Aula Abierta 2011, Vol. 39, núm. 2, pp. 3-34
ICE. Universidad de Oviedo
ISSN: 0210-2773

Diseño de cuadernillos para la evaluación de las Competencias Básicas

Rubén Fernández-Alonso y José Muñiz*
Consejería de Educación y Ciencia del Principado de Asturias, *Universidad de Oviedo

El diseño adecuado de los cuadernillos que han de responder los escolares cuando se evalúan sus competencias académicas es una cuestión clave en la construcción de las pruebas. Un diseño poco equilibrado puede ocasionar una calibración inadecuada de los parámetros de los ítems y, por tanto, generar un sesgo a la hora de estimar el nivel de competencia de la población escolar. El objetivo de este trabajo es mostrar las posibilidades que los constructores de tests tienen a la hora de realizar el diseño de los cuadernillos de evaluación. Se entiende por diseño de cuadernillos la organización, arreglo y distribución de la colección de ítems de acuerdo con unas reglas extraídas del diseño experimental. Se revisan los tipos de diseños de cuadernillos que están empleando las administraciones educativas en España para la realización sus evaluaciones de diagnóstico. Se identifican las principales carencias en este ámbito de la construcción de tests y se presenta una colección amplia y representativa de diseños de cuadernillos que pueden utilizarse en la construcción de los instrumentos de evaluación de las competencias básicas.
Palabras clave: Evaluación de diagnóstico, diseños de cuadernillos, diseños de bloques.

Booklet designs for the evaluation of basic skills. A proper design of the booklets that the students have to answer when assessing their academic skills is a key issue in the development of the tests. Inadequate designs can cause improper calibration of item parameters and therefore generate a bias when estimating the competence level of the school population. The main goal of this paper is to show the possibilities that tests developers have when making the design of the assessment booklets. The design of booklets is defined as the organization, arrangement and distribution of the items according to the rules of the experimental design. We review the kinds of designs of booklets that are using the education authorities in Spain to perform their diagnostic evaluations. Major gaps in this area of test construction are analyzed, and a representative collection of designs of booklets that can be used in the construction of evaluation instruments of basic skills are presented.
Keywords: Diagnostic assessment, booklets designs, block designs.

Las evaluaciones de diagnóstico en España

La Ley Orgánica 2/2006, de 3 de mayo, de Educación (LOE) reconoce en su preámbulo la importancia de la evaluación del sistema educativo como factor de calidad educativa y transparencia del propio sistema. En consonancia con este argumento la LOE introduce como novedad la realización de evaluaciones de diagnóstico de las competencias básicas alcanzadas por el alumnado. Ya dentro de su articulado la LOE distingue dos clases de diagnóstico en función de las finalidades, el alcance y carácter de la evaluación y la responsabilidad de su desarrollo.

En el artículo 144 regula la *Evaluación General de Diagnóstico*. Su finalidad es enjuiciar la situación del sistema educativo español. Dicha evaluación se llevará a cabo

Fecha de recepción: 3-3-2011 • Fecha de aceptación: 13-4-2011
Correspondencia: José Muñiz
Facultad de Psicología
Universidad de Oviedo
Plaza Feijoo, s/n. 33003 Oviedo (Spain)
e-mail: jmuniz@uniovi.es

sobre una muestra representativa de centros y alumnado, tanto a nivel estatal como autonómico, y tendrá un carácter externo para los centros seleccionados. La responsabilidad de liderar la evaluación recae sobre el Instituto de Evaluación del Ministerio de Educación. Las demás administraciones educativas colaborarán en el diagnóstico general del sistema mediante la participación en los órganos rectores y comisiones de asesoramiento técnico del Instituto de Evaluación.

La LOE establece en otros dos artículos (el 21 para educación primaria y el 24 para secundaria obligatoria) la *Evaluación de Diagnóstico*. Su finalidad es que todos los centros evalúen de forma periódica y sistemática las competencias alcanzadas por su alumnado. Se trata pues de una evaluación de alcance censal y de carácter interno y formativo, es decir, será realizada por y para los propios centros educativos. Las Comunidades Autónomas apoyarán a los centros en su diagnóstico, facilitándoles los modelos necesarios para su realización. Por ello, todas las administraciones educativas han creado unidades e instituciones públicas dedicadas a este objetivo. En el anexo I se muestran todas ellas.

Para poner en marcha la Evaluación General de Diagnóstico el Ministerio de Educación encargó a un grupo de expertos la elaboración del marco de la evaluación (Ministerio de Educación, 2009a). Los puntos que estructuran dicho marco responden a las tareas propias de un programa de estas características, desde la definición, finalidades y fundamentos legales, hasta los procedimientos de difusión de resultados y conclusiones. De igual modo, cada Comunidad Autónoma ha comenzado a elaborar su propio marco teórico para fundamentar su Evaluación de Diagnóstico, de tal forma que ya se dispone de un buen número de materiales relevantes: Comunidad de Madrid (2008), Generalitat de Catalunya (2009a), Gobierno de Aragón (2008), Gobierno de Canarias (2009), Gobierno de La Rioja (2009), Govern de les Illes Balears (2009), Gobierno de Navarra (2008), Gobierno del Principado de Asturias (2007a, 2008), Gobierno Vasco (2009), Junta de Andalucía (2008), Junta de Comunidades de Castilla-La Mancha (2009) y Ministerio de Educación (2009b).

La prescripción legal de la evaluación de diagnóstico ha tenido como principal consecuencia que todas las administraciones educativas hayan creado servicios especializados en el tema. El esfuerzo de los mismos en el último lustro permite disponer ahora de un corpus normativo, teórico y técnico sobre todos los aspectos implicados en los procesos de evaluación de sistemas educativos, tan necesarios a la hora de impulsar una estrategia nacional de innovación (Arana, 2010).

Objetivo del trabajo

Este trabajo se centrará en un aspecto muy concreto de toda evaluación de diagnóstico: el diseño de los cuadernos de evaluación, es decir, el modo de organizar y distribuir los ítems dentro de los cuadernillos.

El diseño de los cuadernillos es una tarea vital en la evaluación de las competencias académicas, puesto que un diseño defectuoso puede ocasionar una calibración sesgada de los parámetros de los ítems y, por tanto, una deriva a la hora de estimar el nivel de competencia de la población escolar. El diseño de cuadernillos afecta al corazón mismo de la evaluación de diagnóstico, consistente en estimar las competencias alcanzadas por el alumnado. Un diseño deficiente puede invalidar las inferencias que se realicen sobre el nivel de conocimientos de la población escolar.

Tras revisar los marcos teóricos publicados por las distintas administraciones educativas españolas, se puede observar que el diseño de los cuadernillos es un tema al que se ha prestado poca atención, de hecho en la mayoría de los marcos teóricos consultados las referencias al diseño de los cuadernillos son prácticamente nulas. Los casos que mencionan el tema se limitan a realizar recomendaciones generales, sin entrar a considerar cómo se realiza verdaderamente el diseño de los cuadernillos.

Recientemente Frey, Hartig, y Rupp (2009) después de revisar las revistas especializadas, los informes técnicos de los grandes programas de evaluación y los manuales de referencia llegan a la conclusión de que no existe en la actualidad una teoría sobre el diseño de cuadernos de evaluación, por lo que no se dispone de una documentación técnica que indique a los constructores de tests y pruebas de evaluación cuál es el diseño de cuadernillos óptimo para una evaluación de diagnóstico concreta.

En suma, el diseño de cuadernillos de evaluación es un campo poco explorado pero también una decisión técnica de vital importancia, ya que un diseño defectuoso y poco equilibrado puede llevar a conclusiones sesgadas e imprecisas sobre el nivel de competencia de la población escolar. Pretendemos aportar algo de luz sobre este punto para orientar a los constructores de las pruebas cuando tengan que decidir sobre este aspecto.

El trabajo se ajusta al siguiente esquema. En primer lugar se presentarán los elementos fundamentales en el diseño de los cuadernillos de evaluación. A continuación se revisará el "estado de la cuestión" en las evaluaciones de diagnóstico realizadas por las administraciones educativas españolas. En el tercer punto se mostrará cómo el diseño de los cuadernillos de evaluación sigue las mismas reglas del diseño experimental. Finalmente se mostrará una colección amplia y representativa de los diseños de cuadernillos empleados en la evaluación de sistemas educativos. El trabajo se cierra con unas consideraciones generales sobre la elección de un diseño u otro en función de las ventajas e inconvenientes de cada uno.

Elementos del diseño de cuadernillos

El diseño de cuadernillos es el procedimiento por el cual los ítems de la evaluación se asignan a los cuadernillos con el fin de lograr estimaciones insesgadas de los ítems y de la competencia poblacional. Consideremos, por ejemplo, los programas de evaluación educativa más reputados: *Programme for International Student Assessment* (PISA), *Trends in International Mathematics and Science Study* (TIMSS), *Progress in International Reading Literacy Study* (PIRLS) o *National Assessment of Educational Progress* (NAEP). Todos ellos emplean cientos de ítems, por lo que la arquitectura de sus cuadernillos se complejiza bastante e incluye una serie de elementos que van más allá del simple ítem. Estos elementos que vertebran la organización de los cuadernillos son los siguientes: ítem, unidad de evaluación (*testlet* o también *units*) , cluster de ítems o simplemente *cluster*, cuadernillos de evaluación (*booklet* o *subtest*), y colección de ítems de la evaluación (*item pool*).

El ítem es cada una de las preguntas que conforman la evaluación. El conjunto total de ítems recibe el nombre de colección de ítems de la evaluación (*item pool*). Ahora bien, en las evaluaciones de diagnóstico educativo es muy frecuente que los ítems no aparezcan aislados, ni se redacten de forma independiente. Al contrario, los ítems responden al formato de ítems dependientes del contexto, los cuales están especialmente indicados para evaluar procesos complejos mediante grupos de ítems (Muñiz, Hidalgo, García-Cueto, Martínez, y Moreno, 2005). Este tipo de ítems comienza presentando un estímulo (un texto, un gráfico, una tabla o varios de estos elementos combinados), que describe una situación-problema y que va seguido de una serie de ítems referidos o anidados a dicha situación estimular. Estos ítems dependientes del estímulo o contexto responden fundamentalmente a dos formatos: cerrados de elección múltiple y abiertos, ya sean de respuesta corta o de respuesta construida. En la literatura especializada el conjunto que conforman un estímulo y los ítems dependientes del mismo recibe el nombre de *testlet* (Frey et al., 2009) o *units* (OCDE, 2005). En los marcos teóricos de las administraciones educativas se conocen también como unidades de evaluación (por ejemplo, Ministerio de Educación 2009a, 2009b). El número de ítems de estas unidades de evaluación es bastante variable. Así, por ejemplo en PISA 2006 algunas *units* contienen un único ítem (OCDE, 2005),

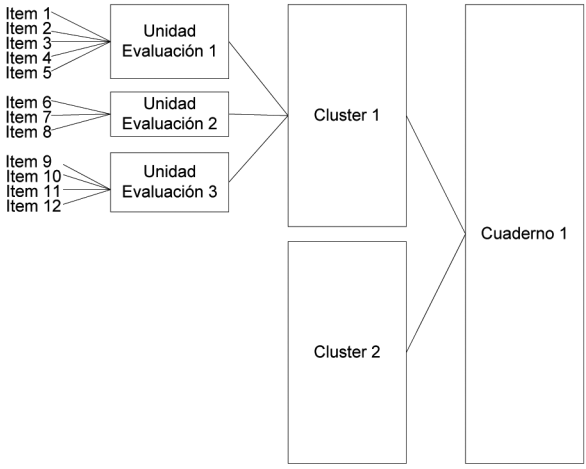

Figura 1. *Elementos estructurales de los cuadernillos de evaluación*

mientras que en PIRLS 2006 cada una de las lecturas propuestas va seguida de unos 12 ítems de promedio (Martin, Mullis, y Kennedy, 2007).

Si se dispone de un número pequeño de unidades de evaluación, éstas pueden tomarse como la base para organizar el diseño de cuadernillos. Sin embargo, cuando la colección de ítems incluye centenares de reactivos el número de unidades de evaluación es, por lo general, bastante elevado. Entonces manejar un gran número de unidades-problema se vuelve una tarea compleja y laboriosa, que compromete la consecución de un diseño limpio y equilibrado. La solución que ofrece el diseño de cuadernillos llegado a este punto es agrupar dos o más unidades de evaluación en *clusters* de ítems (Frey et al., 2009), también denominadas bloques de ítems (*blocks*) siguiendo la nomenclatura empleada en los estudios TIMSS y PIRLS (Olson, Martin, y Mullis, 2008). La Figura 1 muestra los elementos estructurales en el diseño de cuadernillos.

En definitiva, en la arquitectura de un cuadernillo los ítems se agrupan en unidades de evaluación de tamaño variable. Si el número de unidades es elevado éstas se agrupan en clusters o bloques de ítems, que son homogéneos en cuanto a número de ítems o al tiempo necesario para responder a cada cluster. Los cuadernillos de evaluación contienen dos o más clusters de ítems. Por tanto, los elementos sobre los que se asienta el diseño de cuadernillos son los clusters o, en su defecto, las unidades de evaluación. En cualquier caso los ítems individuales nunca deben ser tratados como los referentes para construir un buen diseño de cuadernillos en evaluaciones de la envergadura de un diagnóstico de sistema educativo.

El diseño de cuadernillos en las evaluaciones de diagnóstico en España

Una vez que se ha determinado el número y el tamaño de los bloques o clusters de ítems es necesario elegir un diseño que permita combinar y distribuir los clusters de forma armónica y equilibrada dentro de los cuadernillos. Frey et al. (2009) señalan que la elección del diseño de los cuadernillos está determinada por una serie de condicionantes previos, decisiones que generalmente superan el ámbito técnico pero que limitan el tipo de diseño a emplear. Aquí se identifican cuatro condicionantes: el número de competencias o áreas curriculares que se evaluarán con los cuadernillos; la relación entre el tamaño de la colección de ítems y el

tiempo de evaluación disponible; la necesidad de controlar los efectos que pueden sesgar la estimación de los parámetros de los ítems y de la competencia de la población; y la reutilización de ítems en evaluaciones posteriores.

