* y en la teoría de respuesta al ítem. Los métodos lineal y equipercentil pertenecen a la teoría clásica de los *tests*. Dentro de ellos se presentaron notables avances como los métodos lineales de Tucker, Levine y Braun-Holland, y los métodos equipercentiles encadenado y suavizado. Los cuales tienen como objetivo mejorar la calidad de la equiparación.

Según Arce-Ferrer y Backhoff, citados por Antillón et al. (2006), en un estudio comparativo de los métodos lineal y equipercentil demostraron que el método lineal ignora la variación de la dificultad de un examen a lo largo de la escala de puntajes, es decir mantiene constante la tendencia de la dificultad al realizar la equiparación. Mientras que el método equipercentil no, lo que le permite mostrar cambios en la dificultad a lo largo de la distribución de los puntajes. Además, señalan que ambos métodos presentan mayores discrepancias en el ajuste de valores extremos en las puntuaciones.

El objetivo general de la investigación fue comparar la calidad de ajuste del método equipercentil respecto al método lineal a través del error de equiparación. Para ello, se postula la hipótesis de que el método equipercentil proporciona una tasa de error de ajuste menor que la proporcionada por el método lineal.

La hipótesis se sometió a prueba mediante la evaluación de las puntuaciones en el examen de admisión a la Universidad Nacional Agraria La Molina (UNALM), obtenidas de los postulantes a esta casa de estudios. Los datos utilizados corresponden a los resultados del examen de admisión en los concursos ordinarios 2016-I y 2016-II.

Los objetivos específicos de la investigación cuando se realizó la equiparación de puntuaciones del examen 2016-I al 2016-II con los métodos lineal y equipercentil fueron los siguientes. El primer objetivo fue determinar si los exámenes de admisión 2016-I y 2016-II presentaron la misma dificultad a lo largo de la escala de puntajes en las seis áreas de evaluación de los exámenes. Mientras que el segundo objetivo fue determinar si las áreas de evaluación presentaron la misma dificultad a lo largo de la escala de puntajes en los dos exámenes de admisión.

II. REVISIÓN DE LITERATURA

2.1 Teoría clásica de los test

2.1.1 Definición

En el campo de la psicometría, la teoría de los *test* es aquella que proporciona modelos matemáticos para las puntuaciones de pruebas tomadas a sujetos. Estos modelos sintetizan la puntuación de un sujeto al que se le aplicó un *test* en dos términos: la puntuación verdadera, aquella que mide realmente su habilidad en la prueba, y el error de medición.

Spearman (1904) presentó el primer modelo psicológico que involucraba el error de las medidas realizadas en la aplicación de pruebas. Este modelo se conoce también como "Modelo Lineal de Puntuaciones" o "Modelo Clásico Lineal" y dio origen a la teoría clásica de los *test*.

Chacón y Antonio (2008) comentan que la teoría clásica de los *test* se mantiene vigente ya que aún se producen manuales y trabajos de investigación en revistas reconocidas donde se trabaja del modelo clásico lineal.

2.1.2 El modelo clásico lineal

Meneses et. al (2013), comentan que el modelo estadístico propuesto por Spearman (1904) sistematizado por Gulliksen (1950) y reformulado por Lord y Novick (1968) utiliza tres conceptos fundamentales: la puntuación empírica, que representa a la puntuación obtenida por un sujeto al que se le aplicó un *test*, la puntuación verdadera, que es la puntuación real (libre de cualquier tipo de error) por este sujeto, y el error aleatorio de medida, que va asociado a las mediciones de los *test*

La puntuación empírica (X) se expresa en dos componentes aditivos: la puntuación verdadera (V) y el error aleatorio de medida (e):

$$X = V + e \tag{2.1}$$

Por su parte, Chacón y Antonio (2008), hacen hincapié que en el modelo presentado anteriormente "la puntuación empírica no tiene por qué coincidir con la puntuación real, ya que durante la toma del *test* el sujeto es afectado por múltiples factores no controlados que inciden en su conducta". Esto hace de entendimiento que es poco posible encontrar el cero como un valor factible para el error aleatorio de medición.

El modelo clásico lineal tiene como objetivo estimar el error de medición, ya que con ayuda de los datos de la puntuación empírica se puede realizar la estimación de la puntuación verdadera de un sujeto. (Muñiz, 1996)

2.1.3 Supuestos del modelo

- **Primer supuesto**

Muñiz (1996) para hallar el valor de la puntuación verdadera de un *test*, partió de la idea de medir un *test* infinitas veces y luego obtener una media de las puntuaciones empíricas; esto, en un sentido estricto, no es aplicable. Sin embargo, define la idea basándose en la esperanza matemática tal como se muestra en la siguiente expresión:

$$V = E(X) \tag{2.2}$$

En la ecuación (2.2) se establece que la puntuación verdadera es la esperanza matemática de la puntuación empírica.

- **Segundo supuesto**

Muñiz (1996) establece como segundo supuesto que no existe correlación entre los errores de medida y sus puntuaciones verdaderas. Esto señala que al aplicar un *test* la puntuación verdadera de un sujeto no debe estar sistemáticamente asociada al tamaño del error cometido al medirlo. Esto suprime, por ejemplo, la hipótesis de que los

sujetos que poseen puntajes verdaderos altos obtendrán mayores errores o que obtendrán menores errores. De la misma manera ocurrirá con los sujetos que posean puntajes verdaderos bajos.

$$\rho_{ve} = 0 \tag{2.3}$$

- **Tercer supuesto**

Meneses et. al (2013), indican que si se aplican dos *test* (*i* y *j*) a un grupo de sujetos, los errores de medida obtenidos con cada uno de ellos (e_i y e_j) no están correlacionados. Por su parte Chacón y Antonio (2008) hacen hincapié en que, si un mismo *test* se vuelve a tomar nuevamente a los mismos sujetos, los errores en estas dos mediciones tampoco están correlacionados.

$$\rho_{e_i e_j} = 0 \tag{2.4}$$

En síntesis, los errores de medida de dos *test* diferentes o de uno solo tomado en ocasiones distintas son aleatorios. Por lo que se espera no haya correlación en ellos.

2.1.4 Índice de confiabilidad

El índice de confiabilidad o fiabilidad es la correlación lineal entre las puntuaciones empíricas y verdaderas de un *test*. Según Burga (2006), la importancia de la confiabilidad está en estimar que tan bien representan las puntuaciones observadas a las verdaderas. De tal manera que mientras más fuerte sea la relación lineal entre ambas mejor será la representación.

En la ecuación (2.5) se muestra la fórmula para obtener la correlación entre las puntuaciones observadas y verdaderas.

$$\rho_{XV} = \frac{\sigma_{XV}}{\sigma_X \sigma_V} \tag{2.5}$$

Dado que la estimación de esta correlación no es directa es posible hallarla partiendo de su cuadrado.

$$\rho_{XV}^2 = \frac{\sigma_V^2}{\sigma_X^2} \tag{2.6}$$

El coeficiente de confiabilidad toma valores de 0 a 1. De tal forma que cuanto más se acerque el valor del coeficiente a 1, más confiable será la prueba.

Aliaga (2018) comenta que como no es posible hallar las puntuaciones verdaderas de los sujetos, para el cálculo del coeficiente de confiabilidad se recurren a diversos métodos.

- **Método de formas equivalentes**

Consiste en aplicar dos formas de *test* equivalentes o paralelas al mismo grupo de individuos. Una vez obtenidos los puntajes, el coeficiente de confiabilidad será equivalente al coeficiente de correlación de Pearson obtenido con los puntajes de ambas formas de los *test*.

- **Método del test-retest**

Este método involucra aplicar el mismo test en dos oportunidades a una misma muestra de sujetos, en un lapso de tiempo determinado por el ejecutor. Con los resultados, se calcula el coeficiente de correlación de Pearson y este será el equivalente al coeficiente de confiabilidad.

- **Método de la división por mitades emparejadas**

También denominado *split half method*, trata de aplicar el *test* una sola vez a una muestra de sujetos. Luego, para obtener los puntajes se califican las preguntas pares e impares por separado y se obtiene el coeficiente de correlación ajustado de Pearson – Brown. El valor obtenido por este coeficiente será el equivalente al coeficiente de confiabilidad.

El coeficiente de Pearson – Brown se calcula usando la siguiente expresión:

$$R_{xx} = \frac{2r_{x_1x_2}}{1+r_{x_1x_2}} \tag{2.7}$$

Donde:

$r_{x_1x_2}$: es la correlación entre las puntuaciones obtenidas en las dos mitades del *test*.

- **Método de la equivalencia racional**

Este método se basa en analizar la estructura de un *test*. Debido a que este se encuentra conformado por un conjunto de ítems, son estos ítems considerados "*test* paralelos". Con los resultados de aplicar el *test* a una muestra de sujetos, mediante una ecuación se obtiene la medida para el coeficiente de confiabilidad. Existen varias propuestas para obtener el coeficiente de confiabilidad como el KR_{20}, KR_{21}, y el Alfa de Cronbach, siendo este último uno de los más utilizados.

El Alfa de Cronbach

Según Barbero et. al (2015), el coeficiente Alfa de Cronbach (1951) es un indicador que mide la consistencia interna de un *test*. Este coeficiente representa la confiabilidad de un *test* basándose en el número de ítems y en la proporción de la varianza total del *test* debida a la covariación de los ítems. De tal forma que, si mayor es la covarianza entre los ítems también lo será la confiabilidad del *test*.

El coeficiente de Alfa de Cronbach puede ser calculado con la siguiente expresión.

$$\hat{\alpha} = \frac{k}{k-1}\left(1-\frac{\sum S_j^2}{S_x^2}\right) \tag{2.8}$$

Donde:

k: número de ítems en el *test*.

S_j^2: es la varianza del j-ésimo ítem.

S_x^2: es la varianza de las puntuaciones totales en el *test*.

El coeficiente Alfa de Cronbach puede tomar distintos valores, donde un valor cercano a 1 indicará una alta confiabilidad del *test*. En el cuadro N°1 se muestra un criterio propuesto por George y Mallery (2003) para clasificar la confiabilidad de este indicador.

Cuadro 1: Clasificación del Alfa de Cronbach

Alfa de Cronbach	Clasificación
>0.9	excelente
>0.8	bueno
>0.7	aceptable
>0.6	cuestionable
>0.5	pobre
<0.5	inaceptable

Fuente: George y Mallery (2003)

2.1.5 Validez del instrumento

Barbero et. al (2015) dicen que la validez de un instrumento es el grado en que un *test* mide aquello que pretende medir. Haciendo énfasis en la relación que hay entre el *test* y el constructo definido para el mismo.

Además, señala que el concepto de validez no ha cambiado conforme al tiempo. Pero que la manera de operacionalizar la validez sí, ya que han surgido herramientas estadísticas que han proporcionado distintos indicadores para expresar la validez de un *test*.

La validez puede clasificarse en tres tipos: predicción, constructo y contenido. Sin embargo, un *test* no necesariamente debe poseer los tres aspectos ya que depende del fin con el cual se ha construido. Aiken (1996) citado por Burga (2003) señala que una prueba puede tener los tres tipos de validez, dependiendo de los propósitos específicos con los que se elaboró y de su población objetivo.

- **Validez de predicción**

Burga (2003) define a la validez predictiva como el grado de eficacia con el que se predice una variable criterio a partir de las puntuaciones de un *test*. Además, comenta que el método para calcularlo se basa en la correlación entre los puntajes de un *test* y los obtenidos en el criterio de interés.

Barbero et. al (2015) hacen referencia a las diversas aplicaciones de *test* donde suele ser útil la validez predictiva. Tales como la selección, clasificación o colocación de personal en puestos de trabajo, ya que es posible obtener luego de un tiempo su rendimiento en un programa de formación, trabajo, entre otros. También indican que lo anterior puede traer dificultad y un costo elevado al recoger los datos, pero que es posible recogerlos de forma simultánea, de ser así la validez se denominaría concurrente

- **Validez de constructo**

Barbero et. al (2015) dan importancia a este tipo de validez, ya que se basa en la adecuación de los ítems elegidos como indicadores del constructo (variable latente inobservable). La validación del constructo es el proceso que permitirá obtener evidencia acerca de la capacidad del *test* para medirlo.

Yela (1984) citado por Barbero et. al (2015) señala que la validez del constructo garantiza científicamente lo que este pretende medir en el *test*, si es una variable aceptable, ya que su concepto tendrá un sustento lógico dentro del sistema teórico de la psicología.

- **Validez de contenido**

Aiken (1996) citado por Burga (2003) señala que la validez de contenido se expresa como la medida en qué los ítems de un *test* representan un área de habilidades y comprensiones que la prueba mide.

Por su parte, Barbero et. al (2015) hacen hincapié en que no solo la representatividad de los ítems es importante, sino que también su relevancia. Señalan que la relevancia

involucra una clara y exhaustiva especificación de todas las posibles conductas observables representativas del constructo. Mientras que la representatividad hace referencia a que todas las conductas estén representadas en el *test*.

Si bien la validez de contenido debe hacerse para todo tipo de *test*, ya que los ítems deben estar representados en el dominio de contenido, según lo indican Hernández et. al (2014) citados por Burga (2003). Las pruebas de rendimiento académico y las de ámbito educativo, que se utilizan para saber el grado en que los sujetos dominan un campo de conocimiento, son predilectas para hacer un análisis de contenido debido a la facilidad en la especificación del dominio (campo de conocimiento).

2.1.6 Análisis de ítems

Barbero et. al (2015) explican que el análisis de ítems es un proceso donde los ítems de un *test* son evaluados y examinados críticamente con el fin de identificar y reducir las fuentes de error, aleatorio y sistemático, con el objetivo de eliminar en un futuro próximo aquellos que no reúnen una garantía psicométrica. También señalan que este proceso inicia desde la redacción de ítems, teniendo en cuenta una cierta cantidad de directrices. Meneses et. al (2013) especifican las siguientes características en los ítems al momento de su construcción: elección del contenido, expresión del contenido y construcción de las opciones.

Una vez terminada la elaboración de ítems se obtiene el instrumento final (*test*) el cual, una vez aplicado a los sujetos pasa a ser examinado. Barbero et. al (2015) señalan que el análisis se puede realizar cuantitativamente o mediante juicio de expertos. Siendo específicos para instrumentos que miden variables cognitivas tales como aptitudes, rendimiento, entre otros. Para estas pruebas donde hay respuestas correctas e incorrectas se han desarrollado estadísticos que evalúan su calidad.

La literatura sobre el análisis de ítems es abundante y la clasificación general para los estadísticos abarcan dos grandes ramas: la dificultad y discriminación.
Para la explicación de los índices se utilizarán los conceptos brindados por Barbero et. al (2015).

- **Índice de Dificultad**

Es la proporción de sujetos que han respondido correctamente a un ítem.

$$ID = \frac{A}{N} \tag{2.9}$$

Donde:
A: número de sujetos que aciertan el ítem.
N: número de sujetos evaluados.

Este índice oscila entre 0 y 1. Donde 0 indica que ningún sujeto ha respondido al ítem, por lo que se interpreta que este es difícil. Mientras que 1 indica que todos lo respondieron correctamente, lo que clasificaría al ítem como fácil. En general, se descartarán aquellos ítems que posean índices de dificultad extremos ya que no contribuyen a diferenciar entre sujetos con distinto nivel de conocimiento.

