Revista Iberoamericana de Evaluación Educativa

ISSN: 1989-0397

EQUIPARACIÓN, ALINEAMIENTO Y PREDICCIÓN DE PUNTUACIONES EN MEDICIÓN EDUCATIVA

SCORE EQUATING, ALIGNING AND PREDICTION IN EDUCATIONAL MEASUREMENT

René Gempp

Revista Iberoamericana de Evaluación Educativa 2010 - Volumen 3, Número 2

http://www.rinace.net/riee/numeros/vol3-num2/art6.pdf

Fecha de recepción: 04 de abril de 2010
Fecha de dictaminación: 18 de mayo de 2010
Fecha de aceptación: 18 de mayo de 2010

Las pruebas educativas estandarizadas son dispositivos de evaluación masiva diseñados para proveer evidencia replicable (i.e. "*fiable*") a partir de la cual formular inferencias relevantes y fundamentadas (i.e. "*válidas*") sobre los conocimientos y habilidades de quienes las responden.

El desarrollo de una prueba estandarizada requiere resolver varios problemas técnicos de alta complejidad. Entre ellos, uno de los más cruciales es también el menos conocido fuera del círculo de especialistas en psicometría. Nos referimos al conjunto de problemas asociados a la *comparabilidad* de las mediciones que, paradójicamente, afecta directamente la interpretación y uso de las puntuaciones, que son el resultado más visible de prueba.

Desde el lado de los examinados, los problemas de *comparabilidad* aparecen cuando enfrentan pruebas estandarizadas cuyos resultados se expresan en métricas distintas para evaluaciones que se supone miden lo mismo. Por ejemplo, un estudiante podría obtener 30 puntos en una prueba de inglés contratada y aplicada por su escuela, 250 puntos en otra prueba de inglés rendida en el proceso de selección a un instituto y 400 puntos en una prueba aplicada por alguna agencia nacional o regional que esté realizando un diagnóstico sobre el nivel de inglés de los estudiantes. ¿Qué significan esos resultados dispares? ¿Hay alguna relación entre ellos? ¿En cuál de todas las pruebas el resultado es mejor? ¿Y por qué son diferentes las escalas de puntuación si todas las pruebas evalúan inglés?

Desde el lado de otros usuarios importantes, los investigadores en educación, la situación tampoco es fácil cuando deben enfrentarse a problemas de comparabilidad de mediciones. Para fines ilustrativos, sólo un ejemplo: ¿Cómo se explica que el resultado de un grupo de escuelas en una prueba estandarizada aplicada por el sistema nacional de medición sea de 240 puntos y que las mismas escuelas hayan obtenido 450 puntos en una prueba internacional en la que el país ha participado? ¿Qué significa esa diferencia? ¿Es posible llevar las puntuaciones de la prueba nacional a la escala de reporte de la prueba internacional?

Por otro lado, las instituciones responsables de estas pruebas (por ejemplo, las agencias nacionales de medición educativa o los organismos encargados de los procesos de selección universitaria) enfrentan una cara menos visible pero aún más delicada del problema de comparabilidad al tratar de resolver cuestiones del tipo ¿Cuánto podemos cambiar las especificaciones de la prueba de este año sin comprometer la comparabilidad con los resultados de años anteriores? ¿En qué medida un cambio curricular afecta la capacidad de nuestras pruebas para realizar un monitoreo escolar válido? ¿Cómo hacemos para distribuir los ítems de la prueba en varios cuadernillos distintos sin comprometer su validez ni sesgar los resultados? ¿Qué efecto podría tener en los resultados la técnica de *equiparación* que utilicemos?

Estos y otros muchos problemas asociados a la *comparabilidad* de las mediciones educativas son una fuente continua de dolores de cabeza para los especialistas. La mayoría de ellos puede resolverse mediante metodologías básicas de *equiparación* [del inglés *equating*][1] pero otros requieren de métodos especiales cuyas limitaciones y alcances no siempre son bien comprendidos.

[1] La traducción preferida para el término *equating* está fuertemente condicionada por la geografía. Aunque la voz hispana *equiparación* es, sin duda, la correcta y cuenta, además, con plena aceptación entre los especialistas, en Latinoamérica se escuchan frecuentemente otras alternativas, como "*igualación*", "*equivalencia*" o "*ecualización*". En nuestra opinión, ninguna de ellas supera a "*equiparación*" en fidelidad con el original y pueden, además, inducir a serios malentendidos. Por ejemplo, la acepción castellana de "*igualación*" conlleva la idea equivocada de que es posible lograr que dos pruebas sean iguales, cuando de lo que se trata es apenas de lograr equivalencia entre sus puntuaciones. Por otro lado, "*equivalencia*" respeta el significado del concepto, pero complica la tarea de expresarlo como verbo (i.e. ¿cómo decir que queremos equiparar dos pruebas?). Finalmente, "*ecualización*" es la traducción correcta del término "*equalization*", cuya definición específica en psicometría moderna (Kolen, 1988) es más cercana a su significado en ingeniería acústica que a la idea de *equiparación*.

En nuestra experiencia, por ejemplo, hemos encontrado que un error muy común entre los investigadores es creer que las técnicas basadas en regresión múltiple son un método óptimo para establecer el grado de equivalencia entre dos mediciones. Otra creencia muy arraigada es suponer que cualquier par de mediciones pueden *equipararse* o *igualarse* a todo evento, mediante la técnica de estadística adecuada, ignorando que la *equiparación* es una meta que no siempre se puede lograr empíricamente. Un equívoco relacionado, también muy frecuente, es creer que las técnicas de *equiparación* permiten igualar dos pruebas diferentes, cuando la verdad es que su único propósito es ajustar diferencias de dificultad entre las puntuaciones de ambas pruebas.

Estos y otros malentendidos se originan en una comprensión equivocada de las condiciones que posibilitan que una *equiparación* sea posible y en el desconocimiento de otros procedimientos para establecer *comparabilidad* entre mediciones. Sin embargo, métodos alternativos se han venido aplicando desde principios del siglo XX para, por ejemplo, alinear puntuaciones de pruebas distintas o monitorear progreso educativo. En algunos casos se trata de técnicas relativamente populares (e.g. *Escalamiento Vertical*) y en otros, de procedimientos apenas conocidos por un puñado de especialistas (i.e. *Escalamiento de Baterías*).

Familiarizarse con estas metodologías es necesario por dos razones. Primero, porque facilita visualizar soluciones para problemas de medición educativa que a veces se consideran irresolubles o, peor aún, sobre los cuales no se ha tomado la debida conciencia. Segundo, porque el contraste de estas metodologías con las técnicas de *equiparación* tradicionales ayuda a comprender mejor las condiciones en las cuales una *equiparación* es o no válida.

Lamentablemente, los pocos trabajos que ofrecen exposiciones de metodologías de *comparabilidad* adicionales a la equiparación (Kolen, 2004; Kolen & Brennan, 2004; Holland & Dorans, 2006; Dorans, Pommerich & Holland, 2007) están dirigidos a un público angloparlante y altamente especializado en psicometría. Por ello, y con el propósito de contribuir a la divulgación en nuestro medio de algunas metodologías alternativas a la *equiparación,* en este artículo ofrecemos una revisión sintética de los métodos más importantes para lograr *comparabilidad* entre mediciones. En cada caso, se explica brevemente el propósito de método, se revisa un ejemplo prototípico y se sugieren algunas aplicaciones potenciales en nuestro contexto regional. De acuerdo al objetivo didáctico del trabajo, se optó por no describir los diseños de recogida de información o las técnicas estadísticas para cada procedimiento, aunque se sugieren referencias especializadas para el lector interesado.

La exposición que presentamos a continuación se basa en el marco conceptual propuesto por Holland y Dorans (2006), también descrito por Holland (2007) y por Dorans, Holland & Petersen (2007). Aunque a través de los años diferentes especialistas han desarrollado marcos teóricos para sistematizar los métodos de *comparabilidad* (ver, por ejemplo, los trabajos de Flanagan, 1951; Angoff, 1971; Mislevy, 1992; Linn, 1993; Feuer, Holland, Green, Bertenthal & Hemphill, 1999; Dorans, 2000, 2004), la taxonomía de Holland y Dorans (2006) es la más comprehensiva y mejor fundamentada en el presente.

1. Conceptos básicos: enlace de puntuaciones

Para comenzar, es necesario circunscribir el problema de *comparabilidad entre mediciones* a aquellas situaciones que requieren construir algún tipo de regla de correspondencia que nos permita expresar el resultado de una prueba en la métrica de la otra.

Técnicamente hablando, la solución a estos problemas es aplicar alguna metodología de *enlace de puntuaciones* [del inglés *score linking*][2] es decir, *una transformación estadística que permita encontrar a qué puntuaciones en una prueba equivalen cada una de las puntuaciones de la otra prueba* (Holland & Dorans, 2006).

Nótese que en adelante utilizaremos el término *enlace* en lugar de "*comparabilidad*" [del inglés *comparability*], porque este último no tiene una definición inequívoca en la literatura psicométrica. Por ejemplo, Flanagan (1951) lo utilizó en un sentido más restringido que *enlace,* para describir algunas de las metodologías que hoy en día se agrupan bajo la noción de *alineamiento de puntuaciones*, que explicaremos más adelante. Veinte años después, Angoff (1971) acotó aún más el concepto, utilizándolo sólo para aquellos *enlaces* entre pruebas que midieran constructos diferentes. En la siguiente década (APA, 1985) se lo utilizó para referir a todos los tipos de *enlace* que no alcanzaran el rigor técnico de una *equiparación* propiamente tal [*scaling to achieve comparability*]. Desde entonces se lo emplea de vez en cuando como sinónimo informal de *equiparación* o para referir a problemas de *enlace* en un sentido amplio, pero sin mediar una definición técnica precisa ni mucho menos completamente aceptada. Para complicar aún más las cosas, es cada vez más frecuente utilizar el término *comparabilidad* para denominar al área específica de estudio que se ocupa de evaluar la equivalencia psicométrica entre las versiones informatizada y tradicional de una misma prueba [*"test comparability studies";* ver por ejemplo, ejemplo, Sireci & Zenisky, 2006)].

Por las razones anteriores, las publicaciones especializadas privilegian el uso del término *enlace*, cuya definición técnica es precisa, en lugar de *comparabilidad,* de uso coloquial y significado ambiguo.