A continuación se revisa cómo las administraciones educativas dan respuesta a estos condicionantes en sus cuadernillos de evaluación. Llegados a este punto es necesario realizar una advertencia, el formato de aplicación de la prueba también tiene su importancia en la elección del diseño y, por tanto, puede considerarse un condicionante más del mismo. Sin embargo, el presente análisis se centra sólo en los diseños empleados para pruebas de lápiz y papel aplicadas colectivamente a grupos-aula. Es cierto que otros formatos de prueba y otras condiciones de aplicación son posibles. No obstante, la aplicación colectiva de tests clásicos es, por el momento, el formato mayoritario en las evaluaciones de diagnóstico.

El número de competencias o áreas curriculares a evaluar

La LOE prescribe que el diagnóstico educativo se centrará en evaluar el nivel de alumnado en competencias básicas antes que en áreas curriculares. Esto significa que la evaluación de una competencia (por ejemplo, competencia digital, comunicación lingüística o aprender a aprender) incluye contenidos de varias áreas curriculares. En todo caso, sean competencias o disciplinas curriculares, el número de las mismas que sean objeto de evaluación es uno de los factores que más complejidad introduce en el diseño de cuadernillos. En función del número de competencias evaluadas se pueden distinguir dos tipos de diseños. Por un lado aquellos donde con un único juego de cuadernillos se evalúan dos o más competencias y, por otro, los diseños para evaluar una sola competencia. En evaluación de sistemas educativos ambas opciones están bien representadas. Por ejemplo, PIRLS evalúa exclusivamente comprensión lectora. Por su parte, la evaluación de diagnóstico norteamericana, NAEP ha manejado las dos opciones a lo largo de sus cuatro décadas de existencia. Sin embargo, en la actualidad cada área curricular evaluada tiene su propia colección de cuadernillos.

La segunda posibilidad, es decir, evaluar dos o más competencias con un único juego de cuadernillos, también tiene sus adeptos. Es, por ejemplo, la opción implementada en PISA donde su único juego de cuadernillos incluye ítems de comprensión lectora, matemáticas y ciencias. Similar planteamiento se encuentra en TIMSS donde los cuadernillos presentan ítems de matemáticas y ciencias.

En España la tendencia mayoritaria ha sido evaluar entre dos y cuatro competencias en cada estudio diagnóstico (Generalitat de Cataluña, 2010; Gobierno de Aragón, 2010; Gobierno de Canarias, 2009; Gobierno de Cantabria, 2010; Gobierno de La Rioja, 2009; Gobierno de Navarra, 2010; Gobierno del Principado de Asturias, 2007b, 2007c, 2010; Gobierno Vasco, 2010; Govern de les Illes Balears, 2010; Junta de Andalucía, 2007a, 2007b, 2009, 2010; Junta de Castilla y León, 2009; Ministerio de Educación, 2009a, 2009b; Región de Murcia, 2009; Xunta de Galicia, 2010).

Pese a que todas las comunidades autónomas evalúan más de una competencia, hasta el momento la decisión prácticamente unánime es que cada competencia tenga su propio diseño de cuadernillos. Esta elección supone que habrá tantos juegos de cuadernillos como competencias evaluadas. De la anterior afirmación cabe exceptuar la propuesta de Castilla-La Mancha. Se trata de una experiencia interesante que, en algunos aspectos, se aparta de los planteamientos más ortodoxos que siguen el resto de las administraciones educativas al intentar superar la idea de que detrás de cada competencia subyace la evaluación preferente de un área curricular. Sus cuadernillos se llaman unidades de evaluación y contienen ítems de varias competencias. De esta forma el juego de unidades de evaluación permite evaluar todas las competencias sancionadas legalmente en cada estudio diagnóstico, aunque exige seis horas de trabajo individual (Junta de

Comunidades de Castilla-La Mancha, 2009, 2010).

Las dos opciones barajadas hasta el momento tienen sus ventajas e inconvenientes. Por supuesto el diseño será más simple si cada competencia dispone de su propio juego de cuadernillos. Y al contrario: la organización del diseño se vuelve más compleja a medida en que un único juego de cuadernillos incluya más competencias. En este segundo caso el diseño deberá contemplar la cuestión de la proporcionalidad de cada competencia en la evaluación. Por ejemplo, en cada edición de PISA una de las tres competencias es principal y las preguntas referidas a la misma suponen algo más de la mitad del total de ítems. En cambio en TIMSS los ítems de matemáticas y ciencias se reparten paritariamente.

Por el contrario, si un único juego de cuadernillos contiene más de una competencia puede aligerar aspectos relativos a la logística de la evaluación rebajando el número de cuadernillos, guías y plantillas de corrección y simplificando el proceso de grabación de datos. A lo largo de este trabajo se presentarán algunos diseños que permiten evaluar dos o más competencias con un único juego de cuadernillos.

La relación entre el tamaño de la colección de ítems y el tiempo de evaluación disponible

La segunda decisión que condiciona la organización de los cuadernillos es el número total de ítems y su relación con respecto al tiempo de evaluación disponible. Seguramente es el condicionante más influyente a la hora de decidir el diseño de las pruebas. Tal es así que los diseños que se presentarán más adelante están clasificados por este criterio. Atendiendo a esta relación los diseños se pueden agrupar en dos clases. Por un lado, arreglos donde la colección de ítems equivale al tiempo de evaluación disponible. Por otro, diseños en los que responder a la colección de ítems supera el tiempo disponible. Imagínese que para evaluar una competencia se dispone de una sesión ordinaria de 50 minutos. Si el número de ítems fuese pequeño –por ejemplo, entre 25 y 30- los estudiantes podrán responder a todos ellos dentro del límite temporal de la sesión y, por tanto, todo el alumnado responderá a los mismos ítems. En cambio si la colección contuviera, por ejemplo, 150 ítems será imposible que cada estudiante responda a todos ellos en el tiempo asignado para evaluar la competencia. En este caso cada estudiante sólo contestará a una parte de la colección completa –unos 30 ítems aproximadamente- y no tendrá oportunidad de contestar al resto –en torno a 120 en este ejemplo-. Con esta segunda estrategia los estudiantes no responden a los mismos ítems.

La tendencia mayoritaria en las evaluaciones de diagnóstico en España es que la colección de ítems de cada competencia se pueda responder dentro de los límites temporales asignados para evaluar dicha competencia (Comunidad de Madrid, 2008; Gobierno de Aragón, 2010; Gobierno de Canarias, 2009; Gobierno de Navarra, 2008; Gobierno de la Rioja, 2009; Junta de Andalucía, 2008; Junta de Castilla y León, 2009; Región de Murcia, 2009). Es más, a la vista de los marcos teóricos y de las pruebas liberadas, la opción empleada en estos casos es construir un único modelo de cuadernillo por competencia evaluada. El marco de la evaluación del País Vasco explicita la razón de emplear un cuadernillo único: asegurar la comparación individual de los resultados haciendo que todos los estudiantes respondan a los mismos ítems y en las mismas condiciones (Gobierno Vasco, 2008a).

Por su parte, cuatro administraciones educativas plantean el uso de dos o más cuadernillos por competencia evaluada: el Ministerio de Educación, Asturias, Baleares y Galicia. El marco de la evaluación del Ministerio de Educación es el que mejor explicita las características generales de un diseño matricial de cuadernillos de evaluación. En todo caso, tanto para la Evaluación General de Diagnóstico, como para la evaluación de Ceuta y Melilla, el tema queda despachado con un párrafo dentro del capítulo referido al instrumental de evaluación (Mi-

nisterio de Educación, 2009a, 2009b). En ambos documentos se indica que la colección de ítems de cada competencia se dividirá en diferentes cuadernillos, que se aplicarán a distintos grupos de alumnado con el fin de incluir el mayor número de ítems posibles. A continuación menciona las condiciones generales que deben cumplir el diseño de cuadernillos de la evaluación ministerial: los ítems deben distribuirse equilibradamente a lo largo de los cuadernillos, y organizarse respetando cierta combinación que permita contar con ítems de anclaje a partir de los cuales integrar los resultados de los alumnos en la misma escala.

Sin embargo, en su documento-marco el Ministerio de Educación no precisa ni el número de ítems por competencia ni el diseño de cuadernillos a emplear. Además, como tampoco libera pruebas completas, ya que reserva ítems para estudios de tendencia, no es posible reconstruir el diseño matricial empleado por la administración central de educación. A similar conclusión se llega en el caso de Galicia: inicialmente se anuncia el uso de dos modelos de cuadernillo por competencia (Xunta de Galicia, 2009). Sin embargo en la documentación publicada por la administración gallega no se ha encontrado mención alguna sobre el tipo de diseño elegido ni sobre la distribución de los ítems dentro de los cuadernillos.

Los materiales liberados por Asturias y Baleares han permitido recrear los diseños matriciales que subyacen a la ordenación de los ítems en los cuadernillos. En ambos casos la opción empleada con preferencia ha sido el diseño de anclaje fijo de ítems, es decir, en todos los cuadernillos de evaluación hay un grupo de ítems comunes que se presentan en la misma posición. Este diseño tiene dos virtudes: incluye un buen número de ítems por competencia para cubrir adecuadamente las especificaciones de contenido de la evaluación, y asegura buenas condiciones de comparación de resultados (Gobierno del Principado de Asturias, 2007a; Govern de les Illes Balears, 2009, 2010).

Como ejemplo de este tipo de anclaje se muestra el diseño implementado en Asturias para evaluar la comprensión lectora en el año 2009. En dicha evaluación se elaboraron 7 unidades de lectura de idéntica estructura: un texto seguido de 6 ítems. La colección completa lecturas y preguntas aparejadas equivalían a 105 minutos de evaluación. Sin embargo, el diseño adoptado permitió que cada estudiante sólo invirtiera un máximo de 30 minutos. Como se observa en la Tabla 1, todos los cuadernillos comenzaban con la misma unidad de lectura, que funcionaba como el test de anclaje del diseño. Al colocar el test de anclaje al inicio de todos los cuadernillos se buscaba que las condiciones de aplicación del mismo fuesen lo más homogéneas e idóneas posibles. A continuación del test anclaje se ubicaba una segunda unidad de lectura, que era diferente en cada cuadernillo.

Cualquiera de las dos opciones planteadas tiene ventajas e inconvenientes (Childs

Tabla 1. *Diseño de cuadernillos en la evaluación de Diagnóstico de Asturias. 2009*

Tipo de Unidad	**Nombre de la Unidad**	**Número del Cuadernillo**					
		1	**2**	**3**	**4**	**5**	**6**
Unidad Común (Test de Anclaje)	Lectura – C	1	1	1	1	1	1
Unidad Propia de Comprensión Lectora	Lectura – 1	2					
	Lectura – 2		2				
	Lectura – 3			2			
	Lectura – 4				2		
	Lectura – 5					2	
	Lectura – 6						2

1 Posición de la Unidad en el Cuaderno.
Fuente: elaboración propia.

y Jaciw, 2003; Frey et al., 2009). Si la colección de ítems es pequeña (equivalente al tiempo de evaluación disponible) la planificación del diseño de cuadernillos es relativamente sencilla y es posible encontrar una estructura equilibrada y simétrica con un número pequeño de cuadernos. En cambio, si el tiempo para responder a la colección de ítems supera el límite temporal disponible el diseño se vuelve más complejo. En este caso sólo caben dos posibilidades: aumentar el número de cuadernos o establecer diseños más complejos que permitan manejar un número relativamente pequeño de cuadernos sin que la estimación de los parámetros de los ítems se vea afectada.

Ya se apuntó que, en algunos casos, se considera que evaluar a todo el alumnado con un único modelo de cuadernillo es la opción más equitativa para comparar rendimientos individuales. Sin embargo, es una estrategia con dos riesgos claros. El primero está relacionado con la sustantividad o representatividad del contenido de la propia evaluación. Trabajar con un pequeño número de ítems puede afectar a la validez de contenido de la prueba (Muñiz, 2001). En segundo lugar el empleo un cuadernillo único e idéntico para todo el alumnado puede sesgar la estimación de los parámetros de los ítems y, por ende, el nivel de competencia de la población. En el siguiente apartado se tratará esta idea con más detenimiento.

Por su parte, el empleo de una colección de ítems amplia ofrece más garantías con respecto a la validez de contenido de la prueba y estima con mayor precisión la competencia de la población en la competencia evaluada. En cambio, como los estudiantes no responden a los mismos ítems se corren mayores riesgos a la hora de establecer comparaciones individuales. No obstante es necesario realizar dos matizaciones a la anterior afirmación. La primera es que, por sus características, no está nada claro que una evaluación de diagnóstico educativo pueda ofrecer estimaciones fiables a nivel individual. La segunda cuestión es que, partiendo de un diseño de cuadernillos adecuado, los desarrollos psicométricos actuales permiten construir una única escala de puntuaciones para todo el alumnado, aunque éste no responda exactamente a las mismas preguntas. Para ambas cuestiones véase, por ejemplo, los informes técnicos de PISA (OCDE, 2002, 2005).

Control de los sesgos en la estimación de los parámetros de los ítems

En el apartado anterior se advirtió que el diseño de cuadernillo único puede sesgar la estimación de los parámetros de los ítems y de la competencia de la población. Estos potenciales sesgos ocurren porque una evaluación basada en único modelo de cuadernillo no garantiza el control de los efectos de posición y arrastre de los ítems. Ambos efectos tienen la misma naturaleza ya que están asociados a la ubicación de los ítems en los cuadernillos. Cabalmente se espera que la dificultad de un ítem esté determinada exclusivamente por el tipo de tarea o proceso cognitivo implicado en su solución. Sin embargo, se sabe que esta dificultad puede depender, al menos en parte, del lugar que el ítem ocupa en el cuadernillo. Por ello, un buen diseño de cuadernillos necesita neutralizar estos efectos.