Existen diversas escalas para clasificar a los ítems según su índice de dificultad, hay autores como Backhoff et. al (2000) que utilizan porcentajes del 5% inferior y superior, y otros como Allen y Yen (1979) que involucran un cálculo para el límite según el número de alternativas que tiene cada ítem. Las diferencias radican en la exhaustividad para declarar a un ítem como fácil o difícil. Sin embargo, cuando se trata de la clasificación prioritaria, la que involucra a los ítems que brindan mayor información en la discriminación de sujetos, estos autores coinciden en valores que van entre 0.3 y 0.7 aproximadamente.

- **Índice de Discriminación**

La formulación de este índice se basa en el poder de discriminación que tiene un ítem entre sujetos con mayor y menor competencia. De tal forma que, si un ítem no logra diferenciar a estos sujetos sería candidato a eliminarse.

Existen diversas propuestas para evaluar la discriminación de un ítem. Una de ellas es la basada en la correlación entre las puntuaciones obtenidas en el ítem con las obtenidas en *test*, conocido como índice de correlación biserial-puntual.

Índice de correlación biserial-puntual

Cuando el ítem es una variable dicotómica y el puntaje del *test* es continuo la correlación entre ambos se obtiene usando la siguiente expresión.

$$r_{bp} = \frac{\bar{X}_A - \bar{X}_T}{S_x}\sqrt{\frac{p}{q}} \qquad (2.10)$$

Donde:

$\bar{X}_A$: puntaje medio en el *test* de los sujetos que aciertan el ítem.

$\bar{X}_T$: puntaje medio del *test*.

S_x : desviación típica del puntaje total obtenido con los sujetos evaluados en el *test*

p: proporción de sujetos que aciertan el ítem

q: proporción de sujetos que no aciertan el ítem

De forma similar a como sucede con el índice de dificultad, existen propuestas para clasificar la discriminación de un ítem que no varían mucho. A continuación, se muestra la utilizada por Ortiz et. al (2015) que hace referencia a los estándares internacionales.

Cuadro 2: Clasificación del índice de correlación biserial

Índice de correlación biserial puntal (r_{bp})	**Clasificación**
<0	discrimina negativamente
[0 -0.15>	discrimina pobremente
[0.15 – 0.25]	discrimina regularmente
<0.25 – 0.35>	buen poder discriminativo
>0.35	excelente poder discriminativo

Fuente: Ortiz et. al (2015)

2.2 Equiparación de puntuaciones

2.2.1 Definición

Según Angoff (1984), citado por Navas (2000), la equiparación de puntuaciones es un proceso fundamental cuando se trabaja con distintos instrumentos de medida, ya que representa el medio básico para poder garantizar una comparación adecuada de las puntuaciones obtenidas en ambas pruebas. Equiparar consiste en derivar puntuaciones equivalentes para poder comparar las puntuaciones obtenidas en distintos *test*, que deben medir el mismo constructo o característica.

Por su parte, Kolen y Brennan (2004) indican que la equiparación es un proceso estadístico que se usa para ajustar las puntuaciones en versiones de pruebas para que estas puedan ser intercambiables. La equiparación ajusta las diferencias en dificultad de las versiones de pruebas, las cuales que deben ser construidas con similares condiciones. Gempp (2010) hace hincapié en las aplicaciones de la equiparación de puntuaciones: "Aunque hay muchas situaciones que pueden requerir de métodos de equiparación, las más habituales son cuatro: equiparación de cuadernillos de una misma prueba, equiparación de resultados entre años en una medición estandarizada, desarrollo de cuadernillos de prueba equivalentes para evaluar intervenciones educativas y equiparación de ítems de un banco".

Finalmente, Matus et al. (2012) señalan las condiciones de los *test* para aplicar la equiparación: "Si las pruebas comparten un mismo marco de referencia, es decir miden el mismo fenómeno o constructo, con la misma confiabilidad y fueron construidas bajo las mismas especificaciones técnicas (número de preguntas, tipo de preguntas, etc.) se dice que son intercambiables. En este caso se habla de *equating* o equiparación para determinar la equivalencia de los puntajes".

2.2.2 Propiedades de la equiparación

- **Simetría**

Según Lord, citado por Kolen y Brennan (2004), la propiedad de simetría indica lo siguiente: una función utilizada para transformar una puntuación de la versión X a la escala de la versión Y de una prueba es la inversa de la función utilizada para

transformar una puntuación de la versión Y a la escala de la versión X. Esta propiedad es necesaria para que una función sea considerada una función de equiparación.

- **Similitud de especificaciones**

Kolen y Brennan (2004), señalan que las versiones de las pruebas deben construirse con el mismo contenido y especificaciones estadísticas para hacer la equiparación. De lo contrario, independientemente de los métodos estadísticos usados, los puntajes no pueden utilizarse de forma intercambiable.

- **Equidad**

Lord (1980), citado por Kolen y Brennan (2004), propuso una propiedad equitativa que define lo siguiente para dos versiones de prueba X e Y:

$$G^*[eq_Y(x) \mid \tau] = G(y \mid \tau) \qquad \forall \tau \tag{2.11}$$

Donde:

τ : Puntuación verdadera

X: Variable aleatoria que representa la puntuación en la versión X.

x: Puntuación observada en la versión X.

Y: Variable aleatoria que representa la puntuación en la versión Y.

y: Puntuación observada en la versión Y.

G: Distribución acumulada de los puntajes de la versión Y en la población de evaluados.

$eq_Y(x)$: Función de equiparación usada para convertir puntajes de la versión X a la escala de la versión Y.

*G**: Distribución acumulada de $eq_Y(x)$ para la misma población de evaluados.

Según Navas (1996), "las puntuaciones de los *test*, una vez realizada la equiparación son totalmente intercambiables, es decir, después de la equiparación deben ser idénticas las distribuciones condicionales de las puntuaciones en cada *test*, dado un determinado nivel en el rasgo o característica que éstos evalúan".

Lord, citado por Kolen y Brennan (2004), mostró que, bajo condiciones bastante generales la propiedad especificada en (2.11) es posible sólo si la versión X y la versión Y son esencialmente idénticas. Siendo así innecesario realizar la equiparación de ambas versiones.

Por su parte Morris (1982), citado por Kolen y Brenann (2004), sugirió una propiedad de equidad menos restrictiva que la propuesta por Lord, esta propiedad se llama equidad débil (Yen, 1983). Bajo esta propiedad, los evaluados con una puntuación verdadera dada tienen la misma media en las puntuaciones convertidas de la versión X con las puntuaciones de la versión Y. Luego, una equiparación logra la propiedad de equidad débil si:

$$E[eq_Y(x) \mid \tau] = E(y \mid \tau) \qquad \forall \tau \tag{2.12}$$

- **Puntuación observada equivalente**

Angoff (1971), citado por Kolen y Brenann (2004), señaló que en las puntuaciones observadas equivalentes, las características de las distribuciones de puntuaciones son iguales para una población especifica de evaluados. Por ejemplo: en la equiparación lineal, las puntuaciones convertidas de la versión X tienen la misma media y desviación estándar que las puntuaciones de la versión Y.

- **Invarianza**

Según Kolen y Brenann (2004), la propiedad de invarianza se cumple cuando la ecuación de equiparación es la misma, independientemente del grupo de evaluados usados para derivarla. Por ejemplo, la propiedad de invarianza es válida si se encuentra la misma ecuación de equiparación para hombres y mujeres.

Por otro lado, Gemmp (2010) señala que "la transformación debe ser invariante para cualquier subpoblación de evaluados. Esto supone, en el fondo, que la equiparación no está sesgada (no produce resultados distintos) para diferentes grupos de interés (e. g. etnias, género u otras). En la práctica, esta propiedad también puede evaluarse empíricamente, simplemente subdividiendo la población en varios grupos, estimando

las funciones de equiparación correspondientes, y finalmente verificando si estas no difieren entre sí".

Existe la posibilidad de que algunas propiedades no se cumplan. Según Pacheco-Villamil (2007): "los supuestos necesarios para un proceso de equiparación frecuentemente no se cumplen debido a que la mayoría de pruebas no se diseñan para ser comparadas".

2.2.3 Diseños de equiparación

Kolen y Brenann (2004), indican que existe una gran variedad de diseños de equiparación. El grupo de evaluados debe ser representativo de la población a la cual debería aplicarse la prueba. La elección de un diseño implica cuestiones prácticas como el tiempo de evaluación, la cantidad de individuos a ser evaluados, el tipo de prueba (con ítems comunes o no), etc. En la figura N°1 se muestran cuatro diseños de uso común, donde la versión X es considerada la versión nueva de una prueba Y ya existente. El objetivo es equiparar las puntuaciones de X a la escala de Y.

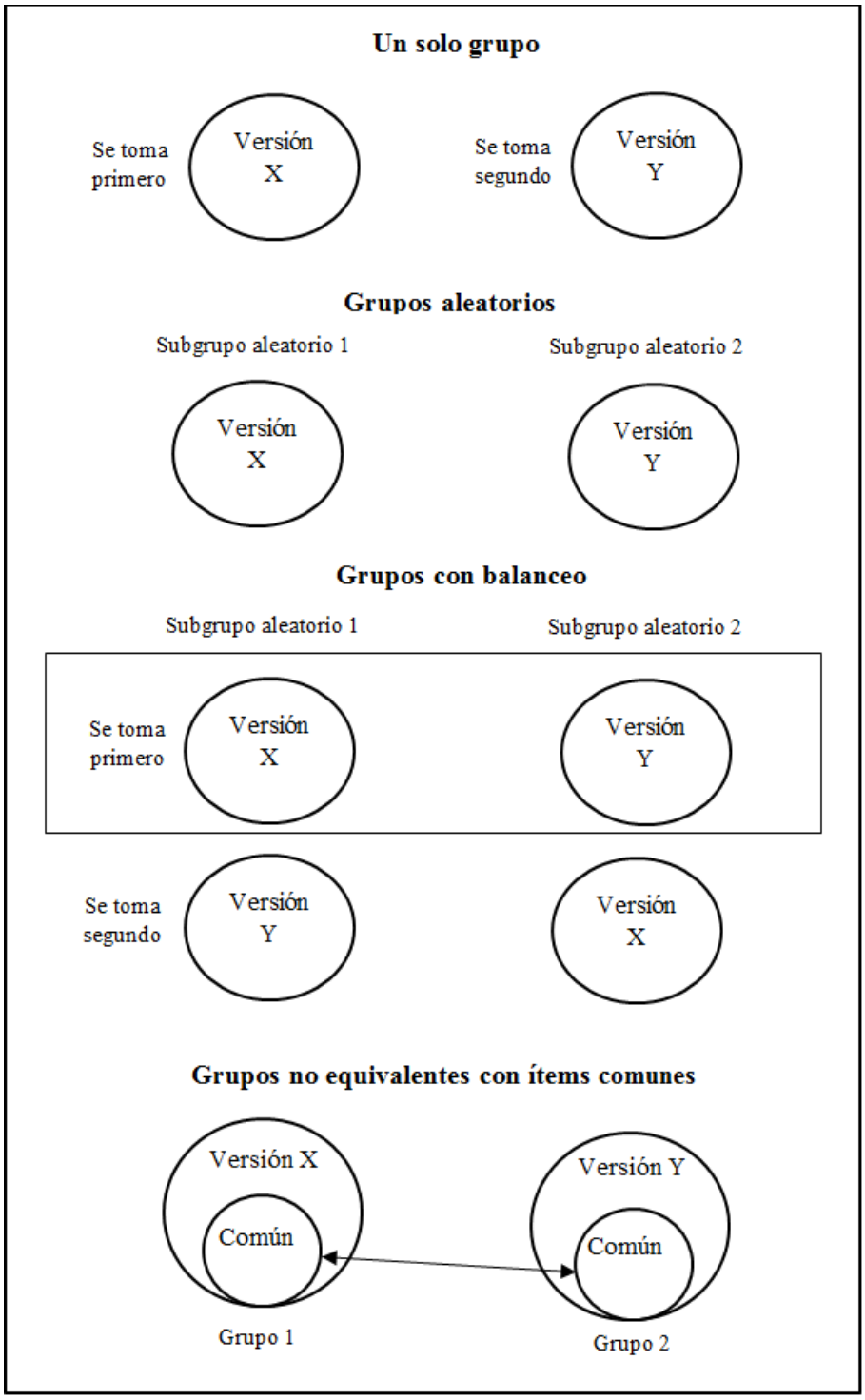

Figura 1: Diseños comunes en equiparación

Fuente: Adaptación de Kolen y Brenann (2004)

La clasificación que presentan estos autores se muestra a continuación:

- **Diseño de un solo grupo**

En este diseño los sujetos son evaluados con la versión X e Y. Un factor importante que se presenta al aplicar este diseño es la fatiga, ocasionada por tomar una versión (X o Y) y luego la otra. Si se toma la versión X primero a todos los sujetos y luego la versión Y, el efecto de la fatiga sobre el rendimiento ocasionará que la versión Y resulte más difícil que la versión X. En contraparte, si el rendimiento aumenta, es posible que la versión Y sea más fácil que la versión X. Es por esta razón que el uso de este diseño no se suele usar en la práctica.

Por su parte Brenann (2006), citado por Ryan y Brockmann (2011), señala que una ventaja del diseño de un solo grupo es que hay poca duda sobre la disimilaridad en las habilidades de los evaluados en la muestra. De tal forma que se consideran de igual habilidad, en términos técnicos esto se conoce como control sobre la competencia diferencial del evaluado.

- **Diseño de grupos aleatorios o equivalentes**

En este diseño, según Barbero et. al (2015) se extrae de la población dos muestras aleatorias de sujetos, y a cada muestra se le aplica una versión de la prueba (X o Y). Por lo que cada sujeto solo responde a una de las versiones.

Por su parte, Kolen y Brenann (2004), indican otra forma de obtener muestras aleatorias y equivalentes, comentan que se puede hacer uso de un proceso espiral para asignar aleatoriamente las pruebas a los sujetos. En dicho proceso, la versión X y versión Y se alternan antes de entregar las pruebas a los sujetos. De tal forma que, una vez que se entregan las pruebas, el primer evaluado recibe la versión X, el segundo la versión Y, el tercero la versión X, y así sucesivamente. Esto garantiza que una diferencia en el desempeño de ambas pruebas equivale a una diferencia en la dificultad de las mismas.

Una ventaja de este diseño frente a otros es que cada sujeto toma solo una versión de la prueba, minimizando el tiempo de evaluación. Otra ventaja es que se evalúe más de una versión de la prueba, incluyéndola en el proceso espiral. Livingston (2014), comenta que una desventaja es que se requiere una mayor cantidad de sujetos para lograr una mayor precisión, entre 5 a 15 veces más sujetos que usando el diseño de un solo grupo. Otra desventaja involucra la seguridad de la prueba, ya que con frecuencia, se evalúa primero la prueba de referencia y los sujetos que rindieron dicha prueba obtienen cierto conocimiento de las preguntas.