La definición del *enlace de puntuaciones* es suficientemente amplia como para describir situaciones muy diversas, que involucran metodologías distintas. En términos generales, las metodologías de *enlace* se diferencian entre sí por la validez, la precisión y la reversibilidad con que se pueden transformar las puntuaciones de una prueba en otra, pudiéndose distinguir tres grandes grupos de métodos, desde los más laxos hasta los más rigurosos en el cumplimiento de esas condiciones: *predicción*, *alineamiento* y *equiparación* de puntuaciones.

2. Predicción de puntuaciones

La *predicción* es la metodología de enlace más antigua que se conoce y una de las más utilizadas en investigación aplicada. Su propósito es predecir la puntuación esperada en una evaluación a partir de otra información relevante. Esa información puede consistir en los resultados de una o más pruebas pero también puede incluir uno o varios antecedentes sociodemográficos de importancia (e.g. género, etnia, nivel socioeconómico). Aunque no es una regla, es habitual que haya una diferencia temporal entre el momento en que se mide esa información relevante y el momento en que se obtiene la puntuación a ser predicha.

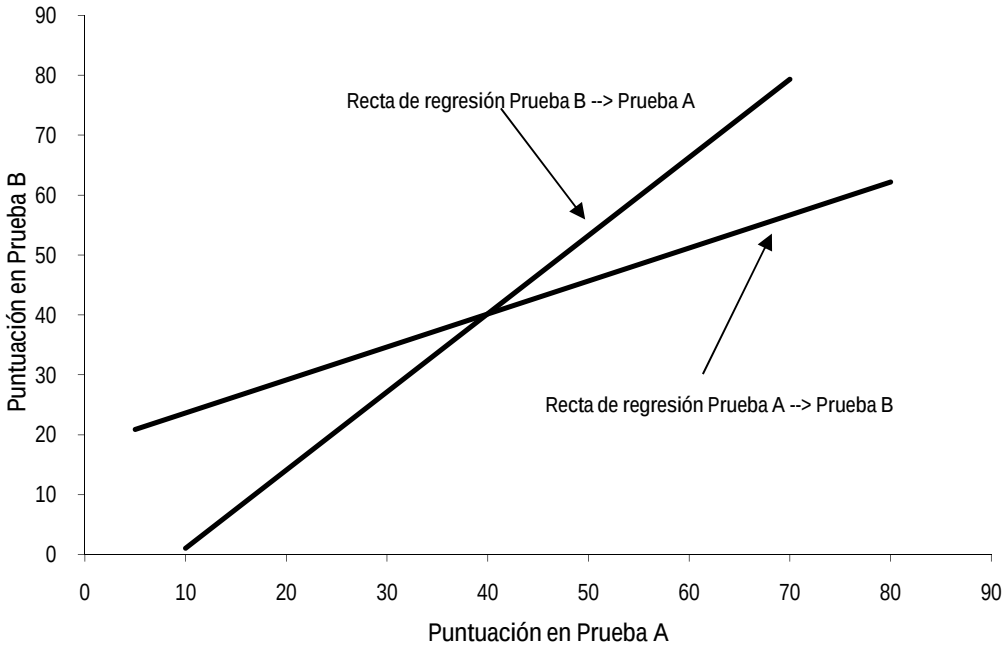
[2] En la psicometría hispanoparlante no existe un equivalente técnico universalmente aceptado para los conceptos psicométricos *link* y *linking*. Hemos optado por la traducción *enlace* porque nos parece más fiel al significado psicométrico del original, en comparación con las otras alternativas disponibles en castellano (i.e. *vínculo*, *conexión*, *unión*, *ligazón*).

El caso más común en que aplican técnicas de *predicción* corresponde a las pruebas de selección universitaria, cuya validez depende de su capacidad para predecir, con la mayor precisión posible, el rendimiento posterior en la Universidad (otra medición basada, por ejemplo, en el promedio de calificaciones del primer semestre). Otro ejemplo clásico es el intento de predecir el resultado en una prueba de selección universitaria rendida al término de la escolaridad (normalmente, al finalizar del 12° grado), a partir de los resultados obtenidos en pruebas de rendimiento educativo aplicadas durante los años previos de escolaridad (por ejemplo, en 10° grado).

En los *enlaces por predicción* de puntuaciones, el énfasis no suele estar puesto en predecir el resultado específico de cada examinado, sino en la puntuación esperada para el alumno "típico" de acuerdo a ciertas variables de contexto o explicativas. Por ejemplo, predecir el rendimiento más probable en la Universidad, a partir de su prueba de selección, para los alumnos de una determinada titulación, en función de que hayan estudiado en escuelas públicas o privadas. En este sentido, los *enlaces* por *predicción* generalmente se utilizan en estudios de *validez predictiva* y en la construcción de modelos de crecimiento escolar [*growth models*].

Una característica crucial de la *predicción* es que existe una asimetría fundamental entre la puntuación predicha y las variables que se utilizan como predictores. Hay razones lógicas y estadísticas para ello. Por una parte, los predictores pueden ser varios y anteceder temporalmente a la puntuación predicha (que es sólo una), por lo cual resulta lógico que la predicción opera en una dirección pero no al revés. Por ejemplo, tiene sentido predecir el rendimiento en el primer semestre de Universidad a partir de los resultados de una o varias pruebas de selección universitaria, pero no parece razonable predecir el rendimiento pasado en esas pruebas a partir del rendimiento presente en la universidad.

FIGURA 1. EJEMPLO DE ASIMETRÍA EN LA REGRESIÓN LINEAL SIMPLE



Existe, además, una poderosa razón estadística para esta asimetría. La técnica preferente para los estudios de predicción es la regresión lineal o alguna de sus variantes. Es bien sabido, desde la invención de esta técnica, a fines del siglo XIX (Galton, 1888), que la recta de regresión para predecir Y a partir de X no es exactamente la inversa de la recta para predecir X a partir de Y. Esta idea se ilustra en la Figura 1, en que se grafican las rectas de regresión entre las pruebas A y B, cuyas puntuaciones tiene una correlación a *r=0.85*. Podemos observar que si predecimos los resultados de B utilizando A y luego intentamos predecir la puntuación de A mediante el valor de B recién obtenido, llegaremos a un resultado distinto del original. Este fenómeno, bien conocido por los estadísticos, pero muchas veces olvidado por los investigadores aplicados, redunda en que las predicciones conseguidas mediante regresión lineal no son reversibles. Como veremos más adelante, la falta de reversibilidad o, lo que es lo mismo, la asimetría de la predicción, supone una distinción fundamental entre el *enlace* de puntuaciones por *predicción* y los *enlaces* por *alineamiento* o *equiparación*. Existen dos subtipos básicos de *predicción*: *predicción de puntuaciones individuales* y *proyección de distribuciones*.

2.1. Predicción de puntuaciones individuales

Este tipo de *enlace* es la predicción por antonomasia y se define por las mismas características antes mencionadas. El ejemplo prototípico, como hemos repetido varias veces, es el de las pruebas de selección, aunque otros usos interesantes también son posibles. Entre ellos, predecir el rendimiento en una prueba de término de grado (e.g. 8° grado) a partir de los resultados de una prueba anterior (e.g. 6° grado), o predecir el resultado de una prueba de selección universitaria, rendida al término de la escolaridad, a partir de pruebas estandarizadas rendidas en los años previos y/o a través de una combinación sensata de antecedentes académicos del estudiante.

Una vez obtenida la función de *enlace* entre ambas puntuaciones (i.e. una ecuación de regresión lineal), ésta puede aplicarse para predecir el rendimiento de otros evaluados que provengan de una población equivalente a aquella con la cual fue realizado el estudio.

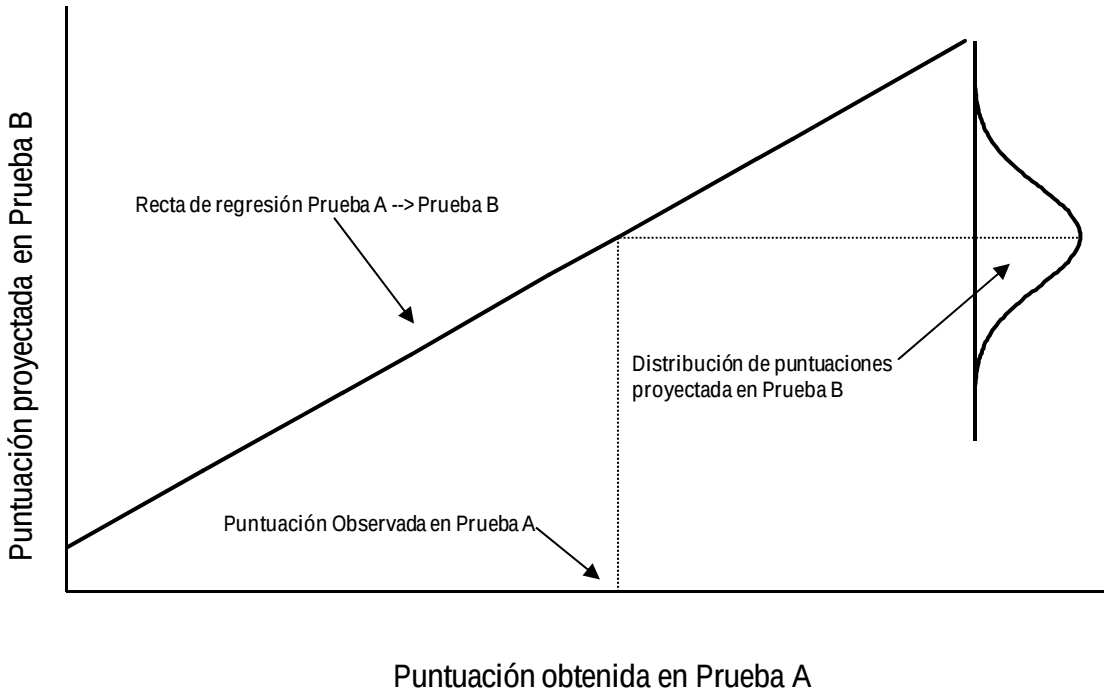
La *predicción de puntuaciones individuales* siempre es unidireccional y tiene una margen de error que viene dado por la fuerza de la asociación entre las variables y el respectivo coeficiente de determinación de la regresión, simple o múltiple (R^2), que permite cuantificar la proporción de varianza explicada por los predictores. Otro punto importante, es que la eficacia de predicción está sujeta al cumplimiento de los supuestos del modelo de regresión, pudiendo ésta ser de tipo lineal, simple o múltiple, aunque también es admisible trabajar con regresión no lineal o con regresión jerárquica (también llamada regresión multinivel o modelos HLM) si los predictores así lo ameritan.

