La educación es de todos

Mineducación

DISEÑO DE ARMADO

para pruebas estandarizadas: usos y metodologías

Subdirección de Diseño de Instrumentos

icfes mejor saber

Presidente de la República
Iván Duque Márquez

Ministra de Educación Nacional
María Victoria Angulo González

Publicación del Instituto Colombiano para la Evaluación de la Educación (Icfes)

Elaborado por
Mónica Liliana Manrique Galindo

Edición
Juan Camilo Gómez-Barrera

Diseño de portada y diagramación
Linda Nathaly Sarmiento Olaya

Directora General
Mónica Patricia Ospina Londoño

Secretario General
Ciro González Ramírez

Directora de Evaluación
Natalia González Gómez

Director de Tecnología
Carlos Alberto Sánchez Rave

Subdirector de Diseño de Instrumentos
Luis Javier Toro Baquero

Subdirectora de Estadísticas
Jeimy Paola Aristizábal Rodríguez

Subdirectora de Análisis y Divulgación
Mara Brigitte Bravo Osorio

ISBN de la versión digital
978-958-11-0891-6

Bogotá, D. C., diciembre de 2020

ADVERTENCIA

TABLA DE CONTENIDO

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

Introducción

El armado es la etapa de la elaboración de una prueba en la que se obtienen las formas finales que se utilizarán en la aplicación. Esta etapa se lleva a cabo cuando se disponen tanto de especificaciones definidas como de resultados del análisis de los posibles ítems[1] que se pueden utilizar. Así mismo, el armado involucra la selección del contenido y la organización de la información que se incluirá en cada uno de los cuadernillos de una prueba o examen. Para guiar las diferentes actividades de esta etapa, se emplean los diseños experimentales, que son estrategias para la distribución de la información en formas finales y permiten controlar variables que pueden tener importancia en el resultado obtenido, se refieran o no al atributo o la característica que se pretende medir.

El objetivo del presente texto, como documento técnico de la Subdirección de Diseño de Instrumentos del Instituto Colombiano para la Evaluación de la Educación (Icfes), es mostrar diferentes diseños experimentales que pueden emplearse en el armado de una prueba. En esa medida, se indican sus principales características y el uso que se les ha dado en diferentes países y en evaluaciones internacionales, con la intención de que esta información sea de utilidad para las personas que realicen actividades de armado de pruebas.

1 Para los propósitos de este documento, se tomarán como equivalentes los términos ítem y pregunta.

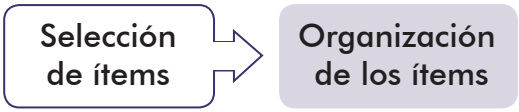
En consecuencia, se enfatizará en los aspectos prácticos de los diseños, particularmente en su utilidad para el control de las variables que no constituyen el objeto de medición, pero que pueden tener influencia sobre su resultado.

El primer capítulo de este documento define de forma general lo que se entiende por armado de prueba, sus requerimientos y necesidades, y los diferentes aspectos que se deben tener en cuenta para su realización. El segundo capítulo presenta información general de los diseños, como los principios a los que deben responder y sus propiedades más relevantes; así mismo, se mencionan algunos diseños que han sido o pueden ser usados en la evaluación educativa con pruebas estandarizadas. Finalmente, en el tercer capítulo, se contextualiza este uso a nivel de país y en aplicaciones internacionales.

1. Armado de pruebas: particularidades y retos

De acuerdo con Brown (1983), el armado de una prueba[2] es la etapa de elaboración en la que, a partir de los resultados obtenidos en los análisis de ítems, se preparan la o las formas finales que van a ser utilizadas en una aplicación. Un armado consta de dos pasos: selección de ítems y organización de los ítems.

Ilustración 1. *Armado de una prueba*



El primer paso consiste en la selección de los ítems que harán parte de las formas. Para esta selección deben tenerse en cuenta diferentes criterios. Los que resultan más básicos son la correspondencia con la estructura de la prueba, de acuerdo con el diseño que se haya empleado para la construcción, y la dificultad, teniendo en cuenta las características de la población a la que va dirigida la prueba (Brown, 1983). En relación con la **estructura**, se debe verificar que cada ítem pueda ser ubicado dentro de esta y que cada elemento relacionado en la estructura (especificaciones de los ítems) se presente en la proporción indicada a lo largo del conjunto de ítems

2 La expresión empleada en la versión en inglés del libro es *Assembling the Test*. En este documento se empleará la palabra **armado** para referir a las pruebas y **ensamblaje** para aludir a los cuadernillos, ya que un cuadernillo puede incluir una o más pruebas.

seleccionado. Esto es muy importante, ya que las proporciones establecidas en la estructura de una prueba se basan en un criterio teórico o práctico, con base en el que también se establecen las interpretaciones que se pueden extraer de los resultados (Icfes, 2018a). En cuanto a la **dificultad**, al margen del modelo de calificación, es posible considerar tanto un rango, dentro del cual deben encontrarse todos los ítems de una prueba, como un estadístico de resumen para cada conjunto de ítems, como el promedio o la desviación. Controlar estos dos factores puede ser beneficioso, especialmente cuando la dificultad se considera como el nivel del atributo o la habilidad que se requiere para contestar, con una determinada probabilidad, una opción particular de un ítem (usualmente la clave)[3] y cuando se busca realizar comparaciones de la medición entre los evaluados o a través del tiempo.

Estos criterios, estructura y dificultad, sin embargo, pueden ser insuficientes para el control que se requiere tener en una determinada prueba, dependiendo de su propósito y de las decisiones que se tomarán con base en sus resultados. Otros

3 La dificultad, así como la discriminación, el pseudo-azar y la información se definen aquí desde la teoría de respuesta al ítem. Usualmente, se obtiene una estimación de la habilidad que posee un individuo, así como de la dificultad de los ítems con respecto a la respuesta correcta; sin embargo, en ítems que no tienen una respuesta correcta ni se orientan específicamente al desempeño o la habilidad (por ejemplo, en escalas de actitudes), se habla más del atributo, y resulta importante obtener los niveles de este asociados con la selección de cada opción de respuesta. Desde la teoría clásica de las pruebas, la dificultad es más bien un índice de facilidad, ya que corresponde a la proporción de personas que respondió correctamente el ítem, con respecto a la cantidad que lo abordó.

aspectos que pueden considerarse son: el flujo de opciones, la discriminación, el pseudo-azar y la función de información que presenta cada uno de los ítems. El flujo de opciones se refiere a la proporción de evaluados que selecciona cada opción disponible, incluyendo los casos en que no se responde o cuando se marca más de una opción; la discriminación se refiere a la medida en que un ítem permite diferenciar entre niveles de habilidad de los evaluados, alrededor de su valor de dificultad[4]; el pseudo-azar se refiere a la probabilidad de responder correctamente el ítem cuando hay un nivel muy bajo del atributo, y la información se relaciona con la precisión en la estimación que tiene un ítem en un determinado valor de habilidad[5]. Aunque estos aspectos son propios de cada ítem, es posible obtener una medida de resumen para toda la prueba[6] o pueden aplicarse condiciones individuales, como que la opción de respuesta correcta tenga siempre un porcentaje de escogencia mayor al de las demás, o un valor mínimo de información.

4 Desde la teoría clásica de las pruebas, se mantiene la definición general, pero no la referencia al valor de dificultad del ítem.

5 La función de información se obtiene para cada ítem; sin embargo, se puede calcular para una agrupación de ítems sumando la información de cada uno de sus elementos para cada uno de los diferentes valores de habilidad.

6 Esto se efectúa de manera similar a como se planteó para la estructura y la dificultad. Todos estos criterios a nivel de prueba o de agrupaciones de ítems se aplican desde el paso siguiente; sin embargo, su planteamiento se realiza de manera previa a la etapa de armado y coinciden con los aplicados en la selección de ítems, por lo que se presentan en este primer paso y se destina el segundo exclusivamente a los diseños experimentales.

Dependiendo de los recursos con los que se cuente, como el tiempo, la cantidad y las características de los ítems disponibles, en ocasiones no será posible aplicar todos estos criterios. Esto se deba a que, a medida que se consideren más criterios, se tendrán menos ítems que los cumplan. En ocasiones se tienen consideraciones adicionales como los contextos que acompañan a los ítems (los textos que deben leer los estudiantes en las pruebas de lectura, por ejemplo). Sin embargo, tres criterios mínimos son la correspondencia con la estructura, la dificultad y la discriminación, toda vez que la estructura es el esqueleto sobre el que se construye la prueba (Herrera-Rojas, 1996), y la dificultad y la discriminación son elementos comunes a diferentes modelos de medición. Criterios como el flujo de opciones y el pseudo-azar requieren análisis adicionales para su interpretación con respecto a cada ítem, particularmente, para ciertos tipos de ítem, y pueden ser orientaciones para modificaciones si no se toman en cuenta para la selección, mientras que la información se considera parcialmente con los criterios mínimos señalados, ya que su valor depende de la dificultad, la discriminación y el pseudo-azar.

Ilustración 2. *Características consideradas para la selección de ítems*

El segundo paso en el armado de pruebas es la organización de los ítems dentro de cada forma. Para ello, se debe tener en cuenta que existen diferentes tipos de agrupaciones de ítems con propósitos de aplicación. A continuación, se indican las características de estos tipos de agrupaciones, ordenados de mayor a menor generalidad:

- Cuadernillo:

  Corresponde al conjunto de ítems a los que se enfrentará un evaluado durante toda una sesión. Estos ítems pueden ser de una o varias áreas, es decir, de una o varias pruebas.

- Forma:

  El conjunto de ítems de una misma área o prueba que enfrentará cada evaluado en un examen, en una o más sesiones. Los cuadernillos son conjuntos de formas[7].

- Bloque:

  Los bloques son agrupaciones de ítems que comparten la longitud (cantidad de ítems) o el tiempo esperado de aplicación y que, al unirse, constituyen una forma (Fernández-Alonso y Muñiz, 2011).

- Contexto:

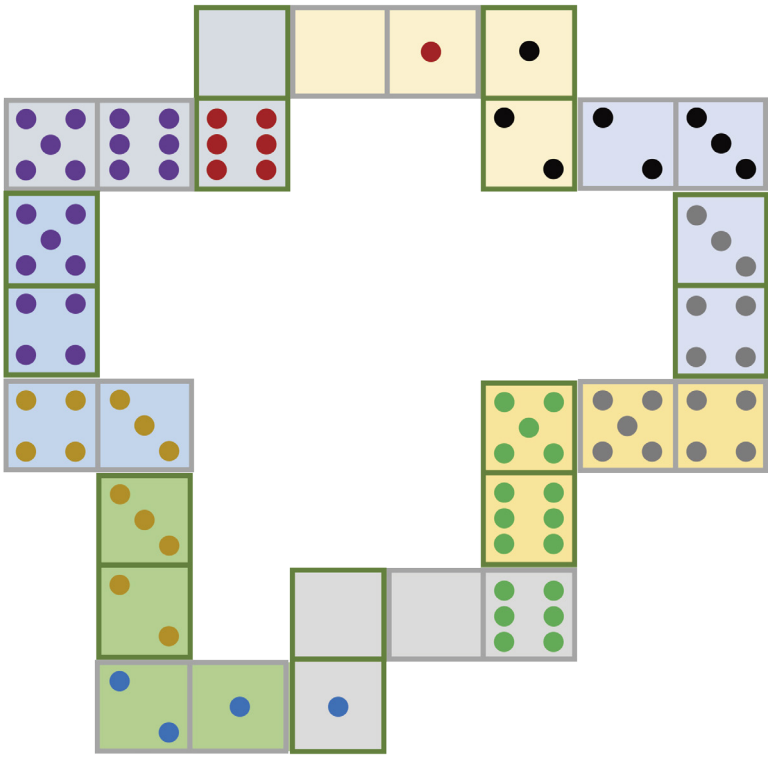
  Aunque los contextos no son en sí conjuntos de ítems, constituyen agrupaciones, ya que la información presentada en cada contexto es necesaria para responder cada uno de los ítems que tiene asociados. Si bien no se tienen que emplear a la vez todos los ítems asociados a un mismo contexto, es recomendable incluir al menos dos o tres ítems por cada contexto que se utilice en una prueba, debido al tiempo y esfuerzo que implica su lectura y análisis.