Finalmente, Barbero et. al (2015) señalan que la equivalencia de ambos grupos se basa en la aptitud que mide la prueba. De tal forma que, si se cumple ello, se evitarán sesgos en el proceso de equiparación.

- **Diseño de grupos con balanceo**

El diseño de grupos balanceados, toma como criterio el balanceo en la asignación de las versiones X e Y a los sujetos. Para realizarlo se utilizan dos grupos de individuos del mismo tamaño. En el primer grupo se alternan las pruebas siguiendo el orden versión X-versión Y mientras que en el segundo se alternan las pruebas siguiendo el orden versión Y-versión X. De modo que cuando se entregan las pruebas del primer grupo, el primer evaluado recibe la versión X, el segundo la versión Y, el tercero la versión X, y así sucesivamente. Cuando se entregan las pruebas del segundo grupo, los sujetos complementan ambas evaluaciones. De esta manera se hace comparable las puntuaciones en las versiones X e Y.

En la Figura N°1, se puede observar que si se consideran las versiones que se toman primero en ambos subgrupos, el diseño es similar al de grupos aleatorios. Sin embargo, si se toma uno de los subgrupos, el diseño es similar al de un solo grupo. Por esta razón, Dorans et al. (2010), considera a este diseño como la combinación del diseño de un solo grupo y de grupos aleatorios. Finalmente, es necesario resaltar que la finalidad de este diseño es aprovechar los resultados tanto de las versiones que se toman primeras como segundas en los subgrupos. Ya que el objetivo es evaluar si el efecto de tomar primero la versión X y luego la versión Y es el mismo que al tomar primero la versión Y y luego la versión X. De ser así, la tendencia obtenida en la

equiparación de las versiones tomadas en primer lugar será la misma que la obtenida en las versiones que se tomaron en segundo lugar.

- **Diseño de grupos no equivalentes con ítems comunes**

Navas (1996), comenta que en este diseño cada grupo de sujetos rinde una versión de prueba diferente. Se le conoce también como diseño de anclaje porque los grupos deben de realizar además de la versión de prueba, un *test* de anclaje (o con ítems comunes). Obteniendo por cada sujeto puntuaciones en la versión X o Y de la prueba y la puntuación en el *test* de anclaje. Los resultados de este último se usan con la finalidad de reducir el error de equiparación por haber trabajado con grupos distintos.

El *test* de anclaje puede ser interno o externo. En el anclaje interno, este *test* se encuentra dentro las versiones de la prueba y es usado para contabilizar el puntaje total obtenido. Mientras que en el anclaje externo el *test* es adicional, considerado independiente, con sus propias instrucciones y no es usado para contabilizar el puntaje total obtenido.

Según Herrera (2013), "cuando se utiliza el *test* de anclaje interno, la principal ventaja es la facilidad en la aplicación ya que se encuentra inmerso en la misma prueba, sin embargo, puede tener dificultades de seguridad por exponer las preguntas en repetidas ocasiones. De esta exposición constante, pueden surgir valores atípicos en el análisis pues suelen ser los datos de aquellas preguntas que se han hecho más fáciles o más difíciles en la nueva forma que lo que eran en el grupo de referencia."

2.3 Métodos de equiparación

2.3.1 Método Lineal

- **Definición**

Según Kolen y Brenann (2004), en la equiparación lineal las diferencias en la dificultad de ambas pruebas varían a lo largo de la escala de puntuación. La ecuación lineal permite que la versión X sea más difícil para los evaluados con la versión Y de bajo rendimiento, pero menos difícil para los evaluados de alto rendimiento.

En la equiparación lineal, las puntuaciones de las versiones X e Y son iguales en la distancia respecto a su media dividida por su desviación estándar. Por lo tanto, este método admite unidades de escala y medias diferentes en ambas versiones.

- **Ecuación lineal**

La ecuación de conversión lineal se define igualando las puntuaciones estandarizadas en ambas pruebas, de tal forma que:

$$\frac{x-\mu(X)}{\sigma(X)}=\frac{y-\mu(Y)}{\sigma(Y)} \tag{2.13}$$

Donde:

$\mu(X)$: Media de los puntajes obtenidos con la versión X

$\mu(Y)$: Media de los puntajes obtenidos con la versión Y

$\sigma(X)$: Desviación estándar de los puntajes obtenidos con la versión X

$\sigma(Y)$: Desviación estándar de los puntajes obtenidos con la versión Y

Lo anterior puede expresarse de la siguiente forma:

$$l_Y(x)=y=\sigma(Y)\left[\frac{x-\mu(X)}{\sigma(X)}\right]+\mu(Y) \tag{2.14}$$

Donde $l_Y(x)$ es la ecuación de conversión lineal para convertir las puntuaciones observadas en la versión X a la escala de la versión Y. Al reordenar los términos, una expresión alternativa para $l_Y(x)$ es:

$$l_Y(x)=y=\frac{\sigma(Y)}{\sigma(X)}x+\left[\mu(Y)-\frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \tag{2.15}$$

La expresión (2.15) es una expresión lineal de la forma pendiente (x) + intercepto con:

$$\text{pendiente} = \frac{\sigma(Y)}{\sigma(X)}, \quad \text{y} \quad \text{intercepto} = \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X). \tag{2.16}$$

- **Propiedades**

Según lo presentado en la ecuación (2.12), y reemplazando la función de equiparación por la ecuación lineal se obtiene:

$$\begin{aligned} E[l_Y(x)] &= E\left[\frac{\sigma(Y)}{\sigma(X)}x + \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \\ &= \frac{\sigma(Y)}{\sigma(X)}E(X) + \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \\ &= \mu(Y) \end{aligned} \tag{2.17}$$

La desviación estándar se obtiene de la siguiente manera:

$$\begin{aligned} \sigma[l_Y(x)] &= \sigma\left[\frac{\sigma(Y)}{\sigma(X)}x + \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \\ &= \frac{\sigma(Y)}{\sigma(X)}\sigma(X) \\ &= \sigma(Y) \end{aligned} \tag{2.18}$$

Finalmente se demuestra que la media y desviación estándar de los puntajes de la versión X equiparados a la escala de la versión Y son iguales a la media y desviación estándar, respectivamente, de las puntuaciones de la versión Y.

Hay que resaltar que la ecuación (2.15) no es una regresión lineal. Ya que para la regresión lineal los términos $\sigma(Y)/\sigma(X)$ se multiplican por la correlación entre X e Y. Además, una ecuación de regresión lineal no califica como una función de

equiparación porque la regresión de X sobre Y es diferente de una regresión de Y sobre X, a menos que el coeficiente de correlación sea 1. Por esta razón, las ecuaciones de regresión no pueden, en general, ser usadas como funciones de equiparación. La comparación entre la regresión lineal y la ecuación lineal se ilustra en la figura N°2, la regresión de Y sobre X es diferente de la regresión de X sobre Y. Se puede observar que sólo hay una relación de equiparación lineal representada gráficamente en la figura. Esta relación se puede utilizar para transformar los puntajes de la versión X a escala de la versión Y o para transformar los puntajes de la versión Y a la versión X.

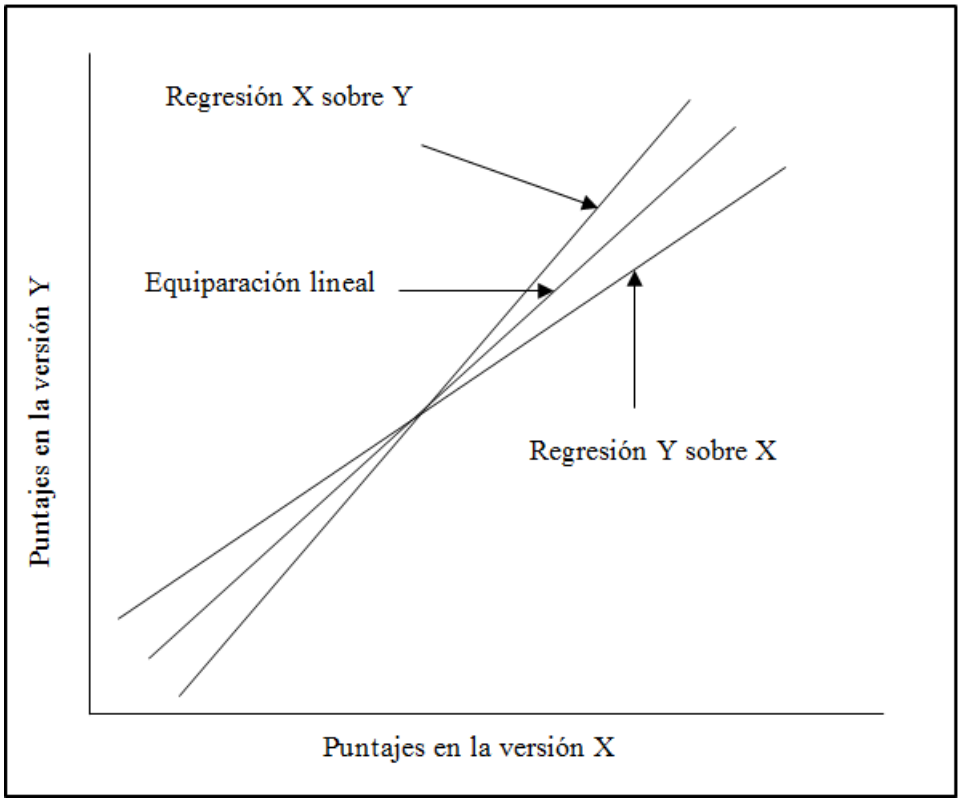

Figura 2: Diferencias entre una regresión lineal y una ecuación lineal

Fuente: Kolen y Brenann (2004)

2.3.2 Método Equipercentil

- **Definición**

Según Kolen y Brenann (2004), en el método equipercentil se usa una curva para describir las diferencias en dificultad de una versión a otra, lo que hace a la función equipercentil más general que la ecuación lineal. La función equipercentil refleja las dificultades de las versiones X e Y en puntajes intermedios.

En la función equipercentil, la distribución de los puntajes de la versión X convertidos a escala de la versión Y tienen la misma distribución que los puntajes de la versión Y en la población. La equiparación usando la función equipercentil se desarrolla mediante la identificación de los puntajes de la versión X que tienen los mismos rangos percentiles que los puntajes de la versión Y.

- **Ecuación equipercentil**

Braun y Holland (1982), citado por Kolen y Brenann (2004), define que $e_Y(x)$ es la ecuación equipercentil si la función de distribución acumulativa de los puntajes de X convertidos a escala de la versión Y es igual a la función de distribución acumulativa de los puntajes de Y. De tal forma que:

$$G^* = G$$
$$e_Y(x) = G^{-1}[F(x)] \qquad (2.19)$$

Donde:

x: Puntaje observado en la versión X.

F: Función de distribución acumulada de los puntajes obtenidos con la versión X.

G: Función de distribución acumulada de los puntajes obtenidos con la versión Y.

G^{-1}: Inversa de la función de distribución acumulada G.

$e_Y(x)$: Función de equiparación simétrica usada para convertir los puntajes de la versión X a escala de la versión Y

*G**: Función de distribución acumulada de e_Y en la misma población. Es decir la función de distribución acumulada de los puntajes de la versión X convertidos a escala de la versión Y.

La ecuación obtenida en (2.19) satisface la propiedad de simetría. Sea e_X una función de equiparación simétrica usada para convertir los puntajes de la versión Y a escala de la versión X y una función distribución acumulada F* de la función e_X en la población. Es decir, la función de distribución acumulada de los puntajes de Y convertidos a escala de la versión X.

Entonces se obtiene que:

$$e_X^{-1}(x) = e_Y(x) \quad y \quad e_Y^{-1}(y) = e_X(y) \tag{2.20}$$

Finalmente:

$$e_X(y) = F^{-1}[G(y)] \tag{2.21}$$

La ecuación (2.21) muestra la función equipercentil para convertir puntajes de la versión Y a escala de la versión X. En esta ecuación F^{-1} representa la función de distribución acumulada inversa de F.

Si las variables *X* e *Y* fueran de naturaleza discreta y se definen ambas variables, por ejemplo, como número de aciertos, la función equipercentil no es la misma que la presentada en (2.19). Para ello será necesario definir lo siguiente:

Sean K_x el número de ítems en la versión X de una prueba y *X* una variable aleatoria que representa los puntajes obtenidos en la versión X teniendo como posibles valores 0,1,…, K_x. Entonces $f(x)$ es una función de densidad discreta Para $X=x$. De tal manera que:

$$
\begin{aligned}
&f(x) \geq 0, \ \forall \ x = 0,1,...,K_X; \\
&f(x) = 0, \ \text{otros casos y} \\
&\sum f(x) = 1
\end{aligned}
$$

(2.22)

Luego F(x) es la función de distribución acumulada discreta. Esto es, F(x) es la proporción de evaluados en la población que tiene un puntaje menor o igual a x. Por lo tanto:

$$
\begin{aligned}
&0 \leq F(x) \leq 1, \ \forall \ x = 0,1,...,K_X; \\
&F(x) = 0, \ \forall \ x < 0; \ \text{y} \\
&F(x) = 1, \forall \ > K_X
\end{aligned}
$$

(2.23)

Si se considera un valor x no entero y se define x^* un valor entero más próximo a x tal que $x^* - 0.5 \leq x < x^* + 0.5$. Entonces el rango percentil será:

$$
P(x) = \begin{cases}
100\{F(x^*-1) + [x - (x^* - 0.5)][F(x^*) - F(x^*-1)]\}, & -0.5 \leq x < K_X + 0.5 \\
0, & x < -0.5 \\
100, & x \geq K_X + 0.5
\end{cases}
$$

(2.24)

La función inversa del rango percentil, también conocida como la función percentil, se simboliza como P^{-1}. De ella, se proponen dos funciones percentil alternas que brindan el mismo resultado, a menos que algunas probabilidades sean cero. Dado un rango percentil, la función inversa es usada para encontrar la puntuación correspondiente a dicho rango. Para encontrar esta función, es necesario resolver la ecuación (2.24) para x. Específicamente, para un rango percentil P^* dado, el percentil es el siguiente.

$$x_U(P^*) = \begin{cases} P^{-1}[P^*] = \dfrac{P^*/100 - F(x_U^* - 1)}{F(x_U^*) - F(x_U^* - 1)} + (x_U^* - 0.5), & 0 \le P^* < 100 \\ K_X + 0.5, & P^* = 100 \end{cases} \tag{2.25}$$

En la ecuación (2.25), para $0 \le P^* < 100$, x_U^* es el menor puntaje entero con un porcentaje acumulado $[100F(x)]$ que es mayor a P^*. La otra expresión alternativa para el percentil es:

$$x_L(P^*) = \begin{cases} P^{-1}[P^*] = \dfrac{P^*/100 - F(x_L^*)}{F(x_L^* + 1) - F(x_L^*)} + (x_L^* + 0.5), & 0 < P^* \le 100 \\ -0.5, & P^* = 0 \end{cases} \tag{2.26}$$

En la ecuación (2.26), para $0 < P^* \le 100$, x_L^* es el mayor puntaje entero con un porcentaje acumulado $[100F(x)]$ que es menor a P^*. Si $f(x)$ es distinto de cero para todos los puntajes 0,1,…, K_X, entonces $x = x_U = x_L$ y cualquier expresión puede ser usada. En general, se suele suponer que $f(x)$ es distinto de cero para el rango de puntajes 0,1,…, K_X. Por esta razón, y para simplificar expresiones se considerará la ecuación (2.25), de tal forma que $x_U = x$.