Finalmente, hay que advertir que los *enlaces por predicción* suelen ser vulnerables a dos problemas metodológicos que requieren de un cuidadoso control. Primero, como típicamente la muestra final está sesgada selectivamente "hacia arriba" (sólo los alumnos que obtuvieron altos rendimientos en la prueba de selección ingresan efectivamente a la Universidad) la *predicción* puede parecer menos robusta de lo que realmente es (i.e. una correlación más baja de lo esperado), debido a problemas de restricción de rango (para una revisión ver, por ejemplo, Sackett & Yang, 2000). Segundo, es necesario determinar si la *predicción* está sujeta al llamado *sesgo externo* o *predicción diferencial* (Camilli, 2006), como se denomina al hecho de que la ecuación de regresión funcione diferencialmente para los examinados de un determinado grupo socioeconómico, por ejemplo. En el contexto de las metodologías de *enlace de puntuaciones*, la *predicción diferencial* atenta contra el requerimiento de *invarianza* de la función de *enlace* (ver sección 4.2).

2.2. Proyección de distribuciones

Un problema relacionado, aunque menos conocido, consiste en proyectar la distribución de puntuaciones en una prueba a partir de los resultados de otra. En este caso, a partir de los resultados obtenidos en una muestra de evaluados que haya respondido ambas pruebas, se intenta predecir la distribución condicional de la prueba B para cada puntuación observada de la prueba A, mediante ciertas variantes de los métodos de regresión lineal. Posteriormente, la función de *enlace* así obtenida se puede utilizar para proyectar la distribución de puntuaciones de la prueba B en otra muestra de estudiantes que sólo haya respondido la prueba A, pero que tenga características comparables a la muestra original (i.e. provenga de la misma población). Una representación gráfica de la idea de *proyección de distribuciones* se presenta en la Figura 2.

FIGURA 2. ENLACE POR PROYECCIÓN DE DISTRIBUCIONES



La característica más importante de la *proyección de distribuciones* es que cada puntuación de la prueba A no predice valores puntuales sino una distribución de puntuaciones en la prueba B; estas distribuciones son posteriormente agregadas para obtener un resultado total, como sucede con los *valores plausibles* utilizados en pruebas como NAEP o PISA. De hecho, las principales aplicaciones de la proyección pueden encontrarse en los esfuerzos para enlazar pruebas que utilizan esa metodología en mediciones a nivel agregado. Un ejemplo emblemático es el estudio realizado por Pashley y Phillips (1993) para proyectar las puntuaciones del *Internacional Assessment of Educationa Progress* (IAEP) en la escala de Nacional *Assessment o Educational Progress* (NAEP) de Estados Unidos.

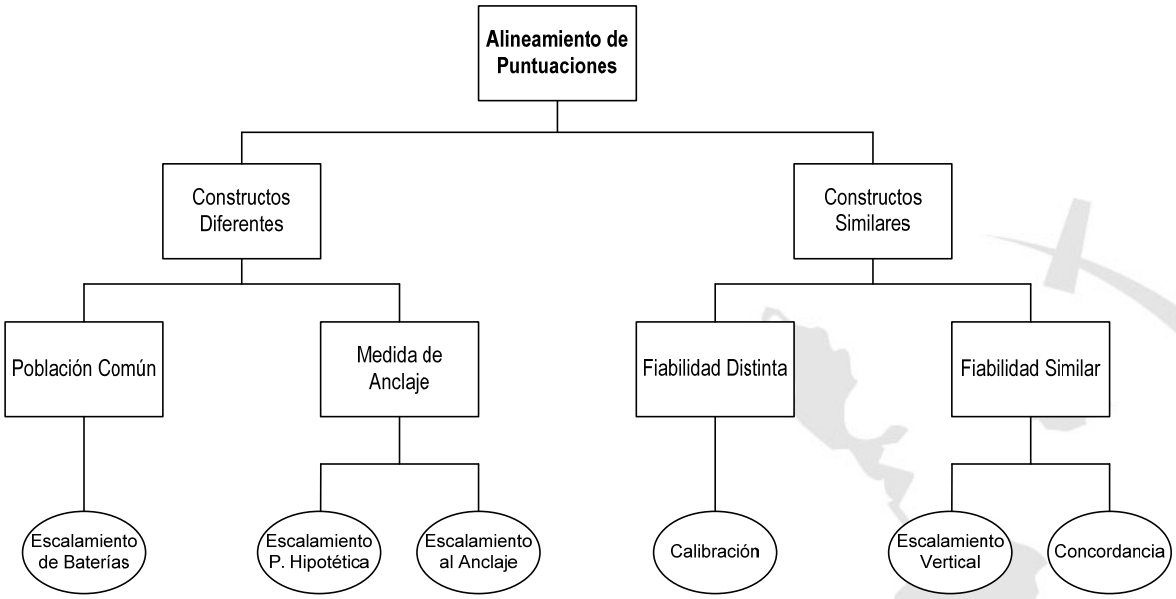
Cabe notar que un área de investigación potencialmente interesante y, hasta donde sabemos, inexplorada en Ibero América, es el enlace entre las mediciones nacionales de aprendizaje de cada país y los resultados obtenidos en las pruebas internacionales en las que se participe. El método de *proyección de distribuciones* es una alternativa metodológica en el caso que la prueba internacional entregue resultados basados en valores plausibles (e.g. PISA).

3. ALINEAMIENTO DE PUNTUACIONES

Bajo el concepto de alineamiento de puntuaciones [del inglés *score aligning*] se agrupan un conjunto de métodos que comparten entre sí el propósito de transformar los resultados de mediciones distintas a una escala común, para obtener así puntuaciones comparables. Mientras la meta de la predicción era lograr la predicción más exacta posible, el objetivo del alineamiento es lograr *puntuaciones comparables*. Ejemplos típicos de metodologías de *enlace* por *alineamiento* son los estudios de *Concordancia* de puntuaciones y la construcción de *Escalas Verticales*, que se explicarán más adelante.

El desarrollo de los métodos de *alineamiento*, iniciado a principios del siglo XX, ha sido bastante desordenado y no sigue una secuencia necesariamente lógica. En su taxonomía de métodos de *enlace* Holland y Dorans (2006) intentan clasificarlos atendiendo a cuestiones tales como el propósito de la evaluación, el diseño empleado y las características psicométricas de las pruebas. Para los fines de esta revisión no haremos una presentación exhaustiva sino que intentaremos describir brevemente los métodos más utilizados, de acuerdo a la organización conceptual presentada en la Figura 3, adaptada de Holland (2007). Como puede observarse, la primera distinción importante es si se pretenden *alinear* pruebas que evalúan constructos similares o distintos. Las restantes distinciones tienen que ver con el diseño de recogida de información y con las características técnicas de la pruebas.

FIGURA 3: TIPOS DE ALINEAMIENTO ENTRE PUNTUACIONES



3.1. Escalamiento de Baterías: Constructos diferentes evaluados en la misma población de examinados

Cuando dos o más pruebas que miden constructos diferentes son administradas a la misma población de evaluados, los resultados de cada una de ellas pueden ser transformadas de manera que tengan una distribución común para esa población, por ejemplo, una distribución normalizada. En otras palabras, ambas pruebas son transformadas a una escala de referencia con características distribucionales conocidas.

La denominación *Escalamiento de Baterías* para este tipo de situaciones puede encontrarse en Kolen (2004), aunque tiene sus orígenes en los trabajos de Kelley (1914, 1923), Flanagan (1951) y Angoff (1971).

Sin entrar en los pormenores estadísticos, un ejemplo de este tipo de alineamiento es el estudio para re-centrar las escalas del SAT, en Estados Unidos (Dorans, 2002), uno de cuyos objetivos era que las escalas Verbal (SAT-V) y Matemática (SAT-M) tuvieran la misma distribución en la población de referencia, definida como la muestra de estudiantes evaluada en 1990. Gracias a este procedimiento un estudiante que obtiene una puntuación más alta en el SAT-V que en el SAT-M puede concluir, con toda propiedad, que su rendimiento fue mejor en el área Verbal que en el área Matemática de la prueba. En este ejemplo, queda muy claro como el Escalamiento de Baterías es una técnica para conseguir escalas con una distribución comparable (¡y no puntuaciones equivalentes!) en pruebas que evalúan constructos distintos en una población común de examinados. No hace falta decir que el enlace entre ambas pruebas no implica que las puntuaciones puedan utilizarse en forma intercambiable y que el error de este tipo de alineamiento es conceptualmente mayor que en los métodos que alinean pruebas dirigidas a evaluar constructos similares. Finalmente, es obvio que el error del alineamiento, en este caso, es inversamente proporcional a la magnitud de la correlación entre ambas pruebas.

3.2. Escalamiento mediante anclajes: Diferentes constructos, evaluados en poblaciones distintas

A veces es necesario alinear pruebas que miden constructos diferentes pero que no han sido respondidas por miembros de la misma población. En este caso, una solución es utilizar una *medida de anclaje* [del inglés *anchor measure*], es decir una tercera medición que sí ha sido administrada a miembros de ambas poblaciones. Esta *medida de anclaje* es utilizada como un puente para *enlazar* las puntuaciones originales.

Un ejemplo sería, por ejemplo, *enlazar* dos pruebas de idiomas, una de Francés y otra de Inglés, aplicadas a estudiantes de enseñanza secundaria, para establecer una métrica comparable. De este modo podríamos expresar en una misma escala ambos resultados y comparar el rendimiento de los estudiantes en una y otra prueba. El problema, es que normalmente cada estudiante responderá sólo aquella prueba que corresponda al idioma que cursa. Esto impide tener una muestra de examinados que hayan rendido ambas pruebas, lo cual imposibilita utilizar un *Escalamiento de Baterías*. La alternativa es emplear como *medida de anclaje* el resultado de una o varias pruebas de otras asignaturas que hayan sido administradas a ambos grupos de evaluados (o incluso variables sociodemográficas de interés) y que correlacionen con el resultado de las pruebas de idiomas. De este modo, es posible *enlazarlas* en una escala con una distribución común. En este sentido, el *escalamiento mediante anclajes* puede concebirse como una aproximación al *escalamiento de baterías* en caso de no contar con muestras equivalentes (i.e. pertenecientes a la misma población).