7 Pueden ser conjuntos parciales si, por ejemplo, se divide la aplicación de una misma forma en dos sesiones.

Otra posible agrupación corresponde a los conjuntos de ítems que vienen de aplicaciones anteriores. Estos conjuntos se denominan **anclas**. Para cualquier aplicación, lo más recomendable es que cada ancla sea un conjunto representativo de la prueba total, es decir, que cumpla con los mismos criterios (de estructura, dificultad y discriminación, o los que se hayan definido), como si se tratara de una pequeña versión de cualquier aplicación de la prueba. Este conjunto puede estar en el interior de las formas que se van a aplicar y ser parte de la calificación (anclaje interno) o puede constituir una forma aparte que no se incluye en la calificación (anclaje externo) (Dorans, Moses y Eignor, 2010). Se puede tomar como ejemplo una prueba en la que cada forma esté compuesta por cuatro grupos de 20 ítems, uno que viene de una aplicación anterior y tres que se conforman para cada nueva aplicación, incluyéndose todos en la calificación. Como todos ellos tienen la misma longitud, se trata de cuatro bloques; sin embargo, el que viene de la aplicación anterior sería un ancla interna. Es necesario aclarar que las anclas no tienen que ser bloques sino que puede tratarse de un conjunto de ítems distribuidos a lo largo de una forma. De igual manera, no es recomendable que las anclas reciban modificaciones, ya que, si son diferentes entre la primera y la segunda aplicación, sus calibraciones[8] pueden cambiar y no se prestarán para la obtención de resultados comparables.

8 Los valores que se estiman para cada una de las características de los ítems, como la dificultad o la discriminación.

Ilustración 3. *Agrupación de ítems*



En la ilustración 3 se presentan dos conjuntos de fichas de dominó. Cada uno corresponde a una prueba diferente y está marcado con un color de borde distinto (verde o gris). Cada ficha es una forma, perteneciente a una única prueba, por lo que cada uno de los dos cuadros que la conforman corresponde a un bloque. Los colores de relleno y los de los puntos indican los cuadernillos, que incluyen una forma de cada prueba. En total, se tienen 28 bloques, 14 por prueba; 14 formas, 7 por prueba, y 14 cuadernillos.

Para el armado de pruebas se pueden emplear las agrupaciones desde el contexto hasta llegar a la forma, o incluso se puede llegar al cuadernillo cuando se compone de una única forma. De igual manera, pueden usarse las anclas cuando se pretenda tener comparabilidad entre distintas aplicaciones de la misma prueba. La necesidad de cumplir en este segundo paso con los criterios señalados en el primero hará que se propongan bloques y formas hasta que se cumplan los criterios definidos para estas agrupaciones y para el total de ítems que se van a utilizar en una prueba en una determinada aplicación.

Cada forma que se proponga será revisada por una serie de expertos en medición y en las áreas que son objeto de evaluación. Una vez se tenga una versión definitiva, se requerirá una nueva revisión de los cuadernillos y de la impresión (Brown, 1983). Esto quiere decir que las formas propuestas al inicio no necesariamente serán las finales, ya que en las revisiones pueden encontrarse aspectos por mejorar. Estos aspectos pueden incluir detalles como errores puntuales en la digitación o la diagramación, o ser más generales, como la repetición de contenido entre dos preguntas o que una responda a otra; en ese caso, se requerirá buscar al menos una pregunta de reemplazo con las mismas características.

1.1 Posibles dificultades

Los anclajes, por su funcionalidad, pueden presentar algunos problemas. Uno de ellos es la exposición de los ítems, ya que se presentan en más de una oportunidad y, en consecuencia, son vistos por un mayor número de personas, lo que puede

dar ventaja a los evaluados de la segunda aplicación. Este riesgo puede reducirse si se aplica el anclaje a una pequeña proporción del total de evaluados. También hay riesgo de copia si los anclajes se presentan en la misma posición en diferentes cuadernillos, ya que dos personas que se sienten juntas podrían tener los ítems compartidos aun si el anclaje no se presenta en todas las formas. Otra posible dificultad se relaciona con la cobertura de la prueba cuando se tienen múltiples formatos de ítem, ya que se quiere representar la totalidad de la prueba y, por fuerza, se requerirá reproducir la proporción en que se presenta cada formato. Adicionalmente, es necesario considerar que, cuando se utilizan anclajes externos, es importante camuflarlos entre las demás agrupaciones de ítems, ya que si los evaluados los identifican como tal y saben que no recibirán calificación, podrían no responderlos, limitando la posibilidad de realizar comparaciones entre los grupos incluidos en las diferentes aplicaciones de la prueba (Dorans, Moses y Eignor, 2010).

Otras posibles dificultades, que no se limitan a los ítems de anclaje, se relacionan con los efectos de contexto, esto es, las diferencias en los resultados de la medición que se dan en función de: a) la posición que ocupan los ítems, b) las condiciones de prueba bajo las cuales se aplican y c) los elementos que se encuentran a su alrededor (Doran, Moses y Eignor, 2010; Robinson, 2016). Los del primer caso (a) se conocen como efectos de localización. Fernández-Alonso y Muñiz (2011) señalan, a partir de una estimación realizada por la OCDE, que los ítems resultan más fáciles si son aplicados al inicio de la prueba, tal vez, debido a factores como el cansancio, que

aumenta a medida que se avanza en la prueba. Los efectos de este tipo pueden controlarse manteniendo los ítems siempre en la misma posición, o variando el orden en que se presentan, de acuerdo con un plan; al respecto, Fernández-Alonso y Muñiz (2011) señalan que deben tomar todas las posiciones posibles. Esto no resulta factible para cada ítem; sin embargo, sí a nivel de los bloques.

El segundo caso (b) sería un reto para cualquier aplicación; cuando se aplican los ítems a través de diferentes medios, es importante realizar comprobaciones para verificar si su comportamiento es similar. Este efecto es importante para la aplicación de pruebas adaptativas, ya que estas consisten en aplicaciones electrónicas que requieren de un número importante de ítems, por lo que es probable que sus calibraciones de base provengan de una aplicación en papel y no de otra aplicación electrónica. El tercer caso (c) puede suceder, por ejemplo, cuando un ítem se encuentra junto a ítems de otro formato —que incluso pertenecen a otra prueba— en el mismo cuadernillo. Este hecho puede confundir a los evaluados en cuanto al formato de respuesta que deben utilizar, o cuando los ítems que lo anteceden presentan información útil para resolverlo. De acuerdo con Fernández-Alonso y Muñiz (2011), esto también puede pasar a nivel de bloques y desafía el supuesto de independencia local de la teoría de respuesta al ítem. Estos efectos pueden controlarse mediante las técnicas y los diseños que se presentarán en el siguiente capítulo.

En síntesis, es muy importante el uso de los diseños que se consideren adecuados en cada caso para la construcción de preguntas[9] y el armado de pruebas, así como la revisión sistemática de las formas de las pruebas y la atención a los criterios de selección de las preguntas y agrupaciones de estas. Ambos aspectos son relevantes porque seguir procedimientos cuidadosos (con criterios bien definidos) en las diferentes etapas por las que debe pasar una prueba es lo que permite que sea estandarizada (Aiken, 1997; Salkind, 1999). Estos procedimientos se siguen con el fin de garantizar una evaluación en igualdad de condiciones para los evaluados, favoreciendo la adecuada interpretación de los resultados. Lo anterior es fundamental cuando se utilizan los resultados para la toma de decisiones (Brown, 1983). Como resulta intuitivo, sería problemático que en un proceso de este tipo se aplicara a un evaluado una forma con un bajo nivel de dificultad y a otro una con uno muy alto, o que fueran diferentes las formas aplicadas en cuanto al contenido o los atributos considerados. Esto no permitiría una adecuada comparación de los participantes; por tanto, no solo se complicarían los procesos orientados a la toma de una decisión, como los de certificación y selección, sino que serían confusos los resultados de cualquier investigación.

9 Ver Icfes (2018a).

2. Diseños experimentales

En este capítulo se abordarán los diseños experimentales que, como señala Kirk (1995), constituyen un plan para la asignación de cada individuo, es decir, de cada evaluado, a una condición experimental o tratamiento, y el análisis estadístico relacionado. Los aspectos de los diseños experimentales que más se relacionan con el armado de pruebas son los del tratamiento y el control del error, que corresponden a la definición de cuántos y cuáles tratamientos se utilizarán y bajo qué regla se hará la asignación de los tratamientos (Hinkelmann y Kempthorne, 2008). En el armado de pruebas, los tratamientos son los bloques, que son las agrupaciones de ítems que se presentan como una unidad a través de las diferentes formas en las que se encuentran presentes. La definición de estos bloques se realiza con base en diferentes elementos, como los factores que se decidan considerar (diferentes tipos de bloques, por ejemplo), la disponibilidad de ítems y las condiciones de aplicación (tiempo disponible, amenazas a la seguridad de la información, cantidad de personas que serán evaluadas). Estos tratamientos pueden combinarse en formas, siguiendo las reglas especificadas para cada diseño[10].

Los diseños experimentales deben seguir tres principios en pro de su validez y sensibilidad[11]: la *replicación*, según la

10 Ver sección 2.2.

11 Se refiere a la capacidad de una prueba para identificar los verdaderos casos positivos. Un verdadero caso positivo en un experimento corresponde a la identificación de un efecto como producto de una intervención cuando realmente lo es.

cual cada tratamiento debe ser asignado a varios individuos; la *aleatorización*, en cuanto a la asignación de los individuos a los tratamientos, y el *control local* o *bloqueo*, que incluye los controles que se realizan para tener grupos homogéneos en cada tratamiento, tanto con respecto a los sujetos como al material que se utilizará (Hinkelmann y Kempthorne, 2008; Fernández-Alonso y Muñiz, 2011; Kuehl, 2001).

Cabe resaltar que, aunque con el término tratamiento se hace referencia a cada bloque, cuando se dice que el tratamiento debe ser asignado a varios individuos significa que cada bloque y cada forma deben ser asignados a varios evaluados. De igual manera, la asignación de los individuos a los tratamientos se refiere a la conformación de los grupos de personas que recibirán cada forma de la prueba (cada combinación de bloques), involucrando un componente aleatorio; en principio, cualquier evaluado puede recibir cualquiera de las formas disponibles.