Finalmente, tomando en consideración lo anterior será necesario que la función equipercentil encuentre un puntaje en la versión Y que tenga el mismo rango percentil que un puntaje en la versión X. Si se define *y* como un puntaje en la versión Y de una prueba, K_Y como el número ítems en la versión Y, $g(y)$ la densidad discreta, $G(y)$ la función de distribución acumulada, $Q(y)$ el rango percentil y Q^{-1} la función inversa del rango percentil de *y*. Entonces la función equipercentil para un puntaje de la versión X equivalente a un puntaje de la versión Y es:

$$e_Y(x) = y = Q^{-1}[P(x)], \quad -0.5 \le x \le K_X + 0.5 \tag{2.27}$$

La ecuación (2.27) indica que para encontrar una función equipercentil equivalente para un puntaje de la versión X a escala de la versión Y, primero se debe encontrar el rango percentil de *x* en la distribución de puntajes de la versión X. Luego encontrar el puntaje *y* que tiene el mismo rango percentil en la distribución de puntajes de la versión Y. La ecuación (2.27) también es simétrica, de tal manera que $e_X(y) = P^{-1}[Q(y)]$

Otra forma de expresar la función equipercentil es la siguiente:

$$e_Y(x) = Q^{-1}[P(x)] = \begin{cases} \dfrac{P(x)/100 - G(y_U^* - 1)}{G(y_U^*) - G(y_U^* - 1)} + (y_U^* - 0.5), & 0 \le P(x) < 100 \\ K_Y + 0.5, & P(x) = 100 \end{cases} \tag{2.28}$$

- **Propiedades**

La función equipercentil brinda resultados equiparados en el rango $-0.5 \le e_Y(x) \le K_Y + 0.5$. Así esta función tiene la propiedad de que los puntajes equiparados se mantendrán en los valores permitidos para percentiles y rangos percentiles. Esto es una ventaja que tiene el método equipercentil frente al método lineal que puede dar puntajes equiparados fuera del rango de puntuaciones posibles.

Por otro lado, al usar el método equipercentil los puntajes equiparados de la versión X tienen la misma distribución que los puntajes de la versión Y, si los valores de los puntajes son continuos. Sin embargo, cuando los valores de los puntajes son discretos, se encuentran diferencias. Las cuales son más grandes mientras las pruebas son más cortas (con menos ítems) y son más pequeñas cuando las pruebas son más largas (con más ítems).

2.4 Indicador de comparación de métodos

Navas (1996), denomina que el estudio de los errores permite determinar la calidad de la equiparación realizada en dos pruebas.

Por su parte, Kolen y Brenann (2004) señalan que hay dos fuentes de error en la equiparación de puntuaciones. Estas son: el error aleatorio y el error sistemático. Hacen hincapié en que el error aleatorio está presente cuando se usan muestras de poblaciones de evaluados para estimar parámetros (medias, desviaciones estándar, etc.) que están implícitos en la estimación de una función de equiparación. Mientras que el error sistemático proviene de la existencia de un sesgo en la estimación de la función de equiparación, donde una fuente es el incumplimiento de supuestos estadísticos al realizar la equiparación.

En diversas fuentes bibliográficas como las de González y Wiberg (2017), así como Kolen y Brenann (2004) demuestran que los valores correspondientes al error estándar aleatorio son muy similares a los del error estándar sistemático cuando los métodos de equiparación empleados son el lineal y equipercentil, y cuando se utiliza un gran número de repeticiones bootstrap. Por esta razón, para realizar la comparación de métodos basta utilizar los errores estándar aleatorios.

Error estándar aleatorio de equiparación

Según Kolen y Brenann (2004), el error estándar de equiparación es la desviación estándar de las puntuaciones equiparadas en base a repeticiones en el proceso de equiparación para varias muestras de una población de evaluados.

Si se define $\hat{eq}_Y(x_i)$ como una estimación de la función de equiparación para un puntaje en la versión X a escala de la versión Y en una muestra y $E[\hat{eq}_Y(x_i)]$ como el valor esperado equivalente, donde E es la esperanza sobre las muestras aleatorias en la población. Para la estimación usando una muestra, el error de equiparación de un puntaje particular de la versión X se define como la diferencia entre el puntaje equiparado a escala de la versión Y en la muestra y el esperado equivalente. De tal forma que el error de equiparación del puntaje x_i para una ecuación dada es:

$$\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)] \tag{2.29}$$

Si se replica el procedimiento un gran número de veces, de tal forma que para cada repetición la ecuación se obtiene a través de muestras aleatorias de la población de evaluados con las versiones X e Y, respectivamente. La varianza del error de equiparación para un puntaje x_i es:

$$V[\hat{eq}_Y(x_i)] = E\{\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)]\}^2 \tag{2.30}$$

Donde la varianza es calculada sobre repeticiones. El error estándar de equiparación se define como la raíz cuadrada de la varianza del error:

$$se[\hat{eq}_Y(x_i)] = \sqrt{V[\hat{eq}_Y(x_i)]} = \sqrt{E\{\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)]\}^2} \tag{2.31}$$

2.4.1 Error estándar mediante Bootstrap

Kolen y Brenann (2004), y Gonzales y Wiberg (2017) señalan el procedimiento para calcular el error estándar mediante Bootstrap para un método de equiparación usando un diseño de grupos aleatorios:

1. Seleccionar una muestra aleatoria Bootstrap con reemplazo de tamaño n_X de la muestra de n_X evaluados (n_X es el número de evaluados con la versión X de la prueba).
2. Seleccionar una muestra aleatoria Bootstrap con reemplazo de tamaño n_Y de la muestra de n_Y evaluados (n_Y es el número de evaluados con la versión Y de la prueba).
3. Estimar la función de equiparación para x_i usando los datos provenientes de las muestras Bootstrap obtenidas en los pasos 1 y 2, obteniendo $\hat{eq}_{Yr}(x_i)$
4. Repetir los pasos 1 a 3 B veces, obteniendo los estimadores Bootstrap: $\hat{eq}_{Y1}(x_i)$, $\hat{eq}_{Y2}(x_i), \ldots, \hat{eq}_{Yb}(x_i)$
5. El error estándar estimado es:

$$SE(x_i) = \hat{se}_{boot}[\hat{eq}_Y(x_i)] = \sqrt{\frac{\sum_b [\hat{eq}_{Yb}(x_i) - \hat{eq}_Y(x_i)]^2}{B-1}} \tag{2.32}$$

Donde: $\hat{eq}_Y(x_i) = \frac{\sum_b \hat{e}_{Yb}(x_i)}{B}$

Este procedimiento puede ser aplicado para cualquier puntaje x_i. Generalmente, las mismas B muestras Bootstrap son usadas para estimar los errores estándar para todos los valores discretos de entre 0 a K_X, porque el objetivo es la estimación de los errores estándar para todo el rango de valores.

III. MATERIALES Y MÉTODOS

3.1 Materiales

- Una computadora Toshiba Intel Corei5 de 64 bits.
- Una Impresora HP Laser Jet P1102w.
- Dos tintas color negro.
- Dos millares de hojas bond A4.
- Programa R versión 3.6.0
- Paquetes de R: CTT versión 2.3.3, knitr versión 1.2.3 y equate versión 2.0-6

3.2 Descripción del caso

3.2.1 Población

Todos los postulantes a la Universidad Nacional Agraria La Molina en el año 2016 que participaron en los concursos ordinarios de admisión 2016-I y 2016-II.

3.2.2 Muestra

En la investigación no se utilizó una muestra, debido a que la aplicación de las técnicas no involucró un proceso de inferencia.

3.2.3 Identificación de las variables

Las variables identificadas en la aplicación de ambas técnicas son:

- X_1 = Puntaje obtenido en Razonamiento Matemático
- X_2 = Puntaje obtenido en Razonamiento Verbal
- X_3 = Puntaje obtenido en Matemática
- X_4 = Puntaje obtenido en Física
- X_5 = Puntaje obtenido en Química
- X_6 = Puntaje obtenido en Biología

El puntaje se define como el número de aciertos obtenidos. De tal forma que las variables independientes X_1, X_2, X_3, X_4, X_5, X_6 son de naturaleza cuantitativa discreta.

3.3 Metodología de la Investigación

3.3.1 Tipo de investigación

El tipo de investigación fue de carácter descriptivo. En primer lugar, debido a que se describieron las puntuaciones y proporciones de aciertos según las diferentes áreas evaluadas en los exámenes de admisión 2016-I y 2016-II. En segundo lugar, debido a que se analizó cuál de los métodos de equiparación de puntuaciones basados en grupos equivalentes: lineal o equipercentil tuvo una menor tasa de error de ajuste sobre los resultados de los exámenes de admisión, así como las razones de ello.

3.3.2 Diseño de la investigación

El diseño de la investigación fue no experimental de tipo transversal descriptivo, debido a que se trabajó con los resultados obtenidos de los exámenes de admisión 2016-I y 2016-II sin involucrar el seguimiento durante el tiempo. Es decir, un grupo de postulantes de la población rindió el examen de admisión 2016-I y otro el examen 2016-II. Cabe la posibilidad que algunos postulantes hayan rendido ambos exámenes. Sin embargo, esto último no afecta los resultados de aplicar los métodos de equiparación en estudio, ya que de acuerdo con Pacheco-Villamil (2007) y Livingston (2014), las dos versiones del examen de admisión contienen diferentes preguntas y fueron tomadas en los meses de marzo y agosto del año 2016, respectivamente, lo que evita la ventaja que tendría algún postulante que haya rendido el examen de admisión por segunda vez, sobre la información del contenido del primer examen de admisión.

3.3.3 Formulación de las hipótesis

La hipótesis general del presente trabajo de investigación fue la siguiente:

El método equipercentil proporciona una tasa de error de equiparación menor en el ajuste de las puntuaciones del examen de admisión de la UNALM que la proporcionada por el método lineal al utilizar el diseño de grupos equivalentes.

Las hipótesis específicas fueron las siguientes:

- Los exámenes de admisión 2016-I y 2016-II presentaron la misma dificultad a lo largo de la escala de puntajes en las seis áreas de evaluación, cuando se realiza la equiparación de puntuaciones del examen 2016-I al 2016-II con los métodos lineal y equipercentil.
- Las seis áreas de evaluación presentaron la misma dificultad a lo largo de la escala de puntajes, cuando se realiza la equiparación de puntuaciones del examen 2016-I al 2016-II con los métodos lineal y equipercentil.

3.4 Metodología aplicada

Los pasos que se realizaron para llevar a cabo este trabajo se detallan a continuación:

1. Análisis descriptivo.
2. Análisis de confiabilidad y validez.
3. Análisis de ítems.
4. Análisis de las estadísticas básicas de equiparación.
5. Elaboración e interpretación de funciones de equiparación.
6. Análisis de las tablas de equivalencia lineal y equipercentil.
7. Comparación de los resultados obtenidos con el método de equiparación líneal y equipercentil.

IV. RESULTADOS Y DISCUSIÓN

4.1 Análisis Descriptivo

Resultados de los exámenes

En el examen de admisión 2016-I la cantidad de postulantes fue de 2747, donde el 12% (319) ingresó a la universidad y el 88% (2428) no ingresó. Mientras que en el examen de admisión 2016-II la cantidad de postulantes fue de 2119, donde el 15% (325) ingresó a la universidad y el 85% (1794) no ingresó.

Estadísticas básicas sobre los puntajes

En el cuadro N°3 se puede observar el número de ítems y las estadísticas básicas del número de aciertos (mínimo, máximo, media, coeficiente de variabilidad y asimetría) obtenidos de los exámenes de admisión 2016-I y 2016-II según área. Al comparar ambos exámenes se obtuvo que el puntaje mínimo fue de 0 en todas las áreas. Mientras que el puntaje máximo obtenido por área fue similar en los dos exámenes. Respecto a la media del puntaje, esta presentó valores similares en ambos exámenes. También se observó que la media del puntaje fue menor al número de ítems, esto indicó que en promedio los postulantes llegaron a responder correctamente a lo más la mitad de ítems por cada área en ambos exámenes de admisión.

Evaluando el coeficiente de variabilidad del puntaje de aciertos por área, se encontró que en ambos exámenes esta fue similar. Sin embargo, al comparar la variabilidad del puntaje de las áreas dentro de cada examen, se encontró que Física fue la que presentó mayor variabilidad que las demás. Mientras que Razonamiento Verbal obtuvo la menor variabilidad.

Respecto a la asimetría, la distribución del puntaje en ambos exámenes de admisión presentó un comportamiento aproximadamente simétrico, ya que el coeficiente de asimetría se encontró entre -0.5 y 0.5, de acuerdo con Bulmer (1979).

Cuadro 3: Estadísticos del número de aciertos en los exámenes de admisión según área

Examen	Área	ítems	mínimo	máximo	media	cv	asimetría
2016-I	Razonamiento Verbal	20	0.00	18.00	10.89	28.62%	-0.36
	Razonamiento Matemático	14	0.00	14.00	5.83	47.82%	0.15
	Matemática	24	0.00	24.00	9.79	59.66%	0.21
	Física	14	0.00	14.00	4.83	74.23%	0.49
	Química	14	0.00	14.00	5.43	64.29%	0.37
	Biología	14	0.00	13.00	5.17	52.30%	0.28
2016-II	Razonamiento Verbal	20	0.00	17.00	9.68	27.12%	-0.49
	Razonamiento Matemático	14	0.00	13.00	5.28	49.41%	0.27
	Matemática	24	0.00	24.00	8.08	56.40%	0.41
	Física	14	0.00	14.00	4.68	73.40%	0.45
	Química	14	0.00	14.00	5.71	59.33%	0.10
	Biología	14	0.00	13.00	4.49	55.57%	0.30

4.2 Análisis de confiabilidad y validez

En el cuadro N°4 se muestra el indicador Alfa de Cronbach para los exámenes de admisión 2016-I y 2016-II. Se observó que para ambos exámenes el indicador fue muy bueno según la escala propuesta por George y Mallery (2003). Esto señala que la consistencia interna de ambos instrumentos es confiable, es decir se presentó una gran ausencia de errores de medida al tomar ambos exámenes.

Cuadro 4: Indicador de confiabilidad

Examen de admisión	Alfa de Cronbach
2016-I	0.928
2016-II	0.899

Debido a que los exámenes de admisión evalúan conocimientos, la validez de estos instrumentos estuvo a cargo del comité de profesores que elaboraron ambas pruebas.

4.3 Análisis de ítems

Índices de Dificultad

Las figuras N°3 y N°4 presentan un gráfico que muestra los índices de dificultad de cada ítem según examen de admisión. En base a la literatura revisada, los ítems que presentan un índice de dificultad entre 0.3 y 0.7 brindan la mayor información para diferenciar a los postulantes, lo que permite clasificar a los ítems como adecuados. También se muestra un límite de rango aceptable, inferior a 0.15 o superior 0.85, para determinar a los ítems como muy difíciles o muy fáciles, respectivamente.