Respecto a los procedimientos estadísticos utilizados es posible diferenciar dos aproximaciones. El *Escalamiento a una Población Hipotética* es similar a la *Proyección* (sección 2.2) en tanto proyecta las distribuciones que corresponderían a cada prueba a partir de los resultados obtenidos en la *medida de anclaje*.

Algo más común es el *Escalamiento al Anclaje*, que algunos autores han denominado *moderación estadística* (McGaw, 1977; Mislevy, 1992; Keeves, 1998). Por ejemplo, Linn (1993) describe un caso en el cual este método fue utilizado para lograr la comparabilidad de puntuaciones en varias pruebas distintas que fueron aplicadas por algunas escuelas de un distrito escolar. El problema era cómo comparar

resultados que se habían obtenido con instrumentos diferentes, y que no evaluaban el mismo constructo. Utilizando como *medida de anclaje* una prueba estandarizada que se administró a todos los estudiantes de ese distrito, se consiguió poner en una escala comparable los resultados de las pruebas específicas de cada escuela.

Aunque no conocemos trabajos de este tipo de Latinoamérica, creemos que estudios de esta naturaleza serían muy interesantes de replicar para, por ejemplo, lograr resultados comparables en aquellos países en que exista alguna prueba estandarizada nacional (o regional) de logro escolar y se desee obtener una escala común para las pruebas aplicadas por agencias privadas bajo el contrato de municipios o redes de colegios. Nótese que en este caso, estamos hablando de enlazar las pruebas específicas entre sí a través de una prueba nacional. Si el objetivo fuera enlazar una prueba específica con la prueba estandarizada nacional, podríamos recurrir a una *Tabla de Concordancia* (ver apartado 3.4), siempre y cuando se cumplieran los requerimientos para ese método.

3.3. Calibración: El mismo constructo evaluado en la misma población con pruebas de diferente fiabilidad

El término *calibración* ha sido utilizado con varias acepciones en la literatura psicométrica. Por ejemplo, Angoff (1971) lo emplea para referirse al tipo de enlace que actualmente se conoce como *Escalamiento Vertical* (ver apartado 3.5), mientras la mayoría de los especialistas en Teoría de Respuesta al Item (TRI) llaman "calibrar" al proceso de obtener los parámetros de los ítems analizados bajo ese enfoque psicométrico. En la taxonomía de Holland y Dorans (2006) *calibración* describe un tipo muy específico de alineamiento entre dos pruebas, que miden el mismo constructo y tienen niveles de dificultad similares, pero que difieren en fiabilidad[3].

Habitualmente, en pruebas que miden el mismo constructo con niveles equivalentes de dificultad, la diferencia de fiabilidades es provocada por diferencias en longitud (número de ítems), por lo que no es sorprendente que el caso más típico de *calibración* sea el alineamiento de los resultados de una prueba con una versión abreviada de la misma.

Aunque es poco común en medición de rendimiento educativo, un ejemplo razonable sería el caso de una prueba de competencias o habilidades intelectuales de 100 ítems, para la cual se construye una forma abreviada de, por ejemplo, 20 ítems para servir como medida de tamizaje [screening]. En este caso, la *calibración* consistiría en construir una escala que permitiera poner el resultado basado en 20 items en la misma escala de la prueba de 100 items. Un problema práctico es que la *calibración,* por sí misma*,* no puede compensar la inevitable pérdida de fiabilidad y de validez causada por la reducción de la prueba original.

3.4. Concordancia: Constructos similares evaluados con pruebas de igual fiabilidad y dificultad en la misma población

Una consecuencia imprevista y cada vez más frecuente de la proliferación de mediciones estandarizadas es que muchas veces un mismo estudiante responde a varias pruebas destinadas a propósitos similares. Por ejemplo, un postulante podría examinarse en dos universidades distintas, cada una de las cuales

[3] Este es otro caso en que las preferencias lingüísticas varían a ambos lados del Atlántico. Mientras el término anglosajón *reliability* ha sido traducido en España como *fiabilidad*, en Latinoamérica se utiliza preferentemente la expresión *confiabilidad*. En este artículo hemos optado por la versión ibérica dado que, en nuestro parecer, refleja con mayor precisión el significado del concepto y remite a un uso técnico del mismo, evitando las malas interpretaciones que surgen cuando se discuten cuestiones científicas con palabras de uso cotidiano.

aplica sus propias pruebas de selección. ¿Es posible establecer alguna equivalencia entre ambas pruebas? ¿Cómo comparar su rendimiento relativo y saber en cuál obtuvo un mejor desempeño?

En casos como estos, estamos frente a dos pruebas diseñadas para evaluar constructos similares, que muy probablemente fueron construidas siguiendo especificaciones técnicas distintas (a veces, sólo levemente distintas). Sin embargo, fuera de algunas diferencias en los subfactores evaluados o en los marcos conceptuales que las fundamentan, ambas pruebas pueden tener muchas similitudes técnicas, como por ejemplo, alta fiabilidad o número equivalente de ítems. Si además están destinadas al mismo propósito y son rendidas por miembros de la misma población, es posible *enlazar* sus puntuaciones a través de un *alineamiento* por *concordancia*. El resultado añade valor a ambas pruebas, al permitir que sus resultados se expresen en una métrica común (¡pero no equivalente!). Un ejemplo clásico de este tipo de problemas y su solución, son los estudios de *concordancia* entre la pruebas SAT y ACT en Estados Unidos (Dorans, 1999).

El producto final de este tipo de *alineamiento* suele ser una *Tabla de Concordancia,* como las que ofrece el College Board en su sitio web (www.collegeboard.com) para demostrar las correspondencias entre las puntuaciones de las pruebas SAT y ACT. Estas Tablas permiten a cualquier estudiante que se haya examinado en ambas pruebas comparar sus resultados en ellas. Por otro lado, aquellos estudiantes que sólo rindieron una de las pruebas pueden formarse una idea aproximada de cuál habría sido su rendimiento en la otra.

Pese a su evidente utilidad los estudios de *concordancia* no están exentos de crítica, pues resulta muy tentador usar sus resultados como si las pruebas fueran equivalentes y sus puntuaciones *intercambiables* (Pommerich, 2007). Como veremos en la sección 3 de este artículo, sólo los procedimientos de *equiparación* pueden garantizar (y sólo bajo el cumplimiento de supuestos bastante estrictos) que los resultados de una prueba sustituyan a los resultados de otra.

Por otro lado, debemos reconocer que la actual abundancia de tests y pruebas estandarizadas amerita ampliamente la realización de estudios serios de *concordancia* entre mediciones, para establecer hasta qué grado son comparables los resultados de pruebas distintas.

En ese sentido, además de la comparabilidad de resultados entre pruebas de selección universitaria administradas por distintas instituciones, otra área que se beneficiaría de estudios de *concordancia* es el alineamiento entre pruebas regionales y pruebas estandarizadas a nivel nacional. ¿Es posible establecer algún tipo de concordancia entre los resultados de la prueba nacional y las pruebas regionales? ¿Cómo evaluar el nivel de alineamiento entre ambas?

Un problema del mismo tipo, que suponemos se incrementará con los años, es la necesidad de establecer *alineamientos* entre las mediciones internacionales de rendimiento educativo (i.e. PISA, TIMSS, SERCE) y las respectivas mediciones nacionales que cada país haya desarrollado. En nuestra experiencia, una pregunta que se formula cada vez con mayor insistencia a los especialistas es ¿A qué puntuación en la prueba internacional X equivale el resultado que obtuvo este año la escuela "A" en nuestra prueba nacional? Cabe observar que en este último caso, la elección entre un estudio de *concordancia* o de *proyección* estaría determinada por el nivel de similitud conceptual entre los constructos evaluados por ambas pruebas y por la metodología y nivel del análisis de los datos (e.g. puntuaciones individuales versus puntuaciones agregadas, mediante valores plausibles, por ejemplo).

Los lectores interesados en profundizar en metodologías de *concordancia* se beneficiarán de la lectura del

capítulo de Pommerich (2007), que presenta una buena discusión sobre las ventajas y limitaciones de las técnicas disponibles. Otra fuente autorizada es el número especial de la revista *Applied Psychological Measurement*, editado por Pommerich y Dorans (2004). Para familiarizarse con los procedimientos estadísticos necesarios, lo mejor es revisar el texto de Kolen y Brennan (2004) y poner especial atención a las técnicas equipercentiles, que se utilizan tanto en estudios de *equiparación* como de *concordancia*.

3.5. Escalamiento Vertical: Constructos similares evaluados con pruebas de igual fiabilidad pero distinta dificultad, en población diferentes

En parte debido a la legislación *No Child Left Behind* (NCLB) de Estados Unidos y su consecuente impulso a los modelos de crecimiento [*growth models*], y en parte gracias a la creciente demanda universal por incrementar el monitoreo del rendimiento de los estudiantes durante toda su escolaridad, en los últimos años ha cobrado fuerza la idea de alinear los resultados de las pruebas de rendimiento de distintos grados escolares, a través de estudios de *Escalamiento Vertical.*

Para contextualizar el problema, recordemos que muchos países aplican pruebas estandarizadas a nivel nacional para evaluar el rendimiento de sus estudiantes al término del año escolar, en distintos grados (por ejemplo, en 2°, 4°, 6° y 8° de primaria). Técnicamente, esas pruebas evalúan constructos similares (no son exactamente los mismos, dado que los cambios curriculares entre un nivel y otro introducen, necesariamente, variaciones en la definición del constructo), con la misma fiabilidad y requerimientos técnicos, pero en poblaciones distintas (estudiantes de distinto grado) y con diferentes rangos de dificultad (las pruebas de los grados superiores son, necesariamente, más difíciles que las de grados inferiores).