El bloqueo, de otro lado, se puede llevar a cabo tomando en consideración uno o más factores en el grupo de evaluados; por ejemplo, puede asegurarse que las diferentes formas sean aplicadas a estudiantes de cada institución, o incluso de cada género. Lo mismo pasa en cuanto a las formas, ya que pueden tenerse en cuenta únicamente sus componentes, es decir, los bloques que involucran, o pueden irse agregando factores como el orden, la extensión, etc. Por ejemplo, si se tienen cuatro bloques, dos cortos y dos largos, pueden crearse formas con todas las combinaciones posibles de a dos bloques, que incluyan un bloque corto y uno largo; con ello, se estarían teniendo en cuenta dos factores: los bloques y su extensión.

Los tres principios, replicación, aleatorización y bloqueo, aplican para todos los experimentos y buscan garantizar unas condiciones mínimas que posibiliten la comparación entre los grupos en cuanto a factores propios de los tratamientos[12]. Sin embargo, también es necesario tener precauciones con respecto a las variables que no son de interés, pero pueden incidir en el resultado, esto es, que pueden ser fuentes de error. Para cada caso, deben definirse estas variables y las técnicas que se utilizarán para su control. Hay cinco técnicas que se pueden utilizar. La primera es la *eliminación*; por ejemplo, prohibir el ingreso de calculadoras y el uso de libros en un examen para que no presten ayuda. La segunda consiste en buscar la *constancia* (McGuigan, 1996), es decir, si no es posible que no afecten a ningún individuo, debe buscarse que los afecten a todos por igual. Esto se puede lograr, por ejemplo, a través de un formato de instrucciones preestablecido, mediante el uso de instrumentos semejantes —es decir, con la misma estructura y dificultad—, o limitando la aplicación a una edad o nivel educativo particular, en una jornada determinada.

La tercera técnica de control es el *balanceo*, que consiste en procurar que las variables afecten de la misma manera a los grupos que reciben los diferentes tratamientos. La manera más simple de asegurar el balanceo es comparar con uno o más grupos de control. Estos grupos tienen la misma conformación,

12 Para el caso de las pruebas, las diferentes formas deben tener características semejantes, por lo que se plantean las comparaciones entre los evaluados con respecto al atributo de interés.

pero no reciben tratamiento o reciben uno parcial. De esta manera, se pueden atribuir las diferencias a los tratamientos (McGuigan, 1996). Sin embargo, en casos como la aplicación de pruebas, en los que todos los grupos deben recibir algún tratamiento y estos deben contar con características semejantes, puede involucrar estrategias alternas como utilizar las diferentes formas en cada institución educativa o asegurar que cada bloque esté presente en una determinada proporción de cuadernillos.

La cuarta técnica es el *contrabalanceo*. Esta técnica de control permite lidiar con los efectos de orden, ya que tiene lugar cuando cada individuo es sometido a dos o más tratamientos. La idea es que el efecto de orden afecte por igual a cada tratamiento, ya que debe ser presentado a la misma cantidad de individuos, así como preceder y seguir a los demás tratamientos el mismo número de veces (McGuigan, 1996; Robinson, 2016). Un ejemplo de tal caso es un diseño de tres bloques, en el que se presentan todas las posibles combinaciones y permutaciones, distribuyéndolas por igual entre todos los individuos. La quinta técnica, la *aleatorización*, ya mencionada como principio, se da cuando a las personas que responderán a una misma forma de una prueba se les cita aleatoriamente en diferentes sitios o jornadas de presentación, lo que implica la asignación de condiciones como la iluminación, posibles ruidos, etc., buscando que tales variables afecten por igual a todos los grupos, sin asociarse sistemáticamente a uno en particular (McGuigan, 1996).

2.1 Características generales de los diseños experimentales

Los diseños que se van a revisar en este texto tienen cuatro características comunes. La primera es la cantidad total de tratamientos (t); para el caso de los diseños con bloques, se trata de la cantidad de bloques y no de la de formas[13]. La cantidad de formas (b) es más bien un resultado, ya que se define a partir de las diferentes características y el tipo de diseño. La segunda característica es la cantidad de tratamientos que hay para cada grupo (k), entendida como la cantidad de bloques que tiene cada forma. La tercera es la cantidad de grupos a los que se aplica cada tratamiento, es decir, la cantidad de repeticiones (r), la cantidad de veces que aparece cada bloque en el total de formas. La cuarta es la cantidad de grupos a los que se aplica cada par de tratamientos (λ); esto es, la cantidad de veces que aparecen juntos dos bloques en el total de formas. Por ejemplo, en un diseño que contemple un total de cuatro bloques[14] (t=4; A, B, C, D), en el que se tengan las cuatro formas (b=4): ABC, BCD, CDA y DAB, es claro que se tienen tres bloques por forma (k=3; A, B y C en la primera forma), tres repeticiones por bloque (r=3; el

13 En el caso de los diseños con bloques, cada individuo recibe una forma, que está conformada por bloques, por lo que es necesario hacer la distinción en este punto.

14 Por practicidad y por su uso frecuente, se emplearán las letras del alfabeto en orden para denominar los bloques; por ejemplo, en un diseño de cinco bloques, estos se denominarán: A, B, C, D y E (Montgomery, 2012; De Mendiburu, 2019).

bloque A, por ejemplo, aparece en tres formas: ABC, CDA y DAB), y cada par de tratamientos aparecerá dos veces ($\lambda=2$; la combinación BC, por ejemplo, aparece en dos formas: ABC y BCD). Estas características permiten individualizar cada diseño; en otras palabras, cuando un valor es asignado a estas características, resulta en un diseño único y diferente de cualquier otro que tome valores distintos.

También existen algunas propiedades que puede o no tener un determinado diseño, y que constituyen cualidades deseables; esto es, que sirven como criterios para determinar cuál es la mejor opción en una determinada situación (Hinkelmann y Kempthorne, 2008; Kuehl, 2001; Olive, 2017). A continuación, se presentan cinco de ellas, que se considera aplican para los diseños de armado, y un ejemplo que las ilustra:

Tabla 1. *Ejemplo de un diseño que presenta propiedades deseables*

Bloque / Forma	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

- *Ortogonalidad:*

  Se refiere a que los diferentes factores del diseño tengan la misma cantidad de niveles, situación que aumenta la precisión en la estimación, ya que puede ser realizada de manera independiente para cada combinación de niveles. Esto se puede ver en la tabla 1: la cantidad de filas (formas) es igual a la de columnas (bloques), y se controla la posición de los bloques, de manera que no se repiten tratamientos (bloques) por fila ni por columna.

- *Balance:*

  Se relaciona con la cantidad de veces que es presentado cada bloque, así como cada par de ellos, en la totalidad de formas. En la medida en que estas cantidades sean las mismas para los diferentes bloques, los tratamientos podrán ser comparados con igual precisión. En relación con la calificación, se tendrá igual cantidad de miradas para cada uno de los ítems; esto se debe tener en cuenta a la luz de los requisitos del modelo de calificación que se vaya a utilizar. En la tabla 1, cada bloque y cada par de bloques se presentan cuatro veces.

- *Eficiencia relativa:*

  Se refiere a la efectividad del control local o bloqueo, es decir, a la medida en que se logra la conformación de grupos homogéneos para los diferentes bloques mediante las reglas establecidas para su organización en formas. La eficiencia relativa de un diseño se obtiene al comparar su varianza al interior de las formas con la de un diseño más sencillo con la misma cantidad de repeticiones.

- *Conexión:*

  Se refiere a la posibilidad de comparar los diferentes pares de tratamientos. Las limitaciones pueden ocurrir según la estructura del diseño, más específicamente, cuando la cantidad de bloques no es la misma para todas las formas o la cantidad de repeticiones es diferente para diferentes bloques o pares de ellos. Un ejemplo de esto último se da cuando cada forma incluye bloques de dos tipos diferentes, por lo que dos bloques del mismo tipo nunca estarán juntos en una misma forma. Esto puede afectar el proceso de calibración de los ítems y el desarrollo de posibles investigaciones que requieran probar diferencias o relaciones entre ítems o agrupaciones. El diseño de la tabla 1 es un diseño conectado, ya que presenta todas las posibles combinaciones de bloques al incluir todos los bloques en cada forma.

- *Simplicidad:*

  Referida a la relación costo-beneficio entre la cantidad de elementos o restricciones contempladas por un diseño (con las correspondientes implicaciones a nivel logístico y de análisis) y la amplitud[15] y representatividad del sector de la población al que va dirigido. Por ejemplo, es más simple un diseño en el que los bloques apenas se agrupan

15 La cantidad de evaluados puede considerarse a la luz de los requerimientos del modelo de calificación, los recursos necesarios para la construcción, validación, aplicación y calificación de la prueba, o el nivel de exposición que vayan a tener los ítems.

en las diferentes formas que uno en el que, además, se controla su orden. Un diseño que contemple tipos de bloques también será de mayor complejidad que uno en el que se contemple un único tipo. Si el grupo al que se dirige la prueba es restringido, no se justificaría emplear un diseño que esté propenso a más errores y pueda ser más difícil de ejecutar y analizar. Sin embargo, cuando se tiene un grupo importante de personas y se corren riesgos en cuanto a la seguridad o hay posibilidad de que el orden u otras variables no relacionadas directamente con el atributo de interés afecten los resultados, se justifica utilizar un diseño más complejo. Por tratarse de un diseño completo, el ejemplo presentado en la tabla 1 constituye un diseño simple.

Es difícil que un determinado diseño cumpla con todas estas propiedades, ya que algunas incluso pueden parecer incompatibles con otros aspectos. Por ejemplo, un diseño en el que se utilice la técnica del contrabalanceo será un diseño menos simple. De igual manera, el tiempo del que se dispone para la aplicación de la prueba y la cantidad de bloques que involucra limitan su nivel de conexión. Lo que se espera es que estas propiedades sean consideradas y se tenga la mayoría posible en un determinado diseño.

2.2 Algunos diseños experimentales

En este apartado se presentan algunos diseños, detallando sus características a la luz de lo expuesto en el apartado anterior. La mayoría de ellos incorpora el uso de bloques[16] para la construcción de las formas.

Una primera clasificación de estos diseños tiene que ver con la proporción de los bloques disponibles que se emplea para construir cada forma. Existen diseños *completos*, que emplean todos los ítems disponibles en cada una de las formas, y diseños *incompletos*, que solo emplean una fracción. En general, se prefieren los diseños completos cuando la cantidad de ítems es reducida, y los incompletos, cuando no (Fernández-Alonso y Muñiz, 2011).

Un segundo criterio de clasificación está relacionado con el balanceo. Se considera *balanceado* un diseño en el que cada bloque, y cada par de bloques, aparece la misma cantidad de veces que los otros, como el diseño presentado en la tabla 1. De otro lado, se encuentran los diseños *parcialmente balanceados*. En estos diseños, todos los bloques aparecen la misma cantidad de veces, pero no sucede lo mismo con los pares de bloques, como en un diseño de cuatro bloques (ABCD) que cuente con cuatro formas (AB, BC, CD y DA). En este diseño,

16 En la literatura de diseños experimentales, en general, se encontrará el término bloque usado para referir a las formas; sin embargo, en cuanto a las pruebas escritas, se identificará con la agrupación más pequeña de ítems, como ya se definió.

