Los programas internacionales de evaluación de los conocimientos en los países pobres: ¿Porque los economistas de la educación deberían ser más cautos?*

GÉRARD LASSIBILLE
Institut de Recherche sur l'Economie de l'Education y Centre National de la Recherche Scientifique
gerard.lassibille@u-bourgogne.fr

RESUMEN

Este artículo revisa los estudios internacionales de evaluación más conocidos que se llevan a cabo en el contexto de los países pobres, y pone de relieve la falta de evidencia empírica que existe hoy día acerca del grado de emparejamiento de las pruebas con los programas escolares. A título de ilustración, el artículo evalúa la sensibilidad de un test internacional de conocimientos, comparando los resultados obtenidos por dos cohortes de alumnos matriculados en dos cursos consecutivos, a una misma batería de pruebas. Utilizando el método del score de propensión, para controlar las diferencias que pueden existir en las características de los alumnos y de los docentes de las dos muestras, se concluye que los tests no son muy sensibles y no llegan realmente a apreciar el valor añadido en cada etapa del cursus escolar de los alumnos. Tal resultado cuestiona la validez de los numerosos trabajos empíricos realizados hasta ahora que se apoyan en instrumentos similares para orientar la política educativa de los países pobres y definir opciones de desarrollo supuestamente destinadas a mejorar la eficacia interna de los sistemas de enseñanza.

Palabras claves: tests internacionales de conocimientos; validez de contenido; países en desarrollo.

* Trabajo realizado en el marco del proyecto de excelencia PO9SEJ4859 de la Junta de Andalucía.

INTRODUCCIÓN

Existe una literatura empírica considerable consagrada a los resultados de los sistemas de enseñanza y a la calidad de la oferta de los servicios educativos (ver por ej., Hanushek 1986, 1995; Pritchett y Filmer 1999). Los estudios internacionales de evaluación de las adquisiciones de conocimientos de los alumnos son la piedra angular de la mayoría de estos trabajos, que tienen como doble finalidad el proporcionar diagnósticos objetivos y alimentar el debate de políticas educativas. Estas evaluaciones plantean un gran número de cuestiones de orden metodológico y práctico, ampliamente comentadas en la literatura (ver por ej., Monseur y Demeuse 2004; Greaney y Kellaghan 2008; Winter 1998). Un aspecto al que no le dan suficiente importancia los utilizadores de los estudios internacionales es la calidad de los instrumentos utilizados y su aptitud para considerar realmente las adquisiones escolares. En efecto, aunque los especialistas de la educación, y particularmente los economistas de la educación, se han preocupado de la especificación de los factores que entran en la determinación de los procesos pedagógicos y últimamente han refinado sus métodos empíricos de estimación, sin embargo han prestado poca atención a la calidad de los tests que emplean para evaluar la eficacia de la escuela. Y qué duda cabe que la facultad que tengan estos instrumentos para reflejar los aprendizajes es crucial para quién quiera medir los resultados de los sectores de enseñanza, y poder formular con su apoyo recomendaciones de políticas educativas bien fundadas.

Este artículo se plantea la capacidad que tienen las evaluaciones internacionales llevadas a cabo en los países pobres para medir sin sesgo los resultados de los alumnos. Una revisión de las principales evaluaciones realizadas muestra que estos programas son relativamente poco discutidos, y que se dispone de escasos elementos empíricos que permitan apreciar la aptitud que tienen los tests para reflejar aquello para lo que se idearon (sección 1). Gracias a unos datos de gran originalidad, colectados en Madagascar, verificamos a posteriori, y a título de ejemplo, la aptitud de los instrumentos concebidos para este país, en el marco del Programa de Análisis de los Sistemas Educativos de la CONFEMEN (PASEC), para recubrir los contenidos de los programas escolares. Nuestros resultados se refieren a una cohorte de aproximadamente 16.000 alumnos repartidos en cerca de 1.000 escuelas primarias públicas (sección 2).Estos alumnos, que siguen las clases de tercero y de cuarto, han pasado todos en el mismo momento un test teóricamente concebido para evaluar los conocimientos adquiridos en el curso de cuarto. Los análisis muestran que los instrumentos del PASEC para el país en cuestión están imperfectamente calibrados, y sin duda no permiten evaluar sin riesgo los resultados del sector de enseñanza (sección 3).