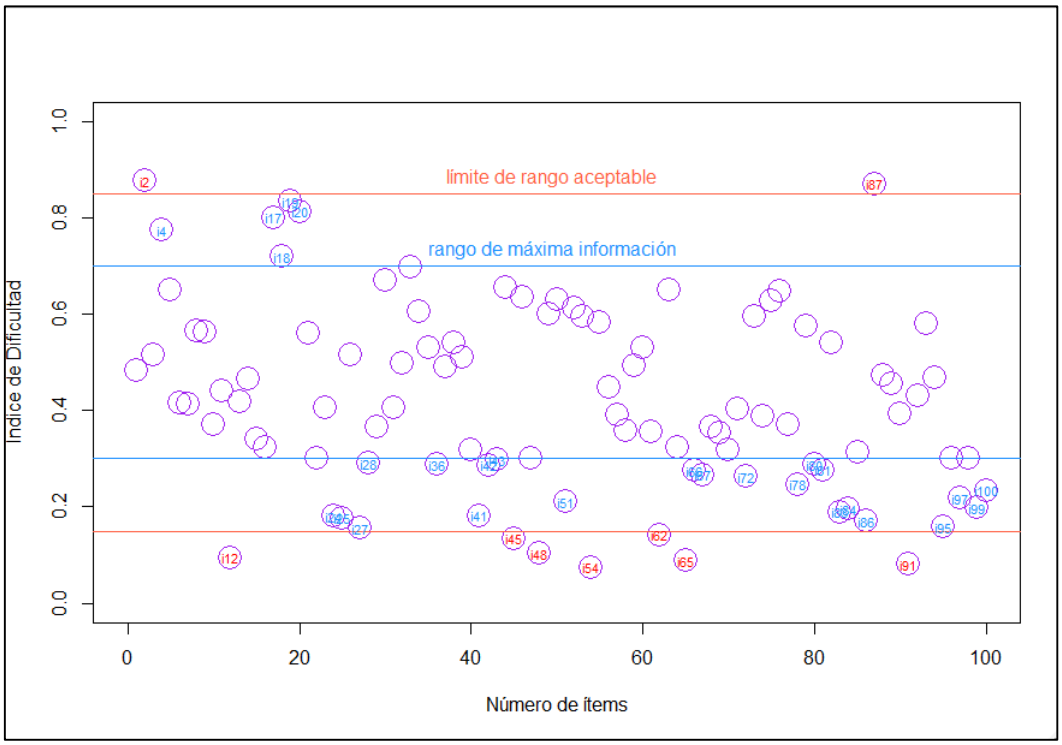

Figura 3: Dificultad de ítems del examen de admisión 2016-I

En el examen de admisión 2016-I, 9 de los 100 ítems fueron considerados extremos debido a que presentaron un índice de dificultad que superó el límite de rango aceptable. De ellos 7 ítems fueron considerados difíciles: 3 ítems de Matemática, 2 de Física, 1 de Razonamiento Verbal y 1 de Biología. Mientras que los 2 restantes fueron considerados fáciles: 1 ítem de Biología y 1 de Razonamiento Verbal. Esto indicó que los ítems señalados no fueron adecuados para diferenciar a los postulantes según su nivel de conocimientos en las áreas mencionadas. Al comparar las áreas según la cantidad de ítems extremos que presentaron,

se encontró que Matemática fue el área que tuvo más ítems extremos y todos ellos fueron considerados como difíciles.

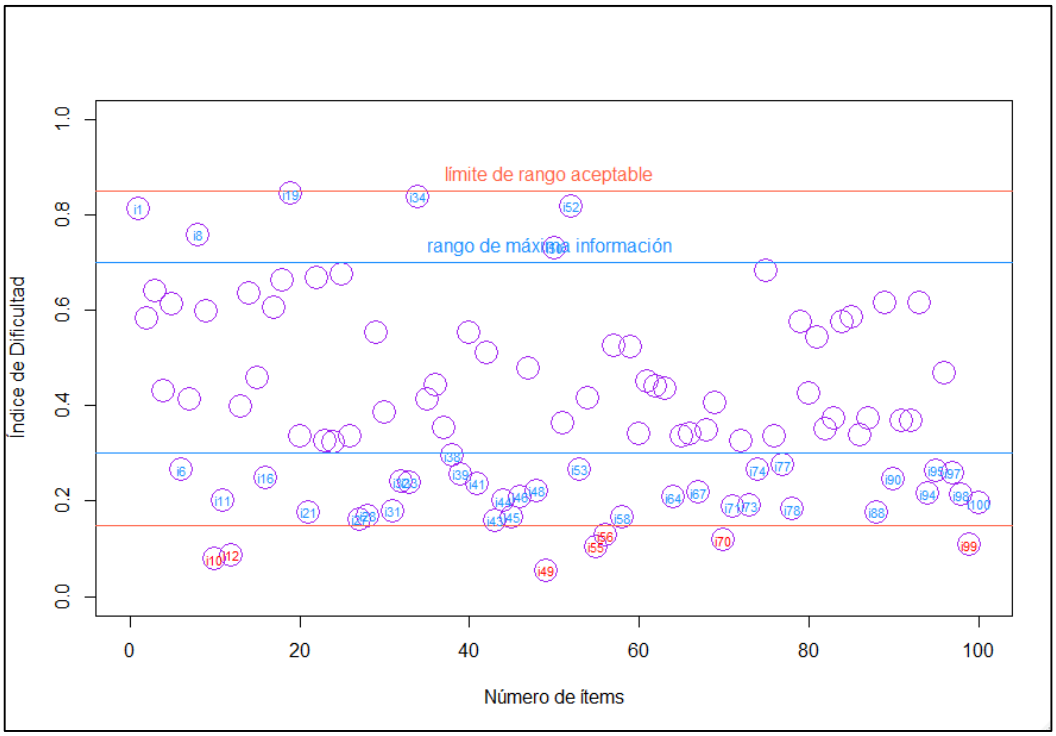

Figura 4: Dificultad de ítems del examen de admisión 2016-II

En el examen de admisión 2016-II, 7 de los 100 ítems fueron considerados extremos. Donde todos fueron difíciles: 3 ítems de Matemática, 2 de Razonamiento Verbal, 1 de Física y 1 de Biología. Estos ítems no fueron adecuados para diferenciar el conocimiento de los postulantes en las áreas señaladas. De las áreas que presentaron ítems extremos en este examen, nuevamente fue el área de Matemática que presentó la mayor cantidad y también fueron considerados difíciles.

Índices de Discriminación

En los cuadros N°5 y N°6, se presentan para los exámenes de admisión 2016-I y 2016-II, los resultados de la evaluación de índices de discriminación y dificultad de los ítems, junto al valor del Alfa de Cronbach que se obtendría si se extrajera el ítem del examen. Los ítems presentados en los cuadros fueron aquellos que tuvieron un valor de correlación biserial-puntual menor a 0.15, es decir con pobre poder discriminativo según Ortiz et. al (2015).

Cuadro 5: Evaluación de ítems del examen de admisión 2016-I

Ítem	Correlación biserial puntual	Índice de Dificultad	Alfa de Cronbach	Área
1	0.121	0.4845	0.9284	Razonamiento Verbal
2	0.1454	0.878	0.928	Razonamiento Verbal
3	-0.0728	0.5166	0.9293	Razonamiento Verbal
6	0.0237	0.415	0.9288	Razonamiento Verbal
8	0.1335	0.5664	0.9283	Razonamiento Verbal
9	0.1146	0.5632	0.9284	Razonamiento Verbal
10	0.0468	0.3713	0.9287	Razonamiento Verbal
13	-0.0165	0.4183	0.929	Razonamiento Verbal
15	-0.0458	0.3404	0.9291	Razonamiento Verbal
16	0.105	0.3236	0.9284	Razonamiento Verbal
17	0.0496	0.8012	0.9284	Razonamiento Verbal
18	0.064	0.7201	0.9285	Razonamiento Verbal
19	0.1099	0.8347	0.9281	Razonamiento Verbal
20	0.0912	0.8133	0.9282	Razonamiento Verbal
25	0.0633	0.1766	0.9283	Razonamiento Matemático
26	0.1078	0.5162	0.9284	Razonamiento Matemático
27	0.04	0.1554	0.9284	Razonamiento Matemático
28	0.1475	0.2901	0.9281	Razonamiento Matemático
54	0.0863	0.0746	0.9281	Matemática
62	0.0682	0.142	0.9283	Física
91	-0.0502	0.0812	0.9285	Biología
95	0.072	0.1584	0.9283	Biología

Cuadro 6: Evaluación de ítems del examen de admisión 2016-II

Ítem	Correlación biserial puntual	Índice de Dificultad	Alfa de Cronbach	Área
1	0.1411	0.8141	0.899	Razonamiento Verbal
2	0.0814	0.5842	0.8997	Razonamiento Verbal
3	0.0353	0.6399	0.9	Razonamiento Verbal
4	0.1057	0.4304	0.8995	Razonamiento Verbal
5	0.0769	0.6135	0.8997	Razonamiento Verbal
6	0.0122	0.2662	0.9	Razonamiento Verbal
7	0.0051	0.4148	0.9003	Razonamiento Verbal
10	-0.054	0.0788	0.8998	Razonamiento Verbal
11	0.0562	0.2015	0.8996	Razonamiento Verbal
12	-0.0417	0.0864	0.8997	Razonamiento Verbal
14	0.0445	0.6371	0.9	Razonamiento Verbal
15	-0.0448	0.4592	0.9008	Razonamiento Verbal
16	0.0893	0.2496	0.8995	Razonamiento Verbal
18	0.0571	0.664	0.8998	Razonamiento Verbal
19	0.1172	0.8447	0.8991	Razonamiento Verbal
20	-0.0586	0.337	0.9007	Razonamiento Verbal
26	0.0871	0.336	0.8996	Razonamiento Matemático
27	-0.0271	0.1619	0.9	Razonamiento Matemático
28	0.0451	0.1694	0.8996	Razonamiento Matemático
33	0.0283	0.2388	0.8999	Razonamiento Matemático
49	-0.0485	0.0547	0.8996	Matemática
55	0.1074	0.1048	0.8991	Matemática
60	-0.0071	0.3403	0.9003	Física
73	0.0665	0.1907	0.8995	Química
91	0.0471	0.3676	0.8999	Biología
94	-0.0199	0.2161	0.9001	Biología
99	0.1002	0.1085	0.8991	Biología
100	0.0648	0.1958	0.8995	Biología

Los resultados evidenciaron que 22 ítems del examen de admisión 2016-I y 28 del examen 2016-II presentaron un pobre poder discriminativo, el resto de ítems presentó un poder discriminativo de regular a excelente. Los valores para el Alfa de Cronbach no presentaron grandes cambios respecto a su valor original (ver cuadro N°4). La distribución de los ítems con pobre poder discriminativo según área se muestra en el cuadro N°7.

Cuadro 7: Distribución de ítems con pobre poder discriminativo

Área	Total de ítems	Examen			
		2016-I		2016-II	
		Ítems de discriminación pobre	Porcentaje	Ítems de discriminación pobre	Porcentaje
Razonamiento Verbal	20	14	70%	16	80%
Razonamiento Matemático	14	4	29%	4	29%
Matemática	24	1	4%	2	8%
Física	14	1	7%	1	7%
Química	14	0	0%	1	7%
Biología	14	2	14%	4	29%

Los resultados obtenidos en el cuadro N°7 indicaron que el área de Razonamiento Verbal fue la que tuvo más ítems con pobre poder discriminativo respecto al total de ítems, tanto en el examen de admisión 2016-I (70%) como 2016-II (80%). Esto quiere decir que los resultados de estos ítems tuvieron una baja correlación con el puntaje final obtenido por los postulantes. Mientras que Química fue el área que presentó menos ítems con pobre poder discriminativo en el examen de admisión 2016-II (7%) y ninguno en el examen 2016-I.

Finalmente, en el examen 2016-I se encontró que 4 ítems sobrepasaron los valores límite para el índice de dificultad y para el de discriminación. De ellos 1 fue de Razonamiento Verbal, 1 de Matemática, 1 de Física y 1 de Biología. Mientras que en el examen 2016-II se encontró que 5 ítems sobrepasaron los valores límite para los índices mencionados. De ellos 2 fueron de Razonamiento Verbal, 2 de Matemática y 1 de Biología. Estos resultados indicaron que dichos ítems deben ser descartados o revisados para su próximo uso en posteriores evaluaciones.

4.4 Estadísticas básicas de equiparación

En el cuadro N°8 se observan estadísticos como la media, desviación estándar, asimetría y kurtosis del examen de admisión 2016-I equiparado al 2016-II con los métodos lineal y equipercentil según área. Estos resultados mostraron que para el método lineal se ajustaron la media y desviación estándar de los puntajes del examen de admisión 2016-I con los obtenidos del examen del 2016-II. Sin embargo, bajo este método el ajuste que presentaron

las características de la distribución tales como la asimetría y la kurtosis no fueron cercanas a las obtenidas por el examen 2016-II. Mientras que para el método equipercentil, a pesar de tener ligeras diferencias al ajustar la media y desviación estándar, el ajuste de las características de la distribución fue más cercano que al utilizar el método lineal para realizar la equiparación.

Cuadro 8: Estadísticas de equiparación

Área		media	desviación estándar	asimetría	kurtosis
Razonamiento Verbal	2016 - II	9.68	2.63	-0.49	3.41
	Método Lineal	9.68	2.63	-0.36	2.87
	Método Equipercentil	9.69	2.61	-0.48	3.38
Razonamiento Matemático	2016 - II	5.28	2.61	0.27	2.57
	Método Lineal	5.28	2.61	0.15	2.5
	Método Equipercentil	5.27	2.6	0.27	2.57
Matemática	2016 - II	8.08	4.56	0.41	2.47
	Método Lineal	8.08	4.56	0.21	1.99
	Método Equipercentil	8.07	4.55	0.4	2.44
Física	2016 - II	4.68	3.44	0.45	2.13
	Método Lineal	4.68	3.44	0.49	2.18
	Método Equipercentil	4.68	3.43	0.45	2.11
Química	2016 - II	5.71	3.39	0.1	2.02
	Método Lineal	5.71	3.39	0.37	2.17
	Método Equipercentil	5.7	3.38	0.09	2.01

4.5 Funciones de equiparación

En las figuras N°5 y N°6 se observan las funciones de los puntajes del examen de admisión 2016-I equiparados al examen 2016-II mediante los métodos lineal y equipercentil para las seis áreas del examen de admisión. Las rectas verticales corresponden a los puntajes mínimos y máximos obtenidos en cada área. La función identidad representa el caso cuando el puntaje (número de aciertos) es el mismo en ambas pruebas.

