Banco Mundial (1993) *El financiamiento de la educación en los países en desarrollo. Opciones de política.* Washington D.C.

Colom Cañelas, A. (1988) "La calidad de la educación desde la teoría pedagógica y la historia", en: Bordón, *Revista de orientación pedagógica*, Madrid, Sociedad Española de Pedagogía.

da Silva, T. (1996) "O projeto educacional da nova dereita e a retórica da qualidade total", en: da Silva, T. y Gentili, P. (comps.) *Escola, S.A.*, Brasilia, SENTE.

Edwards, V. (1991) *El concepto de calidad de la educación*, OREALC, Santiago de Chile.

Elliott, J. (1994) "The teacherís role in curriculum development: an unresolved issue in English attempts at curriculum reform", en: *Curriculum Studies*, Vol.2, N°2.

Fernández Lamarra, N. (1991) "Planificación y calidad de la educación: una perspectiva argentina y latinoamericana", en: AAVV, *Calidad de la educación. Aportes para un debate desde la perspectiva del planeamiento*, Buenos Aires, MCE/BIRF-PNUD.

Foucault, M. (1985) *Vigilar y castigar*, Siglo XXI, México.

Gentili, P. (1994) *El discurso de la "calidad" como nueva retórica conservadora en el campo educativo*, Buenos Aires; CEAL.

Graham, A. (1994) "Consumer choice in education: whatís wrong with parents rights?", en: *Curriculum Studies*, Volume 2, Number 1.

Lapointe, A.; Mead, N. y Philips, G. (1989) *Un mundo de diferencias*, CIDE, Madrid.

Mena, M. (1991) "Conceptualizaciones sobre calidad y variables utilizadas en las evaluaciones efectuadas en América Latina", en: AAVV, *Calidad de la educación. Aportes para un debate desde la perspectiva del planeamiento*, Buenos Aires, MCE/BIRF-PNUD.

Narodowski, M. (1995) *Infancia y poder. La conformación de la pedagogía moderna*, Buenos Aires, Aique.

OCDE (1991) *Escuelas y calidad de la enseñanza*, Madrid, Paidós-Ministerio de Educación y Ciencia.

Popkewitz, TH. (1994) *Sociología política de las reformas educativas*, Madrid, Morata.

Rodríguez Fuenzalida, E. (1994) "Criterios de análisis de la calidad en el sistema escolar y sus dimensiones", en: *Revista Iberoamericana de Educación*, Madrid, N°5, mayo-agosto.

Tenti, E. (1991) "La calidad de la educación como un problema; una lectura sociológica", en: AAVV, *Calidad de la educación. Aportes para un debate desde la perspectiva del planeamiento*, Buenos Aires, MCE/BIRF-PNUD.

Van Vught, Franz (1991) "La calidad de la educación superior en Europa: el siguiente paso", en: *Universidad futura*, vol.3, números 8 y 9, invierno de 1991.

Wilson, J. (1992) *Cómo valorar la calidad de la enseñanza*, Barcelona, Paidós/MEC.

Aportes para su análisis*

ESTELA COLS**
LAURA BASABE***
En colaboración con:
CLAUDIA BROITMAN

La evaluación de los sistemas educativos a escala nacional: dimensiones de abordaje

Evaluar es, en cualquiera de sus niveles y ámbitos de desarrollo, un proceso de complejidad considerable, pues, en su misma constitución, se conjugan lo técnico y lo valorativo. Se trata, como muchos otros problemas propios de la tarea formativa, de una actividad en la que nos enfrentamos, desde el inicio, con problemas de carácter "práctico", en el sentido que Schwab le otorga a la expresión. Los problemas prácticos, según el autor, siempre refieren a situaciones concretas, particulares en las que se carece de una guía o regla única y definida y ante las cuales es preciso, entonces, establecer juicios de valor de carácter contextual y comparativo y tomar decisiones basadas en procesos deliberativos (Schwab, 1973). En esta línea de pensamiento, House define a la evaluación como una compleja combinación de procedimientos de decisión de carácter eminentemente moral (House, 1997).

Estas características describen claramente la naturaleza de las diversas situaciones en las que puede llevarse a cabo la evaluación: tanto si se trata de evaluar los procesos de aprendizaje y de enseñanza a nivel del aula en un marco interpersonal como cuando el propósito es efectuar una evaluación de la calidad de todo un sistema educativo a escala nacional. Pero, cada una de estas instancias tiene, además, rasgos que le son propios y que definen los problemas específicos con los que el evaluador se encuentra y las vicisitudes que caracterizan el desarrollo de todo el proceso.

Cuando lo que se evalúa es la calidad de un programa o un sistema educativo, la complejidad inherente al proceso evaluativo se incrementa al convertirse éste en una cuestión pública estrechamente vinculada con el orden político-institucional, cuyo interés es colectivo y su impacto social. Estas notas configuran un campo de problemáticas y tensiones que es preciso conocer y atender en su especificidad, generando análisis y respuestas que impliquen la consideración de las diferentes dimensiones involucradas.

El panorama que se nos presenta, sin embargo, revela que, si bien esta modalidad de evaluación aumenta día a

* Este artículo ha sido posible gracias al subsidio otorgado a la investigación dirigida por Alicia Camilloni por la Secretaría de Ciencia y Técnica de la UBA (F215).