Aunque nuestros resultados se refieran a un único país, plantean importantes cuestiones sobre la validez del contenido de las baterías de tests llevados a cabo en los países pobres, y por consiguiente sobre las recomendaciones de políticas educativas que pueden desprenderse de su utilización. Nuestras conclusiones pueden ser útiles para los investigadores que emplean este tipo de evaluaciones en sus trabajos, y también pueden serlo para los especialistas de educación que idean estos instrumentos. Si bien se han realizado análisis de emparejamiento a los curricula para apreciar la validez de ciertos tests efectuados en el marco de los países más avanzados (ver por ejemplo, Beaton et al. 1996), este artículo es el primero en aportar cierta

luz sobre la calidad de una batería de instrumentos concebidos en el contexto de los países pobres.

UNA PANORÁMICA DE LAS EVALUACIONES INTERNACIONALES EN LOS PAÍSES MENOS DESARROLLADOS

La gran mayoría de los países menos ricos están excluidos del campo de evaluaciones PIRLS, PISA y TIMSS. En estos países, se han implantado sistemas de evaluación específicos desde principios de los años 1990, a fin de proporcionar comparaciones internacionales y debates sobre las políticas educativas en las diferentes partes del mundo. Aquí revisamos los principales sistemas, poniendo énfasis sobre la dificultad que hay en juzgar la calidad de estos instrumentos de evaluación.

LLECE

Desde 1998, el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) realiza evaluaciones comparativas que nutren el debate de política educativa en la región. El último programa del LLCE se puso en marcha en 2002 (OREALC 2008) en 16 países y se efectuó a cerca de 200.000 alumnos de primaria repartidos en alrededor de 3.000 escuelas de la región. Los tests realizados en el marco de este programa tratan de evaluar los resultados de los alumnos de 3º y de 6º curso de primaria en lectura/escritura y en matemáticas en los 16 países participantes, así como los conocimientos adquiridos en ciencias en 6º curso para un sub- grupo de 10 países. Según los que concibieron el proyecto, los tests se construyen a partir de un análisis previo de los curricula oficiales en vigor en los países concernidos y se refieren a los elementos que son comunes a éstos. Sin embargo, no hay nada en los documentos publicados o colgados en el sitio web de OREALC/UNESCO que permita verificar empíricamente que eso es así.

PASEC

El Programa de Análisis de los Sistemas Educativos de los Países de la CONFEMEN (PASEC) se inauguró en la 43ª sesión ministerial de la Conferencia de Ministros de Educación de los países que tienen el francés como lengua (CONFEMEN), que tuvo lugar en Djibuti en 1991. Las primeras evaluaciones se realizaron en 1993. Los conocimientos adquiridos por los alumnos se midieron en francés (y/o en la lengua nacional si ésta es la lengua de enseñanza) y en matemáticas. Los tests se dirigieron a alumnos de 2º y de 5º curso de primaria de la enseñanza pública y privada. Contrariamente a la mayoría de los otros programas, los resultados de los alumnos se verificaron en dos fases, al principio y al final del curso escolar. Participaron en el programa 18 países de Africa francófona y del Océano Indico. Los tests se dicen elaborados en referencia a los programas escolares en curso en los países participantes, y supuestamente versan sobre los aspectos comunes a éstos, sin que eso sea en realidad confirmado por ningún análisis empírico.