El parámetro de un ítem puede sesgarse por el orden de presentación del reactivo, este hecho se conoce como efecto de posición. La OCDE (2005) estima que un mismo ítem es, de promedio, un 10% más fácil si aparece al inicio del test que si está ubicado al final. En el diseño de cuadernillos el efecto de posición puede controlarse haciendo que todos los ítems aparezcan en todas los órdenes posibles y complementariamente disponiendo de un modelo matemático que tenga en cuenta la posición de los ítems dentro del cuadernillo.

Por su parte, el efecto de arrastre advierte que el índice de dificultad de un ítem puede quedar contaminado por los ítems presentados previamente en el cuadernillo. Por ejemplo, si al final de un cuaderno aparece un conjunto de ítems de geometría estos pueden ser más fáciles si previamente el alumnado ha respondido a otros ítems similares. De igual modo, el efecto de arrastre puede aparecer si

la respuesta un ítem depende información contenida en otros ítems del cuadernillo o de la respuesta dada en ítems previos. El efecto de arrastre supone un problema grave para los modelos psicométricos de Teoría de Respuesta a los Ítems (TRI), ya que viola el principio de independencia local, esto es, que la probabilidad de acertar un ítem cualquiera es independiente de la probabilidad de acierto en el resto de los ítems del cuadernillo (Muñiz, 1997). Al igual que ocurre con el efecto de posición, el modo de controlar el efecto de arrastre en la estimación de los parámetros de los ítems pasa por una estrategia doble: organizar los clusters de ítems de forma balanceada dentro de los cuadernos e incluir el efecto de arrastre de los ítems dentro del modelo de respuesta al ítem que se emplee en la estimación de los parámetros.

Evidentemente cuando la evaluación emplea un único modelo de cuadernillo es imposible controlar los efectos de posición y arrastre, lo que introduce un potencial de sesgo importante. Dentro de las administraciones que emplean más de un cuadernillo de evaluación por competencia hay algunas evidencias de que dichos efectos sí se tienen en cuenta. El Ministerio de Educación lo ha formulado teóricamente en su marco de evaluación, indicando que dentro de los cuadernillos, las unidades de evaluación deben aparecer en diferentes posiciones para evaluar la relación entre la posición del ítem en cada cuadernillo y su dificultad (Ministerio de Educación, 2009a, 2009b).

En los diseños de ítems de anclaje fijo, como el mostrado en la Tabla 1, las posibilidades de controlar los efectos de posición y arrastre son también limitadas. La restricción de que el test de anclaje aparezca en la misma posición en todos los cuadernillos para asegurar iguales condiciones de administración a todo el alumnado cercena esta posibilidad.

Baleares ha empleado un diseño donde el test de anclaje ocupa distintas posiciones lo que permite balancear el efecto de la dificultad de los ítems en función de su ubicación dentro de los cuadernillos. En la Tabla 2 se muestra la organización de los ítems en los cuadernillos de Competencia Matemática de 4º curso de Educación Primaria del año 2009. Esta prueba constaba de 51 ítems pertenecientes a 10 unidades-problemas, las cuales a su vez se distribuyeron en 3 cuadernillos. Había 2 unidades (*London Eye* y *Bitlles*, 14 ítems en total) que se repiten en los 3 cuadernos ocupando distintas posiciones: por ejemplo, la unidad *London Eye* se ubica en la primera posición en el cuaderno 1, en la tercera en el 2 y en el cuarta en el 3.

En definitiva, la revisión de lo publicado hasta el momento por las administraciones educativas indica que sus diseños de

Tabla 2. *Diseño de cuadernillos en la evaluación de Diagnóstico de Islas Baleares. 2008-2009*

Tipo de Unidad	**Nombre de la Unidad**	**Número de ítems en la unidad**	**Número del Cuadernillo**		
			1	**2**	**3**
Anclaje	London Eye	6	1	3	4
	Bitlles	8	2	4	5
Unidad propia de cada cuadernillo	Genograma	7	3		
	Plànol	6	4		
	Excursió	4		1	
	Monuments	3		2	
	IMC	5		5	
	Ferretería	4			1
	Malalties	3			2
	Camp	5			3

1 Posición de la Unidad en el Cuaderno.

Fuente: elaboración propia a partir de los datos contenidos en http://iaqse.caib.es/aval_1.htm.

cuadernillos presentan lagunas importantes en cuanto al control de los sesgos que afectan a la estimación de la competencia poblacional. El control del efecto de posición de los ítems está planteado teóricamente en el marco teórico del Ministerio de Educación y resuelto sólo parcialmente en algunos diseños como el balear. Por su parte, ninguno de los diseños presentados asegura el control del efecto de arrastre de los ítems.

Reutilización de los ítems en evaluaciones posteriores

El último condicionante para establecer el diseño de cuadernillos de evaluación es la posibilidad de reutilizar ítems en evaluaciones posteriores. Se trata de una práctica habitual en los grandes programas de evaluación educativa, tales como PISA, TIMSS, PIRLS o NAEP. El Ministerio de Educación ha realizado comparaciones longitudinales antes incluso de la Evaluación General de Diagnóstico (Ministerio de Educación, Cultura y Deporte, 2001, 2003; Ministerio de Educación y Ciencia, 2005; Ministerio de Educación, 2009c). Por su parte, el País Vasco recoge en su marco teórico la previsión de realizar comparaciones a lo largo del tiempo (Gobierno Vasco, 2008a). Una lectura conjunta de estos documentos permite establecer las finalidades y características generales de un diseño de anclaje longitudinal y los criterios para la selección de los ítems que formarán parte de dicho anclaje.

La finalidad de reutilizar ítems a lo largo del tiempo es valorar evolución dinámica de los resultados tanto del sistema en general como de los centros en particular. De esta forma la comparación longitudinal permitirá comprobar el grado de eficacia de las medidas que vayan tomando y analizar buenas prácticas.

Para formar parte del anclaje longitudinal se seleccionarán ítems que hayan demostrado su fiabilidad, estabilidad y alta significatividad a la hora de evaluar un aspecto concreto de una competencia.

Las características generales de este diseño deben ser las siguientes: los ítems comunes a dos o más evaluaciones deben ser escrupulosamente iguales en su enunciado y alternativas de respuesta; deben ubicarse en las mismas posiciones en distintos modelos de cuadernillos; y estos cuadernillos deben tener una longitud similar.

Por su parte, algunas administraciones educativas han publicado comparaciones de resultados entre diferentes ediciones de sus evaluaciones de diagnóstico (Generalitat de Catalunya, 2010; Gobierno de La Rioja, 2010a, 2010b; Junta de Andalucía, 2010). Sin embargo, en su documentación publicada no se han encontrado referencias a cómo el diseño de cuadernillos prevé la distribución de los anclajes, ni tampoco a los métodos de equiparación de puntuaciones empleados.

A la hora de establecer un buen diseño de anclaje hay dos cuestiones que deben ser tenidas en cuenta. La primera es que el anclaje longitudinal, como cualquier diseño de cuadernillos, también puede verse afectado por los sesgos de posición y arrastre. Por tanto, las reservas ya apuntadas con respecto a los diseños empleados en las evaluaciones educativas son totalmente aplicables a las comparaciones longitudinales. Es necesario disponer de una organización de cuadernillos que mantenga intacta la ubicación de las unidades o clusters a lo largo de serie temporal de evaluaciones.

La segunda consideración cuando se reutilizan ítems a lo largo del tiempo son las cuestiones relativas a la seguridad de los mismos. Aún blindando las pruebas cabe la posibilidad de que los ítems puedan ser memorizados por personas que puedan entrenar a futuros participantes en las evaluaciones. Si esto ocurre los parámetros de los ítems se verán alterados y la competencia poblacional sobreestimada. La probabilidad de memorizar las preguntas aumenta cuando se trabaja con una colección pequeña de ítems, que es presentada en un único cuadernillo. Al contrario, este riesgo se reducirá si el diseño incluye un conjunto amplio de ítems presentados en varios modelos de cuadernillos que contengan distintos clusters en diferentes posiciones.

En resumen, las características generales de los diseños de cuadernillos en el con-

junto de las evaluaciones de diagnóstico en España son las siguientes:

- Si bien las evaluaciones las administraciones educativas evalúan como mínimo dos competencias, cada una de éstas tiene su propio diseño de cuadernillo.
- Existe un amplio número de administraciones educativas que emplean un único modelo de cuadernillo para evaluar cada competencia. En los casos donde hay más de un cuadernillo por competencia se usa con preferencia el diseño de anclaje de ítems fijo: todos los cuadernillos contienen unas preguntas comunes y otras preguntas específicas de cada cuadernillo.
- Los marcos teóricos de las administraciones educativas apenas hacen referencia al control de los efectos de posición y arrastre de los ítems. Sólo el Ministerio de Educación menciona este hecho explícitamente, mientras que los diseños hechos públicos por Asturias y Baleares hay cierta consideración sobre el orden de las unidades de evaluación dentro de los cuadernillos para homogenizar las condiciones de aplicación.
- Se han comenzado a publicar resultados que comparan la competencia poblacional a lo largo del tiempo. Sin embargo, las referencias a los diseños de anclaje longitudinal y a los métodos de equiparación de series temporales son prácticamente nulas. También es cierto que, debido a la novedad que suponen las evaluaciones de diagnóstico en nuestro país, quizá sea un poco prematuro valorar la adecuación de los diseños de cuadernillos para establecer tendencias temporales.

En suma, la revisión de los marcos de evaluación indica que las referencias a los diseños de cuadernillos son bastante escasas y, cuando aparecen se quedan en comentarios y recomendaciones generales no disponiendo, salvo en contadas ocasiones, de un diseño de cuadernillos que cumpla las condiciones necesarias para ser considerado como tal.

El diseño experimental como base para establecer el diseño de cuadernillos de evaluación

Pese a la relativa novedad que suponen los diseños de cuadernillos en las evaluaciones de diagnóstico en España, lo cierto es que los diseños matriciales de ítems aplicados a la evaluación general de los sistemas educativos acumulan décadas de experiencia. Las primeras referencias se deben a Lord (1955, 1962), preocupado por las limitaciones en la equiparación de puntuaciones en los tests clásicos y por estimar la proporción de la población que podría responder correctamente a un ítem, conocidas las respuestas que al mismo había dado una muestra de estudiantes. Lord (1962) acuña el término *matrix sampling designs*, como procedimiento para organizar los ítems en los cuadernillos de evaluación. De esta forma el muestro matricial de ítems surge para dar respuesta a un problema capital en la evaluación de rendimientos académicos: la validez de contenido de la prueba. Para evaluar el nivel de una población en una o más áreas curriculares es necesario disponer de una colección de centenares de ítems que cubran adecuadamente las especificaciones de contenido de la evaluación. Ahora bien, en la práctica, es inviable que una muestra representativa de estudiantes responda a una colección de ítems de tal magnitud. Por tanto, la solución más eficiente es dividir el conjunto total de ítems en partes más pequeñas y que cada estudiante responda sólo a una porción de dicha colección. Operando así las respuestas de las submuestras de estudiantes a submuestras de ítems (subtests) permiten inferir el nivel de la población escolar con respecto al área o material evaluada.

Ahora bien, ¿qué reglas o pautas deben observarse para distribuir la colección completa de ítems a los subtests o cuadernillos de

evaluación? La solución pasa por entender el diseño matricial de ítems como un caso especial del diseño experimental. En otras palabras, haciendo que el diseño y distribución de los ítems a lo largo de la colección de cuadernillos siga las mismas pautas que guían el diseño experimental. En lo que resta de apartado se verá cómo es posible adoptar o "traducir" los conceptos claves que se emplean en diseño experimental (tratamiento, bloqueo, replicación, tamaño del bloque y concurrencia de tratamientos) a los términos del diseño matricial de los cuadernillos de evaluación.

Un *diseño experimental* es un plan mediante el cual una serie de tratamientos son asignados a un conjunto de unidades experimentales. De igual modo, el diseño de cuadernillos es el plan mediante el cual los ítems se asignan a los estudiantes. Los *tratamientos* (t) son las variables experimentales que se someten a prueba. En la evaluación de competencias académicas cada ítem o cluster de ítems puede verse como un "mini-experimento" para evaluar dicha competencia. Por tanto, en el diseño de cuadernillos habrá tantos tratamientos como clusters de ítems disponibles.

En el plan experimental es muy habitual que los tratamientos se organicen en *bloques* (b) con el fin de neutralizar la varianza no deseada. El bloqueo consiste en dividir las unidades experimentales en grupos homogéneos y aplicar aleatoriamente los tratamientos a estos grupos. En diseño de cuadernillos esto supone que la colección de clusters de ítems se organiza y/o divide en varios cuadernillos (subtests), los cuales se administran aleatoriamente a los estudiantes. Con ello se neutralizan fuentes de variación indeseables y se incorpora un factor bloque al diseño.

La metodología experimental distingue dos diseños básicos en función del *tamaño de los bloques* (k): bloques completos e incompletos. Se dice que un bloque es completo si dentro del mismo se incluyen todos los tratamientos disponibles. Por tanto, en los diseños de bloques completos se cumple la siguiente igualdad ($t = k$). En evaluación de competencias curriculares esto quiere decir que cada cuadernillo (b) contiene todos los clusters de ítems disponibles (t). Ahora bien, ocurre con mucha frecuencia que el número de clusters de ítems disponibles supera ampliamente el número de ítems que un estudiante puede responder dentro de los límites temporales de la administración del test. Por tanto, los cuadernillos sólo contienen una parte del total de los ítems disponibles, es decir, el bloque se compone sólo de una parte de los tratamientos disponibles. Cuando $t > k$ se habla de bloques incompletos, ya que cada bloque no incluye todos los tratamientos disponibles.

En un diseño experimental un tratamiento puede aparecer una o varias veces. La frecuencia con la que un tratamiento aparece a lo largo del diseño experimental se denomina *replica o replicación* (r). Análogamente, el número de veces que cada cluster de ítems aparece a lo largo de la colección de cuadernillos también se denomina replicación.