En un típico sistema nacional de evaluación, el resultado de los estudiantes se expresa en escalas de desviación, *relativas al desempeño promedio observado en cada grado*. La manera habitual de construir estas escalas es la siguiente. En la primera ocasión en que se administra la prueba en un grado determinado (por ejemplo 2° de primaria), se transforma la media y desviación estándar de las puntuaciones obtenidas a una media y desviación arbitraria, que definirá la métrica de la escala de evaluación para ese grado escolar. Por ejemplo, muchas pruebas estandarizadas utilizan escalas definidas por una media y desviación estándar de (250,50) puntos o de (500, 100) puntos, respectivamente. Supongamos que un sistema de evaluación elige una escala con media de 250 puntos para definir la métrica de sus pruebas en todos los grados evaluados.

En los años posteriores y bajo condiciones de *equiparación* apropiadas (ver sección 3), los resultados de cada grado representarán la distancia relativa respecto al promedio del año de origen de la escala. Por ejemplo, si el primer año una escuela obtiene un promedio de 270 puntos, puede concluir que se encuentra a 20 puntos por sobre el promedio de rendimiento a nivel nacional. Si al año siguiente obtiene 280 puntos, puede concluir con toda certeza que su promedio creció en 10 puntos y que ahora está a 30 puntos de distancia del promedio de origen (Naturalmente, si ese año el promedio nacional también mejoró y alcanzó 260 puntos, la escuela habrá progresado, pero en términos relativos su rendimiento no será mejor al del año anterior). De este modo, para cada grado evaluado, las escalas de desviación permiten medir el progreso anual de las escuelas respecto al rendimiento promedio de la población en el año en que comenzó el sistema de medición.

Ahora bien, un problema con este tipo de escalas es que los resultados de cada grado no son comparables entre sí, porque están anclados a su propio año de origen. Por lo mismo, su capacidad para monitorear el progreso individual de los estudiantes resulta muy limitada. Por ejemplo, si un estudiante

obtuvo 240 puntos en 2° de primaria y dos años más tarde logró 250 puntos en 4° grado, eso no significa que su rendimiento mejoró en 10 puntos, sino simplemente que en 2° se encontraba bajo el promedio de los otros estudiantes y en 4° tiene un rendimiento equivalente al promedio de los otros niños de su grado.

En otras palabras, la escala de resultados para cada grado no permite cuantificar el progreso relativo cada estudiante entre distintos años de escolaridad. El problema, por lo tanto, es: ¿Cómo *alinear* los resultados de las pruebas de cada grado, para que permitan mostrar el progreso o crecimiento relativo del rendimiento entre años de escolaridad? ¿Es posible *enlazar* las pruebas específicas para cada grado y transformarlas todas a una escala continua de crecimiento o progreso estudiantil?

Frente a este tipo de problema, la solución es llevar a cabo un tipo particular de *alineamiento*, denominado por *Escalamiento Vertical,* cuyo propósito es poner los resultados de cada uno de los grados en una escala común. Esta es llamada *Escala Vertical*[4] y permite monitorear el rendimiento individual de cada alumno a lo largo del tiempo, idealmente a través de todo el ciclo escolar. De hecho, buena parte del impulso que han recibido los métodos de *Escalamiento Vertical* se justifican en la promesa de medir apropiadamente el progreso o crecimiento [*growth*] de los estudiantes, lo que a su vez permite contextualizar mejor los estudios de valor añadido [*value added*] tan en boga últimamente.

A diferencia de las escalas en que se reportan los resultados para grado (llamadas, en este contexto, *Escalas Intragrado* o *Escalas Horizontales*), las *Escalas Verticales* permiten alinear el rendimiento de todos los grados en una única escala global (por ejemplo, de 0 a 100 puntos) y así cuantificar el monto de aprendizaje que los estudiantes *ganan* entre un curso y el siguiente. De este modo, cumplen a lo menos con tres propósitos muy atractivos. Primero, permiten a los padres y profesores monitorear el progreso individual de sus estudiantes y emprender medidas remediales si es necesario. Segundo, ofrecen a los investigadores información muy valiosa para modelar los factores que determinan el crecimiento o progreso educativo de los alumnos [*growth models*]. Tercero, esos mismos datos, una vez procesados, pueden emplearse para cuantificar el valor añadido [*value added*] de los profesores o los colegios sobre el aprendizaje de los estudiantes y así retroalimentar la toma de decisiones y el diseño oportuno y eficiente de políticas educativas.

Naturalmente, el desarrollo de una *Escala Vertical* es bastante complejo técnicamente (mucho más que una *equiparación* tradicional) y requiere de un cuidadoso proceso de planificación. A partir del estudio de fuentes especializadas, de la revisión de *Escalas Verticales* construidas en Estados Unidos y de nuestra propia experiencia, creemos que hay seis puntos críticos a tener en cuenta en el desarrollo de una *Escala Vertical.*

El primero es decidir el nivel de análisis de los datos, esto es, reporte individual (para cada estudiante), reporte a nivel agregado (para investigación sobre crecimiento escolar, análisis del valor añadido de los profesores y escuelas, o para el diseño y evaluación de impacto de políticas educativas, por ejemplo) o ambos. La alternativa elegida condicionará directamente el diseño y los métodos de análisis psicométrico y puntuación e, indirectamente, el error en la escala de reporte de los resultados. Naturalmente, estos usos no son necesariamente contradictorios, pero su elección pasa por decisiones de política educativa y no por cuestiones técnicas, necesariamente.

[4] A veces se utiliza la denominación *Escala Evolutiva* o *Escala de Desarrollo* [*Developmental Scale*]. Un problema con esos términos es que también se utilizan (en inglés y en castellano) para referir a pruebas de habilidades o de inteligencia para niños pequeños.

El segundo punto es que la validez, fiabilidad y precisión de una *Escala Vertical* es inversamente proporcional a la distancia entre los grados que se pretende alinear. La situación ideal es alinear grados adyacentes (5°, 6° y 7°, por ejemplo) o, en su defecto, diseñar un mecanismo para conseguir datos de calidad en grados consecutivos, idealmente con no más de 2 años de diferencia (por ejemplo, utilizar una medición en 6° grado como puente entre pruebas aplicadas en 4° y 8° grado). Si ello es no es posible, resulta necesario aumentar el número y frecuencia de las mediciones, con los costos que ello supone (Haertel, 1991; Huynh & Schneider, 2004).

Un tercer elemento a considerar es que las *Escalas Verticales* son muy sensibles a la dimensionalidad de las pruebas, así que sólo son recomendables para aquellas áreas de aprendizaje como Lectura o Matemáticas (con alguna reserva) en que la habilidad subyacente no sufre modificaciones drásticas a lo largo de la escolaridad. El mismo motivo complica el desarrollo de Escalas Verticales en Ciencias (Lissitz & Huynh, 2003), donde el constructo subyacente aumenta progresivamente su especificidad a través del ciclo escolar, comenzando como "ciencias" en forma global, para llegar a convertirse en "física", "química" y "biología" como componentes diferenciados.

El cuarto punto es que el *Escalamiento Vertical* requiere adoptar previamente una *definición de crecimiento* que opere como marco conceptual para el diseño e interpretación de la escala. Siguiendo a Kolen y Brennan (2004) hay dos enfoques predominantes. En la definición de *dominio de crecimiento* se asume que la habilidad evaluada es un continuo de contenidos y habilidades, que se expresa en distintos niveles de complejidad en las respectivas pruebas para cada grado. Bajo esta definición, la construcción de la *Escala Vertical* requiere de una *prueba de escalamiento* que se aplica a todos los estudiantes de los cursos que se desean alinear (por ejemplo 5° a 8° grado) y sirve para enlazar posteriormente las respectivas *Escalas Horizontales* de cada grado. Por el contrario, en la *definición grado a grado*, se asume que hay contenidos y habilidades propias de cada grado evaluado y el progreso escolar se entiende como el logro sucesivo de las habilidades esperadas para cada grado. Si se adopta esta definición, el diseño de la *Escala Vertical* utilizará alguna variación de los diseños generales de enlace que se emplean para equiparar puntuaciones (Kolen & Brennan, 2004).

Por supuesto, además de su impacto en el diseño, el tipo de definición de crecimiento condiciona las interpretaciones posibles de los resultados y debe estar alineada con los marcos conceptuales que guían el desarrollo de las respectivas pruebas para cada grado. En este sentido, es necesario enfatizar que en toda *Escala Vertical* subyace una definición de crecimiento. A veces, esta definición se encuentra implícita en las decisiones que se toman para el diseño de la escala. Lo óptimo, sin embargo, es que la definición de crecimiento sea explícita al inicio del proceso, de manera tal que sirva de guía para alinear el diseño, análisis, reporte e interpretación de los resultados en forma coherente y válida.

Continuando con la idea, un quinto punto a considerar con detenimiento es el impacto de las decisiones psicométricas sobre los resultados de la Escala. Por ejemplo, cuestiones tales como el tipo de diseño utilizado, el número de ítems y la calidad de los mismos, el modelo de análisis psicométrico para el análisis y la puntuación o hasta el software utilizado pueden repercutir, con mayor o menor fuerza, sobre la validez de los resultados o la magnitud del avance educativo mostrado por los estudiantes evaluados por la *Escala Vertical* (Tong y Kolen, 2007).