MLA

El Monitoring Learning Achievement (MLA), administrado de manera conjunta por la UNESCO y la UNICEF, vió la luz a principios de los años 1990 para la Conferencia Mundial sobre la Educación para Todos que tuvo lugar en Jomtiem. El programa trata de crear un sistema de seguimiento permanente de la calidad de la educación y de reforzar la capacidad de los países para conducir evaluaciones periódicas de sus sistemas de enseñanza. Desde 1992, 72 países, la mayoría de rentas bajas, han participado en el programa (39 países de Africa Sub-sahariana, 13 de Africa del Norte y del Medio Oriente, 10 de Europa y de Asia Central, 5 de Asia y del Pacífico, y 3 de América Latina y Caribe. El último programa MLA evaluó los conocimientos en matemáticas, en ciencias y en la vida corriente de los alumnos que asistían al 8º curso. Al igual que otros programas internacionales, los cuestionarios se administraron simultáneamente a los alumnos, a los docentes y a los directores de centros, de manera a identificar los principales factores que determinan los resultados escolares de los alumnos. La construcción de los tests está poco documentada y aparentemente nada permite apreciar la calidad de los instrumentos desarrollados en el marco del programa.

SACMEQ

El Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) reagrupa varios Ministerios de Educación que, con la asistencia del Institut International de Planification de l'Education (IIPE) de la UNESCO, decidieron compartir sus experiencias en el campo de la evaluación de la calidad de la educación (IIPE 2006). El SACMEQ reagrupa 15 países de África anglófona. Se realizaron tres evaluaciones, efectuadas a 5/6 años de intervalo a fin de apreciar los cambios eventuales en el tiempo. Cada una de ellas se dirige a alumnos de 6º curso. Se han publicado regularmente numerosos informes y análisis realizados por los miembros del proyecto; en ellos se presentan para cada país los principales resultados de las evaluaciones y se hacen recomendaciones en materia de desarrollo de la política educativa. Los métodos de muestreo, los procedimientos de encuestas y de recogida de datos están ampliamente documentados y son fácilmente accesibles para los usuarios. Sin embargo, como lo constata el informe de auditoría que ha sido hecho recientemente sobre el SACMEQ (Ercikan et alli 2008), nada permite evaluar el grado de alineamiento de los tests a los programas escolares. Por esta razón, se ha recomendado vivamente a los responsables del programa de efectuar tests de validez de contenido y de publicar sus resultados, a fin de poder juzgar la pertinencia de los instrumentos que se han elaborado.

DATOS

En el marco de un amplio programa de análisis, el Ministerio de Educación de Madagascar realizó en Junio del 2007 un test de 4° curso de primaria a una cohorte de alumnos que estaban inscritos en 3º curso de este ciclo de estudios durante el año escolar 2005-2006, así como a una de 4º curso, y ello independientemente de la clase en que estuvieron en 2006-2007. Comparando los resultados obtenidos por los alumnos de ambas cohortes, se podrá apreciar la

sensibilidad del test e informar ex-post sobre la forma en que se calibran ciertos tests internacionales.

El test toma los items del pre-test que hizo el PASEC en Septiembre 2004 en Madagascar a los alumnos de 5º curso. Entre los 115 items que constituyen el test original del PASEC, se recoge un conjunto de 80 items: 27 de francés, 25 de malgache y 28 de matemáticas. Como muchos otros tests internacionales, el test del PASEC es un test de respuestas múltiples. Cada item comporta cuatro respuestas a elegir; es decir, que la probabilidad que tiene un alumno de obtener una buena nota por el simple azar es relativamente pequeña. Este conjunto de items se pasó en Junio de 2007 a 15.990 alumnos repartidos en cerca de 1.000 escuelas primarias públicas. Entre estos alumnos, 11.611 estaban en 4º curso y 4.379 en 3º. Como máximo se seleccionaron al azar 25 alumnos en cada escuela, a los que se les hizo el test en las tres materias. En cada escuela había alumnos de 3º y 4º curso.