Finalmente a lo largo de sus replicaciones cada tratamiento se empareja con el resto de tratamientos un determinado número de veces y siguiendo ciertas pautas. Es lo que se denomina *concurrencia de cada par de tratamientos* (λ). En diseño de cuadernillos λ indica el número de veces que dos clusters de ítems aparecen en el mismo cuadernillo. Si el valor de λ es el mismo para cualquier par de tratamientos se dice que el diseño es balanceado. Si por el contrario λ presenta un valor para ciertos pares de tratamientos distinto del valor para el resto de pares de tratamientos se dice que el diseño no es balanceado o es parcialmente balanceado.

Las variables t, b, r, k y λ se denominan parámetros del diseño y para que un diseño tenga cierto equilibrio es necesario que estos parámetros guarden unas determinadas proporciones. Por ejemplo, en todo diseño balanceado deben cumplirse las siguientes relaciones entre los parámetros:

1. $bk = tr$
2. $r(k - 1) = \lambda(t - 1)$
3. De la segunda condición se deduce que: $\lambda = [r(k - 1)] / (t - 1)$, siendo λ un entero positivo.

Las anteriores condiciones indican que el número de diseños balanceados es limitado ya que no cualquier combinación de los parámetros t, b, r, k cumple con las dos primeras igualdades al tiempo que ofrece un valor λ entero. Además, en ocasiones, incluso cuando se encuentra la combinación de parámetros adecuada tampoco es fácil asignar los tratamientos (clusters de ítems) y sus replicaciones dentro de los bloques (cuadernillos) de k tamaño y λ concurrencias por cada par de tratamientos. Por ello, el diseño de cuadernillos además del diseño experimental debe apoyarse en la combinatoria como disciplina que se dedica a la selección, disposición y combinación de una serie de objetos dentro un espacio finito.

En todo caso, cabe recordar a los constructores de tests que los principales diseños experimentales están bien documentados y estudiados en la literatura (Arnau, 1984; Ato y Vallejo, 2007; Box, Hunter, y Hunter, 2005; Cochran y Cox, 1974; Cook y Campbell, 1979; Fisher y Yates, 1963; Kirk, 1995; Vallejo et al., 2010), donde se pueden consultar los diseños que se mostrarán a continuación y muchos más.

Diseños cuando se dispone de pocos ítems: bloques completos

Como se apuntó anteriormente hay un grupo significativo de administraciones educativas que emplean un único modelo de cuadernillo por competencia, de tal modo que todo el alumnado responde a los mismos ítems y exactamente en el mismo orden. Esta estrategia es, en apariencia, más equitativa para comparar desempeños individuales, pero tiene serios riesgos debido a su nulo control del efecto de posición de los ítems, lo que puede invalidar las inferencias de los resultados a la población. Por tanto, las evaluaciones donde los estudiantes responden a los mismos ítems debieran emplear un diseño de cuadernillos que neutralizara los sesgos en la estimación de la competencia poblacional. A continuación se presentan dos diseños especialmente recomendados para aplicaciones en las que todos los estudiantes responden a los mismos ítems.

Diseño de Tratamientos Repetidos

Por sus características el Diseño de Tratamientos Repetidos (DTR) puede ser especialmente recomendado para evaluar competencias como matemáticas o ciencias. A la vista de las pruebas liberadas por algunas administraciones educativas los cuadernillos de evaluación de estas competencias suelen contener entre 20 y 30 ítems, repartidos entre 4 y 6 unidades y se responden en una aplicación que dura entre 50 y 60 minutos. Para organizar un DTR el número de cuadernillos, unidades de evaluación y las posiciones de dichas unidades dentro de los cuadernillos debe ser el mismo, es decir, son diseños que cumplen la siguiente igualdad: t = b = k = r.

La Tabla 3 muestra un ejemplo de DTR pensado para una sesión de evaluación de 50

Tabla 3. *Ejemplo de Diseño de Tratamientos Repetidos*

Nombre del cluster	**Número de cuadernillo**					
	1	**2**	**3**	**4**	**5**	**6**
A	1	2	3	4	5	6
B	6	1	2	3	4	5
C	2	3	4	5	6	1
D	5	6	1	2	3	4
E	3	4	5	6	1	2
F	4	5	6	1	2	3

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 6$; $b = 6$; $k = 6$; $r = 6$; $\lambda = 6$

minutos. La colección completa de ítems está dividida en 6 clusters de ítems de 8 minutos de duración cada uno. En cada uno de los cuadernillos los clusters ocupan una posición diferente. Como se puede ver el diseño presentado es un cuadrado latino 6 x 6 (seis clusters de ítems distribuidos en seis cuadernillos). Sin embargo, la ordenación de los clusters dentro de los cuadernillos está pensada para controlar los efectos de posición y arrastre.

El efecto de posición se controla, como en el cuadrado latino ordinario, gracias a que todos los clusters aparecen en todas las posiciones posibles en los cuadernillos. Así, leyendo la tabla desde las filas, se aprecia que el cluster A es el primero en el cuaderno 1, el segundo en el cuaderno 2, y así sucesivamente.

Además, leyendo esta tabla por columnas, se observa que este diseño controla también los efectos de arrastre de orden superior. En otras palabras: en todos los cuadernillos el cluster A aparece antes que el cluster C; el cluster C antecede al cluster E; el cluster E precede al cluster F; el F al D; y el D al B. Esta regla sólo se incumple cuando el cluster antecedente de la pareja ocupa la última posición del cuadernillo. Entonces el segundo cluster se ubica en la primera posición del cuadernillo. Así en el cuadernillo 6 el cluster A (antecedente del cluster C en el resto de los cuadernillos) aparece en última posición, mientras que los ítems del cluster C son los primeros de dicho cuadernillo.

En definitiva, el DTR tiene dos ventajas que lo hacen muy recomendable, es de fácil construcción y su análisis estadístico es sencillo. Además, presenta el doble control de filas y columnas propio del cuadrado latino, con lo que es bastante efectivo para neutralizar los posibles sesgos en la estimación de la competencia poblacional, cuestión que aún no está bien resuelta en los diseños de cuadernillos de las administraciones educativas.

Diseño de Permutación Completa

Los diseños de permutación completa (DPC) comparten los mismos beneficios que los DTR: son fáciles de construir y, sobre todo, son muy eficaces para controlar fuentes de variación indeseada vinculadas a los efectos de posición y arrastre de los ítems. En los DPC el orden de los clusters se permuta en cada cuadernillo, lo que supone que todos los clusters de ítems aparecen exactamente una vez en cada cuadernillo.

Un modelo muy simple se presenta en la Tabla 4 y podría ser empleado perfectamente en una prueba de comprensión lectora compuesta por tres lecturas y sus correspondientes ítems. El diseño muestra una colección completa de ítems dividida en 3 clusters o unidades de evaluación que ocupan 20 minutos cada una. Por tanto, de nuevo en este diseño la colección completa de ítems equivale al tiempo que un estudiante puede invertir en responder a la prueba.

En principio, como el número total de clusters (t) y el número de clusters a incluir en cada cuadernillo (k) coinciden (es decir, $t = k$), sólo hay una única combinación posible cuando se toman tres clusters de tres posibles (A-B-C). Sin embargo, hay seis posibles modos de ordenar los clusters de esta única selección. Todos ellos se muestran en la Tabla 4. Cada cluster aparece dos veces

Tabla 4. *Ejemplo de Diseño de Permutación Completa (Bloques Completos)*

Nombre del cluster	**Número de cuadernillo**					
	1	**2**	**3**	**4**	**5**	**6**
A	1	1	2	2	3	3
B	2	3	3	1	1	2
C	3	2	1	3	2	1

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 3$; $b = 6$; $k = 3$; $r = 6$; $\lambda = 6$

Tabla 5. *Ejemplo de Diseño de Permutación Completa (Bloques Incompletos Balanceados)*

Nombre del cluster	**Número de cuadernillo**					
	1	**2**	**3**	**4**	**5**	**6**
A	1	1	2	2		
B	2			1	1	2
C		2	1		2	1

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 3$; $b = 6$; $k = 2$; $r = 4$; $\lambda = 2$

en cada una de las tres posiciones posibles. Así, por ejemplo, los cuadernillos 1 y 2 comienzan con el cluster A y en ellos los clusters B y C permutan su posición.

Como se puede ver, los DPC no tienen impuesta la condición rígida que presentan los cuadrados latinos como el DTR sobre la igualdad de parámetros en el diseño, es decir, no es necesario que $t = k = r = b$. Esto ya se observa en la Tabla 4, donde partiendo de tres clusters se construyen seis cuadernillos, es decir, $t = 3$ y $b = 6$.

Sin embargo, también es posible que el número total de clusters (t) no sea igual al número de clusters a incluir en el cuadernillo (k). La Tabla 5 muestra una variación del diseño anterior. En este caso, cada cuadernillo contiene dos clusters o unidades de evaluación en vez de los tres originales. Con ello, el tiempo de aplicación se reduce de 60 a 40 minutos. El diseño mantiene su equilibrio reduciendo el tamaño del bloque o cuadernillo (ahora $k = 2$ y no como antes donde $k = 3$), el número de repeticiones de tratamiento o cluster de ítems por cuadernillo ($r = 4$ y *no* $r = 6$) y la frecuencia con que dos tratamientos o clusters se emparejan a lo largo de la colección de cuadernillos ($\lambda = 2$ y *no* $\lambda = 6$). Nótese sin embargo, que este nuevo diseño los estudiantes ya no responden todos a los mismos ítems. Por ejemplo, los estudiantes que respondan a los cuadernillos del 1 al 4 tendrán la ocasión de responder a los ítems del cluster A. Sin embargo, los estudiantes de los cuadernillos 5 y 6 no responderán a estos ítems. Con ello ya se avanzan ideas que serán tratadas más adelante en el apartado dedicado a los diseños de Bloques Incompletos Balanceados.

Si bien la construcción de los DPC es simple, estos arreglos tienen una indudable carencia: sólo son posibles cuando se trabaja con un número pequeño de unidades o clusters de ítems, ya que de otra manera se vuelven irrealizables por el número de cuadernillos necesarios para dar cabida a las demandas del diseño. Sirva como ejemplo el estudio TIMSS 2007, el cual presenta un diseño de bloques incompletos parcialmente balanceados que se verá más adelante. En TIMSS 2007 se construyeron 28 clusters de ítems, que fueron distribuidos de 4 en 4 a lo largo de 14 modelos de cuadernillos (Ruddock, O'Sullivan, Arora, y Erberber, 2008). Si TIMSS, en vez del diseño empleado, hubiese pretendido cubrir las posibles combinaciones de 28 clusters tomados de 4 en 4 hubiese necesitado 20475 modelos de cuadernillo diferentes. Y si además hubiese querido controlar el orden de presentación de los 4 clusters en cada cuadernillo, entonces hubiera necesitado ¡491400 modelos de cuadernillo diferentes!

Diseños cuando el número de ítems a evaluar supera el tiempo de evaluación disponible: diseño de bloques incompletos

Los dos primeros diseños mostrados hasta ahora son diseños de bloques completos ya que cada modelo de cuadernillo, o bloque en términos experimentales, contiene todos los clusters de ítems disponibles, es decir, todos los tratamientos. Sin embargo, ocurre con mucha frecuencia que la evaluación incluye muchos más ítems de los que puede responder un estudiante en una sesión de evaluación. En otras palabras: cada cuadernillo sólo contiene una parte de los clus-

ters de ítems. Cuando ocurre esto se está ante un diseño bloques incompletos, en los cuales se impone la siguiente condición $t > k$. En este caso, la necesidad de organizar los cuadernillos de un modo eficiente es mucho más perentoria ya que, aunque los estudiantes individualmente tomados respondan a una pequeña parte de la evaluación total, los resultados deben poder generalizarse a la colección completa de ítems.

En diseño experimental hay muchas posibilidades de organizar un diseño de bloques incompletos. Seguramente el criterio más empleado es el que distingue los diseños en función de la precisión de las comparaciones entre cada par de tratamientos, habiendo entonces diseños de bloques balanceados o parcialmente balanceados. No es la única clasificación. También se pueden ordenar en función del arreglo u organización de los tratamientos dentro del diseño, habiendo entonces diseños de bloques incompletos al azar, cuadrados latinos y latices. En este apartado se presentarán distintos diseños que están siendo empleados para la evaluación de competencias académicas.

Diseños de Bloques Incompletos Balanceados (DBIB)

La evaluación del sistema educativo norteamericano, NAEP, lleva empleando diseños matriciales desde sus inicios en el año 1969. Durante los tres primeros lustros el diseño se limitaba a la construcción de una suerte de formas paralelas de tests: la colección de ítems se dividía en diferentes cuadernillos (*packages*), los cuales ocupaban 45 minutos de evaluación. En función del área evaluada se editaban entre 5 y 8 *packages,* lo que suponía que responder a la colección de ítems ocupaba entre 225 y 360 minutos. En el momento de la administración de la prueba cada estudiante contestaba a un cuadernillo. Esto hacía que el diseño matricial de los *packages* fuera bastante eficiente: cada estudiante respondía entre el 20 y el 12.5% del total de ítems. Sin embargo, las respuestas de toda la muestra permitían estimar la proporción de la población que resolvería acertadamente la colección completa de ítems. Desafortunadamente este diseño presentaba claras deficiencias: los cuadernillos no disponían de preguntas comunes por lo que era imposible comparar los resultados de dos estudiantes que hubiesen respondido a cuadernillos diferentes. Incluso era imposible comparar el resultado de los grupos-aula ya que, por las condiciones de administración de los tests, cada grupo-aula respondía al mismo cuadernillo de evaluación.

NAEP implementó el diseño de *packages* hasta mediados de los años ochenta, momento en que fue modificado completamente. El trabajo donde se anuncia la nueva organización de los cuadernillos de evaluación tiene un subtítulo clarificador: "*un nuevo diseño para una nueva era*" (Messick, Beaton, y Lord, 1983). Y, ciertamente, el trabajo inaugura una nueva etapa en el campo de la evaluación educativa ya que presenta dos novedades importantes, que no por casualidad aparecen juntas: el empleo de los modelos matemáticos derivados de la TRI para estimar los resultados, y la construcción de los cuadernillos de evaluación mediante la aplicación de los Diseños de Bloques Incompletos Balanceados (DBIB). Desde entonces NAEP emplea DBIB para evaluar diferentes áreas curriculares como la comprensión lectora (Beaton, 1987), matemáticas (Lazer, 1999) o educación cívica (Weiss y Schoeps, 2001).