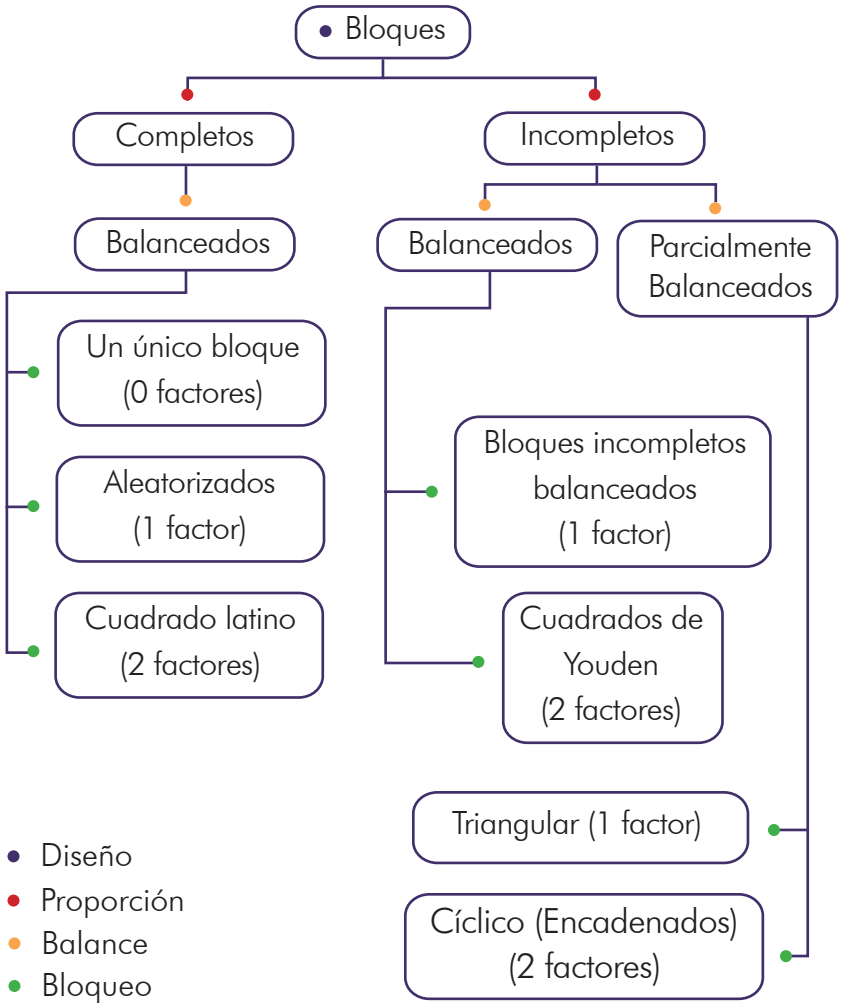
cada bloque (A, B, C, D) aparece dos veces; sin embargo, algunas parejas no aparecen (AC y BD) y otras aparecen una sola vez (Hinkelmann y Kempthorne, 2008; Fernández-Alonso y Muñiz, 2011). El último tipo, de acuerdo con este criterio, es el *no balanceado*, en el que la cantidad de repeticiones no es igual para todos los bloques.

Una tercera posibilidad de clasificación se da de acuerdo con la cantidad de factores de bloqueo que presenten, es decir, la cantidad de aspectos de los bloques que se tomen en cuenta en la matriz de representación de los diseños (como la presentada en la tabla 1). En este sentido, existen diseños que solo agrupan los bloques en formas, mientras que otros involucran otros criterios, como la cantidad de veces que aparecen en cada posición. Estos últimos se conocen como diseños *fila-columna*[17].

Para la presentación de estos diseños se empezará por los diseños completos, balanceados y con un único factor de bloqueo, avanzando hacia los más complejos, como se observa en la ilustración 4:

17 En esta clasificación aparecen diseños con hasta dos factores de bloqueo, una clasificación que tiene en cuenta una mayor cantidad de factores de bloqueo se puede encontrar en Hinkelmann y Kempthorne (2008).

Ilustración 4. *Organización de los diseños con bloques según los criterios de proporción, balance y bloqueo*



2.2.1 Diseños de bloques completos

Este primer grupo de diseños se caracteriza por utilizar todos los tratamientos disponibles con cada individuo (k=t), es decir que en cada forma se emplean todos los bloques (Fernández-Alonso y Muñiz, 2011). Estos diseños son balanceados y gozan de conexión y simplicidad; por esto mismo, tienden a ser los diseños de referencia para evaluar la eficiencia relativa de los diseños incompletos.

- *Diseño de un único bloque*

Un caso particular de los diseños completos es aquel en que se tiene una única forma para todos los evaluados, esto es, un único bloque. En este caso particular, no tendría caso hablar de un control local o de una aleatorización en la asignación de las formas; de hecho, el riesgo de copia sería muy alto (Fernández-Alonso y Muñiz, 2011). Tampoco se podrían tomar acciones con miras a anular o mitigar posibles efectos de contexto diferentes a los de localización. El uso de tal diseño es posible en casos en los que resulte muy baja la cantidad de individuos por forma[18] o cuando la cantidad de ítems por cada uno de los bloques no sea suficiente para que

18 En cuanto a este aspecto, no hay un valor único de referencia; en algunos casos se considera que una muestra es grande cuando está conformada por 30 o más individuos (Lahoz-Beltrá, Ortega-Escobar y Fernández-Montraveta, 1994; Triola, 2004), o que se requiere un mínimo de entre cinco y diez personas por ítem (Nunnally, 2000), mientras que, para un modelo de teoría de respuesta al ítem, un mínimo requerido para la estimación sería de 500 individuos para el caso de dos parámetros y de 1000 para el de tres (Baker y Kim, 2004).

cada forma incluya un conjunto de especificaciones que sea representativo del dominio.

- *Diseño de bloques completos aleatorizados (RCBD[19])*

Este diseño asigna aleatoriamente el orden de aparición de cada tratamiento o bloque. Sin embargo, esta aleatorización es limitada por el hecho de que todos los tratamientos se aplican juntos (Montgomery, 2012). Se considera un único factor de bloqueo, consistente en la presentación de todos los bloques en cada forma. Usualmente, se fija una cantidad de formas; sin embargo, no hay una limitación inicial en este sentido, más allá de las posibles combinaciones. La tabla 2 presenta un ejemplo de un diseño con cinco bloques y tres formas ($t=k=5$, $b=r=\lambda=3$):

Tabla 2. *Bloques, según posición y forma, en un diseño RCBD de cinco bloques*

Posición / Forma	1	2	3	4	5
1	C	A	D	E	B
2	E	C	A	B	D
3	C	D	A	B	E

19 Por su sigla en inglés (Randomized Complete Block Design).

Este diseño, a pesar de la aleatorización, no garantiza que exista contrabalanceo, pues no siempre aparecerán todos los bloques antes y después unos de otros (véase el caso del bloque C en relación con los bloques A, B y D en la tabla 2). Se puede observar que, a medida que se van asignando bloques a las posiciones, se reduce la cantidad de estas que queda disponible para ser ocupada en cada forma.

- *Cuadrado latino (LSD[20])*

En la tabla 3 se presenta un ejemplo de este diseño, un cuadrado latino estándar de cinco bloques (b=t=k=r=λ=5):

Tabla 3. *Bloques, según posición y forma, en un diseño LSD de cinco bloques*

Posición / Forma	1	2	3	4	5
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

20 Por su sigla en inglés (Latin Square Design).

La particularidad de este diseño es que cada bloque aparece una sola vez en cada fila y en cada columna; por tanto, si para la forma 1 se ubicó el bloque C en la tercera posición, este bloque no podrá tomar tal posición en ninguna otra forma (ver las celdas sombreadas en la tabla 3). En este caso, se tienen dos factores de bloqueo, el uso de todos los bloques para la construcción de las formas y su posición. Se trata de un diseño ortogonal, donde la cantidad de formas es igual a la de bloques, los bloques por forma, las repeticiones y la cantidad de apariciones de cada par de bloques (Hinkelmann y Kempthorne, 2008; Montgomery, 2012). Se conoce como cuadrado latino estándar a aquel que sigue el orden alfabético; esto es, parte de la primera letra y se desplaza una posición (a la derecha o a la izquierda) en cada forma. Usualmente, se parte de este diseño, alterándolo para conseguir un cuadrado latino no estándar (Montgomery, 2012). En este diseño sí se puede hablar de contrabalanceo, pudiendo o no haber cierto grado de aleatorización en el ordenamiento de las formas y las posiciones, según se hagan o no alteraciones al cuadrado latino estándar inicial.

2.2.2 Diseños de bloques incompletos

Estos diseños se caracterizan por utilizar en cada forma menos tratamientos o bloques de los que se tienen disponibles; por ejemplo, cuando se tienen cinco bloques para la conformación de una prueba, pero se utilizan únicamente dos por cada cuadernillo (Hinkelmann y Kempthorne, 2008; Montgomery, 2012; Fernández-Alonso y Muñiz, 2011; Hinkelmann y

Kempthorne, 2005). Estos diseños se prefieren cuando se tiene una amplia cantidad de ítems, que no es recomendable aplicar en su totalidad a cada individuo por cuestiones de tiempo, cansancio, exposición de la información, limitaciones en la impresión, posibilidad de solapamiento entre los contenidos de los ítems y demás (Fernández-Alonso y Muñiz, 2011). También resultan de utilidad cuando se tiene una cantidad muy grande de personas, ya que se prestan para obtener una cantidad de formas mayor de las que suelen obtenerse en los diseños completos. En principio, estos diseños implican un factor de bloqueo, consistente en la unión de un conjunto incompleto de bloques para la constitución de cada forma. Sin embargo, pueden involucrar en varios casos un segundo factor de orden, lo que resulta ser una de sus ventajas. En especial, en los que son parcialmente balanceados, ya que cada bloque puede tener contadas apariciones y algunos pares de bloques pueden no aparecer o tener más o menos apariciones que los otros (diseños desconectados), como en el ejemplo presentado en la sección 2.2 (Hinkelmann y Kempthorne, 2008; Montgomery, 2012; Fernández-Alonso y Muñiz, 2011; Hinkelmann y Kempthorne, 2005). Al tener una cantidad de formas diferente a la cantidad total de bloques, estos diseños no resultan ortogonales; no obstante, suelen variar la posición en la que aparecen los bloques, aleatorizar sus formas y mantener uniforme la cantidad de repeticiones por bloque, incluyendo algún nivel de contrabalanceo, aleatorización y balance.

- *Diseño de bloques incompletos balanceados (BIBD[21])*

Se trata de un diseño en el que se garantiza que la cantidad de repeticiones de cada bloque sea igual para los diferentes bloques a lo largo de todas las formas, así como la cantidad de apariciones de cada par; es decir, se trata de un diseño balanceado y conectado. Otra propiedad de este diseño es la equivalencia entre los productos de dos pares de sus parámetros[22] ($r*t=k*b$); la cantidad total de bloques multiplicada por la cantidad de repeticiones por cada bloque debe ser equivalente a la cantidad de bloques por forma multiplicada por la cantidad de formas. Finalmente, debe existir otra relación entre los parámetros: $\lambda(t-1)=r(k-1)$, que posibilita la obtención de la cantidad de apariciones por cada par de bloques: $\lambda=r*(k-1)/(t-1)$.