** Licenciada en Ciencias de la educación y Profesora para la Enseñanza Primaria. Docente de la Cátedra de Didáctica I del Departamento de Ciencias de la Educación de la Facultad de Filosofía y Letras, UBA. Miembro del Programa de Instituciones Educativas del Instituto de Investigaciones en Ciencias de la Educación, Filosofía y Letras, UBA.

*** Licenciada en Ciencias de la Educación. Docente de la Cátedra de Didáctica I del Departamento de Ciencias de la Educación de la Facultad de Filosofía y Letras, UBA.

día, este crecimiento no tiene su correlato en una profundización de la comprensión de los fenómenos y procesos en ella implicados. En palabras del mismo House:

> "El panorama actual de la evaluación se caracteriza por la vitalidad y el desorden. La escala a la que se llevan a cabo las actividades de evaluación, su omnipresencia y diversidad hacen difícil su comprensión, incluso a quienes se mueven en este campo. Más alarmante aún es el hecho de que una mala evaluación puede deteriorar un programa social y ocasionar perjuicios a toda una clase social. La importancia social de la evaluación es enorme; su interpretación, relativamente insignificante." (House,1997: 17)

El trabajo que presentamos procura contribuir al campo con un análisis basado en material empírico y realizado desde una perspectiva técnico-didáctica. Se enmarca en la línea de estudios que viene desarrollando el equipo de investigación de la Cátedra de Didáctica I de la Facultad de Filosofía y Letras (UBA)[1] acerca de la calidad de los programas de evaluación. El interés específico de indagación se centra en la descripción y evaluación de algunos aspectos ligados a la instalación de mecanismos de evaluación a escala nacional por parte de los organismos centrales del Sistema Educativo. En el caso argentino, este hecho se produce en 1993 a partir de la implementación del Sistema Nacional de Evaluación de la Calidad (SNEC)[2] como uno de los elementos centrales de las macropolíticas vinculadas con el sector.

Si se tiene en cuenta la índole particularmente compleja de la evaluación como proceso de intervención social, puede advertirse con claridad que su estudio implica, o al menos puede implicar, la consideración de más de una dimensión de análisis. Mencionemos aquí sólo algunas de ellas que, por su relevancia, merecen ser destacadas.

En primer lugar, y en relación con lo planteado anteriormente, el caso puede ser legítimamente planteado en términos políticos, en tanto se trata de un sistema de evaluación de la calidad a escala nacional que forma parte de un conjunto de políticas de gobierno en función de las cuales cobra sentido y adquiere cierta legalidad. Existen algunos desarrollos teóricos recientes pertenecientes a este enfoque que focalizan su interés en la descripción e interpretación del marco más general de cuestiones de orden político y social en el que este fenómeno se inscribe[3].

En este sentido, recordemos que la preocupación por la evaluación de la calidad no es un fenómeno nuevo en nuestros países, dado que estuvo presente como elemento esencial dentro del movimiento de promoción de los procesos de planeamiento institucional de la década de los sesenta y setenta. Sin embargo, es interesante señalar que se han producido modificaciones a través del tiempo en el status otorgado a los sistemas de evaluación en el conjunto de los mecanismos de decisión política. En efecto, puede afirmarse que el papel que actualmente se le atribuye excede en gran medida al de décadas anteriores, en el que constituía un mecanismo auxiliar o de apoyo de otros niveles de decisión y definición de políticas a mediano y largo plazo.

Si bien, entonces, se continúa sosteniendo un discurso de planificación racional, la diferencia está en el valor sustantivo atribuido a la evaluación, que asume hoy una función supervisora y reguladora expost (Camilloni, 1994).

En segundo término, y ligado a este punto, también es posible efectuar un análisis priorizando la dimensión institucional del problema, considerando en ese caso al SNEC en tanto dispositivo instituido de evaluación con carácter regulador de las prácticas de enseñanza y evaluación. De acuerdo con Barbier, el Programa en su conjunto puede definirse como un caso de evaluación instituida, es decir, un "acto deliberado y socialmente organizado dirigido a generar juicios de valor" (Barbier, 1993: 37). Como tal, presenta ciertas características propias y tiene por principal función la de objetivar el proceso de evaluación, realizando la operación con cierta independencia en relación con los actores que la practican. En este último sentido, el impacto -y principalmente la naturaleza de este impacto- que una medida de esta naturaleza pueda tener en los diversos agentes del sistema educativo constituiría un tema central dentro de esa perspectiva de análisis.

Desde un planteo más organizacional, podríamos preguntarnos, por ejemplo, por el tipo de funcionamiento y dinámica que la instauración de prácticas de este tipo puede generar en los establecimientos escolares y por las nuevas fuentes de tensión que puede contribuir a instalar. De igual modo, el seguimiento del proceso desde una perspectiva histórica podría abrir una línea interesante de indagación centrada en la identificación, la descripción y el análisis de las formas específicas que asume esta regulación, por ejemplo, al nivel de los modelos de desempeño docente y de las prácticas de enseñanza y de evaluación[4].