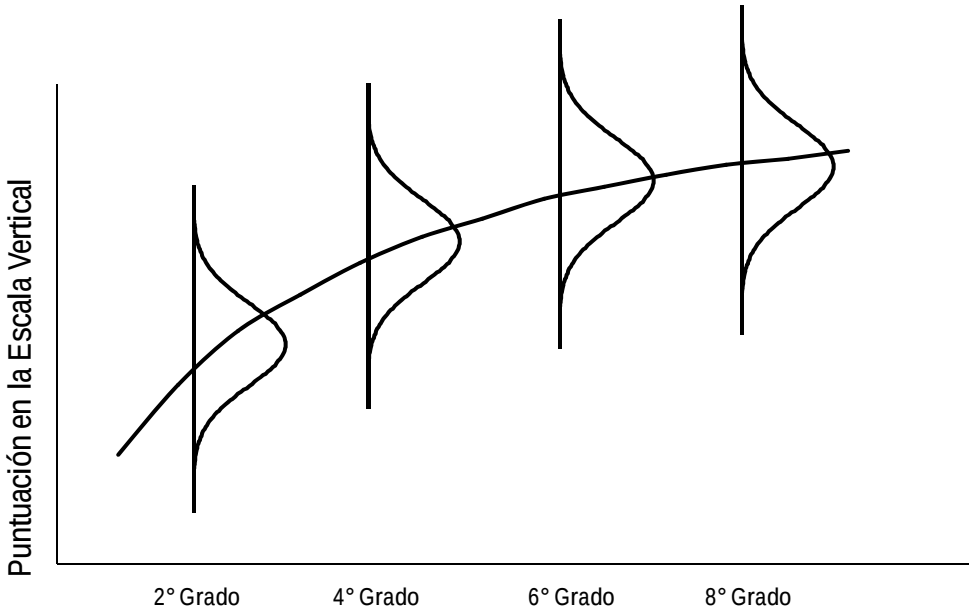
Finalmente, un último punto relevante es la articulación de la *Escala Vertical* con las respectivas *Escalas Horizontales* y con los estándares de aprendizaje, si éstos existen. La alternativa más recomendada es mantener ambas escalas separadas, esto es, reportar los resultados de las pruebas para cada grado en su

escala habitual (por ejemplo, con media 250 y desviación estándar 50) y definir una métrica independiente para la Escala Vertical (por ejemplo, una escala en el rango entre 0 a 100 puntos). La otra alternativa es situar todas las pruebas en una misma escala (por ejemplo, entre 200 y 800 puntos, donde la centena equivalga al grado evaluado y las decenas y unidades a la escala de rendimiento por grado, entre 0 y 100 puntos).

Cualquiera sea la opción que se escoja, debe cuidarse que la elección de la escala evite los resultados paradójicos, como que un desempeño muy deficiente en un grado superior se traduzca en la misma puntuación que un buen rendimiento en grados inferiores.

Este punto, que quizá es el más crítico para la finalidad última de una *Escala Vertical*, puede ilustrarse en la Figura 4. Allí se grafican los resultados hipotéticos obtenidos por estudiantes de 2°, 4°, 6° y 8° grado en una *Escala Vertical*. Asumiendo una distribución normal de las puntuaciones en la *Escala Horizontal* de cada grado, es obvio que los alumnos de más bajo rendimiento en 8° tendrán, en la *Escala Vertical*, en una puntuación similar a los alumnos de rendimiento medio en 2° grado. Por el contrario, los estudiantes de alto rendimiento en 2° grado podrían lograr puntuaciones similares a los examinados de rendimiento promedio en 8° grado. Un resultado de este tipo, que podría prestarse a muchos malentendidos, puede evitarse, o al menos minimizarse, si se pone el suficiente cuidado en la elección de la métrica para el reporte de los resultados.

FIGURA 4. EJEMPLO DE LA DISTRIBUCIÓN DE RENDIMIENTO POR GRADO EN UNA ESCALA VERTICAL



Hemos dedica algunas líneas adicionales al *Escalamiento Vertical* porque se trata de un tópico candente en la investigación y práctica psicométrica contemporáneas, que además creemos interesante de abordar en Latinoamérica. Para los lectores interesados en profundizar, una buena idea es partir documentándose sobre algunas de las *Escalas Verticales* desarrolladas en Estados Unidos, como por ejemplo, el *Stanford Achievement Test Series* (Harcourt Educational Measurement, 2003), los *Metropolitan Achievement Tests* (Harcourt Educational Measurement, 2002), el *TerraNova* (CTB/McGraw-Hill, 1996), o el *Florida*

Comprehensive Assessment Test (Hoffman, Wise, Thacker & Ford, 2003). Una vez comprendidos los alcances del problema, la mejor introducción técnica al *Escalamiento Vertical* es la sección 9.8 del texto de Kolen y Brennan (2004), que, por lo demás, es una referencia obligada sobre métodos de *enlace* de puntuaciones. Adicionalmente, en Young (2006) puede encontrarse una buena presentación de los diseños de recogida de información, y en Patz y Yao (2006) revisarse algunas de las técnicas de análisis disponibles para llegar a puerto. Para contar con una mirada crítica, el trabajo de Harris (2006) ofrece una buena revisión de algunos de los problemas prácticos que supone el construir una *Escala Vertical* (incluyendo la pregunta de si merece la pena el intento), mientras Briggs y Weeks (2009) demuestran empíricamente el impacto que pueden causar pequeñas decisiones sobre los resultados arrojados por la *Escala Vertical*.

4. EQUIPARACIÓN DE PUNTUACIONES

La *equiparación* de puntuaciones es el más importante y estadísticamente robusto de todos los tipos de *enlace*. Su propósito básico es poner en una misma escala los resultados de una o más pruebas que evalúan el mismo constructo, con la misma fiabilidad y bajo especificaciones técnicas similares, con el propósito de que éstos pueden ser utilizados en forma intercambiable.

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. A continuación describiremos brevemente cada uno de ellos.

4.1. Problemas típicos cuya solución requiere aplicar técnicas de equiparación

La finalidad más habitual de las técnicas de equiparación es poner en una misma métrica distintos *cuadernillos* de prueba. Este problema surge porque en muchos sistemas de medición se acostumbra subdividir las pruebas en varios cuadernillos diferentes (también llamados *formas* o *modelos* de pruebas) de modo que no todos los evaluados respondan exactamente los mismos ítems. Esta práctica puede cumplir varios propósitos. En las pruebas con resultados individuales de alto impacto para los respondientes (e.g. pruebas de selección universitaria, evaluación académica, certificación laboral o de acreditación de competencias) se construyen cuadernillos diferentes en un intento de controlar o al menos minimizar los efectos de la copia entre los estudiantes.

Por otro lado, en pruebas estandarizadas destinadas al monitoreo de resultados escolares a nivel agregado (e.g. NAEP) o en estudios internacionales de rendimiento escolar comparado (e.g. PISA, TIMSS, SERCE) un único cuadernillo de prueba es insuficiente para muestrear todos los aprendizajes esperados de los alumnos, lo que hace necesario distribuir el total de preguntas en varios cuadernillos diferentes, para optimizar los tiempos y costos de aplicación. De esta manera, cada alumno responde una parte del total de la prueba y la agregación de todos los alumnos permite reconstruir la prueba total. El resultado, en uno u otro caso, es que los alumnos respondan cuadernillos de prueba distintos, lo que obliga a desarrollar metodologías para asegurar que el resultado de la prueba no sea afectado por cuál haya contestado el estudiante.

Otro problema prototípico que requiere técnicas de *equiparación* se presenta en los sistemas nacionales de medición educativa que tienen entre sus propósitos el monitoreo del sistema escolar a lo largo de

varios años (e.g. NAEP de Estados Unidos). Diversas razones (cambios curriculares, seguridad de las pruebas, publicación de ítems, entre otros) aconsejan actualizar periódicamente el *Banco de Ítems* y obligan al sistema a desarrollar pruebas cuyo contenido (ítems) cambia con el paso de los años. En la práctica, esto exige desarrollar metodologías especiales para garantizar que los resultados sean perfectamente comparables entre un año y el que sigue, para así evaluar precisión cuánto ha mejorado el rendimiento (del país, región, estado, escuela) desde la última medición.

En otro contexto, la *equiparación* entre años es necesaria en caso que un mismo resultado requiera utilizarse en más de una ocasión. Un ejemplo típico de este problema afecta a las pruebas de selección universitaria, cuyos resultados habitualmente son válidos sólo para el año y el período de postulación en que fueron rendidas. Para extender la vigencia de sus resultados a varios períodos de postulación, es necesario garantizar, previamente, que los resultados de cada año son comparables (misma validez, misma fiabilidad y misma métrica) con los de otros años.

Otra variación de los problemas de *equiparación* entre ocasiones de medición se presenta en el contexto de estudios de impacto, experimentales o cuasi experimentales, con diseño pre-test y post-test, en que los que se mide el estado inicial de una variable antes de una intervención escolar y luego se mide el estado final de la misma variable, post intervención. A continuación, la magnitud del cambio relativo antes – después, contrastado con uno o más grupos de control, se utiliza para evaluar la efectividad de la intervención. Un problema bien documentado de estos estudios es que el uso de la misma medida en el pre test y el post test test, limita la validez de la investigación (Shadish, Cook & Campbell, 2002) debido al efecto de la *instrumentación* (sesgo de los resultados por memorización de preguntas, por ejemplo). Para evitar este problema se requiere contar con al menos dos versiones de la prueba que tengan un contenido diferente pero que al mismo tiempo sean completamente paralelas, es decir, midan con igual validez y fiabilidad. Para ello, la aplicación de procedimientos de equiparación es fundamental.

Finalmente, otra situación típica que demanda el uso de técnicas de equiparación es el desarrollo de *Bancos de Items*. La mayoría de los sistemas de evaluación estandarizada operan a través de un banco de preguntas que se actualiza periódicamente y a partir del cual se obtienen los ítems que integran las pruebas. Normalmente, estos *Bancos de Items* contienen ítems analizados con Teoría de Respuesta al Item y es necesario, antes de ingresarlos al *Banco*, que sus parámetros sean puestos en la misma métrica de los restantes ítems.

Es importante advertir que una misma prueba estandarizada puede enfrentar simultáneamente todas esas demandas y que la mayoría de los sistemas de medición requieren resolver problemas de equiparación bastante complejos, lo que exige aplicar varias técnicas de equiparación simultáneamente. En nuestra experiencia, no es exagerado afirmar que ningún sistema de evaluación estandarizado puede cumplir correctamente con su misión si no destina suficientes recursos a resolver los problemas de equiparación entre sus pruebas.

Nótese que los métodos de *equiparación* pueden clasificarse en varias subcategorías (e.g. Kolen & Brennan, 2004; Holland & Dorans, 2006), pero para comprenderlas es necesario manejar primero cuestiones técnicas referidas al diseño de recogida de datos y a los métodos estadísticos para *equiparar* las pruebas. Como ello escapa al alcance de este trabajo, obviaremos su presentación, aunque el lector interesado puede consultarlas en las fuentes antes señaladas.