RESULTADOS

Dado que los alumnos que repitieron el curso de 3º en 2006-2007 no tienen forzosamente las mismas características que los que pasaron a 4º curso, utilizamos la técnica del score de propensión para comparar los resultados obtenidos por ambos grupos al mismo test de 4º curso. Con este método, nos aseguramos que los alumnos de 4º (grupo de tratamiento) y los de 3º (grupo de control) difieren solamente porque los primeros siguieron el programa de 4º curso, y no así los segundos. El score de propensión (Rosenbaum y Rubin 1985) se define como la probabilidad condicional de haber seguido el programa de 4º dadas un conjunto de covariables o de características de pre-tratamiento. Siguiendo una práctica corriente, el score de propensión se estima con la ayuda de un modelo probit. La probabilidad de pertenecer a un grupo más que a otro se predice en base a las variables individuales y familiares siguientes: edad y sexo del alumno, nivel de educación del padre y de la madre, y nivel de riqueza del hogar (aproximado por el hecho de que el hogar disponga o no de electricidad). En la medida en que el progreso de los alumnos puede también estar influenciado por las características de los profesores, el status del maestro se incluye entre el conjunto de variables susceptibles de determinar la pertenencia a un grupo más que a otro. Esta variable opone los docentes funcionarios a los contractuales. En Madagascar, como en muchos otros países pobres, el status en el empleo es generalmente una buena proxy del nivel de cualificación del docente. En efecto, en Madagascar los maestros contractuales tienen en promedio un nivel de educación general superior a los otros, cuando su nivel de cualificación profesional es netamente inferior al de los funcionarios. Las dos muestras de alumnos fueron constituidas emparejando individualmente cada alumno tratado a uno no tratado que tuviera el score de propensión más próximo. Cuando un alumno tratado no podía ser emparejado a uno no tratado, se excluía del análisis. Como muestra la Tabla 1, operando de esta manera es posible de constituir un grupo de control que tiene características totalmente similares a las del grupo de tratamiento.

Tabla 1: Características de las muestras no emparejadas y emparejadas de alumnos

	Muestra	Media		% sesgo	% reducción del sesgo	t-test	
		Alumnos de 4° (grupo tratado)	Alumnos de 3 (grupo de control)			t	p > t
Características de los alumnos							
Edad (año)	No emparejada	11,601	11,275	21,1		11,70	0,000
	Emparejada	11,597	11,596	0,1	99,6	0,06	0,950
Niñas (%)	No emparejada	48,437	43,663	9,6		5,40	0,000
	Emparejada	48,559	48,682	-0,2	97,4	-0,19	0,853
Madre alfabetizada (%)	No emparejada	90,224	87,885	7,5		4,28	0,000
	Emparejada	90,26	90,444	-0,6	92,1	-0,47	0,638
Padre alfabetizado (%)	No emparejada	87,339	84,605	7,9		4,53	0,000
	Emparejada	87,571	87,632	-0,2	97,8	-0,14	0,888
Hogar que dispone de electricidad (%)	No emparejada	5,124	4,156	4,6		2,54	0,011
	Emparejada	5,133	4,817	1,5	67,4	1,10	0,273
Características de los docentes							
Funcionarios (%)	No emparejada	51,589	50,24	2,7		1,52	0,128
	Emparejada	51,546	51,546	0,0	100,0	-0,00	1,000

En la medida en que los alumnos de 3º curso no han seguido el programa de 4º, deberían obtener normalmente malos resultados en el test de 4º. ¿Qué muestra la evidencia empírica a este respecto? La Tabla 2 reproduce los porcentajes de respuestas correctas a los diferentes items, calculados para los alumnos de cada nivel de estudios que componen la muestra emparejada. Hay varios hechos que merecen señalarse. Como se puede esperar, los alumnos de 3º tienen en media peores resultados que sus homólogos. Sin embargo, las diferencias entre cada categoría de alumnos son relativamente pequeñas. En matemáticas y en malgache, los alumnos de 4º tienen una ventaja sobre los de 3º de 18 y 14% respectivamente; en francés, esta diferencia no es más que del 4%. Si se considera ahora los scores combinados en las tres materias, los alumnos de 4º tienen en promedio una ventaja de 12 puntos solamente sobre los alumnos de 3º curso.

Tabla 2: Respuestas correctas al test de 4°

	Número de alumnos	Respuestas correctas (%)			
		Matemáticas	Francés	Malgache	Total
Alumnos de 3°	4.249	35,8	24,8	34,3	31,6
Alumnos de 4°	11.417	54,7	29,2	48,0	44,0
Diferencia	—	18,9*	4,4*	13,7*	12,4*

* = significativo al 1%.