La necesidad y el lugar de un análisis técnico de la cuestión

Hasta el presente, los planteos y las opiniones vertidas relativos al SNEC por parte del propio sistema o de especialistas se han situado principalmente en perspectivas predominantemente políticas, siendo escaso el desarrollo de estudios que incluyan la dimensión técnica del problema. En tal sentido, la realización de este trabajo se vincula principalmente con el reconocimiento de la importancia que reviste el estudio de estos aspectos no sólo al nivel del Programa en general, sino también, en lo relativo a los instrumentos que lo componen; porque sólo en la medida en que ello estén adecuadamente resueltos pueden comenzar a plantearse las cuestiones y los interrogantes vinculados con las posibilidades de aprovechamiento de la información que un programa de este tipo produce.

Sin embargo, queda claro que el análisis técnico no puede efectuarse en abstracto, desgajado del marco político y pedagógico general que lo contiene. Ello significa que en la valoración de procesos de esta naturaleza es preciso incluir criterios que excedan los parámetros clásicos instituidos por los enfoques positivistas o ampliar su significación

dotando de nuevos sentidos a las tradicionales categorías de análisis. En esta línea de pensamiento, pero profundizando aún más la cuestión, House enfatiza la necesidad de incorporar en los estudios de este tipo principios éticos y políticos como los de justicia y verdad. Sin duda, la reflexión sobre estas cuestiones, lejos de excluir la posibilidad del análisis técnico, permite completarlo, abriendo líneas de interpretación más globales.

En síntesis, y teniendo presentes estas consideraciones, este estudio se propone, además de reconstruir el marco político y pedagógico en el que surge el SNEC -cuestión que es objeto de otro trabajo[5]- focalizar el análisis en la dimensión técnica del asunto, describiendo las estrategias utilizadas y evaluando la calidad técnica de los instrumentos[6].

Esta última tarea supone, en principio, establecer cuáles serán los criterios o parámetros teórico-técnicos a tener en cuenta en el análisis. Ello, a su vez, está condicionado por el modo en que se conceptualice el problema y por las dimensiones consideradas.

Un primer aspecto a considerar, en este sentido, es el hecho de que por las características mismas del objeto de estudio, es posible situar el análisis en por lo menos dos niveles. Justamente, por tratarse de un Sistema Nacional constituido por Operativos que se llevan a cabo año tras año, pueden, por ejemplo, estudiarse algunos aspectos técnicos del programa en su conjunto, entendiéndolo como la aplicación sucesiva de pruebas de rendimiento, pero puede también considerarse la calidad técnica de los instrumentos empleados en cada operativo en particular. En este caso, lo que se está evaluando, en sentido estricto, es el valor de los resultados obtenidos en el proceso de medición del rendimiento.

Desde esta perspectiva, es central el análisis del concepto de calidad que el Sistema sustenta. Se trata de un término sobre el que abunda bibliografía específica y en virtud del cual se han desarrollado un sin número de trabajos de índole teórico y empírico. Es, por otra parte, un término polisémico y, el hecho de definirlo es, en sí, una de las decisiones políticas, teóricas y técnicas centrales de un Programa que tiene por objeto su evaluación. El análisis del material contenido en los documentos permite señalar, en relación con este punto, que, en este caso, se ha efectuado una significativa reducción del concepto a una de sus dimensiones, esto es, a la medición de los logros académicos de los estudiantes a través de la cantidad de respuestas correctas alcanzadas en las pruebas objetivas de rendimiento. Si bien es cierto que se diseñaron instrumentos de carácter complementario para obtener información acerca de otros aspectos asociados al rendimiento, ellos han adquirido, progresivamente un carácter secundario en el conjunto del programa, tanto en lo que se refiere a la cantidad de casos con los que se trabajó como en lo que respecta a la difusión de los resultados del análisis.

En estrecha relación con la conceptualización y dimensionalización de las variables principales, una cuestión capital desde el punto de vista metodológico, pero que es indudable y primeramente un problema de carácter político y teórico, es el de la selección de un diseño para llevar a cabo la evaluación de la calidad. Ello implica elegir una estrategia, definir dimensiones y variables relevantes y buscar indicadores válidos para los conceptos que se desean medir -en este caso, la calidad educativa. Además de la operacionalización de los conceptos, la definición de un plan de recolección de la información y de un plan de análisis.

Es por ello que Barbier señala que un punto crucial en todo proceso evaluativo es el constituido por las operaciones relativas al proceso de representación. En la medida en que la evaluación es un proceso que se inicia con un material de trabajo (la realidad o las realidades sobre las cuales o a partir de las cuales se efectúa un proceso de transformación) y culmina en un producto (la realidad nueva aparecida al término de la actividad o del proceso de transformación y que puede considerarse como el resultado específico de esa actividad o ese proceso), puede ser considerada como "un proceso de transformación de representaciones, cuyo punto de partida sería una 'representación factual' de un objeto y el punto de llegada una 'representación normalizada' (se puede hablar también de juicio de valor) de ese mismo objeto..." (Barbier, 1993: 64).