Todo DBIB presenta cuatro condiciones:

- Cualquier tratamiento o cluster (t) debe aparecer al menos en un bloque o cuadernillo (b).
- Todos los bloques o cuadernillos son de igual longitud (k), siendo $t > k$.
- Cada tratamiento o cluster de ítems tendrá igual número de repeticiones (r).
- Cada par de tratamientos o clusters aparecen conjuntamente igual número de veces (λ) a lo largo de los bloques.

Las proporciones que se acaban de mencionar hacen que los DBIB sean los arreglos

experimentales más equilibrados y eficientes estadísticamente de cuantos se emplean en las evaluaciones educativas con grandes muestras. Por todo ello su empleo es altamente recomendable siempre que sean viables y lo permitan los condicionantes de la evaluación.

Pero, pese a su indudable robustez, no siempre es posible organizar los cuadernillos de acuerdo a las pautas de un DBIB. De hecho, los diseños del programa NAEP que se acaban de citar se antojan imposibles de replicar por las administraciones educativas españolas, ya que exigen la construcción de un enorme número de cuadernillos. Por ejemplo, el primer diseño de comprensión lectora de NAEP contenía 19 clusters de ítems que se replicaban 9 veces, y cada cuadernillo contenía 3 clusters. Para distribuir los clusters según un DBIB fue necesario construir 57 modelos de cuadernillo diferentes (Beaton, 1987). Tal número de cuadernillos sólo está al alcance de un programa como el norteamericano que maneja tamaños muestrales por encima del cuarto de millón de escolares.

No obstante existen DBIB que, por proporciones y tamaño, pueden ser perfectamente aplicables por las administraciones educativas en España, incluso por las más pequeñas. Para ilustrar esta afirmación se presenta a continuación un ejemplo ficticio, aunque viable, extraído de Messick, Beaton, y Lord (1983). Supóngase que se dispone de una muestra de 100 ítems de matemáticas organizados en 5 clusters de 20 ítems. Cada cluster de ítems tiene una duración de 25 minutos, lo que significa que la evaluación completa ocupa más de dos horas. No obstante el tiempo de evaluación disponible es de 50 minutos por lo que cada estudiante responderá a 2 de los 5 clusters. Esto permite lograr una evaluación viable y eficiente. Viable, porque cada estudiante responderá dentro del tiempo de evaluación disponible; y eficiente porque, si bien cada estudiante sólo responde al 40% de la colección de ítems, al final se dispondrá de parámetros poblacionales para la colección completa. La representación gráfica del diseño puede verse en la Tabla 6.

A continuación se detallan las características del diseño:

- Es un diseño incompleto ya que cumple la siguiente condición: $t = 5 > k = 2$. En este caso cada cuadernillo sólo incluye dos quintas partes de los ítems.
- Todos los cuadernillos contienen el mismo número de clusters de ítems ($k = 2$).
- Cada cluster aparece o se replica en igual número de cuadernillos ($r = 4$).
- Se verifica la siguiente igualdad: $b \cdot k = t \cdot r$, propia de cualquier DBIB. Es decir, el producto del número bloques por el tamaño del bloque es igual al producto del número de tratamientos por sus replicaciones. Por tanto, conocidos tres parámetros del diseño ($t = 5$, $k = 2$ y $r = 4$), el número de cuadernillos necesarios para un DBIB se calcula despejando

Tabla 6. *Ejemplo de Diseño de Bloques Incompletos Balanceados*

Nombre del cluster	**Número de cuadernillo**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	1				2	1			2	
B	2	1					1			2
C		2	1			2		1		
D			2	1			2		1	
E				2	1			2		1

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 5$; $b = 10$; $k = 2$; $r = 4$; $\lambda = 1$

Tabla 7. *Número del cuadernillo en el que aparecen conjuntamente dos clústers de ítems en un DBIB*

Nombre de los Clusters	A	B	C	D	E
A	*				
B	1	*			
C	6	2	*		
D	9	7	3	*	
E	5	10	8	4	*

* No aplica: un clúster no puede aparecer dos veces en el mismo cuadernillo

esta ecuación: $b = (tr) / k$. En este caso, b = 10.

— El diseño está arreglado para controlar el efecto de posición de los clusters dentro de los cuadernillos. En este ejemplo cada cluster ocupa la primera posición en dos cuadernillos y la segunda posición en otros dos.

— El diseño presenta un balanceo completo. Cualquier cluster a lo largo de sus cuatro apariciones se empareja una única vez con el resto de los clusters, es decir, $\lambda = 1$. La Tabla 7 muestra estas coincidencias entre pares de clusters a lo largo de los cuadernillos. Por ejemplo, los clusters A y B aparecen conjuntamente en el cuaderno 1, los clusters B y C lo hacen en el cuaderno 2 y así sucesivamente.

En suma, los DBIB tienen una serie de propiedades estadísticas y unas relaciones entre sus parámetros que los hacen especialmente recomendables. Además existen DBIB que pueden aplicarse en estudios con muestras pequeñas o muy pequeñas. Cochran y Cox (1974) compilan una relación exhaustiva de diseños organizados en bloques incompletos al azar que serían aplicables a cualquier evaluación de diagnóstico educativo a partir de 4 modelos de cuadernillo distintos.

Diseños de Cuadrados Latinos Incompletos Balanceados

Ya se apuntó en los DTR que el cuadrado latino ordinario es un arreglo que tiene dos características básicas. En primer lugar, cumple la siguiente igualdad: $t = b = k = r$. Además, las repeticiones de los tratamientos (clusters de ítems) se agrupan tanto por filas como por columnas de tal modo que cada fila (cuadernillo) y cada columna (orden de los clusters) son repeticiones completas. Este agrupamiento doble permite neutralizar las diferencias entre filas (cuadernillos) y entre columnas (posición de los clusters de ítems dentro del cuadernillo).

Sin embargo, esta organización se torna problemática cuando se manejan muchos clusters, ya que el número de repeticiones exigidas hace impracticable el diseño. Por ello, los cuadrados latinos ordinarios sólo se emplean cuando la colección de ítems es pequeña o equivalente al tiempo total de una sesión de evaluación, cuestión también advertida al tratar los DTR.

Ahora bien, existe una solución para el constructor de tests que, disponiendo de un número elevado de clusters de ítems, no quiere perder las ventajas del doble control que supone el cuadrado latino ordinario. Esta solución se denomina Cuadro Latino Incompleto Balanceado (CLIB). En realidad, el CLIB es una clase especial dentro de los DBIB y, por tanto, comparte con éstos las cuatro condiciones vistas en el apartado anterior. Además, como se trata de un diseño incompleto, el CLIB relaja la condición de igualdad de tratamientos y replicaciones propia del cuadrado latino ordinario. Por tanto, en los diseños CLIB: $t \neq r$.

En principio, construir un CLIB es bastante sencillo, ya que se trata de omitir ciertas columnas (repeticiones) de un cuadrado latino ordinario permitiendo mantener el diseño

balanceado. Por ejemplo, si se cuenta con una colección de entre 4 y 11 clusters arreglados según un cuadrado latino bastaría con omitir la última columna, es decir, la última replicación del cluster de ítems para disponer de un CLIB (Cochran y Cox, 1974). Sin embargo, esta estrategia sigue siendo insuficiente si el constructor del tests cuenta con un número elevado de clusters de ítems. Supóngase que se pretende evaluar la competencia científica con una colección de 11 clusters de ítems de 10 minutos cada uno y se dispone para ello de una sesión ordinaria de 60 minutos. Eliminando la última columna del cuadrado latino los cuadernillos constarían de 10 clusters de 10 minutos, lo que excedería el tiempo de de evaluación disponible.

Por suerte, existen arreglos de cuadrados latinos ordinarios que permiten eliminar más de una repetición y que, manteniendo el doble control característico del cuadrado latino incompleto, hacen posible la aplicación al diseño de cuadernillos de evaluación. Siguiendo con el ejemplo anterior es posible construir un CLIB donde los 11 clusters de ítems se repitan sólo 5 ó 6 veces a lo largo de la colección de cuadernillos, haciendo entonces viable el diseño.

Esta clase especial de CLIB que se vienen mencionando y que ha hecho posible la aplicación de cuadrados latinos incompletos al diseño de cuadernillos en las evaluaciones de competencias escolares recibe el nombre de Diseño Youden (DY). El DY respeta todas las condiciones de los DBIB y de los CLIB, pero añaden una restricción adicional: cada tratamiento o cluster de ítems debe aparecer en cada posición del bloque o cuadernillo con igual frecuencia. Esto determina bastante las relaciones entre los parámetros del diseño ya que obliga a establecer dos condiciones adicionales en el plan de construcción de cuadernillos:

— Debe haber tantos cuadernillos (bloques) como clusters de ítems (tratamientos), es decir, se impone que: $t = b$.
— El número de veces que un cluster de ítems aparece a lo largo de la colección de cuadernillos debe ser igual al número de clusters de ítems incluidos en los cuadernillos, es decir: $r = k$.

Estas restricciones hacen que las posibilidades de encontrar un DY no sean muchas. Siguiendo a Cochran y Cox (1974) apenas existen una veintena de DY para evaluaciones que manejen entre 7 y 91 clusters de ítems. Sin embargo, cuando son posibles, la solución que plantean es elegante, equilibrada y estadísticamente muy eficiente.

En realidad los DY se han empleado con cierta frecuencia en las evaluaciones de diagnóstico: NAEP ha construido algunos de sus cuadernillos de acuerdo a este arreglo. Sin embargo, considerando las circunstancias de las evaluaciones de diagnóstico en España la propuesta de DY que mejor puede replicada es la que implementa PISA. PISA propuso, por primera vez, un DY en su segunda edición de 2003, y desde entonces el modo de organizar los ítems dentro de los cuadernillos de evaluación permanece invariante (OECD, 2005, 2009).

Para mostrar cómo la arquitectura de los cuadernillos de la evaluación PISA se ajusta al DY se recrea el diseño de la evaluación del 2006. En dicha edición PISA desarrolló una colección de 218 ítems (139 de ciencias, 48 de matemáticas y 31 de comprensión lectora). Estos ítems se agruparon en 76 unidades de tamaño variable: 37 de ciencias, 31 de matemáticas y 8 de comprensión lectora. Por último, las unidades fueron anidadas a 13 clusters de ítems: 7 de ciencias, 4 de matemáticas y 2 de comprensión lectora. Cada cluster ocupaba unos 30 minutos de evaluación, por lo que la colección completa de ítems de PISA 2006 equivalía a 6 horas y media de evaluación. Para hacer viable la administración del test el cuadernillo de evaluación sólo contenía 4 de los 13 clusters y, por tanto, se respondía en 2 horas. Es decir, cada estudiante respondía a algo menos de un tercio de la colección completa de ítems (OECD, 2009). La tabla 8 muestra la distribución de los 13 clusters de ítems a lo largo de los 13 cuadernillos de evaluación.

Tabla 8. *Diseño de cuadernillos en las evaluaciones PISA 2003, PISA 2006 y PISA 2009*

Tipo de Cluster	Nombre del Cluster	Número de cuadernillo												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Ciencias	S1	1								2	4		3	
	S2	2	1						3			4		
	S3		2	1		4				3				
	S4	3		2	1		4							
	S5				3	1						2	4	
	S6					2	1		4		3			
	S7	4				3		1						2
Matemáticas	M1			4					1				2	3
	M2				4			3	2	1				
	M3		3		2						1			4
	M4			3				4			2	1		
Lectura	R1		4				3	2					1	
	R2						2			4		3		1

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 13$; $b = 13$; $k = 4$; $r = 4$; $\lambda = 1$

En la tabla se puede observar cómo la distribución de los clusters cumple todas las condiciones impuestas, tanto para los DBIB en general como las restricciones propias del DY:

— Es un DBIB porque $t > k$, y también un DY ya que: $t = b$ y $k = r$. Ya que: $t = b = 13$, se cumple que hay tantos cuadernillos como clusters de ítems. Y también: $k = r = 4$, es decir, la longitud de los cuadernillos es igual al número de replicaciones de cada cluster en la colección de cuadernillos.
— Control del efecto de posición. En las cuatro replicaciones cada cluster aparece en las cuatro posiciones posibles del cuadernillo. Por ejemplo, el cluster S1 está al inicio del cuaderno 1; es el segundo en el cuaderno 9; el tercero en el 12; y el cuarto en el 10.
— Balanceo completo. Cualquier cluster se emparejará una única vez con el resto de los cluster a lo largo de sus cuatro replicaciones. Así el cluster S1 aparece conjuntamente con los clusters S2, S4 y S7 en el cuaderno 1 y no vuelve a coincidir con estos tres clusters en otro cuadernillo. Lo mismo ocurre entre el cluster S1 y los clusters S3, M2 y R2, que aparecen conjuntamente en el cuaderno 9. Es resto de los emparejamientos se pueden verificar en la tabla 8.

Tal y como está presentado en la Tabla 8 puede ser difícil percibir que el diseño PISA no es más que un cuadrado latino 13 x 13 del que sólo se toman las cuatro primeras repeticiones.

Para mostrar con mayor calidad esta característica la Tabla 9 organiza los clusters de ítems cómo se presentarían en un cuadrado latino incompleto: los bloques (cuadernillos) en las filas, las repeticiones en las columnas, y los clusters de ítems en las cedas. De esta forma es fácil comprobar cómo el diseño PISA es un cuadrado latino ordinario del que sólo se han tomado las cuatro primeras replicaciones.