El Gráfico A1 del anexo ilustra mejor las diferencias de los resultados a los tests, al reproducir la distribución de los porcentajes de respuestas correctas obtenidas por las dos categorías de alumnos en las tres materias. Como se puede apreciar claramente las diferentes distribuciones se solapan de forma considerable, dejando pensar a priori que el test es poco sensible al contenido de los programas (ver la discusión abajo). En matemáticas y en malgache, se estima que cerca del 60% de las dos distribuciones se solapan. La cobertura es aún mayor en francés, puesto que según nuestros cálculos el 80% de las dos distribuciones se solapan[1].

¿Cómo pueden explicarse estos resultados tan intrigantes? Cada alumno de 4º curso se emparejó con uno de 3º que tenía el mismo score de propensión. Como lo mostró la Tabla 1, el emparejamiento que se hizo es de buena calidad y las dos categorías de alumnos tienen características prácticamente idénticas. Es decir, que las diferencias de resultados que observamos entre los alumnos de 3º y los de 4º no se pueden explicar por efectos de alumnos o de profesores, que serían aleatorios. Más simplemente, se podrían atribuir los buenos resultados de los alumnos de 3º a la suerte que tuvieron cuando hicieron el test. Sin embargo, esta hipótesis no es plausible. Como se ha indicado, cada item de elección múltiple del PASEC comporta en promedio cuatro respuestas posibles, y por consiguiente la probabilidad de obtener un buen resultado en el test sin ningún conocimiento previo es muy pequeña. Podrían explicar en parte los resultados obtenidos posibles diferencias en dotaciones de las escuelas en recursos físicos y pedagógicos directamente ligados a la enseñanza en clase. Pero tampoco esta hipótesis es verosímil. Como muestra la Tabla 3, los dos grupos de alumnos asisten a centros de características similares. En cada sub-muestra, las escuelas están dotadas en promedio de un mismo ratio alumnos-profesor. En cuanto a los recursos pedagógicos, los alumnos de 3º y los de 4º disponen del mismo número de manuales escolares. También la calidad de las infraestructuras escolares es parecida en cada sub-muestra de alumnos.

Tabla 3: Recursos de las escuelas. (media)

	Alumnos de 3°	Alumnos de 4°	Diferencia
Ratio alumnos/docente	43,9	43,7	-0,2 (0,5)
Libros por alumno	1,0	1,0	0,0 (0,0)
Indice de equipamiento de la escuela[a/]	99,8	100,9	1,1 (1,5)

Nota: Desviaciones estándar entre paréntesis

a/ La calidad de los equipamientos se mide a través de un índice construido por medio de un análisis de componentes principales; el índice toma en cuenta los elementos siguientes: la estructura de la escuela es permanente, el número de clases es suficiente, la escuela dispone de electricidad, de agua, de letrinas y de sillas para todos los alumnos. El índice varía entre 126 para las escuelas que están dotadas de todos estos equipamientos, y 75 para las que no disponen de ninguno.

[1] En la medida en que la mayoría de los tests se refieren a grandes conjuntos de conocimientos, las distribuciones se solapan siempre. Sin embargo, en este caso, y es sobre lo que queremos insistir, el solapamiento es muy grande.

Los buenos resultados obtenidos por los alumnos de 3º podrían explicarse también por el hecho de que la enseñanza dada en clase de 4º no esté alineada con el programa de 4º. Sin embargo, no hay ninguna razón objetiva para pensar que los docentes de 4º curso siguen los programas de manera menos estricta que sus compañeros de 3º. Además, los responsables de las sub-circunscripciones escolares inspeccionan frecuentemente las escuelas en Madagascar y siguen de cerca los aspectos ligados a la pedagogía. Según la encuesta efectuada por el Ministerio de Educación a las escuelas implicadas en el test, aproximadamente el 90% fueron inspeccionadas durante el año previo a la encuesta; alrededor del 75% de estas inspecciones tuvieron por objeto aspectos pedagógicos.