Los datos de referencia de la evaluación constituyen este componente factual, que puede definirse, siguiendo al autor, como aquello a partir de lo cual se produce el juicio de valor. En cuanto a su status o naturaleza, los datos de referencia pertenecen al orden de las representaciones de los hechos. En este sentido y en la medida en que estos datos no nos son dados y no existen en estado natural, decíamos que un punto crítico en este proceso se refiere justamente a la constitución de los datos de referencia, dentro del cual pueden recortarse, a su vez, dos conjuntos de operaciones : el proceso de producción de indicadores del objeto o de la realidad evaluada, y la producción de informaciones propiamente dichas para la evaluación, es decir, la cuestión de las técnicas de evaluación.

La validez: una cuestión central

Cuando se trata de estimar la calidad técnica de una prueba o programa y de los resultados obtenidos, la condición de validez tiene una relevancia central puesto que define la pertinencia y la adecuación del dispositivo a los propósitos del proceso de evaluación. Aún más, es posible afirmar que la validez es el determinante central, básico de la calidad de una evaluación.

Existen diferentes modos de concebir la idea de validez, dependiendo de cuál sea el enfoque en el que el evaluador se sitúa, pero, en términos muy generales, el concepto está asociado a la posibilidad de que una evaluación posea o no la cualidad de " merecer el reconocimiento" -en términos de House-. De algún modo, la validez da cuenta de los principios que permiten dar legitimidad a un proceso de evaluación específico.

Simplificando quizá demasiado la cuestión podemos decir que existen posiciones de corte objetivista, sustentada por quienes asumen que la validez descansa en los

principios de legitimación provistos por la tradición empírica y el modelo de pensamiento propio del método científico -más específicamente, de las ciencias naturales-. Como contraparte, encontramos definiciones de validez subjetivistas, de carácter fenomenológico, en las que ésta se fundamenta apelando a las experiencias y percepciones personales o colectivas. En distintos puntos de este continuum se ubican las distintas posturas o enfoques que los evaluadores o teóricos de la evaluación pueden adoptar en relación con este punto[7].

Más allá de estas diferenciaciones de orden epistemológico, existen algunos aspectos inherentemente vinculados al requisito de validez que aparecen de modo recurrente en los trabajos sobre el tema y que es oportuno retomar aquí.

De acuerdo con Ebel, el término designa el grado de exactitud con que un conjunto de puntajes de prueba miden lo que tendrían que medir. (Ebel, 1977). Podríamos decir que un instrumento es válido en la medida en que mide aquello para lo cual ha sido construido. De modo semejante, Gronlound señala que la validez determina el punto hasta el cual los resultados de un proceso de evaluación sirven para aquellos usos particulares para los que se obtuvieron. Afirma, además, que "la validez se refiere siempre al uso específico que ha de hacerse de los resultados y al grado de veracidad de nuestras interpretaciones propuestas." (Gronlund, 1973: 86). Por esta razón, no se trata de una cualidad de tipo general, sino que, es siempre un atributo específico en función de los propósitos en virtud de los cuales se hace uso de un conjunto de resultados. De ello se desprende que la validez nunca puede ser absoluta sino que se trata de una cuestión de grado. Tal como explica Ebel, las pruebas no son válidas o no válidas sino que pueden ser más o menos válidas (Ebel, 1977).

La estimación de la validez de un instrumento es una tarea compleja en gran medida porque requiere de la construcción de un juicio basado en la combinación de varios criterios. Además, la validez de la prueba no es determinada exclusivamente por la prueba misma: depende de la finalidad para la que se la utilice, el grupo con el cual se la usa y la manera como se la administra y puntúa.

El mismo hecho de que el grado de validez de un programa o instrumento dependa de criterios externos en función de los cuales se juzga su adecuación, impone la necesidad de que éstos también se legitimen a través de algún proceso de validación a través de criterios más amplios. Llegados a este punto del planteo, es interesante el aporte de Camilloni, quien propone a la teoría didáctica como fuente para sustentar la validez de las decisiones tomadas en el proceso de evaluación:

> "En la Didáctica, ese proceso de validación al infinito podría ser reemplazado por una teoría de bases firmes, tanto desde un punto de vista científico como filosófico. [...] No se trata, en consecuencia, de validar instrumentos y criterios en un juego de permanente retroceso hacia los principios, sino de apoyarse en principios didácticos fuerte, con cierto grado de generalidad y que permitan fundamentar racionalmente las decisiones de diseño del programa y de los instrumentos." (Camilloni, 1998: 77).

Este conjunto de supuestos teóricos señalará los principales sentidos atribuidos a la idea de validez. Porque cuando se habla de validez, puede estar haciéndose referencia a distintos tipos o modalidades, en función de cuál sea el criterio tenido en cuenta . De este modo, es posible hablar, por ejemplo, de validez de contenido, validez predictiva, validez de convergencia, validez de retroacción, validez de significado, validez manifiesta[8].

El relevamiento de información realizado a través del SNEC estuvo estructurado sobre la base de la administración de dos tipos de instrumentos: pruebas de rendimiento académico y encuestas. Estas últimas estaban dirigidas a docentes, directivos, alumnos y padres y procuraban obtener material referido a una serie de variables posiblemente asociadas con el rendimiento. Los documentos indican que las encuestas constituyen instrumentos de carácter secundario o complementario, convirtiéndose las pruebas, entonces, en el dispositivo central.