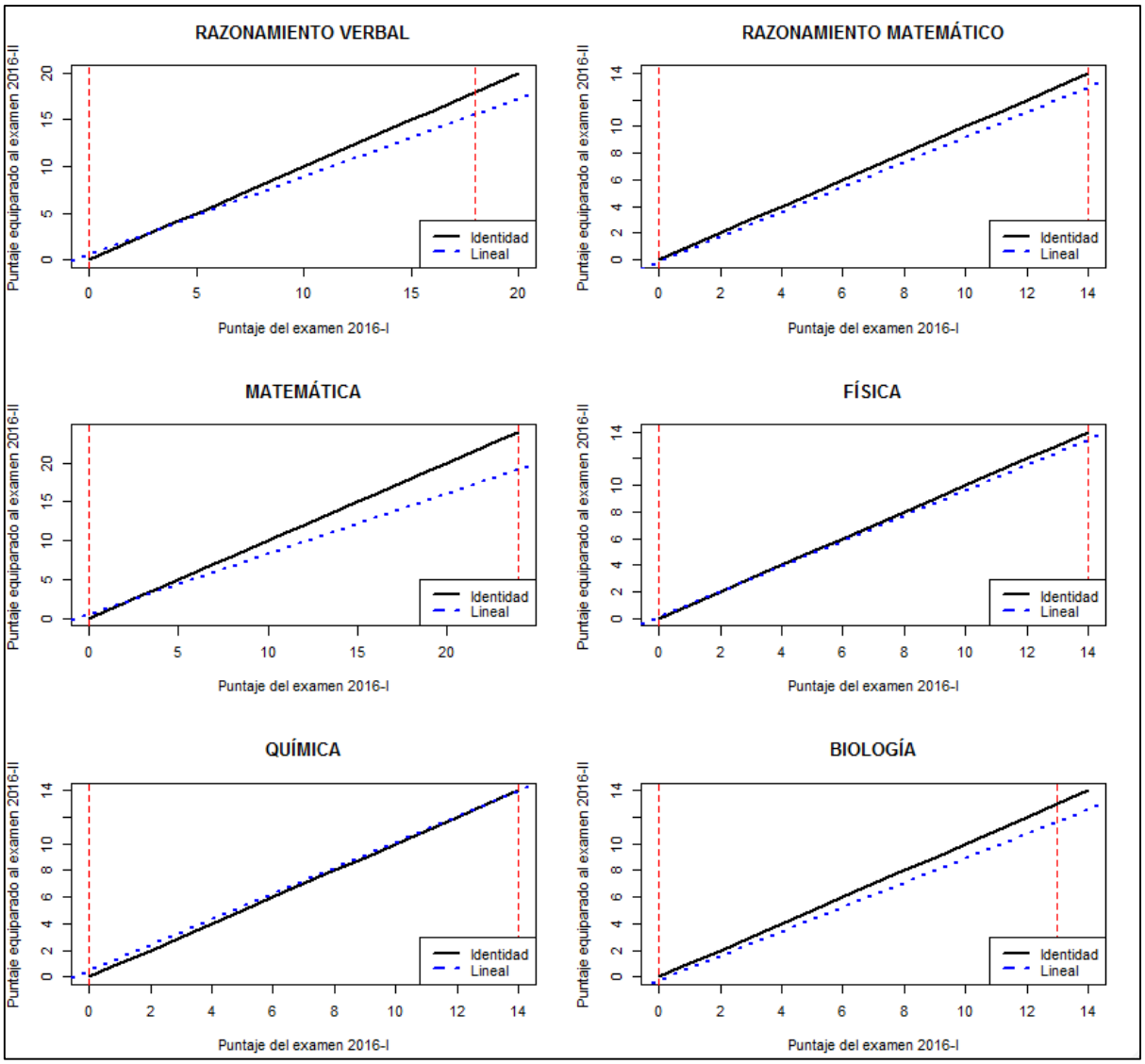

Figura 5: Función de equiparación lineal según área

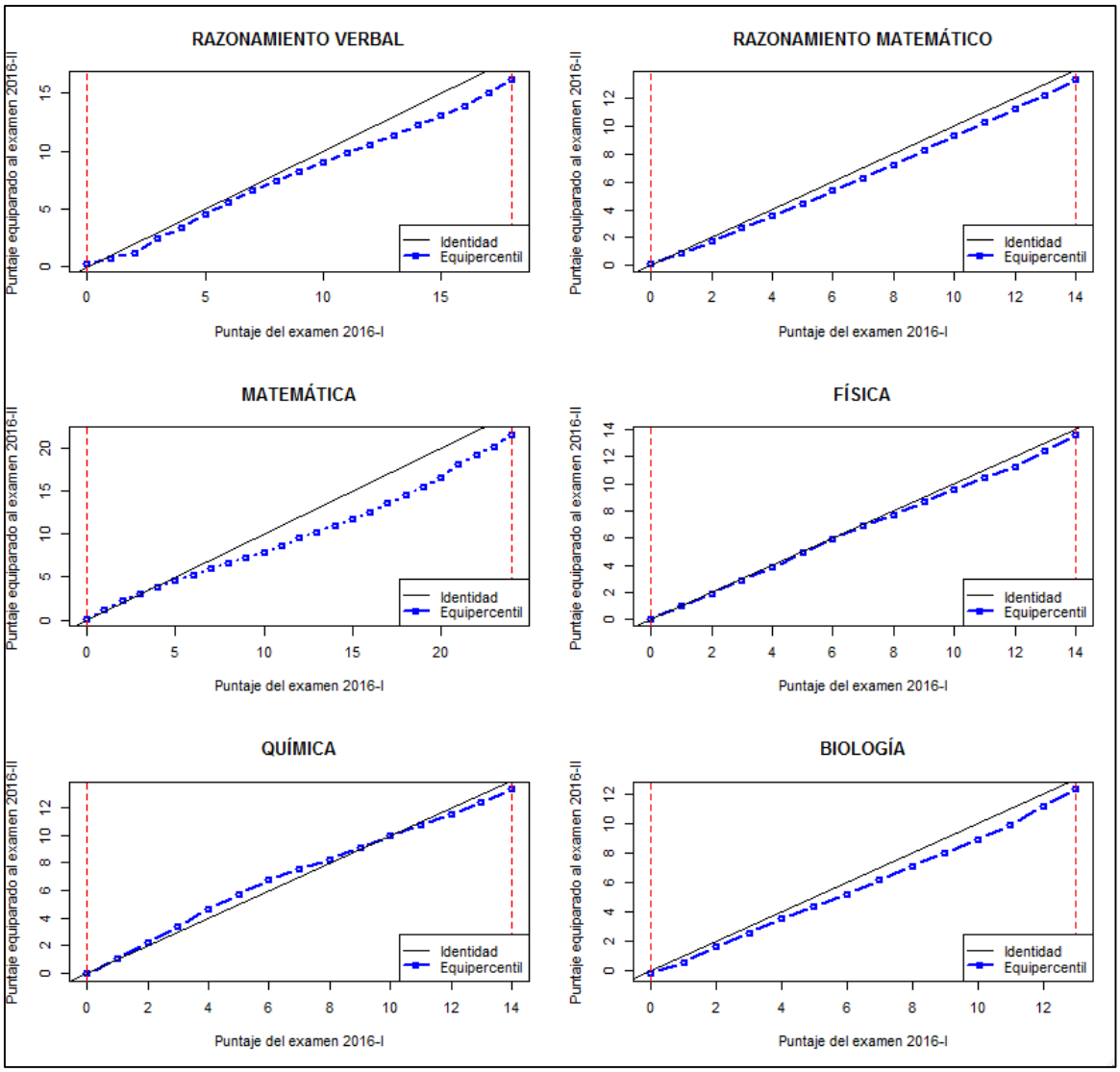

Figura 6: Función de equiparación equipercentil según área

En las figuras N°5 y N°6 se evidenció que las funciones lineal y equipercentil se encontraron por debajo de la función identidad en las áreas de Razonamiento Verbal, Razonamiento Matemático, Matemática y Biología. De las áreas mencionados anteriormente, la gráfica para el área de Matemática presentó mayores diferencias en las puntuaciones en la parte extrema superior. Esto indicó que esta área presentó una mayor dificultad en el examen 2016-II frente al examen 2016-I. Las gráficas para las áreas de Razonamiento Verbal, Razonamiento Matemático y Biología presentaron también diferencias en las puntuaciones, pero fueron menores a las de Matemática.

Por otro lado, en los gráficos de las áreas de Química y Física, se encontraron que las funciones lineal y equipercentil estuvieron alrededor de la función identidad. Esto, que fue

consecuencia de que las puntuaciones fueron similares en toda la distribución de puntajes para ambos métodos, evidenció que la dificultad fue similar en las dos áreas para ambos exámenes de admisión.

4.6 Tablas de equivalencia

Se obtuvieron los puntajes equiparados del examen 2016-I al 2016-II mediante las técnicas lineal y equipercentil para las seis áreas utilizando el paquete y función *equate*. Los resultados se dispusieron en los cuadros N°9, 10, 11 y 12. Donde la primera columna indica el puntaje obtenido en el examen de admisión 2016-I, la segunda y tercera muestran el puntaje equiparado para los métodos lineal y equipercentil, respectivamente.

Cuadro 9: Equivalencia de puntajes para Razonamiento Verbal según método

Puntaje 2016-I	Lineal	Percentil
0	0.51	0.27
1	1.35	0.77
2	2.20	1.20
3	3.04	2.48
4	3.88	3.35
5	4.72	4.54
6	5.57	5.63
7	6.41	6.58
8	7.25	7.44
9	8.09	8.23
10	8.94	9.07
11	9.78	9.82
12	10.62	10.56
13	11.46	11.36
14	12.31	12.22
15	13.15	13.05
16	13.99	13.93
17	14.83	15.01
18	15.68	16.18

En el cuadro N°9 se observan los resultados en el área de Razonamiento Verbal luego de aplicar los métodos en estudio. En estos resultados se encontró que para un puntaje de 8 a 12 en el examen 2016-I, los puntajes equiparados con ambos métodos disminuyeron aproximadamente 1 punto. Mientras que para un puntaje superior a 12, los puntajes equiparados disminuyeron en aproximadamente de 2 puntos. Finalmente, ambos métodos

mostraron que en esta área hubo una dificultad mayor en el examen de admisión 2016-II frente al 2016-I.

En el cuadro N°10 se observan los resultados para el área de Razonamiento Matemático. En ellos se observó que para un puntaje de 5 a más en el examen 2016-I, los puntajes equiparados con ambos métodos disminuyeron en aproximadamente 1 punto. Por último, para esta área en los dos métodos también se presentó una mayor dificultad en el examen de admisión 2016-II frente al 2016-I.

Cuadro 10: Equivalencia de puntajes para Razonamiento Matemático según método

Puntaje 2016-I	Lineal	Percentil
0	-0.18	0.13
1	0.75	0.89
2	1.69	1.73
3	2.62	2.68
4	3.56	3.55
5	4.49	4.44
6	5.43	5.37
7	6.37	6.26
8	7.30	7.23
9	8.24	8.27
10	9.17	9.28
11	10.11	10.25
12	11.05	11.23
13	11.98	12.23
14	12.92	13.35

En el cuadro N°11 se observan los resultados para el área de Matemática. En ellos se encontró que la tendencia en las diferencias de los puntajes equiparados por ambos métodos fue distinta. Para el método lineal, conforme fue incrementándose el puntaje de 5 a más en el examen 2016-I, los puntajes equiparados fueron disminuyendo aproximadamente de 1 a 5 puntos. Mientras que para el método equipercentil, conforme fue incrementándose el puntaje de 6 a 19, los puntajes equiparados también fueron disminuyendo aproximadamente de 1 a 4 puntos. Sin embargo, cuando se incrementó el puntaje de 20 a 24, los puntajes equiparados disminuyeron de 3 a 2 puntos. Finalmente, ambos métodos también mostraron que en esta área hubo una dificultad mayor en el examen de admisión 2016-II frente al 2016-I.

Cuadro 11: Equivalencia de puntajes para Matemática según método

Puntaje 2016-I	Lineal	Percentil
0	0.44	0.16
1	1.22	1.23
2	2.00	2.25
3	2.78	3.13
4	3.56	3.91
5	4.34	4.59
6	5.12	5.29
7	5.90	5.97
8	6.68	6.58
9	7.46	7.23
10	8.24	7.91
11	9.02	8.64
12	9.81	9.52
13	10.59	10.27
14	11.37	10.97
15	12.15	11.70
16	12.93	12.55
17	13.71	13.54
18	14.49	14.59
19	15.27	15.41
20	16.05	16.60
21	16.83	18.12
22	17.61	19.15
23	18.39	20.19
24	19.17	21.57

En el cuadro N°12 se observan los resultados para las áreas de Física, Química y Biología. Donde se encontró que, para el área de Física, en ambos métodos las diferencias en la equiparación no llegaron a 1 punto. Por lo que puede considerarse que para esta área los dos exámenes de admisión presentaron una dificultad similar.

Para el área de Química, el método lineal no presentó diferencias en la equiparación superiores a 1 punto. Sin embargo, en el método equipercentil, conforme fue incrementándose el puntaje de 4 a 7 en el examen 2016-I, el puntaje equiparado se incrementó en aproximadamente 1 punto. Esta ligera tendencia fue compensándose mientras se iba incrementando el puntaje en dicho examen. Por último, se pudo observar que el método lineal no reflejó diferencias al equiparar mientras que el método equipercentil sí, ya que fue susceptible a los cambios en el puntaje del examen 2016-I. Dado que las diferencias

para el método equipercentil no fueron mayores a 1 pero sí constantes conforme a un aumento o disminución del puntaje en el examen 2016-I, se consideró que para esta área los dos exámenes presentaron también una dificultad similar.

Para el área de Biología, con un puntaje de 5 a más en el examen 2016-I, los puntajes equiparados con ambos métodos disminuyeron en aproximadamente 1 punto. Finalmente, ambos métodos indicaron que en esta área hubo una dificultad mayor en el examen de admisión 2016-II frente al 2016-I.

Cuadro 12: Equivalencia de puntajes para Química, Física y Biología según método

Puntaje 2016-I	Física		Química		Biología	
	Lineal	Percentil	Lineal	Percentil	Lineal	Percentil
0	0.05	0.04	0.44	-0.05	-0.27	-0.22
1	1.01	0.99	1.41	1.02	0.65	0.57
2	1.97	1.93	2.38	2.22	1.57	1.59
3	2.92	2.90	3.35	3.39	2.50	2.56
4	3.88	3.83	4.32	4.60	3.42	3.50
5	4.84	4.90	5.29	5.71	4.34	4.35
6	5.80	5.97	6.26	6.73	5.26	5.20
7	6.76	6.90	7.23	7.55	6.19	6.17
8	7.72	7.74	8.20	8.26	7.11	7.11
9	8.67	8.64	9.17	9.06	8.03	8.00
10	9.63	9.57	10.14	9.94	8.95	8.89
11	10.59	10.45	11.11	10.74	9.88	9.90
12	11.55	11.30	12.08	11.49	10.80	11.13
13	12.51	12.46	13.05	12.38	11.72	12.31
14	13.46	13.65	14.02	13.32	-	-

4.7 Comparación de métodos de equiparación

Como indicador de comparación se utilizó error estándar aleatorio de equiparación obtenido mediante la técnica Bootstrap. Efron y Tibshirani (1993), señalan que para estimar errores estándar se debe usar entre 25 y 200 repeticiones Bootstrap. Sin embargo, en la práctica es común usar hasta 1000 repeticiones.