Cada diseño se desarrolla a partir de los valores de tres parámetros iniciales (cantidad total de bloques, bloques en cada forma y repeticiones por bloque). Debido a las dos últimas propiedades mencionadas, existe la limitación de que no es posible cualquier combinación. Algunos diseños resultan irresolubles, es decir, no se puede llegar a una constitución de las formas. Un ejemplo de esto es un diseño de ocho bloques en total, en el que se utilicen tres para cada forma, cada uno con dos repeticiones en el total de formas.

21 Por su sigla en inglés (Balanced Incomplete Block Design).

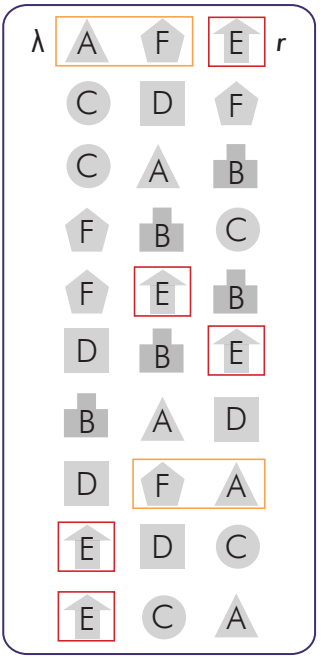
22 En el sentido de una función, son los argumentos a partir de cuyos valores se determina el diseño y se obtiene una solución u organización de los bloques.

En este ejemplo, reemplazando los valores en la expresión previamente presentada, se tiene: 2*8=3*b. Para que esto tuviera solución, se requeriría que b fuera igual a 16/3, lo que no es un número entero y no representa una cantidad de formas. También sería irresoluble un diseño en el que se obtenga un valor que no sea entero para λ. De igual manera, este diseño no garantiza que haya un contrabalanceo, ya que la solución resultante puede contener bloques que siempre estén en una misma posición con respecto a otros, o puede haber posiciones que no sean ocupadas por un determinado bloque. Esto implica que el diseño se ve expuesto a posibles efectos de contexto.

A continuación, se presenta un ejemplo de un diseño con seis bloques, en el que se utilizan tres para cada forma, con cinco repeticiones por bloque y dos por cada par (b=10, t=6, k=3, r=5, λ=2):

Ilustración 5. *Bloques, según posición, forma y apariciones, en un BIBD de seis bloques*

Forma	Bloque 1	Bloque 2	Bloque 3
1	A	F	E
2	C	D	F
3	C	A	B
4	F	B	C
5	F	E	B
6	D	B	E
7	B	A	D
8	D	F	A
9	E	D	C
10	E	C	A

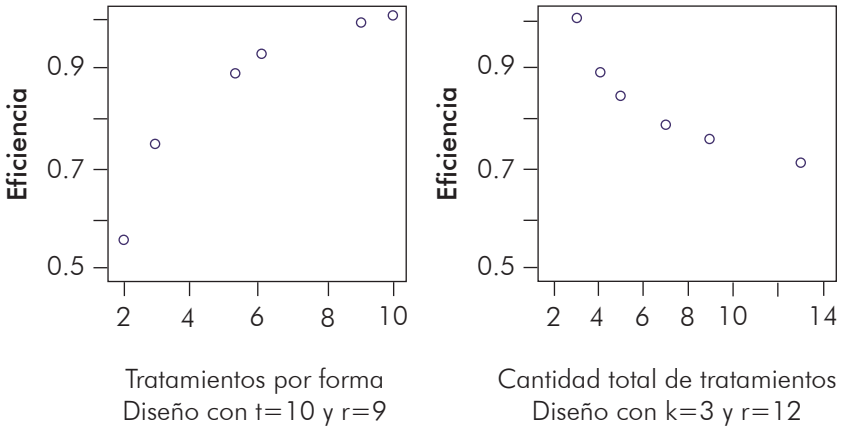


En la ilustración 5 se puede observar la disposición de los bloques en formas, tanto a través de figuras como en una tabla. Se señalan con un recuadro rojo las cinco apariciones del bloque E y con uno amarillo las dos apariciones del par A-F. Se puede comprobar que λ=5*(3-1)/(6-1)=5*2/5=2. También se puede comprobar que r*t=k*b, ya que 5*6=3*10. Esto, porque ambos lados de la igualdad corresponden a la cantidad de celdas que tiene la matriz, 30 en este caso.

Cada bloque ha ocupado cada una de las posiciones posibles; sin embargo, no siempre es así, ya que se realizan procesos de aleatorización para la asignación a cada posición. Podría, por ejemplo, haber resultado una distribución con tres apariciones del bloque A en la primera posición y dos en la tercera, sin apariciones en la segunda. También se observa que, a pesar de esto, cada bloque no es precedido y seguido por todos los demás; el bloque A, por ejemplo, siempre es precedido por el C (ver formas 3 y 10).

Estos diseños tienen la particularidad de que se pierde eficiencia relativa a medida que disminuye la cantidad de bloques por forma o aumenta la cantidad total de bloques. Esto se puede ver en la ilustración 6:

Ilustración 6. *Eficiencia de los diseños BIBD, según la cantidad de tratamientos*



En la gráfica de la izquierda puede verse cómo aumenta la eficiencia a medida que aumenta la cantidad de bloques por forma, manteniendo la cantidad total de tratamientos y las repeticiones. De manera inversa, en la gráfica de la derecha se observa que, para cantidades fijas de bloques por forma y repeticiones, la eficiencia disminuye a medida que aumenta la cantidad total de bloques. La ausencia de puntos para ciertos valores del eje horizontal indica que se trata de diseños irresolubles.

- *Cuadrados de Youden*

Este diseño corresponde a un cuadrado latino al que se le retiran una o varias filas, columnas o diagonales. En todo caso, para que sea un diseño incompleto, se requiere que la cantidad de bloques en cada forma sea inferior a la cantidad total de bloques; es decir, que se retire al menos una columna o diagonal (de acuerdo con la presentación de la tabla 4). Realmente, se trata de un rectángulo. Se realiza un control por filas y columnas; esto es, cada bloque aparece como máximo una vez en cada forma y una vez en cada posición (Kabe y Gupta, 2007). Es un diseño balanceado, ya que los diferentes pares de bloques aparecen la misma cantidad de veces; por esto, no basta con retirar cualquier conjunto de elementos a un cuadrado latino (Fernández-Alonso y Muñiz, 2011). En esta medida, es un diseño conectado en el que puede o no aplicarse el contrabalanceo, dependiendo de la diferencia entre el total de bloques y los bloques que aparecen en cada forma. Es un diseño más complejo que un diseño completo, por tanto, será menos eficiente; sin embargo, no se diferenciará de un BIBD,

ya que también es incompleto, conectado y balanceado. A continuación, se presenta un ejemplo de un diseño de este tipo, con cuatro bloques y tres repeticiones (b=t=4, k=r=3, λ=2):

Tabla 4. *Bloques, según posición y forma, en un cuadrado de Youden, con cuatro bloques y tres repeticiones*

Posición / Forma	1	2	3
1	D	C	A
2	C	D	B
3	B	A	D
4	A	B	C

En los cuadrados de Youden, la máxima cantidad de repeticiones posible es igual a la cantidad total de bloques, caso en el que constituiría un cuadrado latino. Así, siendo estrictos, solo se podría construir un cuadrado de Youden con una cantidad máxima de repeticiones de t-1, 3 en el ejemplo, ya que si se tuvieran t=4 repeticiones, se estaría utilizando cada bloque en todas las formas, con lo que se tendría un diseño completo.

- *Diseños de bloques incompletos parcialmente balanceados (PBIBD[23])*

Existe una gran cantidad de diseños parcialmente balanceados. Algunas clasificaciones se pueden encontrar en Hinkelmann y Kempthorne (2005). Su característica principal es que diferentes pares de tratamientos aparecen en una cantidad diferente de veces a lo largo de todas las formas; inclusive, pueden existir pares que no aparezcan en ninguna forma, por lo que se tiene más de un valor para el parámetro λ (Hinkelmann y Kempthorne, 2005). Sin embargo, la cantidad de repeticiones por bloque es igual para los diferentes bloques a lo largo de todas las formas (r). A continuación, se presentan dos diseños de este tipo: triangular y cíclico.

- *Diseño triangular*

Este diseño parte de una tabla cuadrada, en la que no se toca la diagonal, es decir, las celdas cuyas filas y columnas están en la misma posición. En las celdas restantes, se asignan los tratamientos simétricamente con respecto a esta diagonal, partiendo de la fila o columna más externa, en orden, por columnas debajo de la diagonal, y por filas encima (Hinkelmann y Kempthorne, 2005). El resultado final se fija cuando se presenta la tabla sin la diagonal, como una matriz. A continuación, se muestra un ejemplo, un diseño triangular con seis bloques ($b=4$, $t=6$, $k=3$, $r=2$, $\lambda_1=0$, $\lambda_2=1$):

23 Por su sigla en inglés (Partially Balanced Incomplete Block Design).

Tabla 5. *Bloques, según posición y forma, en un diseño triangular de seis bloques*

Cuadrado			
	B	C	D
B		F	E
C	F		A
D	E	A	

Posición / Forma	1	2	3
1	B	C	D
2	B	F	E
3	C	F	A
4	D	E	A

En este ejemplo se ha señalado la tabla inicial como Cuadrado, las celdas más externas se rellenaron con amarillo y las que conforman la diagonal con azul. El grupo de celdas que se encuentra a la derecha corresponde al resultado final. También podría decirse que se parte de organizar los tratamientos por debajo de la diagonal, por columnas, para después hacer corresponder cada columna debajo de la diagonal con cada fila por encima. Como se puede observar, resulta un diseño con una cantidad de bloques mayor a la de formas, en el que existe poca conexión, ya que algunas parejas de bloques nunca aparecen, como A-B, C-E y D-F. Tampoco se trata de un diseño ortogonal ni se aplica el contrabalanceo; son diferentes las cantidades de filas y columnas y se observan bloques que aparecen siempre en la misma posición: el bloque B siempre aparece de primeras, el F de segundas y el A de terceras, con los posibles efectos de orden que esto conlleva.