Específicamente, nuestro trabajo de análisis técnico[9] se focalizó en los instrumentos de carácter principal, es decir, las pruebas objetivas de rendimiento, sobre las que descansa todo el dispositivo técnico de evaluación de los logros. Por su propia naturaleza, las condiciones de validez de estos instrumentos deben ser juzgadas de acuerdo con los criterios propios de la tradición objetivista antes mencionada.

Entre los múltiples significados atribuidos a la validez, se ha priorizado el estudio de la validez de construcción, en la medida en que ella condiciona cualquier otro tipo de validez que pueda plantearse. Ella está asociada a dos tipos de consistencia: por un lado, la coherencia entre el dispositivo técnico y los principios teóricos que sostienen un proyecto pedagógico, y, por otra parte, la coherencia del programa y los instrumentos con las normas y criterios teóricos y técnicos. Existen una serie de recaudos técnicos referidos a cada tipo de instrumento de evaluación y su observancia contribuye, ciertamente, a incrementar la validez en la medida en que se evitan problemas de ambigüedad, vaguedad, redacción inadecuada, etc..

En el caso del SNEC, dentro de los diversos tipos de instrumentos de evaluación de los resultados del aprendizaje, la decisión adoptada para medir el rendimiento, fue la administración de pruebas objetivas de selección múltiple. Se trata de un formato de prueba específico que reúne ciertas ventajas evidentes en lo que se refiere a los aspectos prácticos de la administración, la corrección y la puntuación posibilitando -por la naturaleza misma del ítem- un control objetivo de estos aspectos. Pero se trata también de ítems cuya construcción reúne un número importante de requisitos técnicos, que deben ser adecuadamente atendidos, si se desea garantizar la validez y confiabilidad de los resultados.

CUADRO 1:
Cantidad de ítems observados en Matemática y Lengua. Operativos 1993 y 1994.

	MATEMÁTICA		LENGUA	
	1993	1994	1993	1994
Cantidad de ítems observados	29	31	33	28
Porcentaje de ítems observados	72,5	77,5	82,5	70

Cantidad total de ítems: 40

Es por eso que, la cuidadosa redacción de los ítems de una prueba de opción múltiple y el respeto por ciertos criterios teórico-técnicos presentes en la literatura referida al tema, es un aspecto que no puede quedar fuera de consideración si se desea estimar la pertinencia del instrumento y avanzar en la identificación de las posibles fuentes de invalidación de los resultados.

El trabajo apuntó, entonces, a realizar una descripción y evaluación de este aspecto en cuatro de la serie de pruebas disponibles correspondientes al nivel primario: Matemática y Lengua, Operativos 1993 y 1994.

Se llevó a cabo el análisis de cada uno de los ítems que componen la prueba, en función de un patrón de referencia teórico elaborado previamente para tal fin, constituido por una serie de pautas relativas a la construcción de ítems objetivos de selección múltiple[10]. Cada uno de los ítems fue evaluado en función de los distintos criterios establecidos y, como resultante de este paso, se identificó cuáles eran los ítems de prueba que tenían observaciones en uno o varios de los principios técnicos de construcción.

Con respecto al área de Matemática, se ha podido constatar que, en el caso de la prueba correspondiente al Operativo 1993, 29 de los 40 ítems que la integran sufrieron algún tipo de observación, es decir que el 72,5% de los ítems presenta algún tipo de problema de construcción. Tal como puede verse en el Cuadro 1, la cantidad de ítems con observaciones asciende a 32 en la prueba de Matemática tomada en 1994, lo que representa el 75% de los ítems que la componen.

En el caso de Lengua, en cambio, mientras que en el Operativo 1993 aparecen 35 ítems con al menos una observación en cuanto a su construcción -cifra que representa el 87,5 % del total de ítems de la prueba-, en 1994 ese valor desciende a 28 ítems, es decir, el 70% del total.

Como puede advertirse, las modificaciones registradas en las dos áreas tienen sentido opuesto y distinta magnitud. En el área de Lengua, la tendencia consiste en una disminución de los ítems defectuosos, ya que en el 94 hay un total de cinco ítems mejor construidos que en el 93. En Matemática, en cambio, se observa en el segundo Operativo un leve aumento del número de ítems con problemas de construcción -la diferencia es de sólo dos ítems-. Considerando que en Lengua el número inicial de ítems observados es muy alto, la disminución de ítems con problemas de construcción en Lengua conduce a un nivel más o menos equivalente entre las áreas en el 94.

CUADRO 2:
Cantidad de observaciones por ítem en Matemática y Lengua. Operativos 1993 y 1994.[11]

ITEM N°	MATEMATICA		LENGUA	
	1993	1994	1993	1994
1	1	2	1	1
2	1	1	1	1
3	1	1	2	0
4	1	2	3	3
5	0	1	1	1
6	3	1	2	2
7	1	2	1	1
8	1	0	1	2
9	1	1	3	2
10	1	2	1	0
11	3	1	0	0
12	0	1	3	0
13	2	1	2	1
14	2	1	1	1
15	0	1	3	2
16	1	3	1	2
17	1	3	2	2
18	1	1	0	0
19	1	0	3	0
20	3	2	1	1
21	2	0	2	2
22	2	1	4	3
23	5	0	1	0
24	0	1	0	1
25	1	1	1	1
26	0	1	3	2
27	0	1	3	1
28	4	0	2	1
29	1	2	1	1
30	2	0	2	2
31	2	0	0	1
32	1	1	1	0
33	0	1	1	1
34	0	1	1	2
35	1	1	1	1
36	0	1	0	2
37	2	0	0	0
38	0	2	0	0
39	1	1	1	0
40	0	0	1	0
Total observ.	49	42	59	43

De todos modos, más allá de estas modificaciones, encontramos que en los distintos Operativos de ambas áreas más de la mitad de los ítems que componen la prueba tiene observaciones en por lo menos uno de los criterios teórico-técnicos que rigen su construcción.