Con toda verosimilitud, los buenos resultados que obtienen los alumnos de 3º curso al test de 4º se explican por el hecho de que este test está mal calibrado y que su aptitud para recoger los aspectos para los que teóricamente fue concebido es débil. En otros términos, todo hace pensar que el test está imperfectamente alineado al contenido de los programas escolares enseñados en clase de 4º curso, y que apenas permite apreciar los conocimientos que han sido adquiridos realmente en esta etapa del sistema de educación primaria. Ciertamente, se podría objetar a esta conclusión que a) algunas nociones se enseñaron a la vez en los dos niveles de estudios, naturalmente el test de 4º puede incluir ciertos aspectos del programa de 3º; b) los alumnos de 3º – al menos los mejores – pueden responder correctamente a algunos items del test de 4º. Si estos argumentos explican bien porqué ciertos alumnos de 3º pueden obtener resultados tan buenos – ver mejores – que otros alumnos de 4º, ciertamente no pueden justificar el solapamiento tan importante (hasta el 80%) que se observa en las distribuciones de las respuestas correctas aportadas por las dos categorías de alumnos.

CONCLUSIÓN

En este artículo hemos pasado revista brevemente a las evaluaciones realizadas en los países en desarrollo o de rentas intermedias, que en la gran mayoría de casos están excluidos del campo de las evaluaciones « gold standard » como son PISA, TIMSS o PIRLS. Esta revisión ha mostrado que la construcción de instrumentos de evaluación en estos países está muy mal documentada y que nada permite apreciar la aptitud de los tests para aprehender los aspectos para los que han sido teóricamente concebidos. Incluso los programas más elaborados no se preocupan de verificar que los instrumentos estén alineados a los curricula, los utilizadores de estas evaluaciones deben contentarse con los comentarios publicados en los informes técnicos que aseguran, pero sin justificarlo, que los tests han sido construidos después de hacer análisis detallados de los programas y que coinciden con el contenido de los mismos.

Gracias a unos datos inéditos, hemos verificado ex-post la sensibilidad de un instrumento concebido en el marco del PASEC. Con la ayuda del método del score de propensión, hemos comparado los resultados obtenidos por alumnos de 3º y de 4º curso de primaria a un test de 4º curso que se realizó a todos al final del mismo año escolar. Los resultados han mostrado un solapamiento hasta del 80% de las distribuciones de las respuestas correctas aportadas por las dos categorías de alumnos. El hecho de que un número tan grande de alumnos de 3º curso obtengan tan buena nota en un test concebido para medir los conocimientos adquiridos a la

terminación de un programa de 4º plantea importantes cuestiones sobre la validez del contenido del test. Nuestras conclusiones ponen profundamente en cuestión los resultados de un gran número de trabajos empíricos que analizan la eficacia interna de los sistemas de enseñanza apoyándose en este tipo de datos, y plantean serias dudas sobre las recomendaciones en materia de desarrollo de la política educativa que han podido formularse a partir de ellas. Como nuestros resultados se limitan a un sólo país en desarrollo, naturalmente no son generalizables a otros contextos, o a otros instrumentos internacionales de evaluación. Sin embargo, hay dos lecciones que sacar del test que hemos efectuado aquí. Por una parte, los economistas de la educación, que utilizan de manera intensiva las evaluaciones internacionales en sus trabajos, deberían ser más prudentes a la hora de interpretar sus resultados. Por otra parte, los ideadores de tests deberían asegurarse de que sus instrumentos están bien emparejados con las intenciones de los curricula. A fin de evitar dudas y errores que pueden tener graves consecuencias para la definición de las políticas educativas, deberían hacerse sistemáticamente estudios de alineamiento y sus resultados deberían publicarse y ser accesibles a los utilizadores potenciales de estas baterías de tests.

REFERENCIAS

Beaton, A.E., Martin, M.O., Mullis, I.V.S, Gonzalez, E.J., Smith, T.A., y Kelly, D.L. (1996). *Science Achievement in the Middle School Years: IEAs' Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College, TIMSS International Study Center.

Ercikan, K., Arim, R., Oliveri, M. y Sandilands, D. (2008). Evaluation of Dimensions of the Work of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and of its Programme of Cooperation with the International Institute for Educational Planning (IIEP). Paris: UNESCO.