▸ *Diseño cíclico*

Estos diseños parten de un conjunto de bloques o forma inicial, cuyo patrón se sigue para la creación de los siguientes. Por ejemplo, si se inicia con la forma A-C-G en un diseño de nueve bloques (como el de la tabla 6), la forma que se encontrará a una distancia de 3 bloques de ella será D-F-A. Este diseño resulta muy práctico cuando se tiene una cantidad importante de bloques, pero se pueden utilizar muy pocos en cada forma, y se acomoda a diferentes estructuras cuando se utilizan apenas dos bloques por forma, como las cadenas (Diseño de bloques encadenados) o las estrellas (Fernández-Alonso y Muñiz, 2011; Clatworthy, 1955). En la tabla 6 se presenta un diseño cíclico, con nueve bloques, en el que se utilizan tres bloques por forma ($b=9$, $t=9$, $k=3$, $r=3$, $\lambda_1=0$, $\lambda_2=1$):

Tabla 6. *Bloques, según posición y forma, en un diseño cíclico de nueve bloques*

Forma	A	B	C	D	E	F	G	H	I
1									
2									
3									
4									
5									
6									
7									
8									
9									

Bloque		
1	2	3
A	C	G
B	D	H
C	E	I
D	F	A
E	G	B
F	H	C
G	I	D
H	A	E
I	F	B

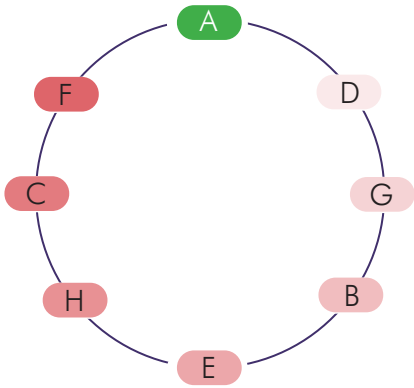
En la tabla 6 se puede observar una disposición inicial de los bloques (a la derecha), tal como se derivan de la forma inicial (ACG); a esto también corresponden las celdas sombreadas, en las que el amarillo pálido indica la primera posición, el color oro la segunda y el rojo pálido la tercera. Estos bloques pueden reorganizarse internamente dentro de cada forma, de manera aleatoria, o puede controlarse la posición en que aparecen, generando una nueva disposición final. De igual manera, las formas se pueden aleatorizar para lidiar con posibilidades de copia si el diseño es de conocimiento público. Como se puede observar, en este caso la cantidad de formas es igual a la cantidad total de bloques. Sin embargo, el esquema final no necesariamente asegura un contrabalanceo (el bloque B, por ejemplo, podría no aparecer en la última posición por efecto de la aleatorización). Tampoco se trata de un diseño ortogonal ni conectado; sin embargo, es un diseño fácil de utilizar y relativamente simple cuando no se controla la posición del bloque. Es necesario indicar, sin embargo, que un diseño de este tipo puede tener algún grado de conexión en los casos en que todos los valores λ_i sean mayores que cero, y no necesariamente tiene que utilizarse una distancia de 1. Es posible, por ejemplo, utilizar un diseño de ocho bloques y emplear únicamente las formas 1, 3, 5 y 7, dejando una distancia de 2. Cuando cada forma se genera tomando una distancia mayor a un bloque de la anterior, se tiene un diseño cíclico generalizado. Estos diseños no se consideran en este documento, pero se puede encontrar más información en John y Williams (1995).

▸ *Diseño de bloques encadenados o circuito cerrado*

Se trata de un caso especial del diseño cíclico, en el que cada forma está compuesta por dos bloques y las formas constituyen eslabones de una cadena que se cierra (Fernández-Alonso y Muñiz, 2011). Esto quiere decir que los pares de formas compartirán bloques entre ellas, que no se encontrarán en ninguna otra forma. Este tipo de diseño permite hacer una especie de contrabalanceo, ya que se aplica una corrección fila-columna, en la que aparece cada bloque en las diferentes posiciones disponibles (véase el bloque A, sombreado con amarillo). A continuación, se presenta un ejemplo para un total de ocho bloques ($b=8$, $t=8$, $k=2$, $r=2$, $\lambda_1=0$, $\lambda_2=1$):

Ilustración 7. *Bloques, según posición y forma, en un diseño de bloques encadenados (ocho bloques)*

Posición / Forma	1	2
1	D	G
2	G	B
3	B	E
4	E	H
5	H	C
6	C	F
7	F	A
8	A	D



Como se puede ver, la cadena se cierra, ya que el segundo bloque de la forma final es el primero de la inicial. Cada bloque aparece en dos formas diferentes y se obtiene un diseño con ocho formas.

Este diseño puede involucrar más de dos bloques en cada forma, caso en el que se mantendría la igualdad entre la cantidad total de bloques y de formas, pero aumentaría la cantidad de repeticiones por bloque y por pares de bloques. Ejemplo: un diseño de siete bloques en total, con tres en cada forma: A-B-D, B-C-E,..., G-A-C.

3. Ejemplos del uso de los diseños para el armado de pruebas

En este capítulo se realizará una breve contextualización del uso de diferentes diseños que involucran bloques en la evaluación educativa. Primero, se abordarán las pruebas internacionales, como el Programa Internacional para la Evaluación de Estudiantes (PISA), el Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS) y el Estudio Internacional del Progreso en Comprensión Lectora (PIRLS). En seguida, las pruebas de uso en los diferentes niveles educativos al interior de cada país, incluyendo a Colombia.

3.1 Pruebas de uso a nivel internacional

- *Programa Internacional para la Evaluación de Estudiantes (PISA)*

Este programa de la Organización para la Cooperación y el Desarrollo Económico (OCDE) busca medir qué tan bien preparados están los estudiantes de 15 años para enfrentar los desafíos de las sociedades de hoy, teniendo en cuenta que se encuentran cerca de terminar el ciclo de estudios obligatorios. La OCDE ha reportado el uso del diseño de bloques incompletos balanceados para PISA en el año 2012 y otros previos (OCDE, 2012). Tal es el caso de los años 2003, 2006 y 2009, de acuerdo con Fernández-Alonso y Muñiz (2011). Para este examen, se describe un conjunto de 13 bloques (siete de Matemáticas, tres de Lectura y tres de Ciencias para 2012, y siete de Ciencias, cuatro de Matemáticas y dos de

Comprensión lectora para los otros tres años), cuatro de los cuales se emplean para la construcción de cada forma. Cada bloque tiene una duración esperada de 30 minutos.

En este caso, se partió de los 13 bloques en total para el diseño, ya que en cada forma se mezclaban bloques de diferentes áreas, por lo que cada forma resultó ser un cuadernillo; podían encontrarse formas con ninguno o con tres bloques de matemáticas, por ejemplo. Para estas aplicaciones, se obtuvieron 13 formas o cuadernillos, con cuatro repeticiones de cada bloque y una aparición de cada par. Sin embargo, tal como lo ilustran Fernández-Alonso y Muñiz (2011), este diseño también corresponde a un cuadrado de Youden, ya que cada tratamiento aparece una sola vez en cada forma y en cada posición.

Además de este diseño, en otros momentos, PISA ha hecho uso de bloques incompletos parcialmente balanceados. Tal es el caso del estudio principal del año 2000. Para esta aplicación se crearon nueve bloques de lectura con una duración de 30 minutos, seis de matemáticas de 15 minutos y seis de ciencias de 15 minutos. Estos bloques se agruparon en nueve cuadernillos, con una duración de dos horas. Estos cuadernillos eran producto de la combinación de dos diseños y otros bloques adicionales. Del primero de estos diseños, un diseño de bloques encadenados con un total de siete bloques de lectura, surgieron siete formas, cada una incluida al inicio de los primeros siete cuadernillos. El segundo diseño es un cuadrado latino estándar con dos bloques de lectura en total, que se aplicaban al final de los cuadernillos 8 y 9, y eran

diferentes de los de los primeros cuadernillos. Los seis primeros cuadernillos eran complementados por tres combinaciones de bloques de matemáticas y tres combinaciones de ciencias. El séptimo cuadernillo finalizaba con el octavo bloque de lectura, y los cuadernillos 8 y 9 iniciaban por dos bloques de matemáticas seguidos por dos bloques de ciencias, o dos bloques de ciencias seguidos de dos bloques de matemáticas. Se incluyó un cuadernillo adicional para población con discapacidad. Esta forma era más corta y fácil que las demás (OCDE, 2002).

En 2015 y 2018 se emplearon 18 bloques en el dominio principal (Ciencias y Lectura, respectivamente) y seis en cada uno de los restantes. La aplicación del año 2018 se realizó de forma electrónica[24] e involucró pruebas adaptativas y, para dos bloques, una versión estándar y una facilitada. En los *Reportes técnicos* de la OCDE (2017; *s.f.*), se observan combinaciones entre diseños de bloques encadenados, cuadrados latinos de dos bloques, combinaciones entre los diferentes pares de pruebas y, adicionalmente, cuadernillos de solo lectura en el año 2018 (OCDE, 2017; OCDE, *s.f.*).

24 Solo un pequeño grupo de países no la implementó de esta forma, sino a través de cuadernillos de lápiz y papel.

- *Estudio Internacional del Progreso en Comprensión Lectora (PIRLS)*

Este estudio evalúa el rendimiento en lectura de estudiantes de cuarto de primaria, momento en el que se considera que han aprendido a leer y leen para aprender. Fue creado como complemento al Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS) (Mullis y Martin, 2015). De acuerdo con Mullis y Martin (2015), este examen utiliza un diseño de bloques incompletos parcialmente balanceados por limitaciones de tiempo, ya que la aplicación de todas las preguntas tendría una duración de ocho horas, pero solo se dispone de un tiempo de aplicación de 1 hora y 20 minutos. Se parte de doce bloques, que se agrupan de a dos en cada cuadernillo, uno de tipo informativo y otro de tipo literario, de acuerdo con el tipo de texto que contienen (la lectura como experiencia literaria, y para la adquisición y el uso de información). Esto quiere decir que no se encuentran todas las combinaciones posibles entre pares de bloques, sino solo de cada bloque con tres más ($r=3$), resultando en 15 formas. Dado que estas agrupaciones no siguen un patrón claro y al ser la cantidad de bloques por forma menor que la de repeticiones por bloque, no se trata de un diseño cíclico. A pesar de que se varía la posición en el tipo de bloque (informativo vs. literario), no se efectúa el mismo control por cada uno de los doce bloques. También se agrega una forma con dos bloques adicionales, que no se combinan con los demás. Esta última forma no sigue el diseño del resto del examen, se trataría de un segundo diseño, una extensión de un diseño de un único bloque, dado que incluye dos. Este tipo de formas, que no comparten ítems con

las demás, resultan de utilidad cuando se tiene el propósito de realizar estudios o investigaciones, empleando una muestra con las mismas cualidades de una aplicación regular.

- *Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS)*

De acuerdo con Martin y Kelly (1996), este es el estudio comparativo más amplio y ambicioso que se ha realizado sobre el logro en la enseñanza y el aprendizaje de las matemáticas y las ciencias, integrando más de cincuenta países. En esta evaluación se contemplan tres aspectos centrales relacionados con el currículo: el contenido, las expectativas de desempeño con respecto a un conjunto de procesos cognitivos, y las perspectivas, que involucran actitudes, intereses y hábitos, entre otros. En este estudio se cobijaron tres poblaciones, diferenciadas por la edad o el grado (9 años, 13 años de edad y final de la secundaria), y el tiempo máximo de aplicación (70 min, 90 min y 90 min). Cada población tenía un diseño diferente.

Para las poblaciones 1 y 2, se tenían ocho formas[25] distribuidas aleatoriamente entre los evaluados, que compartían un único bloque, ubicado siempre en la segunda posición. Aunque los autores señalan que las diferentes formas eran aproximadamente semejantes en cuanto a dificultad y contenido,

25 Estas formas involucraban contenidos tanto de ciencias como de matemáticas; sin embargo, se tratan como formas, no como cuadernillos, debido a que reflejan diseños específicos mencionados en la anterior sección.

los bloques no tenían el mismo tiempo de aplicación esperado ni aparecían la misma cantidad de veces. Esto porque se trató de una combinación de diseños. Cada forma incluía el bloque compartido (diseño de un único bloque), tres bloques producto de un diseño de bloques encadenados (de siete bloques posibles y que solo aplica para las primeras siete formas), dos bloques producto de uno de dos diseños de bloques encadenados (de cuatro bloques posibles cada uno), y un bloque adicional para cada forma que no aparece en ninguna otra. La octava forma incluía para la primera población: el bloque compartido, uno de los siete bloques del primer diseño encadenado, una forma de uno de los otros diseños encadenados y tres bloques que no aparecían en ninguna otra forma. Esto refleja el interés por realizar una mayor cobertura del contenido en la última forma, tanto en ciencias como en matemáticas, ya que la gran mayoría de las otras formas se enfoca solo en una de las áreas. En la segunda población, esta octava forma no incluía la forma del diseño de bloques encadenados señalada para la primera población, es decir, que solo incluía cinco bloques. En conjunto, cada forma incluía cinco o siete de 26 bloques posibles; es decir, se trataba de un diseño incompleto. En la tercera población, de otro lado, se contó con nueve formas, cada una conformada por entre dos y tres bloques de un total de 12. Se tenían cuatro conjuntos de formas: las dos formas del primer conjunto compartían dos bloques, aunque en un orden de presentación distinto, como un cuadrado latino, y tenían uno adicional que solo aparecía en cada forma. Las tres formas del siguiente conjunto compartían un bloque y cada una tenía un bloque adicional que no estaba en ninguna otra forma. Sucedía lo mismo para las tres formas del siguiente conjunto.