El diseño de PISA no se reproduce sólo por su equilibrio, elegancia y posibilidades

Tabla 9. *El Diseño de cuadernillos en PISA 2006 es un Cuadrado Latino Incompleto*

		Repeticiones: Orden de los clusters			
		1º	**2º**	**3º**	**4º**
	1	S1	S2	S4	S7
	2	S2	S3	M3	R1
	3	S3	S4	M4	M1
	4	S4	M3	S5	M2
Bloques	**5**	S5	S6	S7	S3
(Número	**6**	S6	R2	R1	S4
de cuadernillo)	**7**	S7	R1	M2	M4
	8	M1	M2	S2	S6
	9	M2	S1	S3	R2
	10	M3	M4	S6	S1
	11	M4	S5	R2	S2
	12	R1	M1	S1	S5
	13	R2	S7	M1	M3

de análisis estadístico, otra razón que justifica esta recreación es la viabilidad del diseño para cualquier administración educativa. Se estima que el diseño de PISA se puede implementar en evaluaciones que manejen muestras a partir 6000-7000 estudiantes. Por ejemplo, con una muestra de 6000 estudiantes cada cuadernillo sería respondido por 461 estudiantes de promedio. Y como cada ítem aparece en cuatro cuadernillos distintos será respondido por unos 1850 estudiantes, número más que suficiente para calibrar y ajustar los reactivos a los modelos derivados de la TRI.

En todo caso cabe recordar que existen otros muchos diseños que permiten organizar los clusters de ítems como un cuadrado latino incompleto. De nuevo, Cochran y Cox (1974) recogen una relación exhaustiva de los planes disponibles para organizar este tipo de arreglo en colecciones de ítems de incluyan entre 3 y 91 clusters o unidades de evaluación.

Diseños de Bloques Incompletos Parcialmente Balanceados

Hasta el momento se han visto diseños balanceados, es decir, arreglos donde cada par de clusters de ítems aparecen juntos el mismo número de veces a lo largo de la colección de cuadernillos. Esto hace que cada par de clusters se compare con igual precisión y que, por tanto, los diseños balanceados presenten unas propiedades estadísticas que les hacen especialmente relevantes para su implementación en pruebas de evaluación de rendimiento.

Sin embargo, hay situaciones prácticas que desaconsejan la elección de un DBIB. La primera emana de la propia condición de balanceo: la exigencia de que cualquier cluster empareje al resto con igual frecuencia hace que muchas veces (como ya se advirtió en la evaluación norteamericana) el número de cuadernillos necesarios se dispare haciendo inviable la logística de la evaluación. En segundo lugar, hay situaciones donde ni siquiera es recomendable o posible que un par de clusters determinados aparezcan juntos. Eso puede ocurrir, por ejemplo, ante la sospecha de efecto de interacción entre clusters de ítems. En este caso se estaría ante "clusters enemigos", debido a que la información que contiene uno de ellos sirve para responder ítems en el otro.

Para situaciones como las anteriores son más efectivos los Diseños de Bloques Incompletos Parcialmente Balanceados (DBI-PB). La diferencia fundamental entre los DBIB y DBI-PB es que los segundos relajan la condición de que el parámetro λ sea igual

Tabla 10. *Ejemplo de Diseño de Bloques Incompletos Parcialmente Balanceados*

Cluster	Número del cuadernillo				
	1	**2**	**3**	**4**	**5**
A	1				2
B	2	1			
C		2	1		
D			2	1	
E				2	1

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 5$; $b = 5$; $k = 2$; $r = 2$; $\lambda_1 = 1$ y $\lambda_2 = 0$

para todos los pares de clusters, es decir, en la ecuación $\lambda = [r(k - 1)] / (t - 1)$, el resultado deja de ser un entero positivo. Si bien es cierto que en los DBI-PB cada par de clusters no se compara con igual precisión, no lo es menos que esta estrategia permite rebajar el número de replicaciones por tratamiento y, en consecuencia, simplificar la logística de la evaluación, al disminuir el número de cuadernillos o bloques necesarios en el diseño.

Para ilustrar el modo de construir un DBI-PB se vuelve sobre el ejemplo sobre competencia matemática presentado al tratar los DBIB y recogido en la Tabla 6. Allí se vio que para lograr un DBIB con 5 clusters distribuidos de 2 en 2 eran necesarias 4 repeticiones por cluster y 10 cuadernillos para emparejar cada cluster con el resto. Es posible disponer de un diseño que siga siendo eficiente pero que reduzca a la mitad el número de cuadernillos. En este caso concreto, se puede lograr el DBI-PB simplemente disminuyendo las repeticiones de cada cluster de 4 a 2. El nuevo diseño se presenta en la Tabla 10. Si se comparan las tablas 6 y 10 se observa que en ambos casos los cinco primeros cuadernillos son exactamente iguales. Esto es así porque estos cinco cuadernillos contienen las dos primeras replicaciones de cada cluster de ítems en el diseño balanceado.

En concreto, el ejemplo de la Tabla 10 muestra una clase particular de DBI-PB, que recibe el nombre de Bloques Encadenados o Diseño de Circuito Cerrado. Se trata de un arreglo que produce una estructura de considerable consistencia. Los cuadernillos van enlazando los clusters sucesivamente entre sí mediante ítems comunes, construyendo así una cadena que se cierra en el último cuadernillo cuando se enlazan el primer y el último cluster. En el diseño se aprecia el rasgo distintivo de todo diseño parcialmente balanceado: el parámetro λ presenta más de un valor (en este caso, $\lambda_1 = 1$ y $\lambda_2 = 0$). Estos dos valores indican que en este diseño algunos tratamientos o clusters de ítems aparecen juntos en algún cuadernillo, mientras que otros no lo hacen. Por ejemplo, el cluster A aparece conjuntamente con el cluster B en el cuaderno 1 y con el cluster E en el cuaderno 5. Sin embargo, el cluster A no aparece conjuntamente con los clusters C y D. La Tabla 11 recoge las coincidencias entre pares de clusters del DBI-PB que se viene mostrando.

Cuando, como en este caso, el parámetro λ toma dos valores se dice que el DBI-PB tiene dos factores asociados, aunque también es posible disponer de DBI-PB con tres o más factores asociados, es decir, diseños donde λ toma tres o más valores.

Muchas evaluaciones han empleado DBI-BP. De hecho, los diseños de Islas Baleares y Asturias mostrados en las tablas 1 y 2 son DBI-PB. PISA también siguió este tipo de diseño en su primera edición (Wu, 2002) y NAEP lo lleva empleando tanto tiempo como los diseños de bloques balanceados (Beaton, 1987).

Sin embargo, posiblemente sean las evaluaciones TIMSS y PIRLS las que hayan explotado con mayor profusión las posibilidades de los DBI-PB. Así, TIMSS siempre ha

Tabla 11. *Número del cuadernillo en el que aparecen conjuntamente dos clusters de ítems en un DBI-PB*

Nombre de los Clusters	A	B	C	D	E
A	*				
B	1	*			
C	-	2	*		
D	-	-	3	*	
E	5	-	-	4	*

* No aplica: un cluster no puede aparecer dos veces en el mismo cuadernillo

utilizado 28 clusters de ítems, para lograr dos finalidades: cubrir adecuadamente las especificaciones de las áreas evaluadas y establecer estudios de tendencia. Se aprecia que a lo largo de sus sucesivas evaluaciones el diseño para organizar los clusters ha ganado en simetría y equilibrio. Así, en las dos primeras ediciones (1995 y 1999) se utilizó un diseño de anclaje fijo de ítems, que era bastante complejo: utilizaba 14 modelos de cuadernillos e incluía hasta cinco pares de clusters en función del grado de asociación entre ellos (Adams y Gonzalez, 1996; Garden y Smith, 2000). En la evaluación del 2003 el diseño ganó en simetría, abandonado el diseño de anclaje fijo de ítems y distribuyendo los clusters en 12 cuadernillos de un modo más compacto (Smith-Neidorf y Garden, 2004). Por último, la organización de cuadernillos del año 2007, que será idéntica en 2011, es la más lograda y evolucionada. Se trata de un diseño de cadenas de bloques –similar en su arquitectura al mostrado en la tabla 10- y construido para albergar unos 400 ítems en 14 cuadernillos de evaluación (Ruddock et al., 2008). La lectura conjunta de todas estas referencias permite entender cómo los diseños de TIMSS van reservando clusters de ítems de evaluación en evaluación para establecer estudios de tendencias. Esta estrategia permitirá que los sistemas educativos que hayan participado regularmente desde la primera edición de TIMSS dispongan en el año 2015 de tendencias de resultados que abarcarán dos décadas.

Por su parte, los estudios PIRLS han implementado siempre un diseño DBI-PB basado en el diseño de cadenas de ítems, aunque con ligeras variaciones para cubrir la tabla de especificaciones y establecer tendencias de rendimiento. Cualquier administración que quiera evaluar la comprensión lectora encontrará los modelos para organizar sus cuadernillos y replicar el diseño de PIRLS en Campbell, Kelly, Mullis, Martin, y Sainsbury (2001), Kennedy y Sainsbury (2007), Mullis, Kennedy, Martin, y Sainsbury (2006), Mullis, Martin, Kennedy, Trong, y Sainsbury (2009) y Sainsbury y Campbell (2003).

Diseños en Látices Cuadrados Parcialmente Balanceados

La distribución de los ítems en los cuadernillos de la evaluación de diagnóstico Asturias 2010 se ajusta a un *Látice Simple Parcialmente Balanceado* (LS-PB). Este diseño pretende dar respuesta a las decisiones tomadas por la Consejería de Educación, las cuales transcendían el ámbito técnico. En primer lugar, se decidió que en dicho año se evaluarían dos competencias: matemática y conocimiento e interacción con el mundo físico. Ahora bien, para simplificar la logística de la prueba, así como las tareas de corrección y grabación de datos se determinó que cada estudiante sólo respondería a un cuadernillo. Por tanto, las dos competencias debían convivir en un único diseño de cuadernillos. Por otro lado, la evaluación debería cubrir ampliamente las especificaciones de contenido de las dos competencias, planteándose desde el principio trabajar con unos 200 ítems en total. Sin embargo, el tiempo de evaluación por estudiante no podría superar los 120 minutos, es decir, dos sesiones de 60 minutos partidas por un des-

Tabla 12. *Parámetros t, k, r y b en los diseños Látices Balanceados*

Tratamientos o número de clusters (*t*)	9	16	25	49	64	81
Tamaño del bloque o clusters por cuadernillo (*k*)	3	4	5	7	8	9
Replicaciones o veces que aparece el clúster en los cuadernillos (*r*)	4	5	6	8	9	10
Número de bloques o número de cuadernillos (*b*)	12	20	30	56	72	90

canso. Esto significa que la colección completa de ítems era muy superior al número de preguntas que cada estudiante podría responder en el tiempo asignado a la evaluación. Finalmente se pretendía que el nuevo diseño mejorara el diseño de anclaje fijo empleado hasta entonces, y que fuese original, equilibrado y con vocación de estabilidad en el tiempo, para permitir, llegado el caso, establecer tendencias de rendimiento.

A partir de estos condicionantes la propuesta técnica fue construir los cuadernillos de evaluación basándose en un diseño en Látice Cuadrado, que es uno de los principales arreglos dentro los DBIB. A continuación se describen las características generales de los diseños de látices cuadrados y se finalizará presentando el diseño concreto empleado en la evaluación Asturias 2010.

Como es bien sabido, un diseño arreglado en látice tiene dos características distintivas: a) El número de tratamientos o clusters de ítems (*t*) debe ser un cuadrado exacto. Por tanto, en el látice el parámetro *t* sólo puede tomar valores como 4, 9, 16,..; y b) El tamaño del bloque, es decir, el número de clusters por cuadernillo (*k*), es la raíz cuadrada del número de tratamientos, en todo diseño látice se cumple la siguiente igualdad: $t = k^2$. Esta doble condición hace que los otros dos parámetros del diseño, repeticiones (*r*) y número de bloques (*b*), queden rígidamente determinados. La Tabla 12 recoge los cuatro parámetros de los primeros látices posibles, siendo en todos ellos $\lambda = 1$.

En la Tabla 12 se observa que ya a partir del tercer látice ($t = 16$) el número de cuadernillos necesarios para disponer de un diseño balanceado se aumenta de tal forma que compromete a viabilidad de la evaluación. Como ocurre con cualquier diseño balanceado, los látices que manejan muchos tratamientos exigen un número de bloques tan elevado que suponen un grave inconveniente para la construcción de los cuadernillos. Para soslayar este problema cabe la posibilidad de trabajar con látices parcialmente balanceados. Para ello es suficiente con tomar menos repeticiones de las pautadas en el diseño completamente balanceado. Así, el látice simple parcialmente balanceado toma las dos primeras repeticiones del látice balanceado elegido; el látice triple, se logra tomando las tres primeras repeticiones; el latice de cuatro repeticiones se obtiene, bien por duplicación del látice simple o bien por el uso de un látice cuádruple, es decir, tomando las cuatro primeras replicaciones del látice elegido. Cochran y Cox (1974) ofrecen las indicaciones para organizar látices parcialmente balanceados de hasta doce repeticiones.

A continuación se recrean los números de la evaluación Asturias 2010 y el diseño elegido. Asturias 2010 contaba con 192 ítems, la mitad para la competencia matemática y la otra mitad para evaluar el conocimiento e interacción con el mundo físico. La colección total se organizó en 33 unidades que contenían entre 3 y 8 ítems de ambas competencias. Finalmente las unidades se agruparon en 16 clusters de ítems que contenían 2 o 3 unidades. Cada cluster equivalía a 30 minutos de evaluación, por lo que la colección total de ítems suponía unas 8 horas de evaluación total. Como el tiempo de evaluación era de 2 horas eso suponía que cada estudiante debería responder a 4 de los 16 clusters, es decir, al 25% del total de la colección.

Con estos guarismos, el látice balanceado de 16 tratamientos era la solución más adecuada. Sin embargo, elaborar 20 modelos de cuadernillo, tal y como exige este di-

Tabla 13. *Diseño de cuadernillos en la evaluación Asturias 2010*

Nombre del clúster	Número de cuadernillo 1	2	3	4	5	6	7	8
A	1				4			
B	2					3		
C	3						2	
D	4							1
E		3			2			
F		4				1		
G		1					4	
H		2						3
I			4		1			
J			3			2		
K			2				3	
L			1					4
M				2	3			
N				1		4		
O				4			1	
P				3				2

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 16$; $b = 8$; $k = 4$; $r = 2$; $\lambda_1 = 0$ y $\lambda_2 = 1$

seño hacía demasiado compleja la logística de la prueba. Por ello, se decidió trabajar un látice parcialmente balanceado, tomando sólo las dos primeras repeticiones del látice de 16 tratamientos y reduciendo el número de cuadernillos de 20 a 8. El diseño en cuestión se presenta en la Tabla 13.