En el cuadro N°13 se muestran los coeficientes de variabilidad de los errores estándar Bootstrap para cada área según método de equiparación y tamaño de réplica utilizado. Los resultados confirmaron el uso de un tamaño de 200 réplicas Bootstrap ya que se encontró mayor reducción en el coeficiente de variabilidad. También se apreció que los errores estándar obtenidos por el método lineal (39% a 42%) tuvieron una menor variabilidad que los obtenidos por el método equipercentil (58% a 62%).

Cuadro 13: Coeficiente de variabilidad de errores estándar Bootstrap de los métodos lineal y equipercentil

Método	**Tamaño de réplicas Bootstrap (B)**				
	50	**100**	**200**	**500**	**1000**
Lineal					
Razonamiento Verbal	0.4378	0.4493	0.3923	0.3904	0.3736
Razonamiento Matemático	0.3715	0.3728	0.3876	0.3938	0.4075
Matemática	0.4058	0.4395	0.3791	0.3930	0.3941
Física	0.3539	0.3872	0.3642	0.3923	0.3794
Química	0.4210	0.4068	0.3902	0.3942	0.4081
Biología	0.4489	0.3813	0.4295	0.3770	0.3915
Equipercentil					
Razonamiento Verbal	0.6668	0.6124	0.6154	0.6251	0.6021
Razonamiento Matemático	0.6305	0.6154	0.5903	0.6254	0.6145
Matemática	0.6091	0.6122	0.6038	0.6074	0.6138
Física	0.6711	0.6362	0.5835	0.6228	0.6186
Química	0.7624	0.5796	0.6106	0.6081	0.6234
Biología	0.6309	0.6077	0.6257	0.6295	0.6226

En la figura N°7 se muestra la distribución de los errores estándar Bootstrap de equiparación para ambos métodos a lo largo del puntaje obtenido en el examen de admisión 2016-I.

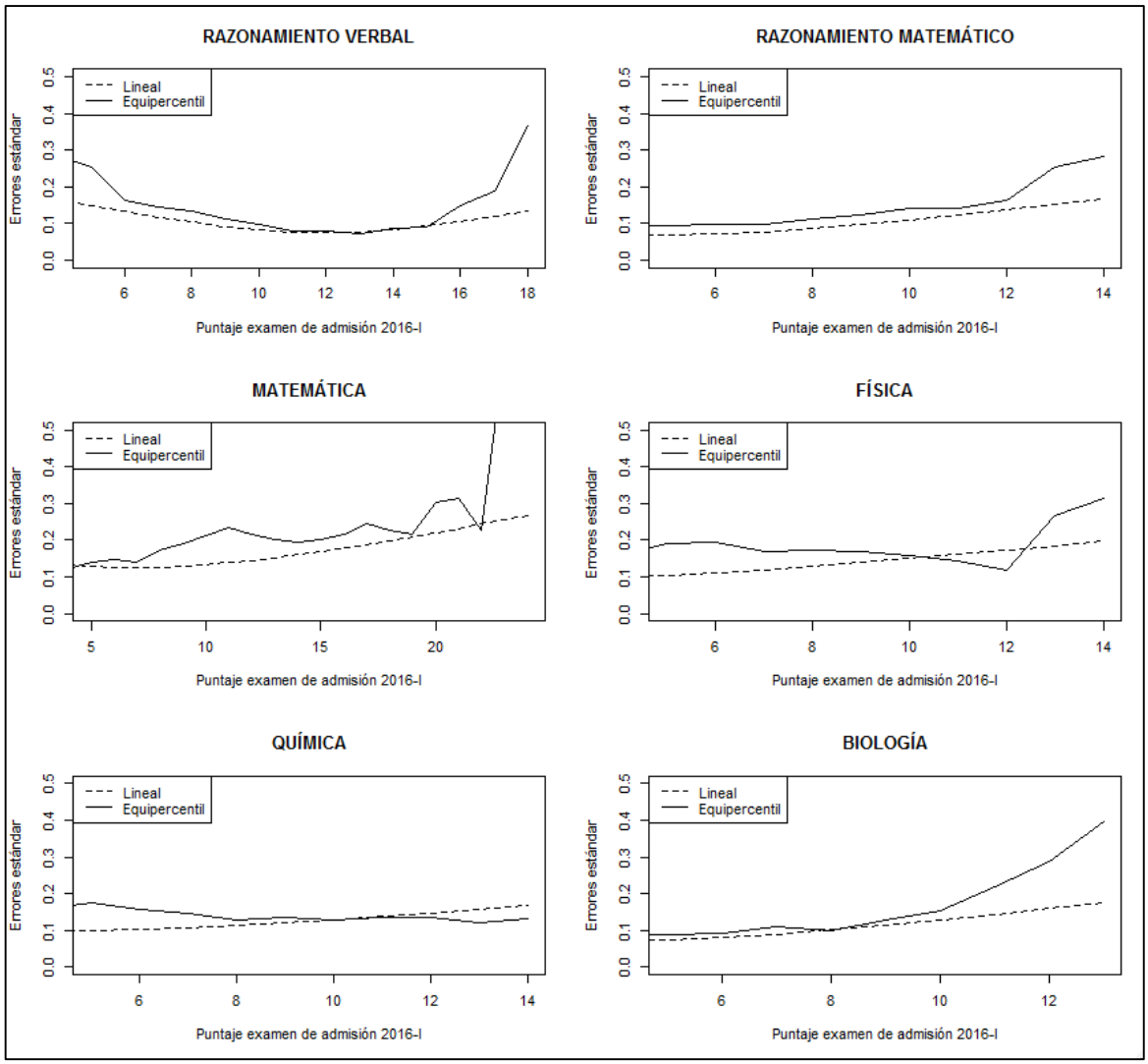

Figura 7: Errores estándar Bootstrap de equiparación según método

Para el área de Razonamiento Verbal, en ambos métodos el menor error de equiparación se presentó a un puntaje de 10 a 14. Además, el método equipercentil presentó errores más elevados en los extremos que el método lineal.

Para las áreas de Razonamiento Matemático y Biología la tendencia de los errores fue creciente para ambos métodos, encontrándose menores errores en menores puntajes. Los errores en los extremos superiores fueron mayores con el método equipercentil que con el método lineal, siendo esta ocurrencia más pronunciada en el área de Biología.

Para el área de Matemática la tendencia de los errores fue creciente para ambos métodos, pero más variable con el método equipercentil que con el lineal. Además, el método equipercentil obtuvo errores más elevados a lo largo de toda la distribución de puntajes.

Para las áreas de Física y Química los errores presentaron una tendencia ligeramente creciente. Siendo para el área de Física más variable con el método equipercentil que con el lineal. Para esta área los mayores errores se obtuvieron con el método equipercentil y se ubicaron en los extremos. De forma similar ocurrió para el área de Química, solo con la diferencia de que los mayores errores se ubicaron únicamente en el extremo inferior.

V. CONCLUSIONES

1. El método de equiparación lineal obtuvo un mejor ajuste de puntuaciones que el método equipercentil. Esto debido a que el método lineal presentó menores errores estándar para las seis áreas evaluadas y fueron menos variables que los obtenidos con el método equipercentil.

2. Los resultados al aplicar los dos métodos de equiparación indicaron que el examen de admisión 2016-II presentó una mayor dificultad que el examen 2016-I. Las áreas que influenciaron en esta dificultad fueron Razonamiento Verbal, Razonamiento Matemático, Matemática y Biología.

3. El área de Matemática tuvo una mayor dificultad en el examen 2016-II frente al 2016-I al compararlo con las demás áreas. Esto debido a que su función de equiparación obtenida con ambos métodos estuvo más alejada de la función identidad al compararla con las de las otras áreas.

4. Las áreas que presentaron una dificultad similar en los exámenes de admisión 2016-I y 2016-II fueron Física y Química. Esto debido a que la función de equiparación obtenida con ambos métodos para cada área estuvo alrededor de la función identidad.

5. Los resultados obtenidos con el análisis de ítems guardaron coherencia con los obtenidos mediante la equiparación de puntuaciones. Esto debido a que las áreas que evidenciaron tener ítems no adecuados para diferenciar el conocimiento en los postulantes fueron, en su gran mayoría, las que reflejaron una mayor dificultad al equiparar los puntajes del examen de admisión 2016-I al 2016-II.

VI. RECOMENDACIONES

Se recomienda:

1. Utilizar otros indicadores de ajuste diferentes al error estándar de equiparación.

2. Comparar los resultados de equiparación de puntuaciones aplicando técnicas de suavizamiento.

3. Comparar los resultados obtenidos por los métodos de equiparación de la teoría clásica de los *test* (TCT) con los obtenidos mediante los métodos de la teoría de respuesta al ítem (TRI).

4. Indicar a la Comisión Permanente de Admisión, desagregar las calificaciones del área de Matemática en las materias de Algebra, Aritmética, Geometría y Trigonometría, para realizar un análisis más exhaustivo en las puntuaciones de esta área.

5. Continuar con este tipo de investigaciones que son importantes para examinar la calidad del examen de admisión, ya que permitirán encontrar mejoras para su diseño y aplicación.

VII. REFERENCIAS BIBLIOGRÁFICAS

Aiken, L. 1996. Tests psicológicos y Evaluación. México, Pearson. 544 p.

Aliaga, J. 2018. Psicometría: Disciplina de la Medición en Psicología y Educación. Lima, PE, Fondo editorial UIGV. 500 p.

Allen, MJ; Yen, WM. 1979. Introduction to measurement theory. Illinois, US. Waveland Press. 320 p.

Angoff, W. 1971. Scales, norms and equivalentscores. Washington, US, Educational measurement. p. 508-597

Angoff, W. 1984. Scales, norms and equivalent scores. Princeton, NJ, Educational Testing Service. 145 p.

Antillón, L; Larrazolo, N; Backhoff, E. 2008 Igualación Lineal de Tres Versiones del Examen de Conocimientos y Habilidades EXHCOBA. (en línea). Consultado 1 abr. 2016. Disponible en https://revistas.uam.es/index.php/riee/article/view/4674/5111

Antillón, L; Larrazolo, N; Backhoff, E. 2006 Igualación equipercentil del examen de habilidades y conocimientos básicos EXHCOBA (en línea). Consultado 15 nov. 2016. Disponible en http://www.uv.es/RELIEVE/v12n2/RELIEVEv12n2_2.htm

Backhoff, E; Larrazolo, N; Rosas, M. 2000. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos EXHCOBA (en línea). Consultado 2 de jun. 2019. Disponible en http://redie.uabc.mx/index.php/redie/article/view/15

Barbero, M; Vila, E; Holgado, F. 2015. Psicometría. Madrid, ES, Sanz y Torres. 489 p.

Burga, A. 2003. Módulos de cálculos psicométricos. Consultado 20 may. 2019. (en línea). Lima, Perú. Disponible en http://www.fafich.ufmg.br/ladi/files/Manual%20MCP.doc

Burga, A. 2006. Teoría clásica de los tests (TCT): índice y coeficiente de confiabilidad. Consultado 20 may. 2019. (en línea). Lima, Perú. Disponible en http://www2.minedu.gob.pe/umc/admin/images/publicaciones/artiumc/3.doc

Braun, H; Holland, P. 1982. Test equating. New York, US, Academic. p. 9-49

Brennan, RL. 2006. Educational Measurement, Westport, CT, Praeger Publishers. 779 p.

Bulmer, M. 1979. Principles of Statistics. 2 ed. US, Dover. 252 p.

Castro, S. 2016. Classical Test Theory (en línea). Consultado 27 may. 2019. Disponible en https://rpubs.com/castro/141954

Chacón, S; Antonio, J. 2008. Diseño y medición de programas de intervención neuropsicológica: aspectos fundamentales (en línea). Sevilla, España. Consultado 15 may. 2019. Disponible en http://innoevalua.us.es/wakka.php?wakka=disenoymedicionenprogramasdeintervencionneuropsicologica/files&get=tema2_tct.pdf

Dorans, N; Moses, T; Eignor, D. 2010. Principles and Practices of Test Score Equating. New Jersey, US, Educational Testing Service. 41 p.

Efron, B; Tibshirani, RJ. 1993. An introduction to the bootstrap (Monographs on Statistics and Applied Probability 57). New York, US, Chapman & Hall. 430 p.

Gempp, R. 2010. Equiparación, alineamiento y predicción de puntuaciones en medición educativa. Revista Iberoamericana de Evaluación Educativa 3(2):104-126.

George, D; Mallery, P. 2003. SPSS for Windows step by step: A simple guide and reference. 11.0 update. 4 ed. Boston, US, Allyn & Bacon. s.p.

Gonzales, J; Wiberg, M. 2017. Applying Test Equating Methods using R. Cham, Suiza, Springer. 212 p.

Gulliksen, H. 1950. Theory of mental tests. New York, US, Willey. 516 p.

Hernández, R.; Fernández, C; Baptista, P. 2014. Metodología de la investigación. 6 ed. México, Mcgraw - Hill. 599 p.

Herrera, A. 2013. Métodos de equiparación de puntuaciones: Los examines de estado en población con y sin limitación visual. Tesis Mag. Bogotá D.C, CO, Universidad Nacional de Colombia. 142 p.

Kolen, M; Brennan, R. 2004. Test Equating, Scaling and Linking. New York, US, Springer. 567 p.

Livingston, S. 2014. Equating Test Scores (without IRT). US, Educational Testing Service. 73 p.

Lord, F. 1980. Aplications of ítem response theory to practical testing problems. Hillsdale, US, Erlbaum. 274 p.

Lord, F. M; Novick, MR. 1968. Statistical theories of mental test scores. Addison-Wesley. 568 p.

Matus, C; Stevenson, M; Valencia, M; Guzman, E. 2012. Alineamiento de las puntuaciones SIMCE 2008 y PISA 2009 en muestras de estudiantes de 2° Medio. Lectura y Matemática (en línea). Consultado 19 de nov. 2016. Disponible en http://www.agenciaeducacion.cl/wp-content/files_mf/semina2.pdf

Meneses, J; Barrios, M; Bonillo, Albert; Cosculluela, A; Lozano, LM; Turbany, J; Valero, S. 2013. Aproximación histórica y conceptos básicos de la psicometría. Catalunya, ES, UOC. 50 p.

Muñiz, J. 1996. Teoría Clásica de los Tests. Madrid, ES, Ediciones Pirámide. s.p.

Navas, MJ. 1996. Equiparación de puntuaciones. En J. Muñiz (Ed.), Psicometría. Madrid, ES, Universitas. p. 293-369.

Navas, MJ. 2000. Equiparación de puntuaciones: exigencias actuales y retos de cara al futuro (en línea). Consultado 10 may. 2016. Disponible en http://www2.uned.es/490015/CV/AEMCCO2000.pdf

Ortiz, GM; Díaz, PA; Llanos, OR; Pérez, SM; González, K. 2015. Dificultad y discriminación de los ítems del examen de Metodología de la Investigación y Estadística (en línea). Consultado 2 de jun. 2019. Disponible en http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S207728742015000200003&lng=es&tlng=es.