La novena forma, que representaba el cuarto conjunto, incluía uno de los bloques compartidos de cada uno de los otros conjuntos, es decir, estaba compuesta por tres bloques.

Para la aplicación del 2007, de otro lado, se empleó el diseño de bloques encadenados, con 14 bloques de matemáticas y 14 de ciencias, la mitad de anclaje con la aplicación del 2003, a partir de los que se construyeron 14 formas o cuadernillos, que incluían dos bloques de cada área (Olson, Martin y Mullis, 2008; Fernández-Alonso y Muñiz, 2011). La aplicación se realizó en dos sesiones. En cada una se aplicaba un par de bloques de la misma área, intercambiando el orden de aparición de las áreas para las diferentes formas. Es decir, mientras unas iniciaban con dos bloques de ciencias, otras iniciaban con dos de matemáticas, con dos repeticiones por bloque y un tiempo de resolución de 72 minutos para el grado cuarto y de 90 minutos para octavo, a partir de un total de 353 ítems en cuarto y 429 en octavo (Olson, Martin y Mullis, 2008). Esta aplicación se realizó como si se tuvieran dos diseños de 14 bloques, uno para cada área, variando no solo el orden de aparición de cada bloque, sino también de cada área.

- *Tercer Estudio Regional Comparativo y Explicativo (TERCE)*

Este estudio fue realizado en 2015 con el objetivo de obtener información de tipo nacional (15 países) y regional sobre los logros alcanzados por los estudiantes, y los factores asociados a ellos, en las áreas de lectura, escritura y matemáticas en los grados tercero y sexto, y en ciencias naturales en el grado

sexto (LLECE, 2015). Para cada prueba se emplearon seis bloques, dos de los cuales eran bloques de ancla obtenidos del Segundo Estudio Regional Comparativo y Explicativo (SERCE). Cada forma estaba compuesta por dos bloques, en una posición diferente cada vez, para un total de seis formas. Así mismo, cada bloque incluía, para lectura, 11 preguntas en tercero y 16 en sexto; para matemáticas, 12 y 13 en tercero, y 16 y 17 en sexto (bloques nuevos y de anclaje, respectivamente). Para ciencias naturales en sexto se incluyeron 15 y 16 ítems (bloques nuevos y de anclaje) (LLECE, 2015; UNESCO-OREALC, 2016). Como se puede ver en UNESCO-OREALC (2016), esto da cuenta de un diseño encadenado, en el que se tienen dos repeticiones por bloque, y existen pares que no aparecen mientras que otros lo hacen únicamente una vez.

- *Encuesta de Habilidades para la Empleabilidad y la Productividad (STEP)*

El STEP es un programa del Banco Mundial que se aplica en doce países, con niveles de ingresos bajos a medios, con el objetivo de entender la interacción entre las habilidades y la empleabilidad y la productividad (ETS, 2014). Para la evaluación de la competencia lectora, en este programa se utiliza un grupo central de ocho ítems que sirve de tamizaje, es decir, como valoración inicial para determinar si se va a realizar una evaluación completa. Esta prueba, que involucra un diseño de un único bloque, toma aproximadamente siete minutos para su aplicación y, si es superada, da paso a la asignación aleatoria de uno de cuatro cuadernillos de

ejercicios. Cada uno de estos cuadernillos se encuentra conformado por dos bloques de nueve ítems, siguiendo un diseño cíclico, con 28 minutos de tiempo esperado para su resolución (ETS, 2014). Este diseño cuenta con dos repeticiones para cada bloque, con cero a una aparición de cada par, sin que se indique con precisión en el documento informativo si se realiza o no un control sobre el orden de aparición de los bloques.

3.2 Pruebas de uso nacional en diferentes países

- *Evaluación Nacional del Progreso Educativo (NAEP) (Estados Unidos)*

Este programa evalúa el desempeño académico de los estudiantes estadounidenses en varias áreas del conocimiento y presenta resultados a nivel grupal (por género, grupo étnico, ubicación del colegio, etc.) (NCES, *s.f.a*). Este examen, existente desde el año 1969, reportó para el periodo 1983-1984 el uso de diseños de bloques incompletos balanceados y no balanceados, con 24 bloques para la construcción de un total de 63 formas (Beaton, 1987). En sus inicios, el diseño era prácticamente de un único bloque, basado en la construcción de formas paralelas, cambiando al diseño de bloques incompletos balanceados desde mediados de los años ochenta (Fernández-Alonso y Muñiz, 2011). En 1996, la prueba de matemáticas de este examen, para el grado octavo, estuvo compuesta por 13 bloques, combinados en 26 cuadernillos, con tres bloques por cuadernillo, seis repeticiones

por bloque y una aparición de cada par de bloques (Van der Linden, Veldkamp y Carlson, 2004). Para este examen también se han empleado cuadrados de Youden y bloques incompletos parcialmente balanceados (Fernández-Alonso y Muñiz, 2011), encadenados en particular. Esto se reporta desde 2003, de acuerdo con el Centro Nacional sobre las Estadísticas de la Educación (NCES) (*s.f.b*).

- *Evaluación Periódica de la Educación (PPON) y Evaluación Anual de Niveles Educativos (JPON) (Holanda)*

Estas dos evaluaciones buscan brindar información de utilidad para legisladores, educadores y público en general sobre las tendencias en el logro de los estudiantes. Sin embargo, mientras que el JPON evalúa el dominio de la lengua holandesa y las matemáticas en los grados cuarto y octavo, con la intención de dar una retroalimentación regular con respecto a las reformas en la educación primaria, el PPON se realiza cada cinco años, busca cubrir un mayor espectro del currículo (se incluyen áreas como historia, biología, inglés, música y educación física), especialmente en estudiantes de octavo (solo ocasionalmente se incluyen los de cuarto), y se enfoca en cambios más generales a través del tiempo (Van der Linden, Veldkamp y Carlson, 2004). Las pruebas aplicadas en estos programas siguen el diseño de bloques incompletos balanceados, con el propósito de cubrir una amplia variedad de ítems sin sobrecargar a los estudiantes (Van der Linden, Veldkamp y Carlson, 2004; Nusche, Braun, Halász y Santiago, 2014).

- *Evaluación Censal de Estudiantes (ECE) (Perú)*

El Ministerio de Educación del Perú (Minedu, 2016), ha reportado el uso tanto de formas únicas, algunas con versiones diferentes dependiendo de la lengua empleada (esto para las instituciones con educación intercultural bilingüe —EIB—, en las que se maneja la lengua castellana como un segundo idioma), como de formas inspiradas en un diseño de bloques incompletos, aunque con modificaciones. En el correspondiente *Reporte técnico* (Minedu, 2018), se puede observar que estas modificaciones corresponden a la unión de bloques compartidos, ya sea con formas de un único bloque o con bloques encadenados, de la siguiente manera:

- *Lectura:*

  Evaluada en segundo de primaria mediante una forma única, repartida en dos cuadernillos de 25 ítems. En cuarto se tienen seis cuadernillos, cada uno de los cuales incluye un bloque compartido y uno adicional, que suman 24 ítems. Como los bloques compartidos son dos, cada uno está contenido en tres cuadernillos, resultando en dos grupos, uno para cada sesión. Mientras que los bloques compartidos se aplican de primeras en cada sesión, los otros seis, que no tienen ítems en común, se aplican al final. Cada evaluado recibe dos cuadernillos, un total de 48 ítems. Para cuarto EIB en castellano, se tiene una única forma de dos cuadernillos, conformada por 44 ítems; mientras que se tiene una forma de 42 ítems para cada una de seis lenguas originarias, también presentadas mediante dos cuadernillos. Para segundo de secundaria, se tienen

dos conjuntos de a cinco cuadernillos, uno para cada sesión; cada cuadernillo está compuesto por 25 ítems e incluye un bloque compartido y dos adicionales; mientras que cada bloque compartido está en todas las formas de cada conjunto y se aplica al inicio de cada sesión, los diez restantes se dividen en dos grupos, formándose en cada grupo un diseño de bloques encadenados.

- *Matemáticas:*

  Se evalúa en los mismos grados que lectura, excepto por cuarto grado EIB. Se tienen diseños semejantes para segundo de primaria (23 ítems por cuadernillo) y cuarto de primaria (27 ítems por cuadernillo); el de segundo de secundaria es el mismo empleado en cuarto de primaria, pero con 25 ítems por cuadernillo.

- *Historia, geografía y economía:*

  Estas se evalúan solo en segundo de secundaria e involucran seis formas, de 29 ítems cada una, que corresponden a un diseño de bloques encadenados (seis bloques, dos por forma). Cada evaluado responde los 29 ítems en una única sesión.

Pruebas nacionales (Cuba)

Los Operativos Nacionales de Evaluación de la Calidad de la Educación empezaron a realizarse en Cuba con el objetivo de evaluar la calidad de la educación y el efecto de las políticas educativas, a partir de la participación de este país en el Primer Estudio Internacional Comparativo sobre

Lenguaje, Matemáticas y Factores Asociados, conducido por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) en 1997 (Torres-Fernández, 2008). Para las pruebas de rendimiento cognitivo que involucran estos operativos se emplearon previamente formas paralelas basadas en la teoría clásica de las pruebas, pero, desde el año 2004, se inició el uso de los bloques incompletos balanceados para primaria (Torres-Fernández, 2008).

- *Pruebas del Sistema Nacional de Evaluación de la Educación Básica (SAEB) (Brasil)*

El SAEB es un programa del Instituto Nacional de Estudios e Investigaciones Educativas (INEP), entidad adscrita al Ministerio de la Educación de Brasil, que aplica pruebas cada dos años a los estudiantes de educación fundamental y media, con el propósito de efectuar un acompañamiento longitudinal, alternando los grados e incluyendo las áreas de lengua portuguesa, matemáticas, física, química, biología, historia y geografía (Gajardo, 2002). De acuerdo con Gajardo (2002), en la aplicación del año 2001 se evaluaron las áreas de lengua portuguesa y matemáticas en los grados cuarto y octavo de la educación fundamental y tercero de la media, utilizando 169 ítems por área, agrupados en 26 cuadernillos de 39 preguntas cada uno, conformados a través del diseño de bloques incompletos balanceados. Esto daría cuenta de la utilización de 13 bloques en total, cada uno compuesto por 13 ítems, con tres bloques por forma o cuadernillo, seis repeticiones por bloque y una aparición de cada par.