4.2. Características definitorias de los enlaces por equiparación de puntuaciones

En el apartado anterior mencionamos algunas situaciones prototípicas que requieren aplicar técnicas de *equiparación.* Sin embargo aún si en esos casos utilizamos un diseño correcto y aplicamos las técnicas estadísticas apropiadas, no lograremos obtener una *equiparación* a no ser que se cumplan ciertas condiciones mínimas.

De hecho, en el desarrollo de la tipología de métodos de *enlace* que hemos venido presentando, la *equiparación* aparece como una forma especial de *alineamiento*, pero con supuestos mucho más estrictos. Estos supuestos se traducen en ocho requisitos fundamentales. Algunos son cualitativos, así que su evaluación es materia del juicio experto, lo que a veces complica decidir su cumplimiento. Otros son cuantitativos o pueden evaluarse parcialmente a través de procedimientos estadísticos. En uno u otro caso, es necesario insistir en que aún si aplicamos metodologías estadísticas altamente sofisticadas para equiparar dos pruebas, el resultados no será una equiparación válida si alguno de estos ocho supuestos no se cumple apropiadamente.

Los cuatro criterios básicos fueron sugeridos por Lord (1980):

4.2.1. El mismo constructo

Ambas pruebas deben medir el mismo constructo, definido de la misma manera. De lo contrario, el enlace no puede considerarse una *equiparación* legítima. La evaluación de este criterio es principalmente cualitativa, pero puede asistirse con técnicas como Modelos de Ecuaciones Estructurales para, por ejemplo, modelar la invarianza de una solución factorial. Dependiendo del diseño de equiparación, se podría evaluar invarianza total o parcial.

4.2.2. Equidad

Una vez que dos pruebas han sido equiparadas sus puntuaciones pueden utilizarse en forma intercambiable entre sí, de manera que sea indiferente para el resultado cuál de las dos responda el evaluado. Como puede apreciarse, este es requerimiento es teórico y, hasta cierto punto, una consecuencia de los demás. También es un prerrequisito para la quinto condición.

4.2.3. Simetría de la transformación

La equiparación debe llevarse a cabo con una técnica que garantice la reversibilidad de los resultados. Esto es, que la función que permite transformar la puntuación de A en B, sea exactamente la inversa de la que permita transformar B en A. Esta condición es fácil de evaluar empíricamente. Por ejemplo, si de acuerdo a la función de equiparación 10 puntos de la prueba A equivalen a 8 puntos de la prueba B, entonces 8 puntos de la prueba B deben equivaler a 10 puntos de la prueba A, en caso que la función se aplique en sentido inverso. Volveremos a insistir en que el requerimiento de *simetría* es esencial en la definición de equiparación. Además, que impide que los enlaces por *predicción* puedan ser considerados una forma de equiparación y descarta el uso de técnicas de regresión tradicional del arsenal de procedimientos estadísticos para equiparar puntuaciones.

4.2.4. Invarianza de la transformación en subpoblaciones de interés

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). Otra forma de plantearlo, es que la *invarianza* exige que la función de transformación para equiparar ambas pruebas sea válida para cualquier examinado con independencia

de variables no relacionadas con el constructo que evalúa la prueba. En la práctica, este criterio también puede evaluarse empíricamente, simplemente subdividiendo la población en varios grupos, estimando las funciones de equiparación correspondientes, y finalmente verificando si éstas no difieren significativamente entre sí.

Además de estos cuatro criterios Holland y Dorans (2006) consideran un quinto que, desde su punto de vista, debería ocupar el segundo lugar en orden de importancia:

4.2.5. Igual fiabilidad

Las puntuaciones de ambas pruebas deben tener la misma fiabilidad. Aunque se trata de una condición implícita en el criterio 2 (*Equidad*) y por lo tanto resulta algo redundante, su importancia es que puede evaluarse fácilmente, simplemente comparando los respectivos coeficientes de fiabilidad. Holland y Dorans (2006) han demostrado que no basta con que ambas pruebas tengan igual fiabilidad, sino que además ésta debe elevada.

4.2.6. Las mismas inferencias

Las dos pruebas deben tener el mismo propósito y estar diseñadas para obtener el mismo tipo de conclusiones. Aunque este criterio está implícito en el primero (*mismo constructo*) es útil explicitar que el propósito debe ser el mismo si dos pruebas van a ser equiparadas (e.g. selección universitaria, monitoreo escolar).

4.2.7. La misma población

Ambas pruebas deben estar diseñadas para la misma población objetivo.

4.2.8. Las mismas propiedades psicométricas

Las dos pruebas deben estar construidas con la misma tabla de especificaciones, ser administradas en idénticas condiciones y ser equivalentes en todas sus propiedades métricas. Nuevamente, este criterio es relativamente redundante con los requisitos primero y quinto señalados anteriormente.

Si alguna de estas ocho condiciones no se cumple, el enlace entre las puntuaciones no podrá ser considerado una *equiparación* sino un *alineamiento*. Obviamente, en algunos casos se trata de condiciones dicotómicas, que además son fácilmente comprobables empíricamente (e.g. *misma fiabilidad* o *invarianza*). El problema es que en otros casos (e.g. *mismas propiedades psicométricas*) se trata más bien de un juicio cualitativo respecto a cuánto nos podemos acercar al cumplimiento del requisito.

Típicamente, la medición de temperatura ambiental, con termómetros graduados en escalas Celsius versus Fahrenheit, se utiliza como ejemplo idealizado de *equiparación.* Se trata del mismo constructo, la temperatura (condición 1), que es medido con dos instrumentos que sirven al mismo propósito (condición 6), que se aplican en problemas equivalentes (condición 7), y que son idénticos en todas sus características técnicas (condición 8), incluyendo su precisión (condición 5), y su validez en diferentes situaciones (condición 4). Se trata, en el fondo, de un mismo instrumento (termómetro) con versiones distintas, que difieren entre sí únicamente en la escala de medida. Por ello, ambas escalas de temperatura pueden *equipararse* mediante las transformaciones *F=(9/5)C+32* o, en forma equivalente, *C=(5/9)(F-32)*. Como puede verse, la transformación es simétrica (condición 3) y, una vez conocida, permite que ambos instrumentos se utilicen en forma intercambiable, sin que ello afecte los resultados (condición 2).

A modo de síntesis, recordemos que todos los autores especializados en la materia insisten en que la

equiparación simplemente ajusta las diferencias en dificultad pero no en contenido de dos pruebas y, por lo tanto, equiparar no es sinónimo de igualar (Kolen y Brennan, 2004). Esta aseveración insiste, una vez más, en que los procedimientos estadísticos para equiparar las puntuaciones de las pruebas no garantizan por sí mismos que sea válido utilizar sus resultados en forma intercambiable, a no ser que primero se cumplan las condiciones antes señaladas.

5. ¿EQUIPARACIÓN, ALINEAMIENTO O PREDICCIÓN?

Aunque conceptualmente la distinción entre las tres metodologías de *enlace* es bastante clara, no siempre es fácil elegir cuál de ellas aplicar frente a un problema empírico. A continuación sugerimos cuatro criterios que pueden ayudar a tomar una decisión.

En primer lugar, la diferencia fundamental entre los tres tipos de *enlace* se desprende de la finalidad de cada uno de ellos. Esencialmente, el propósito de la *Equiparación* es *obtener puntuaciones intercambiables* entre medidas equivalentes de un mismo constructo. En cambio, el objetivo de los métodos de *Alineamiento* es transformar los resultados de medidas distintas a una *escala común*, para *obtener puntuaciones comparables*. Mucho menos ambiciosos, la meta que persiguen los estudios de *Predicción* es simplemente *predecir el resultado más probable en una medición*, para un individuo o grupo, a partir de otra información relevante (e.g. su puntuación en otra prueba). Por lo tanto, el punto de partida es preguntarse ¿Qué pretendemos lograr con nuestro *enlace?*

Una vez definido el objetivo y optado tentativamente por una metodología, la segunda consideración critica es si se cumplen o no las condiciones técnicas que el método exige. Por ejemplo, en la sección 4.2 definimos la *equiparación* en función de ocho requerimientos básicos y comentamos que si alguno de ellos no se cumple, el *enlace* de las puntuaciones sólo alcanzaba el estatus de *alineamiento*. En el Cuadro 1 hemos organizado estos criterios en perspectiva comparada con las otras metodologías de *enlace*, en un intento de sintetizar las diferencias entre ellas.

CUADRO 1. DIFERENCIAS ENTRE EQUIPARACIÓN, ALINEAMIENTO Y PREDICCIÓN DE PUNTUACIONES

REQUERIMIENTOS TECNICOS	METODOLOGÍA DE ENLACE DE PUNTUACIONES		
	Equiparación	Alineamiento	Predicción
Constructos	Iguales	Similares, distintos o iguales	Similar o distintos
Equidad	Necesaria	No existe	No existe
Simetría	Necesaria	Necesaria	No existe
Invarianza	Necesaria	Necesaria	Deseable
Fiabilidad	Igual	Similar o igual	Similar (deseable)
Inferencias	Iguales	Similares o distintas	Distintas o similares
Población objetivo	Igual	Similar, igual o distinta	Similar o distinta
Propiedades Psicométricas	Iguales	Similares	Similares o distintas

Podemos ver que la *predicción* difiere de las otras dos metodologías en que no hay *simetría* en la función de enlace y, por lo tanto, tampoco es posible alcanzar el requerimiento de *equidad* entre las pruebas. Sin embargo, las otras condiciones pueden ser similares a las que exige un *alineamiento*. Este, por su parte, se diferencia fundamentalmente de la *equiparación* en que los *constructos* evaluados por ambas pruebas no requieren ser los mismos, lo cual supone que las características técnicas de las pruebas (*fiabilidad, inferencias, propiedades psicométricas*) tampoco sean iguales. De allí se desprende que el requerimiento de *equidad* es imposible de lograr.

Aplicando estos ocho criterios a la metodología o método específico que se ha escogido, es posible contar

con un segundo criterio para decidir si se trata de una opción recomendable.