A continuación se detallan las cuatro características básicas del diseño:

En primer lugar es un diseño incompleto. Se observa que los 16 clusters se distribuyen a lo largo de 8 modelos de cuadernillo con las siguientes condiciones: cada cluster se replicará 2 veces a lo largo de la colección de cuadernillos, los cuales a su vez contienen, como ya se ha mencionado, 4 clusters cada uno.

También se puede apreciar cómo el diseño controla el efecto de posición de los ítems. Si bien cada cluster se replica sólo dos veces, su posición está equilibrada a lo largo de los cuadernillos. Así los clusters que están en el inicio de un cuadernillo aparecen al final de otro y viceversa. Por ejemplo, el cluster A ocupa la primera posición en el cuaderno 1 y la cuarta en el cuaderno 5. De igual modo, los clusters que ocupan la segunda posición en un cuadernillo ocupan la tercera en la segunda aparición. Por ejemplo, el cluster B es segundo en el cuaderno 1 y tercero en el cuaderno 6.

El tercer rasgo es balanceo parcial. En el diseño el parámetro λ presenta dos valores (0 y 1), es decir, es un diseño con dos clases de clusters asociados. Cualquier cluster es primer asociado (es decir, comparte cuadernillo) con otros 6 clusters y es segundo asociado (no comparte cuadernillo) con otros 9 clusters. Por ejemplo, el cluster A aparece conjuntamente con los clusters: B, C y D en el cuadernillo 1 y con E, I y M en el 5. Para todos estos pares $\lambda_1 = 1$. Por otro lado, el cluster A no comparte cuadernillo con el resto de los 9 clusters. Para estos segundos asociados del cluster A $\lambda_2 = 0$.

A continuación se profundiza en la estructura parcialmente balanceada. Para mantener la simetría y posibilidades estadísticas del diseño las asociaciones entre los clusters deben estar bien equilibradas. Considérense dos clusters que sean primeros asociados, por ejemplo, A y B. En el caso del cluster A los primeros asociados –además de B- son los cluster C, D, E, I y M, siendo el resto segundos asociados. Por su parte los primeros asociados del cluster B son –además de A-

Tabla 14. *Relaciones de los Cluster A y B (primeros asociados) con el resto*

		Relación del cluster A con el resto	
		1º Asociado	**2º Asociado**
Relación del cluster B con el resto	**1º Asociado**	C, D	F, J, N
	2º Asociado	E, I, M	G, H, K, L, O, P

Tabla 15. *Distribución del número de emparejamientos de dos clusters que son primeros asociados*

		Relación del cluster A con el resto	
		1º Asociado	**2º Asociado**
Relación del cluster B con el resto	**1º Asociado**	2	3
	2º Asociado	3	6

Tabla 16. *Relaciones de los Cluster A y F (segundos asociados) con el resto*

		Relación del cluster A con el resto	
		1º Asociado	**2º Asociado**
Relación del cluster F con el resto	**1º Asociado**	B, E	G, H, J, N
	2º Asociado	C, D, I, M	K, L, O, P

Tabla 17. *Distribución del número de emparejamientos de dos clusters que son segundos asociados*

		Relación del cluster A con el resto	
		1º Asociado	**2º Asociado**
Relación del cluster F con el resto	**1º Asociado**	2	4
	2º Asociado	4	4

los clusters C, D, F, J, y N, siendo el resto de los clusters segundos asociados. Si estas relaciones se representan en una tabla bivariada se obtiene el resultado que aparece en la Tabla 14.

Como se puede ver, para cada par de clusters que ocurren juntos (por ejemplo, A y B):

- Hay otros dos clusters que son primeros asociados de ambos (en este caso C y D).
- Hay otros 6 clusters que son primeros asociados de un cluster y segundos asociados del otro. Por ejemplo, E, I y M son primeros asociados de A y segundos asociados de B. La situación inversa ocurre con los clusters F, J y N.
- Finalmente, hay otros seis clusters con los que A y B no tienen ninguna relación a lo largo de la colección de cuadernillos. Son pues segundos asociados de A y B.

Por tanto, en este diseño se cumple la siguiente condición: dados dos clusters que son primeros asociados, la relación que mantienen con el resto de los clusters es la de la tabla 15.

Este equilibrio también se mantiene cuando se comparan las relaciones entre dos clusters que son segundos asociados (por ejemplo A y F). En este caso la tabla bivaria-

Tabla 18. *Diseño de anclaje temporal sobre la base de un LS-PB de 16 tratamientos*

Nombre del cluster	Número de cuadernillo 1	2	3	4	5	6	7	8
A	1				4			
E	2					3		
N1	3						2	
N2	4							1
N3		3			2			
N4		4				1		
F		1					4	
B		2						3
G			4		1			
C			3			2		
N5			2				3	
N6			1					4
N7				2	3			
N8				1		4		
D				4			1	
H				3				2

1 Posición del clúster en el cuaderno.
Parámetros del diseño: $t = 16$; $b = 8$; $k = 4$; $r = 2$; $\lambda_1 = 0$ y $\lambda_2 = 1$

da que representa las relaciones entre ambos clusters con el resto aparece en la Tabla 16.

Por tanto, en el diseño ahora expuesto la relación que mantienen dos clusters que son segundos asociados con el resto de los clusters se ajusta a la distribución de la Tabla 17.

En definitiva, el diseño presenta una buena simetría, que está pensada para controlar los efectos de posición de los ítems y lograr estimaciones de error de los parámetros de los ítems similares.

La cuarta característica del diseño es su capacidad para incorporar de modo sencillo ítems nuevos y combinarlos con ítems de anclaje aplicados previamente. El diseño está pensado para facilitar el establecimiento de tendencias de rendimiento a lo largo del tiempo. A continuación se muestra un ejemplo de cómo es posible mantener el diseño constante en diferentes ediciones de la evaluación al tiempo que se introducen nuevos clusters de ítems que son combinados con clusters ya aplicados sin alterar las tres características apuntadas previamente.

Supóngase que en una segunda edición de la evaluación de diagnóstico se emplea el diseño de la Tabla 13 y se decide que los ítems de anclaje temporal entre las dos evaluaciones serán los contenidos en los clusters del A al H. Por tanto, para la segunda evaluación no serán necesarios más que 8 clusters nuevos –numerados desde N1 a N8. En la Tabla 18 se muestra la propuesta de diseño para el anclaje temporal.

Este segundo diseño cumple con las siguientes condiciones:

- Cada cuadernillo contiene dos clusters nuevos o de reposición y dos clusters ya empleados o de anclaje.
- Los clusters de anclaje mantienen sus posiciones dentro de los cuadernillos. Así, los clusters A, D, F y G siguen estando al principio y al final de sus cuadernillos. De igual modo B, C, E y H ocupan los lugares segundo y tercero en sus respectivos cuadernillos.
- A cada cluster de anclaje le corresponden otros 6 clusters que son primeros asociados. Pues bien, en el diseño se cumple que 2 de estos 6 primeros asociados son siempre clusters de anclaje y 4 son clusters nue-

vos. Por ejemplo, el cluster A se empareja con los clusters de anclaje E en el cuaderno 1 y con G en el cuadernillo 5. Los otros 4 primeros asociados de A son clusters nuevos: N1 y N2 en el cuaderno 1, y N3 y N7 en el cuaderno 5. El resto de los emparejamientos se pueden verificar en la tabla 14.

— Por su parte, de los 6 primeros asociados que le corresponden a cualquier cluster nuevo, 4 son clusters de anclaje y 2 clusters nuevos. Así el cluster N1 se empareja con A y E en el cuaderno 1 y con F y D en el cuaderno 7. Esta forma de distribuir los clusters nuevos y repetidos permitirá establecer comparaciones temporales robustas al tiempo que el diseño mantiene un equilibrio y simetría suficientes para hacer un buen análisis estadístico.

En suma, el diseño LD-PB que se acaba de presentar cumple con las siguientes condiciones: permite incluir una colección amplia de ítems, contiene más de una competencia o área de evaluación en un único diseño y está preparado para establecer tendencias de rendimiento.

Consideraciones finales

Se han presentado los principales diseños de cuadernillos que pueden ser empleados por las administraciones educativas para realizar sus evaluaciones de diagnóstico. Presentar un abanico amplio de arreglos se debe a que sobre el papel no hay ninguno que sea superior al resto. Además, la elección del diseño suele estar condicionada por una serie de decisiones que, en muchas ocasiones, superan el plano meramente técnico. Por tanto, antes de elegir el diseño concreto el constructor del test deberá sopesar las ventajas e inconvenientes de cada diseño y el contexto de su evaluación concreta.

Los diseños completos, es decir, aquellos que emplean pocos ítems (RTD y CPD) tienen indudables ventajas. En primer lugar, son diseños sencillos de construir, además, como necesitan un número pequeño de ítems las tareas relacionadas con el desarrollo de los ítems, tales como escritura, revisión, pilotajes, análisis de datos y desarrollo de las guías de corrección, tienen un coste menor. Los diseños donde todos los estudiantes responden a los mismos ítems presentan ventajas en cuanto a la comparación de resultados y a la divulgación de los mismos. El principal inconveniente de los diseños con pocos ítems tiene que ver con la validez y fiabilidad de las pruebas. Una colección pequeña de ítems no suele cubrir adecuadamente las especificaciones de la competencia a evaluar y por tanto, la prueba tiene bastante comprometida su validez de contenido. Además, en la evaluación de las competencias académicas, como ocurre con cualquier otra variable, la estimación de las puntuaciones mejora al aumentar el número de ítems.

Las ventajas e inconvenientes de los diseños que manejan muchos ítems son la otra cara de la situación descrita más arriba. Cuando la colección de ítems es muy grande aumentan los costos de desarrollo y la logística y administración de la prueba se vuelve más compleja. Por el contrario, una colección amplia de ítems ofrece muchas más garantías con respecto a la validez de contenido y permite establecer estimaciones más fiables con respecto a la competencia académica de la población estudiada.

En definitiva, las posibilidades de las administraciones educativas con respecto al diseño de cuadernillos a emplear son muy amplias. Y la elección del arreglo concreto debe hacerse después de una valoración cuidadosa de los condicionantes de la evaluación y las posibles fortalezas y debilidades de cada diseño.

Referencias

Adams, R.J., y Gonzalez, E.J. (1996). *The TIMSS Test Design*. En M. O. Martin y D. L. Nelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Arana, L. (2010). La importancia de la educación en la estrategia estatal de innovación. *Aula Abierta, 38*(2), 41-52.

Arnau, J. (1984). *Diseños experimentales en psicología y educación*. México: Trillas.

Ato, M., y Vallejo, G. (2007). *Diseños experimentales en psicología*. Madrid: Pirámide.

Beaton, A. (1987). *Implementing the new design: The NAEP 1983-84 Technical Report*. Princeton, NJ: National Assessment of Educational Progress / Educational Testing Service.

Box, G.E., Hunter, J.S., y Hunter, W.G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd Edition). Hoboken NJ: Wiley.

Campbell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M. O., y Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001*. Chestnut Hill, MA: Boston College.

Campbell, D.T., y Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Childs, R. A., y Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation, 8*(16), http://pareonline.net/getvn.asp?v=8&n=16.

Cochran, W.G., y Cox, G.M. (1974). *Diseños experimentales*. México: Trillas. (orig. 1957).

Comunidad de Madrid (2008). *Evaluación de diagnóstico, Lengua y Matemáticas, 4º de Educación Primaria*. Madrid: Viceconsejería de Organización Educativa.

Comunidad de Madrid (2010a). *Evaluación de Diagnóstico 4º Educación Primaria. Instrucciones de aplicación*. Madrid: Dirección General de la Mejora de la Calidad de la Enseñanza.

Comunidad de Madrid (2010b). *Evaluación de Diagnóstico 2º ESO. Instrucciones de aplicación*. Madrid: Dirección General de la Mejora de la Calidad de la Enseñanza.

Cook, T.D., y Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for the field settings*. Boston: Houghton Mifflin.

Fisher, R.A., y Yates, F. (1963). *Statistical Tables for biological, agricultural and medical research* (6ª ed.). Edimburgo: Oliver & Boyd.

Frey, A., Hartig, J., y Rupp, A.A. (2009). An NCME instructional Module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Garden, R.J., y Smith, T.H. (2000). TIMSS Test Development. En M.O. Martin, K.D. Gregory, y S.E. Stemler (Eds.), *TIMSS 1999 Technical Report* (pp. 49-67). Chestnut Hill, MA: Boston College.

Generalitat de Catalunya (2009a). *Prova d'avaluació de sisè curs d'educació primària 2009*. Barcelona: Consell Superior d'Avaluació del Sistema Educatiu. Consultado el 08/02/2010 en http://www20.gencat.cat/portal/site/Educacio.

Generalitat de Catalunya (2009b). *Resultats de l'avaluació de sisè d'educació primària 2009*. Barcelona: Consell Superior d'Avaluació del Sistema Educatiu. Barcelona: Consultado el 08/02/2010 en http://www20.gencat.cat/ portal/site/Educacio.

Generalitat de Catalunya (2009c). *L´avaluació de sisè d´educació primària 2009. Avançament de resultats*, *Quaderns d´avaluació, 15*. Barcelona: Consell Superior d'Avaluació del Sistema Educatiu. Consultado el 08/02/2010 en http://www20.gencat.cat/portal/site/Educacio.

Generalitat de Catalunya (2010). *L´avaluació de sisè d´educació primària 2010. Síntesi de resultats, Quaderns d´avaluació, 18*. Barcelona: Consell Superior d'Avaluació del Sistema Educatiu. Consultado el 08/02/2010 en http://www20.gencat.cat/portal/site/Educacio.