Un segundo tipo de análisis permite alcanzar un nivel de discriminación mayor y consiste en identificar la cantidad de observaciones registradas para cada ítem de prueba en ambos operativos.

Una lectura del cuadro permite describir la evolución de cada área en particular y señalar, además, nuevas diferencias entre ambas. Para el área de Matemática, mientras el primer cuadro nos mostró que la cantidad de ítems defectuosos era mayor en el 94 que en el 93, a través de este análisis observamos que la cantidad total de observaciones disminuye -en siete observaciones- de un año a otro. Es decir que hay una mayor cantidad de ítems con problemas de construcción, pero en términos generales, hay una menor cantidad de observaciones por ítem. Puede verse, por ejemplo, que en 1993 hay un ítem con cinco observaciones, otro con cuatro, tres ítems con tres observaciones y varios con dos; en el Operativo 94 el máximo número de observaciones por ítems es igual a tres y hay una gran cantidad de ítems con una observación.

En el caso de Lengua, la información que aporta este análisis refuerza la tendencia señalada en principio acerca de un mejoramiento general de la prueba. Por un lado, la cantidad de ítems observados disminuye de 33 -en 1993- a 28 - en el 94- y, además, la cantidad total de observaciones disminuye en una proporción aún mayor: hay un total de 16 observaciones menos en el Operativo 94. En este caso, la diferencia obedece tanto a la disminución general de la cantidad de ítems defectuosos como a la disminución de la cantidad de ítems con varias observaciones. Por ejemplo, mientras que en 1993 hay un ítem con cuatro observaciones y siete ítems con tres; en el 94 no aparecen ítems con cuatro observaciones y hay sólo dos ítems que registran tres observaciones.

Además, hay varios ítems que pasan de tener una o más observaciones a no tenerlas en el 94 o a la inversa. Sin embargo, ello no significa un aumento o descenso en la calidad del ítem en particular porque, como se dijo, no hay correspondencia entre los números de los ítems en ambos Operativos.

En síntesis, en Lengua no sólo disminuye el total de ítems "defectuosos" de un año a otro, sino que se registra una diferencia importante en términos de la cantidad de observaciones de los ítems, que puede hacernos pensar en un incremento en la calidad de su construcción. Sin embargo, en Matemática la tendencia es diferente ya que la prueba empeora ligeramente en cuanto a la cantidad de ítems con problemas, aunque hay evidentemente un mejoramiento en la atención a algunos de los criterios de construcción.

Un tercer tipo de análisis, consiste en describir el tipo de problemas de construcción, lo cual permite caracterizar más minuciosamente las diferencias entre áreas y, además, analizar la variación interna dentro de cada una de ellas de un año a otro. Ello implica describir cuáles fueron los requisitos más desatendidos en cada uno de los casos, información que aparece en los cuadros 3a y 3b en forma sintética.

Veamos primero el caso de Matemática.

CUADRO 3a

Cantidad de ítems observados por criterio. Pruebas de Matemática. Operativos 1993 -1994.

Criterios	Cantidad de ítems observados	
	1993	1994
Con respecto al núcleo:		
Redactar el núcleo de manera clara, sencilla y precisa.	14	13
El núcleo debe formular claramente un problema.	6	4
Con respecto a las alternativas de respuesta:		
Una de ellas debe ser inequívocamente correcta o mejor que las otras.	3	1
Las respuestas de distracción deben tener la misma probabilidad de ser elegidas.	24	24
Deben evitarse asociaciones entre términos del núcleo y la respuesta correcta.	1	0

Este análisis nos permite observar que del conjunto de criterios considerados, cinco fueron los que no se han contemplado rigurosamente. Además, hay dos requisitos de construcción que fueron más frecuentemente desatendidos en ambas pruebas. Uno de ellos se refiere a la claridad y precisión en la formulación de la consigna, otro criterio se refiere a la adecuación de las alternativas de respuesta elegidas en cada caso. Puede verse en el cuadro que tiene 24 observaciones en ambos Operativos.

Esta constatación reviste un carácter sumamente crítico, en especial si se piensa que en este criterio reside el aspecto más estrechamente ligado a la validez de construcción de este tipo de instrumentos . Para explicarlo mejor, recordemos que la pauta indica que los distractores no deben constituir alternativas poco probables o plausibles y, menos aún, respuestas casi disparatadas que puedan ser descartadas de entrada disminuyendo, por ende, el abanico de opciones. En caso de ser así, como se enfatiza en la bibliografía, lo que se estaría evaluando no es la competencia del estudiante relativa a un contenido específico, sino una suer-

te de habilidad para desestimar alternativas poco probables. Entonces, si bien es cierto que el conjunto de criterios de construcción es importante desde una perspectiva técnica, podría decirse que es en torno a este punto que pivotea la cuestión de la validez.