En tercer lugar, es una buena idea evaluar empíricamente la magnitud de la asociación entre las puntuaciones de las pruebas que se desean *enlazar*. En este sentido, Dorans (2004) argumenta que se requieren correlaciones de al menos *r=.866* para que la *concordancia* o la *equiparación* entre dos pruebas tenga un nivel de incertidumbre aceptable para fines prácticos. Con correlaciones de menor magnitud es cuestionable la exactitud de *enlace* obtenido, aunque ello dependerá de la fiabilidad de las pruebas y de la amplitud de la escala de puntuaciones. Sin embargo, es difícil defender algún tipo de *alineamiento* con correlaciones inferiores a *r=.70* (es decir, un 49% de varianza común), así que en esos casos sería razonable optar por algún método de *predicción.* Una discusión interesante sobre el uso de coeficientes de correlación para orientar la decisión entre métodos de *enlace,* puede consultarse en un par de trabajos de Dorans (2000, 2004).

Por último, diferentes trabajos (Dorans, 2000; Dorans & Holland, 2000; Yin, Brennan & Kolen, 2004) han enfatizado la importancia de la *invarianza* del *enlace* como condición indispensable, tanto para la *equiparación* como para cualquier tipo de *alineamiento de puntuaciones*. En esos artículos pueden encontrarse algunos procedimientos estadísticos, de distinto nivel de complejidad, para evaluar el cumplimiento de la *invarianza*, así como algunas recomendaciones para su interpretación.

6. Comentarios finales

El propósito de este trabajo ha sido presentar una síntesis de los distintos métodos disponibles para el *enlace* de puntuaciones, a partir de la taxonomía propuesta por Holland y Dorans (2006). Para terminar, creemos importante señalar los temas que no hemos abordado pero que resulta necesario conocer si se quiere profundizar en estas materias.

En concreto, omitimos el tratamiento de los aspectos metodológicos propiamente tales para cada tipo en *enlace*. Brevemente, cualquier metodología de *enlace de puntuaciones* tiene dos componentes básicos: un diseño de recogida de la información (de los examinados y de las pruebas, es decir, las características de las muestras utilizadas y las especificaciones técnicas y propiedades psicométricas de las pruebas aplicadas) y un conjunto de técnicas estadísticas para el análisis de los datos. Estas incluyen, a su vez, tres grupos de procedimientos que son aplicados en forma secuencial: técnicas para evaluar el cumplimiento de los supuestos de la metodología de *enlace*, técnicas para construir una función de transformación entre las pruebas (es decir, el *enlace* propiamente tal) y técnicas para evaluar el ajuste, calidad y precisión del *enlace* resultante.

En general, los diseños se dividen en dos grupos; aquellos que asumen que los examinados provienen de una *misma población* y aquellos que utilizan alguna *medida de anclaje* como en los ejemplos que ya se comentaron en la sección 3.2. A partir de esa distinción, la variedad de diseños utilizados en aplicaciones prácticas es inmensa. Kolen y Brennan (2004) ofrecen una descripción de los diseños más comunes para distintos métodos y Young (2006) sintetiza muy bien los diseños disponibles para el *Escalamiento Vertical*.

Por otro lado, las técnicas estadísticas están estrechamente ligadas al diseño y condicionadas por el modelo psicométrico con que se trabaje (Teoría Clásica de los Test o Teoría de Respuesta al Item). Generalizando, las técnicas clásicas construyen los *enlaces* a partir de las puntuaciones en el total de la prueba o en un *anclaje*, utilizando la distribución completa de puntuaciones (e.g. técnicas equipercentiles)

o alguna medida sintética (e.g. *equiparación* lineal por la media). Las técnicas basadas en Teoría de Respuesta al Item construyen los enlaces a partir de todos o algunos de los ítems, mediante técnicas que transforman los parámetros de los ítems o que los intervienen directamente en la calibración. En cualquier caso, la presentación más completa y didáctica de las técnicas estadísticas también puede consultarse en el texto de Kolen y Brennan (2004), aunque Holland et al. (2007) ofrecen una síntesis recomendable para los lectores con mayor formación estadística.

En castellano, los textos introductorios a la psicometría de Martínez, Hernández y Hernández (2006) y de Muñiz (2001) contienen una buena presentación de los diseños y técnicas clásicas de *equiparación*, que son generalizables a los demás problemas de *enlace*. Finalmente, el capítulo de Navas (1996) ofrece un tratamiento completo de las metodologías de *equiparación*, incluyendo los diseños y las aproximaciones estadísticas clásicas y basadas en Teoría de Respuesta al Ítem.

Referencias bibliográficas

American Psychological Association. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Angoff, W.H. (1971). Scales, norms and equivalent scores. En R.L. Thorndike (Ed.) *Educational Measurement (2dn Ed).* Washington, DC: American Council on Education.

Briggs, D.C. & Weeks, J. (2009). The impact of Vertical Scaling decisions on growth interpretations. *Educational Measurement: Issues and practice*, *28*(4), 15-26.

Camilli, G. (2006). Test Fairness. En R.L. Brennan (Ed.) *Educational Measurement (4th Ed)*. Wesport, CT: Praeger Publishers.

CTB/McGraw-Hill. (1996). *TerraNova prepublication technical bulletin.* Monterray, CA: Author.

Dorans, N.J. (1999). *Correspondences between ACT and SAT I scores*. College Board Report 99-1. New York. The College Board.

Dorans, N.J. (2000). *Distinctions among classes of linkages*. College Board Research Note RN-11. New York. The College Board.

Dorans, N.J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, *39*, 59-84.

Dorans, N.J. (2004). Equating, concordance and expectation. *Applied Psychological Measurement*, *28*(4), 227-246.

Dorans, N.J. & Holland, P.W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281-306.

Dorans, N.J., Pommerich, M. & Holland, P.W. (2007). *Linking and aligning scores and scales.* New York: Springer.

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W. & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council. Washington, DC: National Academy Press.

Flanagan, J.C. (1951). Units, scores and norms. En E.F. Linquist (Ed.) *Educational Measurement (1st Ed)*. Washington, DC: American Council on Education.

Galton, F. (1888). Co-relations and their measurements, chiefly from anthropological data. *Proceedings of the Royal Society of London, 45,* 135-145.

Haertel, E. (1991). *Report of TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress*. ERIC Clearinghouse Document Reproduction Service N° ED404367. Washington, DC: National Center for Education Statistics.

Harcourt Educational Measurement. (2002). *Metropolitan Achievement Tests: Technical Manual (8th Ed)*. San Antonio, TX: Author.

Harcourt Educational Measurement. (2003). *Stanford Achievement Tests Series: Spring technical data report (10th Ed)*. San Antonio, TX: Author.

Harris, D.J. (2007). Practical issues in Vertical Scaling. En N.J. Dorans, M. Pommerich & P.W. Holland (Eds.) *Linking and aligning scores and scales.* New York: Springer.

Hoffman, R.G., Wise, L.L., Thacker, A.A. & Ford, L.A. (2003). *Florida Comprehensive Assessment Tests: Technical Report on vertical scaling for reading and mathematics.* A HumRRO Report under subcontract to Harcourt Assessment, San Antonio, TX.

Holland, P.W. (2007). A framework and history for score linking. En N.J. Dorans, M. Pommerich & P.W. Holland (Eds.) *Linking and aligning scores and scales.* New York: Springer.

Holland, P. W. & Dorans, N.J. (2006). Linking and equating. En R.L. Brennan (Ed.) *Educational Measurement (4th Ed)*. Wesport, CT: Praeger Publishers.

Holland, P. W., Dorans, N.J. & Petersen, N.S. (2007). Equating test scores. En C.R. Rao & S. Sinharay (Eds.) *Handbook of Statstics, Vol 26.* Amsterdam: Elsevier.

Huynh, H. & Schneider, C. (2004). *Vertically moderated standards as an alternative to vertical scaling: Assumptions, practices and a odyssey through NAEP*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Keeves, J. (1988). Scaling achievement test scores. En T. Husen & T.N. Postlethwaite (Eds.) *International Encyclopedia of Education.* Oxford: Pergamon.

Kelley, T.L. (1914). Comparable measures. *Journal of Educational Psychology*, *5*, 589-595.

Kelley, T.L. (1923). *Statistical Methods*. New York: MacMillan.

Kolen, M.J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*(2), 97-110.

Kolen, M.J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*, 219-226.

Kolen M.J. & Brennan, R.L. (2004). *Test equating, scaling and linking. Methods and practice (2th Ed)*. New York: Springer.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83-102.

Lissitz, R.W. & Huynh, H. (2003). Vertical equating for state assessment: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research and Evaluation, 8*(10).

Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems.* Hilsdale, NJ: Erlbaum.

Martínez, M.R., Hernández, M.J. & Hernández, M.V. (2006). *Psicometría*. Madrid: Alianza Editorial.

McGaw, B. (1977). The use of rescaled teacher assessments in the admission of student to tertiary study. *Australian Journal of Education*, *21*, 209-225.

Mislevy, R.J. (1992). *Linking Educational Assessments: Concepts, issues, methods and prospects* (Policy Information Report). Princeton, NJ: Educational Testing Service.

Muñiz, J. (2001). *Teoría Clásica de los Test*. Madrid: Pirámide.

Navas, M.J. (1996). Equiparación de Puntuaciones. En J. Muñiz (Ed.). *Psicometría*. Madrid: Universitas

Pashley, P.J. & Phillips, G.W. (1993). *Toward world-class standards: A research study linking international and national assessments.* Princeton, NJ: Educational Testing Service.

Patz, R.J. & Yao, L. (2007). Methods and models for Vertical Scaling. En N.J. Dorans, M. Pommerich & P.W. Holland (Eds.) *Linking and aligning scores and scales.* New York: Springer.

Pommerich, M. (2007). Concordance: The Good, the Bad and the Ugly. En N.J. Dorans, M. Pommerich & P.W. Holland (Eds.) *Linking and aligning scores and scales.* New York: Springer.

Pommerich, M. & Dorans, N.J. (Eds.), (2004). Concordance [Special Issue]. *Applied Psychological Measurement, 28*(4).

Sackett, P.R. & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112-118.

Shadish, W., Cook, T. & Campbell, D. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton Mifflin.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. En S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Testing*. Mahwah, NJ: Lawrence Erlbaum.

Tong, Y. & Kolen, M.J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*(2), 227-253.

Yin, P., Brennan, R.L. & Kolen, M.J. (2004). Concordance betweed ACT and ITED scores from different populations. *Applied Psychological Measurement*, *28*(4), 274-289.

Young, M.J. (2006). Vertical Scales. En S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Testing*. Mahwah, NJ: Lawrence Erlbaum.