Gobierno de Aragón (2008*). La evaluación de diagnóstico. Comunidad Autónoma de Aragón. Curso 2008-2009*. Zaragoza: Departamento de Educación, Cultura y Deporte.

Gobierno de Aragón (2010). *Evaluación censal de diagnóstico en Aragón. 2009*. Zaragoza: Departamento de Educación, Cultura y Deporte.

Gobierno de Canarias (2009). *La evaluación educativa institucional: Plan de Evaluación Diagnóstica de Canarias*. Las Palmas de Gran Canaria: Instituto Canario de Evaluación y Calidad Educativa. Consultado el 08/02/2010 en http://www.gobiernodecanarias.org/educacion.

Gobierno de Cantabria (2010). *Informe evaluación de diagnóstico 2008-2009*. Cantabria, Santander: Consejería de Educación de Cantabria.

Gobierno de La Rioja (2009). *Marco Teórico de la Evaluación Autonómica de Diagnóstico*. La Rioja, Logroño: Servicio de Innovación Educativa y Formación del Profesorado.

Gobierno de La Rioja (2010a). *Informe de resultados de la Evaluación autonómica de diagnóstico. 4º Educación Primaria*. Logroño: Servicio de Innovación Educativa y Formación del Profesorado.

Gobierno de La Rioja (2010b). *Informe de resultados de la Evaluación autonómica de diagnóstico. 2º Educación Secundaria Obligatoria*. Logroño: Servicio de Innovación Educativa y Formación del Profesorado.

Gobierno de Navarra (2008*). Marco Teórico de la evaluación diagnóstica. Educación Primaria*. Pamplona: Departamento de Educación - Servicio de Inspección Educativa. Consultado el 08/02/2010 en http://dpto.educacion.navarra.es/publicaciones/pdf/Marcoteorico1.pdf.

Gobierno de Navarra (2010). *Evaluación Diagnóstica 2009/10. Educación Primaria: informe final*. Pamplona: Servicio de Inspección Educativa – Sección de evaluación. Consultado el 08/02/2010 en http://www.educacion.navarra.es/portal/digitalAssets/48/48632_EP_informe_final_2009_10.pdf.

Gobierno del Principado de Asturias (2007a). *Evaluación de diagnóstico Asturias 2006. Marco de la evaluación*. Oviedo: Servicio de Evaluación, Calidad y Ordenación Académica.

Gobierno del Principado de Asturias (2007b). *Evaluación de diagnóstico Asturias 2006. Informe de resultados: 4º Educación Primaria*. Oviedo: Servicio de Evaluación, Calidad y Ordenación Académica.

Gobierno del Principado de Asturias (2007c). *Evaluación de diagnóstico. Asturias 2006. Informe de resultados: 2º Educación Secundaria Obligatoria*. Oviedo: Servicio de Evaluación, Calidad y Ordenación Académica.

Gobierno del Principado de Asturias (2007d). *Evaluación de diagnóstico. Asturias 2007*. Oviedo: Servicio de Evaluación, Calidad y Ordenación Académica.

Gobierno del Principado de Asturias (2008). *Evaluación de diagnóstico. Asturias 2008*. Oviedo: Servicio de Evaluación, Calidad y Ordenación Académica.

Gobierno Vasco (2008a). *Evaluación de diagnóstico en la comunidad autónoma vasca: características de las pruebas de rendimiento*. Bilbao: Instituto Vasco de Evaluación e Investigación Educativa. Consultado el 08/02/2010 en http://www.isei-ivei.net/

Gobierno Vasco (2009a). *20 preguntas básicas sobre los informes de la evaluación diagnóstica*. Bilbao: Instituto Vasco de Evaluación e Investigación Educativa. Consultado el 08/02/2010 en http://www.isei-ivei.net.

Gobierno Vasco (2009b). *Evaluación Diagnóstica e-dossier.* Bilbao: Instituto Vasco de Evaluación e Investigación Educativa. Consultado el 08/02/2010 en http://www.isei-ivei.net.

Govern de les Illes Balears (2009). *Avaluacions de diagnòstic 2009-2010. Aspectes generals de l'Avaluació de Diagnòstic*. Palma de Mallorca: Institut d'Avaluació i Qualitat del Sistema Educatiu. Consultado el 08/02/2010 en http://iaqse.caib.es.

Govern de les Illes Balears (2010*). Avaluació de diagnòstic 2008-2009. Informe executiu*. Palma de Mallorca: Institut d'Avaluació i Qualitat del Sistema Educatiu. Consultado el 08/02/2010 en http://iaqse.caib.es/documents/informe_AADD_2008_2009.pdf.

Junta de Andalucía (2006). *Evaluación de diagnóstico. Informe preliminar avance de resultados. Curso 2006-2007*. Sevilla: Consejería de Educación.

Junta de Andalucía (2007a). *Evaluación de diagnóstico. Informe curso 2006-2007*. Sevilla: Consejería de Educación.

Junta de Andalucía (2007b). *Avance de la evaluación de diagnóstico. Curso 2007-2008*. Sevilla: Consejería de Educación.

Junta de Andalucía (2008). *El modelo de evaluación de diagnóstico en Andalucía*. Sevilla: Consejería de Educación.

Junta de Andalucía (2009). *Evaluación de diagnóstico. Curso 2008-2009*. Sevilla: Consejería de Educación.

Junta de Andalucía (2010). *Evaluación de diagnóstico. Curso 2009-2010. Avance*. Sevilla: Agencia Andaluza de Evaluación Educativa.

Junta de Castilla y León (2009). *Dossier informativo: evaluación de diagnóstico. 4º de educación primaria*. Valladolid: Dirección General de Calidad, Innovación y Formación del Profesorado Consultado el 08/02/2010 en http://www.educa.jcyl.es.

Junta de Comunidades de Castilla-La Mancha (2009). *Evaluación de diagnóstico de las competencias básicas. Castilla La-Mancha 2009-2011. Marco Teórico*. Toledo: Oficina de Evaluación de la Consejería de Educación. Consultado el 08/02/2010 en http://www.educa.jccm.es/educa-jccm/cm.

Junta de Comunidades de Castilla-La Mancha (2010). *Evaluación de diagnóstico censal de Castilla La-Mancha 2009-2011. Presentación Fase 2010*. Toledo: Oficina de Evaluación de la Consejería de Educación y Ciencia. Consultado el 23/11/2010 en http://www.educa.jccm.es/educa-jccm/cm.

Kennedy, A.M., y Sainsbury, M. (2007). Developing the PIRLS 2006 Reading Assessment and Scoring Guides. En M.O. Martin, I.V.S. Mullis, y A.M. Kennedy (Eds.), *PIRLS 2006 Technical Report* (pp. 9-22). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kirk, R.E. (1995). *Experimental designs: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks-Cole.

Lazer, S. (1999). Assessment instruments. En N.L. Allen, J. E. Carlson, y C.A. Zelenak (Eds.), *The NAEP 1996 Technical Report* (pp. 77-81). Washington, DC: U.S. Department of Education / National Center for Education Statistics.

Lord, F.M. (1955). Equating test scores: A maximum likelihood solution. *Psychometrika, 20*, 193-200.

Lord, F.M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22,* 259-267.

Martin, M.O., Mullis I.V.S., y Kennedy, A.M. (2007). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: Boston College.

Messick, S., Beaton, A., y Lord, F.M. (1983*). National Assessement of Educational Progress reconsidered: A new design for a new era*. Princeton, NJ: Department of Education.

Ministerio de Educación (2009a). *Evaluación General de Diagnóstico 2009. Marco de la Evaluación*. Madrid: Instituto de Evaluación.

Ministerio de Educación (2009b). *Evaluación de Diagnóstico Ceuta y Melilla 2009. Marco de la Evaluación*. Madrid: Instituto de Evaluación.

Ministerio de Educación (2009c). *Evaluación general del sistema educativo. Educación Primaria 2007*. Madrid: Instituto de Evaluación.

Ministerio de Educación, Cultura y Deporte (2001). *Evaluación de la Educación Primaria 1999*. Madrid: Instituto Nacional de Calidad y Evaluación.

Ministerio de Educación, Cultura y Deporte (2003). *Evaluación de la Educación Secundaria Obligatoria 2000*. Madrid: Instituto Nacional de Evaluación y Calidad del Sistema Educativo.

Ministerio de Educación y Ciencia (2005). *Evaluación de la Educación Primaria 2003*. Madrid: Instituto Nacional de Evaluación y Calidad del Sistema Educativo.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., y Sainsbury, M. (2006). *PIRLS 2006 Assessment Framework and Specifications*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Trong, K.L., y Sainsbury, M. (2009). *PIRLS 2011 Assessment Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y, Arora, A., y Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston Collage.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gonzalez, E.J., Chorostowski, S.J., y O´Connor, K.M. (2002). *TIMSS assessment framewoks and specificacions 2003 (2 ed.)*. Chestnut Hill, MA: Boston College.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los items*. Madrid: Pirámide.

Muñiz, J. (2001). *Teoría clásica de los tests*. Madrid: Pirámide.

Muñiz, J., Hidalgo, A.M., García-Cueto, E., Martínez, R., y Moreno, R. (2005). *Análisis de los ítems*. Madrid: La Muralla.

OECD (2005). *PISA 2003 Technical Report*. París: Organisation for Economic Co-operation and Development.

OECD (2009). *PISA 2006 Technical Report*. París: Organisation for Economic Co-operation and Development.

Olson, J.F., Martin, M.O., y Mullis, I.V.S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Región de Murcia (2009). *Prueba de evaluación de diagnóstico. 4º Primaria. Competencia Lingüística y Competencia Matemática*. Murcia: Consejería de Educación, Formación y Empleo. Consultado el 08/02/2010 en http://www.carm.es/neweb2/servlet/integra.servlets.ControlPublico?IDCONTENIDO=5214&IDTIPO=100&RASTRO=c860$m.

Ruddock, G.J., O'Sullivan, C.Y., Arora, A., y Erberber, E. (2008). Developing the TIMSS 2007 Mathematics and Science Assessments and Scoring Guides. En J.F. Olson, M.O. Martin, e I.V.S. Mullis (Eds.), *TIMSS 2007 Technical Report* (pp. 13-44). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Sainsbury, M., y Campbell, J.R. (2003). Developing the PIRLS Reading Assessment. En M.O. Martin, I.V.S. Mullis, y A.M. Kennedy (Eds.), *PIRLS 2001 Technical Report* (pp. 13-27). Chestnut Hill, MA: Boston College.

Smith-Neidorf, T.H., y Garden, R.J. (2004). Developing the TIMSS 2003 Mathematics and Science Assessment and Scoring Guides. En

M.O. Martin, I.V.S. Mullis, y S.J. Chrostowski, (Eds.), *TIMSS 2003 Technical Report* (pp. 23-65). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Vallejo, G., Arnau, J., Bono, R., Fernández, P., y Tuero, E. (2010). Selección de modelos anidados para datos longitudinales usando criterios de información y la estrategia de ajuste condicional. *Psicothema, 22*, 323-333.

Weiss, A.R. y Schoeps, T.L. (2001), Assessment frameworks and instruments for the 1998 Civics assessment. En N. L. Allen, J. R. Donoghue, y T. L. Schoeps (Eds), *The NAEP 1998 Technical Report* (pp. 255-268). Washington, DC: U.S. Department of Education / National Center for Education Statistics.

Wu, M. (2002). Test Design and Test Development. En R. Adams, y M. Wu (Ed.), *PISA 2000 Technical Report* (pp. 21-31). París: Organisation for Economic Co-operation and Development.

Xunta de Galicia (2009). *Avaluación de Diagnóstico Galicia*. Santiago de Compostela: Consellería de Educación e Ordenación Universitaria.

Anexo I

Servicios y unidades responsables de la evaluación de diagnóstico en las comunidades autónomas.

Ministerio de Educación (y Ceuta y Melilla): Instituto de Evaluación (IE).
http://www.institutodeevaluacion.educacion.es/

Andalucía. Agencia Andaluza de Evaluación Educativa (AGAEVE).
http://www.juntadeandalucia.es/educacion/agaeve/web/agaeve

Aragón. Servicio de Equidad y Evaluación – Unidad de Evaluación.
http://evalua.educa.aragon.es/

Asturias. Servicio de Ordenación Académica, Formación del Profesorado y Tecnologías Educativas.
http://www.educastur.es/

Baleares. Institut d'Avaluació i Qualitat del Sistema Educatiu de les Illes Balears (IAQSE).
http://www.iaqse.caib.es/

Canarias. Instituto Canario de Evaluación y Calidad Educativa (ICEC).
http://www.gobiernodecanarias.org/educacion/Portal/WebICEC/

Cantabria. Unidad Técnica de Evaluación y Acreditación.
http://www.educantabria.es/evaluacion_educativa/

Cataluña. Consell Superior d'Avaluació del Sistema Educatiu.
http://www20.gencat.cat/portal/site/Educacio/

Comunidad de Madrid. Subdirección General de Evaluación y Análisis.
http://www.madrid.org/cs/

Comunidad Valenciana. Instituto Valenciano de Evaluación y Calidad Educativa (IVECE).
http://www.edu.gva.es/eva/index.asp

Castilla y León. Servicio de Supervisión de Programas, Calidad y Evaluación.
http://www.educa.jcyl.es/

Castilla-La Mancha. Oficina de Evaluación.
http://www.educa.jccm.es/educa-jccm/cm/educa_jccm/

Extremadura. Agencia Extremeña de Evaluación Educativa.
http://www.juntaex.es/consejerias/educacion/aeee/

Galicia. Subdirección Xeral de Inspección, Avaliación e Calidade do Sistema Educativo.
http://www.edu.xunta.es/web/

La Rioja. Dirección General de Ordenación Académica e Innovación.
http://www.educarioja.org/

Navarra. Servicio de Inspección Educativa – Sección de Evaluación.
http://www.educacion.navarra.es/portal/

País Vasco. Instituto Vasco de Evaluación e Investigación Educativa (ISEI-IVEI).
http://www.isei-ivei.net

Región de Murcia. Servicio de Evaluación y Calidad Educativa.
http://www.carm.es/