En el contexto de estas consideraciones se comprende mejor aún la importancia de la cuestión ya que la cantidad de observaciones referidas a este criterio no se modifica año a año y es elevado el número en ambos casos. Se trata, exactamente, del 60% de los ítems que componen la prueba.

Si a estos 24 le sumamos, además, los 4 ítems que tienen observaciones en los otros dos criterios relativos a la adecuación de las alternativas de respuesta, nos encontramos con un 70% de los ítems de la prueba en los que no se puede afirmar con un grado de certeza considerable que la respuesta correcta del estudiante garantice el dominio de la competencia que se pretende evaluar.

El siguiente cuadro nos permite analizar el caso de Lengua.

CUADRO 3b

Cantidad de ítems observados por criterio. Pruebas de Lengua. Operativos 1993 -1994.

Criterios	Cantidad de ítems observados	
	1993	1994
Con respecto al núcleo:		
Redactar el núcleo de manera clara, sencilla y precisa.	16	14
El núcleo debe formular claramente un problema.	3	0
No utilizar enunciados negativos	4	1
Cuando el núcleo es una frase incompleta, cada solución debe constituir un final de frase gramaticalmente correcto.	3	0
No debe contener informaciones no pertinentes.	1	1
Con respecto a las alternativas de respuesta:		
Una de ellas debe ser inequívocamente correcta o mejor que las otras.	13	9
Las respuestas de distracción deben tener la misma probabilidad de ser elegidas.	14	15
Deben evitarse asociaciones entre términos del núcleo y la respuesta correcta.	1	1

En primer lugar, puede verse, en comparación con Matemática que el tipo de requisitos desatendidos es más variado. Mientras que en el primer caso los problemas se concentraban en torno a cinco criterios, aquí aparecen otros tres: el empleo de enunciados negativos; la existencia de problemas gramaticales entre el núcleo y las alternativas de respuesta y la presencia de "pistas" bajo la forma de asociaciones entre el núcleo y la respuesta correcta. De todos modos, estos dos últimos criterios sólo tienen una observación en cada prueba.

También aquí los requisitos más frecuentemente desatendidos en 1993 y 94 fueron el relativo a la claridad y precisión de la consigna (16 y 14 observaciones respectivamente) y el referido a la probabilidad de las alternativas de respuesta (14 y 15 observaciones respectivamente). Si bien se trata de cifras importantes -alrededor del 30% de la prueba- son en los dos Operativos menores en Lengua que en Matemática.

Un aspecto que llama la atención es la cantidad de ítems que en ambos operativos son problemáticos en cuanto a la existencia de una única alternativa inequívocamente correcta o mejor que las otras, situación que no se presentaba en las pruebas de Matemática. Estamos hablando de 13 ítems con este tipo de problemas en el 93 -lo que representa el 32,5% del total de ítems que componen la prueba- y 9 ítems en 1994, lo que equivale al 22,5%.

Si sumamos la cantidad de ítems que presentan defectos de construcción relativos a la elección de las alternativas de respuesta (por no existir una alternativa inequívocamente correcta; por existir alternativas menos probables que las demás o poco probables; o por presentar asociaciones entre el núcleo y la respuesta correcta), podemos constatar que 28 del total de ítems de prueba del 93 -es decir, el 70%- y 25 del año 1994 -es decir, el 62,5%- se encuentran en esta situación.

A la luz de la información ahora disponible, puede efectuarse también una nueva lectura de la afirmación inicial referida al mejoramiento general de las pruebas de Lengua. Si bien es cierto que este hecho se constata claramente en nuestra descripción, puede discriminarse ahora en qué aspectos se produjo el avance. Por lo que puede observarse en el cuadro, no hay modificaciones sustantivas en tres de los requisitos más frecuentemente desatendidos, y, cuya importancia es central desde el punto de vista de la validez. La disminución de observaciones de un año a otro, se explican en términos de las variaciones que se refieren a otros criterios que podríamos considerar más básicos, como el empleo de enunciados negativos y la existencia de desajustes gramaticales entre el núcleo y las alternativas de respuesta (en el 94 hay un total de diez observaciones menos atribuibles a estos factores).

La posibilidad de establecer comparaciones consistentes

La confiabilidad es la segunda de las dimensiones importantes a tener en cuenta para estimar la calidad

técnica de un programa de evaluación y sus instrumentos y, por ende de los resultados obtenidos. Ella, según Gronlound se refiere al grado de consistencia de una medición, esto es , a cuán consistentes son los resultados de las pruebas u otros resultados de evaluación entre una medición y otra (Gronlund,1973). En este sentido, la estimación del grado de confiabilidad -al permitirnos calcular el margen de error dentro del que nos movemos al efectuar una medición- nos indica cuánta confianza podemos tener en un cuerpo determinado de resultados.