Pacheco-Villamil, S. 2007. La equiparación de puntuaciones en procesos de comparación de pruebas diferentes. Revista avances de medición 5(1):153-156.

Ryan, J; Brockmann, F. 2011. A practitioner's Introduction to Equating. Washington DC, US, Counsil of Chief State School Officers. 100 p.

Spearman, C. 1904. The proof and measurement of association between two things. American Journal of Psychology 18:161-169.

Yela, M. 1984. Introducción a la teoría de los tests. Madrid, España, Universidad Complutense. s. p.

Yen, WM. 1983. Tau-equivalence and equipercentile equating. Psychometrika 48(3): 353-369.

VIII. ANEXOS

ANEXO 1: Funciones y procedimientos en R

Función para Gráfico de ítems

```
library(CTT)
library(knitr)
item.difficulty <-  function(responses,nombre){

   # CRITICAL VALUES
   cvpb = 0.20
   cvdl = 0.15
   cvdu = 0.85

   require(CTT, warn.conflicts = FALSE, quietly = TRUE)
   ctt.analysis <- CTT::reliability(responses, itemal = TRUE, NA.Delete = F)

   test_difficulty <- data.frame(item = 1:ctt.analysis$nItem ,
                                 difficulty = ctt.analysis$itemMean)

   plot(test_difficulty,
       main = nombre,
       type = "p",
       pch = 1,
       cex = 2.8,
       col = "purple",
       ylab = "Índice de Dificultad",
       xlab = "Número de ítems",
       ylim = c(0, 1),
       xlim = c(0, ctt.analysis$nItem))
```

```
abline(h = cvdl, col = "tomato")
abline(h = cvdu, col = "tomato")

abline(h = .3, col = "dodgerblue")
abline(h = .7, col = "dodgerblue")

text(diff(range(test_difficulty[, 1]))/2, 0.7,
    "rango de máxima información",
    col = "dodger blue",
    pos = 3)

text(diff(range(test_difficulty[, 1]))/2, cvdu,
    "límite de rango aceptable",
    col = "tomato",
    pos = 3)

outlier <- data.matrix(subset(cbind(test_difficulty[, 1], test_difficulty[, 2]),
                              subset = (test_difficulty[, 2] < cvdl |
                                        test_difficulty[, 2] > cvdu)))

text(outlier, paste("i", outlier[,1], sep = ""), col = "red", cex = .7)

outlier2 <- data.matrix(subset(cbind(test_difficulty[, 1],
                                     test_difficulty[, 2]),
                               subset = ((test_difficulty[, 2] > cvdl &
                                          test_difficulty[, 2] < .3) |
                                         (test_difficulty[, 2] < cvdu &
                                          test_difficulty[, 2] > .7))))

text(outlier2, paste("i", outlier2[,1], sep = ""),
    col = "dodgerblue",
    cex = .7)
```

```
  return(test_difficulty[order(test_difficulty$difficulty),])
 }
```

Fuente: Adaptado de Castro (2016)

Función para Evaluar ítems

```
item.analysis <-
 function(responses){
  # CRITICAL VALUES
  cvpb = 0.15
  cvdl = 0.15
  cvdu = 0.85

  require(CTT, warn.conflicts = FALSE, quietly = TRUE)
  (ctt.analysis <- CTT::reliability(responses, itemal = TRUE, NA.Delete = F))

  # Mark items that are potentially problematic
  item.analysis <- data.frame(item = seq(1:ctt.analysis$nItem),
                              r.pbis = ctt.analysis$pBis,
                              bis = ctt.analysis$bis,
                              item.mean = ctt.analysis$itemMean,
                              alpha.del = ctt.analysis$alphaIfDeleted)

  # code provided by Dr. Gordon Brooks
  if (TRUE) {
   item.analysis$check <-
    ifelse(item.analysis$r.pbis < cvpb |
           item.analysis$item.mean < cvdl |
           item.analysis$item.mean > cvdu, "+++", "")
  }

  return(item.analysis)
 }
```

```
kable(item.analysis(responses),
      align = "c",
      caption = "Item Analysis")

item.difficulty <-
  function(responses){
    # CRITICAL VALUES
    cvpb = 0.20
    cvdl = 0.15
    cvdu = 0.85

    require(CTT, warn.conflicts = FALSE, quietly = TRUE)
    ctt.analysis <- CTT::reliability(responses, itemal = TRUE, NA.Delete = F)

    test_difficulty <- data.frame(item = 1:ctt.analysis$nItem ,
                                  difficulty = ctt.analysis$itemMean)

    plot(test_difficulty,
         main = "Test Item Difficulty",
         type = "p",
         pch = 1,
         cex = 2.8,
         col = "purple",
         ylab = "Item Mean (Difficulty)",
         xlab = "Item Number",
         ylim = c(0, 1),
         xlim = c(0, ctt.analysis$nItem))

    abline(h = cvdl, col = "tomato")
    abline(h = cvdu, col = "tomato")

    abline(h = .3, col = "dodgerblue")
    abline(h = .7, col = "dodgerblue")
```

```
  text(diff(range(test_difficulty[, 1]))/2, 0.7,
       "maximum information range",
       col = "dodger blue",
       pos = 3)

  text(diff(range(test_difficulty[, 1]))/2, cvdu,
       "rule of thumb acceptable range",
       col = "tomato",
       pos = 3)

  outlier <- data.matrix(subset(cbind(test_difficulty[, 1], test_difficulty[, 2]),
                                subset = (test_difficulty[, 2] < cvdl |
                                          test_difficulty[, 2] > cvdu)))

  text(outlier, paste("i", outlier[,1], sep = ""), col = "red", cex = .7)

  outlier2 <- data.matrix(subset(cbind(test_difficulty[, 1],
                                       test_difficulty[, 2]),
                                 subset = ((test_difficulty[, 2] > cvdl &
                                            test_difficulty[, 2] < .3) |
                                           (test_difficulty[, 2] < cvdu &
                                            test_difficulty[, 2] > .7))))

  text(outlier2, paste("i", outlier2[,1], sep = ""),
       col = "dodgerblue",
       cex = .7)

  return(test_difficulty[order(test_difficulty$difficulty),])
}
```

Fuente: Adaptado de Castro (2016)

Procedimiento para calcular CV Bootstrap

```
set.seed(2019)
repe<-c(50,100,200,500,1000)
cv_lineal<-matrix(0,6,5)
cv_equi<-matrix(0,6,5)
for(j in 1:5){
for(i in 1:6){
  rx <-freqtab(bd161[,1])
  ry <-freqtab(bd162[,1])
  eql1_rv <- equate(rx, ry, type = "l",boot=TRUE,reps=repe[j])
  eql2_rv <- equate(rx, ry, type = "equipercentile",boot=TRUE,reps=repe[j])
  cv_lineal[i,j]<-sd(eql1_rv$con[,4])/mean(eql1_rv$con[,4])
  cv_equi[i,j]<-sd(eql2_rv$con[,4])/mean(eql2_rv$con[,4])
}
}
```

Procedimiento para obtener gráficas de funciones de equiparación

```
## Función Lineal
par(mfrow=c(3,2))
###RV
plot(0:20,0:20, type='l',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="RAZONAMIENTO VERBAL")
abline(a=mean(bd162$RV)-(sd(bd162$RV)/sd(bd161$RV))*mean(bd161$RV),b=sd(bd162$RV)/sd(bd161$RV),lwd=2,lty=3,col="blue")
abline(v=0,lty=2,col="red")
abline(v=18,lty=2,col="red")
legend("bottomright",lty=c(1,2), c("Identidad","Lineal"),lwd=c(2,2),col=c("black","blue"))
###RM
plot(0:14,0:14, type='l',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="RAZONAMIENTO MATEMÁTICO")
```

```
abline(a=mean(bd162$RM)-
(sd(bd162$RM)/sd(bd161$RM))*mean(bd161$RM),b=sd(bd162$RM)/sd(bd161$RM),lw
d=2,lty=3,col="blue")
abline(v=0,lty=2,col="red")
abline(v=14,lty=2,col="red")
legend("bottomright",lty=c(1,2), c("Identidad","Lineal"),lwd=c(2,2),col=c("black","blue"))
###MAT
plot(0:24,0:24, type='l',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-
II",xlab="Puntaje del examen 2016-I",main="MATEMÁTICA")
abline(a=mean(bd162$MAT)-
(sd(bd162$MAT)/sd(bd161$MAT))*mean(bd161$MAT),b=sd(bd162$MAT)/sd(bd161$M
AT),lwd=2,lty=3,col="blue")
abline(v=0,lty=2,col="red")
abline(v=24,lty=2,col="red")
legend("bottomright",lty=c(1,2), c("Identidad","Lineal"),lwd=c(2,2),col=c("black","blue"))
###FIS
plot(0:14,0:14, type='l',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-
II",xlab="Puntaje del examen 2016-I",main="FÍSICA")
abline(a=mean(bd162$FIS)-
(sd(bd162$FIS)/sd(bd161$FIS))*mean(bd161$FIS),b=sd(bd162$FIS)/sd(bd161$FIS),lwd
=2,lty=3,col="blue")
abline(v=0,lty=2,col="red")
abline(v=14,lty=2,col="red")
legend("bottomright",lty=c(1,2), c("Identidad","Lineal"),lwd=c(2,2),col=c("black","blue"))
###QUI
plot(0:14,0:14, type='l',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-
II",xlab="Puntaje del examen 2016-I",main="QUÍMICA")
abline(a=mean(bd162$QUI)-
(sd(bd162$QUI)/sd(bd161$QUI))*mean(bd161$QUI),b=sd(bd162$QUI)/sd(bd161$QUI),l
wd=2,lty=3,col="blue")
abline(v=0,lty=2,col="red")
abline(v=14,lty=2,col="red")
legend("bottomright",lty=c(1,2), c("Identidad","Lineal"),lwd=c(2,2),col=c("black","blue"))
```

```
###BIO
plot(0:14,0:14, type='l',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="BIOLOGÍA")
abline(a=mean(bd162$BIO)-(sd(bd162$BIO)/sd(bd161$BIO))*mean(bd161$BIO),b=sd(bd162$BIO)/sd(bd161$BIO),lwd=2,lty=3,col="blue")
abline(v=0,lty=2,col="red")
abline(v=13,lty=2,col="red")
legend("bottomright",lty=c(1,2), c("Identidad","Lineal"),lwd=c(2,2),col=c("black","blue"))

## Función Equiparación
par(mfrow=c(3,2))
#rv
rx <-freqtab(bd161[,1])
ry <-freqtab(bd162[,1])
eql2<- equate(rx, ry, type = "equipercentile")
xa<-eql2$con[,1]
yo<-eql2$con[,2]
plot(xa,yo, type='b',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="RAZONAMIENTO VERBAL",col="blue")
abline(a=0,b=1)
abline(v=0,lty=2,col="red")
abline(v=18,lty=2,col="red")
legend("bottomright",lty=c(1,1),
c("Identidad","Equipercentil"),lwd=c(1,2),col=c("black","blue"),pch=c(NA,1))
#rm
rx <-freqtab(bd161[,2])
ry <-freqtab(bd162[,2])
eql2<- equate(rx, ry, type = "equipercentile")
xa<-eql2$con[,1]
yo<-eql2$con[,2]
plot(xa,yo, type='b',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="RAZONAMIENTO MATEMÁTICO",col="blue")
```

```
abline(a=0,b=1)
abline(v=0,lty=2,col="red")
abline(v=14,lty=2,col="red")
legend("bottomright",lty=c(1,1),
c("Identidad","Equipercentil"),lwd=c(1,2),col=c("black","blue"),pch=c(NA,1))
#MAT
rx <-freqtab(bd161[,3])
ry <-freqtab(bd162[,3])
eql2<- equate(rx, ry, type = "equipercentile")
xa<-eql2$con[,1]
yo<-eql2$con[,2]
plot(xa,yo, type='b',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-
II",xlab="Puntaje del examen 2016-I",main="MATEMÁTICA",col="blue")
abline(a=0,b=1)
abline(v=0,lty=2,col="red")
abline(v=24,lty=2,col="red")
legend("bottomright",lty=c(1,1),
c("Identidad","Equipercentil"),lwd=c(1,2),col=c("black","blue"),pch=c(NA,1))
#FIS
rx <-freqtab(bd161[,4])
ry <-freqtab(bd162[,4])
eql2<- equate(rx, ry, type = "equipercentile")
xa<-eql2$con[,1]
yo<-eql2$con[,2]
plot(xa,yo, type='b',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-
II",xlab="Puntaje del examen 2016-I",main="FÍSICA",col="blue")
abline(a=0,b=1)
abline(v=0,lty=2,col="red")
abline(v=14,lty=2,col="red")
legend("bottomright",lty=c(1,1),
c("Identidad","Equipercentil"),lwd=c(1,2),col=c("black","blue"),pch=c(NA,1))
#QUI
rx <-freqtab(bd161[,5])
ry <-freqtab(bd162[,5])
```

```
eql2<- equate(rx, ry, type = "equipercentile")
xa<-eql2$con[,1]
yo<-eql2$con[,2]
plot(xa,yo, type='b',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="QUÍMICA",col="blue")
abline(a=0,b=1)
abline(v=0,lty=2,col="red")
abline(v=14,lty=2,col="red")
legend("bottomright",lty=c(1,1),
c("Identidad","Equipercentil"),lwd=c(1,2),col=c("black","blue"),pch=c(NA,1))
#bio
rx <-freqtab(bd161[,6])
ry <-freqtab(bd162[,6])
eql2<- equate(rx, ry, type = "equipercentile")
xa<-eql2$con[,1]
yo<-eql2$con[,2]
plot(xa,yo, type='b',lwd=2.0,lty=1,ylab="Puntaje equiparado al examen 2016-II",xlab="Puntaje del examen 2016-I",main="BIOLOGÍA",col="blue")
abline(a=0,b=1)
abline(v=0,lty=2,col="red")
abline(v=13,lty=2,col="red")
legend("bottomright",lty=c(1,1),
c("Identidad","Equipercentil"),lwd=c(1,2),col=c("black","blue"),pch=c(NA,1))
```

Procedimiento para gráfica de errores de equiparación

```
par(mfrow=c(3,2))
nombre<-c("RAZONAMIENTO VERBAL","RAZONAMIENTO MATEMÁTICO","MATEMÁTICA","FÍSICA","QUÍMICA","BIOLOGÍA")
max<-c(18,14,24,14,14,13)
for(i in 1:6){
  rx <-freqtab(bd161[,i])
  ry <-freqtab(bd162[,i])

  eql1_rv <- equate(rx, ry, type = "l",boot=TRUE,reps=200)
  eql2_rv <- equate(rx, ry, type = "equipercentile",boot=TRUE,reps=200)

plot(0:max[i],eql1_rv$con[,4],type="l",lty=2,xlim=c(5,max[i]),ylim=c(0,0.5),xlab="Puntaje examen de admisión 2016-I",ylab="Errores estándar",main=nombre[i])
  lines(0:max[i],eql2_rv$con[,4],lty=1)
  legend("topleft",c("Lineal","Equipercentil"),lty=c(2,1))
}
```