- *Exámenes de Estado de la Calidad de la Educación - Saber 11, Saber TyT y Saber Pro (Colombia)*

El Instituto Colombiano para la Evaluación de la Educación (Icfes), entidad encargada de la evaluación de la calidad de la educación en Colombia, desarrolla el examen de Estado Saber 11 para diagnosticar el desarrollo de competencias en estudiantes que han completado o están por completar sus estudios de educación media. Este examen, además, se exige como requisito para ingresar a la educación superior (Icfes, 2018b).

De acuerdo con Icfes (2018b), con respecto al año 2017, este examen incluye pruebas en las áreas de matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas, e inglés, brindando resultados individuales y por diferentes niveles de agregación. Las pruebas se aplican en dos sesiones, excepto por lectura crítica e inglés, las cuales se aplican en una sesión, por lo que cada evaluado recibe dos cuadernillos, de 120 y 124 ítems, respectivamente. Para el armado se sigue el diseño de bloques incompletos balanceados, con un total de ocho bloques, cuatro de los cuales se incluyen en cada forma, variando la cantidad de ítems para cada prueba. Esto no es así para inglés, donde el armado se realiza considerando una agrupación de ítems denominada *parte*, que responde al Marco Común Europeo de Referencia para las lenguas.

Esta entidad también desarrolla exámenes para la evaluación de la calidad de la educación superior (Saber TyT y Saber Pro). Cada área corresponde a un módulo, de competencias

genéricas o específicas, dependiendo de si deben ser desarrolladas por todos los estudiantes o si son propias de cada área de formación. Para el armado de los módulos de competencias genéricas se emplea el diseño de bloques incompletos balanceados, pero con un total de cuatro bloques, tres por forma de prueba. Estos módulos no incluyen ciencias naturales, pero sí un área adicional de comunicación escrita, con respecto a Saber 11. Este módulo se aplica a través de una única pregunta a cada evaluado (Icfes, 2018c; Icfes, 2019; Icfes, 2020), con lo que podría decirse que sigue un diseño de un único bloque, en el que cada pregunta constituye un bloque.

REFERENCIAS

Aiken, L.R. (1997). *Questionnaires and Inventories. Surveying Opinions and Assessing Personality*. Estados Unidos: John Wiley & Sons, Inc.

Baker, F.B. y Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques* (2a. Ed.). Estados Unidos: Marcel Dekker, Inc.

Beaton, A.E. (1987). *Implementing the New Design: The NAEP 1983-84 Technical Report*. Disponible en: https://files.eric.ed.gov/fulltext/ED288887.pdf

Brown, F.G. (1983). *Principles of Educational and Psychological Testing*. Holt, Rinehart and Winston.

Centro Nacional sobre las Estadísticas de la Educación (NCES). (*s.f.a*). *About NAEP. A Common Measure of Student Achievement*. Disponible en: https://nces.ed.gov/nationsreportcard/about/

Centro Nacional sobre las Estadísticas de la Educación (NCES). (*s.f.b*). *NAEP Technical Documentation. Bundling of the Student Booklets*. Disponible en: https://nces.ed.gov/nationsreportcard/tdw/instruments/cog_bundle.aspx

Clatworthy, W.H. (1955). Partially balanced incomplete block designs with two associate classes and two treatments per block. *Journal of Research of the National Bureau of Standards*, 54(4), 177-190.

De Mendiburu, F. (2019). *Package 'agricolae'*. Disponible en: https://cran.r-project.org/web/packages/agricolae/agricolae.pdf

Dorans, N.J., Moses, T.P. y Eignor, D.R. (2010, diciembre). *Principles and Practices of Test Score Equating*. Educational Testing Service. Disponible en: http://www.ets.org/research/contact.html

Educational Testing Service (ETS). (2014). *A Guide to Understanding the Literacy Assessment of the STEP Skills Measurement Survey*. Disponible en: https://microdata.worldbank.org/index.php/citations/9683

Fernández-Alonso, R. y Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas. *Aula Abierta*, 39(2), 3-34.

Gajardo, M. (2002). Sistemas probados: evaluación de los aprendizajes en Chile y en Brasil. *Formas y Reformas de la Educación. Serie Mejores Prácticas, 11*. Chile: Preal.

Herrera-Rojas, A.N. (1996). *Algunas Consideraciones Técnicas sobre la Construcción de Ítems de Pruebas Objetivas según la Clasificación de Objetivos Educativos de Bloom*.

Hinkelmann, K. y Kempthorne, O. (2005). *Design and Analysis of Experiments. Volume 2. Advanced Experimental Design*. Estados Unidos: John Wiley & Sons, Inc.

Hinkelmann, K. y Kempthorne, O. (2008, 2a. Ed.). *Design and Analysis of Experiments. Volume 1. Introduction to Experimental Design*. Estados Unidos: John Wiley & Sons, Inc.

Instituto Colombiano para la Evaluación de la Educación (Icfes). (2018a). *Guía Introductoria al Diseño Centrado en Evidencias*. Disponible en: https://www.icfes.gov.co/documents/ 20143/516332/Guia+introductoria+al+ dise%C3%B1o+ centrado+en+evidencias+2018.pdf

Instituto Colombiano para la Evaluación de la Educación (Icfes). (2018b). *Guía de Diseño, Producción, Aplicación y Calificación del Examen Saber 11*. Disponible en: https://www.icfes.gov.co/documents/20143/193560/Guia%20de%20diseno%20produccion%20apliccion%20y%20calificacion.pdf

Instituto Colombiano para la Evaluación de la Educación (Icfes). (2018c). ¿Qué diseño de armado se emplea en el Icfes para medir las pruebas Saber? *Saber al Detalle. Edición 02*. Disponible en: https://www.icfes.gov.co/documents/20143/528208/Boletin+2+-+que+diseno+del+armado+se+emplea+en+el+icfes+para+medir+las+pruebas+saber.pdf

Instituto Colombiano para la Evaluación de la Educación (Icfes). (2019). *Guía de Orientación Saber Pro 2019. Módulos de Competencias Genéricas*. Disponible en: https://www.icfes.gov.co/documents/20143/1518930/Guia+de+orientacion+modulos+de+competencias+genericas+saber+pro+2019.pdf

Instituto Colombiano para la Evaluación de la Educación (Icfes). (2020). *Guía de Orientación Saber TyT 2020-I. Módulos de Competencias Genéricas*. Disponible en: https://www.icfes.gov.co/documents/20143/1708309/Guia+de+orientacion+modulos+de+competencias+genericas+ Saber+TyT+2020-1.pdf

John, J.A. y Williams, E.R. (1995). *Cyclic and Computer Generated Designs* (2a. Ed.). Gran Bretaña: Chapman & Hall.

Kabe, D.G. y Gupta, A.K. (2007). *Experimental Designs: Exercises and Solutions*. Estados Unidos: Springer.

Kirk, R.E. (1995). *Experimental Design. Procedures for the Behavioral Sciences* (3ra. Ed.). Pacific Grove, CA: Brooks/Cole.

Kuehl, R.O. (2001). *Diseño de Experimentos. Principios Estadísticos de Diseño y Análisis de Investigación* (2a. Ed.). México: Thomson Learning, Inc.

Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE). (2015). *Informe de Resultados Tercer Estudio Regional Comparativo y Explicativo. Cuadernillo No. 2. Logros de Aprendizaje*. Disponible en: http://umc.minedu.gob.pe/wp-content/uploads/2015/09/TERCE-Cuadernillo2-Logros-aprendizaje-WEB.pdf

Lahoz-Beltrá, C., Ortega-Escobar, J. y Fernández-Montraveta, C. (1994). *Métodos Estadísticos en Biología del Comportamiento*. España: Editorial Complutense.

Martin, M.O. y Kelly, D.L. (1996). *Third International Mathematics and Science Study. Technical Report. Volume I: Design and Development*. Estados Unidos: International Association for the Evaluation of Educational Achievement (IEA). Disponible en: https://timss.bc.edu/timss1995i/TIMSSPDF/TRall.pdf

McGuigan, F.J. (1996). *Psicología Experimental. Métodos de Investigación* (6a. Ed.). Prentice Hall.

Ministerio de Educación del Perú (Minedu). (2016). *Marco de Fundamentación de las Pruebas de la Evaluación Censal de Estudiantes*. Perú: Minedu.

Ministerio de Educación del Perú (Minedu). (2018). *Reporte Técnico de la Evaluación Censal de Estudiantes (ECE 2016). 2° Grado y 4° Grado de Primaria (ERB y EIB), 2° Grado de Secundaria*. Perú: Minedu.

Montgomery, D.C. (2012). *Design and Analysis of Experiments* (8a. Ed.). Estados Unidos: John Wiley & Sons, Inc.

Mullis, I.V.S. y Martin, M.O. (Eds.). (2015). *PIRLS 2016. Marco de la Evaluación* (2ª. Ed.). IEA. Disponible en: https://books.google.com.co/books?id=77tPDQAAQBAJ&printsec=frontcover&hl=es&source=gbs_ge_summary_ r&cad=0#v=onepage&q&f=false

Nunnally, J. (2000). *Teoría Psicométrica*. México: Trillas.

Nusche, D., Braun, H., Halász, G. y Santiago, P. (2014). OECD *Reviews of Evaluation and Assessment in Education: The Netherlands*. OECD. Disponible en: http://www.oecd.org/education/school/OECD-Evaluation-Assessment-Review-Netherlands.pdf

Oficina Regional de Educación para América Latina y el Caribe (UNESCO-OREALC). (2016). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE*. Santiago, Chile. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000247123

Olive, D.J. (2017). *Linear Regression*. Springer International Publishing.

Olson, J.F., Martin, M.O. y Mullis, I.V.S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Estados Unidos. Disponible en: https://timssandpirls.bc.edu/TIMSS2007/PDF/TIMSS2007_TechnicalReport.pdf

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2002). PISA 2000. *Technical Report*. Disponible en: https://www.oecd.org/pisa/data/33688233.pdf

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2012). *PISA 2012. Technical Report*. Disponible en: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2017). *PISA 2015. Technical Report*. Disponible en: https://www.oecd.org/pisa/data/2015-technical-report/

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (s.f.). PISA 2018. *Technical Report*. Disponible en: https://www.oecd.org/pisa/data/ pisa2018technicalreport/

Robinson, M.A. (2016). Quantitative research principles and methods for human-focused research in engineering design. En: Cash, P., Stanković, T. y Štorga, M. (Eds.). *Experimental Design Research. Approaches, Perspectives, Applications*. Suiza: Springer.

Salkind, N.J. (1999). *Métodos de Investigación*. Pearson Educación.

Torres-Fernández, P. (2008, octubre). El complejo camino de la evaluación. La experiencia cubana. *Aula Urbana*, 69, 6-7.

Triola, M.F. (2004). *Probabilidad y Estadística* (9a. Ed.). México: Pearson Educación.

Van der Linden, W.J., Veldkamp, B.P. y Carlson, J.E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, 28(5), 317-331.

Calle 26 N.° 69-76, Torre 2, Piso 15, Edificio Elemento, Bogotá, D. C., Colombia • www.icfes.gov.co
Líneas de atención al usuario: Bogotá Tel.: (57+1) 484-1460 PBX: (57+1) 484-1410 -
Gratuita nacional: 018000-519535