Greaney, V. y Kellaghan T. (2008). *Assessing National Achievement Levels in Education*. Washingthon DC: The World Bank..

Hanushek, E. (1986). The economics of schooling: production and efficiency in public schools. *Journal of Economic Literature*, 24, 1141–77.

Hanushek, E. (1995). Interpreting recent research on schooling in developing countries. *World Bank Research Observer*, 10 (2), 227–246.

IIPE (Institut International de Planification de l'Education de l'UNESCO) (2006). Lettre d'information de l'IIPE, n°1, janvier-mars.

Monseur, C. y Demeuse, M. (2004). Quelques réflexions méthodologiques à propos des enquêtes internationales dans le domaine de l'éducation. *Politiques d'Education et de Formation*, 11(2), 37-54.

OREALC (Oficina Regional de Educación para América Latina y el Caribe) (2008). Primer Reporte de Resultados del Segundo Estudio Regional Comparativo y Explicativo (SERCE). Santiago: UNESCO.

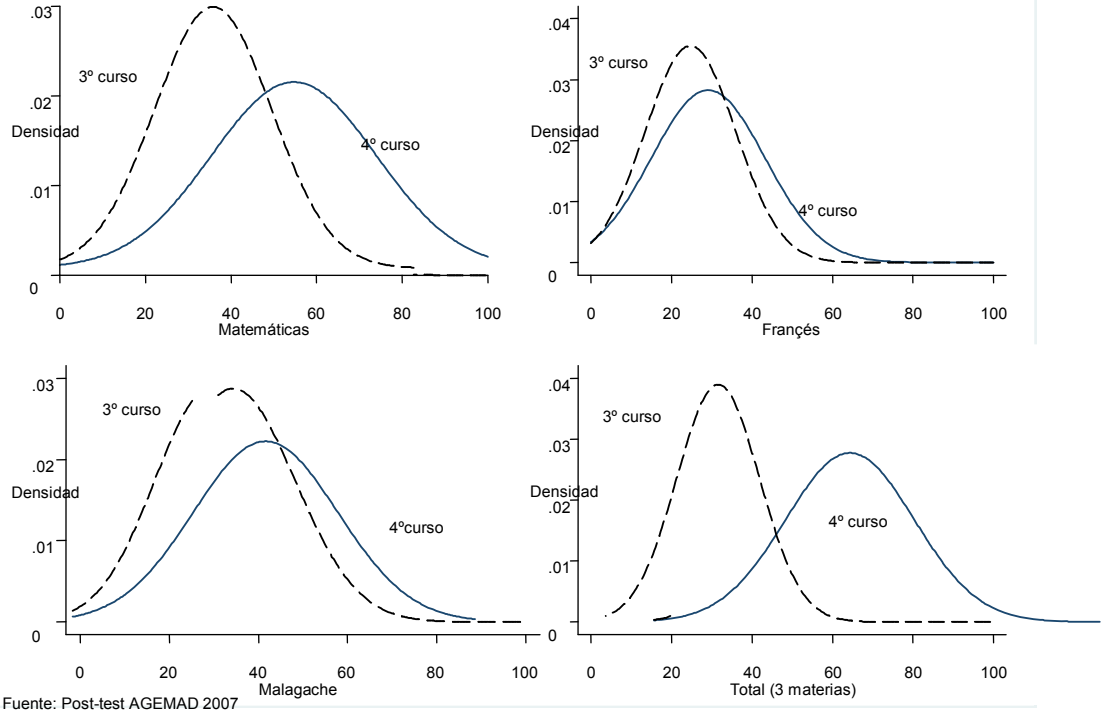
Pritchett, L. y Filmer, D. (1999). What education production functions really show: a positive theory of education expenditures. *Economics of Education Review*, 18 (3), 223–239.

Rosenbaum P.R. y Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39 (1), 33–38.

Winter, S.J. (1998). International comparisons of student achievement: can they tell us which nations perform best and which education systems are the most successful? *Education 3 to 13*, 26(2), 26-32.

ANEXO

Gráfico A1: Distribución de los % de respuestas correctas, por curso y por materia



Fuente: Post-test AGEMAD 2007