En un clásico trabajo sobre el tema, Sachs Adams señala tres tipos de factores ligados al tema de la confiabilidad. El primero de ellos tiene que ver con el hecho de que la tarea sea apropiada y bien definida. El segundo se refiere a la estabilidad o constancia de la capacidad que tiene un estudiante para cumplir con las tareas incluidas en la prueba. Por último, el tercer aspecto es la coherencia y objetividad de la persona que puntúa la prueba (Sachs Adams, 1970).

Otros autores definen el concepto de modo análogo, pero enfatizando la idea de que la confiabilidad nos indica la certeza, exactitud y precisión de nuestras mediciones y el grado en que ellas pueden variar a través del tiempo, en muestras diferentes del mismo comportamiento o en función de los distintos evaluadores (Sachs Adams,1970; Thorndike y Hagen,1989 ; Ebel,1977).

Por ende, la confiabilidad, al igual que la validez, no puede predicarse en términos genéricos porque se refiere siempre a un tipo particular de consistencia: consistencia en el tiempo de las mediciones, consistencia de los puntajes con independencia de quién califique las pruebas, consistencia interna de un instrumento, etc. Cada una de estas interpretaciones del concepto requiere procedimientos y análisis diferentes.

Por otra parte, y a diferencia de lo que sucede en los análisis de validez, los procedimientos para estimar el grado de confiabilidad de una medición son de tipo estadístico y correlacional y se traducen en índices y coeficientes. El término en sí mismo remite a un concepto de carácter estrictamente estadístico.

Estas cuestiones están claramente referidas a la fiabilidad de un instrumento de prueba en forma aislada. Pero, como ya se adelantó, es posible también analizar el problema a otro nivel si se piensa en la implementación de un programa de evaluación a escala mayor -distrital, municipal, provincial , regional o nacional-. En el caso de nuestro trabajo, si se tiene en cuenta como objeto de estudio, al sistema de evaluación en su conjunto y no a las pruebas consideradas en forma independiente, surgen nuevas cuestiones de interés.

La puesta en práctica de una estrategia de evaluación de la calidad de este tipo implica la administración de pruebas destinadas a medir el rendimiento año tras año en los distintos operativos, con el propósito de efectuar un seguimiento de la evolución de la calidad y de efectuar comparaciones entre los resultados correspondientes a uno y otro operativo. En situaciones como esta, desde una perspectiva estrictamente técnica, existen algunos recaudos específicos que deben ser contemplados.

En este sentido, Ebel señala que cuando la prueba está destinada a un empleo reiterado, tiene que estar provista de formas equivalentes, es decir, versiones de la misma que puedan ser empleadas de forma intercambiable. (Ebel, 1977). Dado que la posibilidad de establecer comparaciones entre resultados con un grado adecuado de consistencia interna proviene de administrar o bien el mismo instrumento de prueba cada vez, o bien, formas equivalentes de éste .

En el caso del SNEC, la información disponible proveniente de diversas fuentes[12] evidencia que se han empleado los resultados con fines comparativos, pero no sabemos aún si las pruebas empleadas en cada caso son equivalentes -ya que, de hecho, no son las mismas-. Este punto se constituyó, justamente, en un aspecto prioritario en el marco de este trabajo.

Existen algunos procedimientos específicos de carácter correlacional que permiten no sólo producir formas equivalentes de pruebas, sino también, determinar cuándo es posible predicar esta propiedad acerca de dos o más pruebas. La realización de este tipo de estudios, en el marco de este trabajo, se torna sumamente difícil, porque ello requeriría poder contar con las matrices de datos de respuesta de los alumnos. Es por ello que la estrategia metodológica consistió en efectuar un análisis del contenido de cada ítem de prueba y comparar las pruebas correspondientes a distintos Operativos. A partir de allí, sería posible determinar si las pruebas administradas en los distintos operativos son comparables desde el punto de vista del objeto que se evalúa[13].

Se desarrolló, entonces un primer nivel de descripción centrado en la identificación de los temas incluidos en uno y otro caso y de la cantidad de ítems dedicados a cada uno de ellos. En este caso, se trabajó con las pruebas de matemática correspondientes a los operativos 1993 y 1994[14]. Ello permite alcanzar una caracterización global en la medida en que las categorías temáticas son muy genéricas o abarcativas.

CUADRO 4:

Comparación de temas evaluados en Operativos 93 y 94[15]

TEMA	Cantidad de ítems en los que se evalúa	
	1993	1994
Números naturales	10	11
Fracciones	10	5
Números decimales	4	7
Porcentajes	1	3
Proporcionalidad	3	1
Medición	7	7
Gráficos	2	1
Geometría	2	5

Una primera lectura del cuadro permite observar que la mayoría de los temas fueron evaluados en ambos operativos, aunque con desigual proporción de ítems en un año y otro en algunos casos. En algunos temas esta diferencia es mínima y en otros es más significativa. Tal es el caso de los siguientes contenidos: números fraccionarios, geometría.

Ello permite suponer la existencia de algunos temas que fueron evaluados en el Operativo 93 y no en 1994 y, a la inversa, algunos contenidos evaluados en 1994 no habían sido incluidos en el primer Operativo.

Si bien este tipo de análisis permite establecer algunas comparaciones de interés como las anteriormente citadas, considera a los núcleos temáticos en forma muy general y dice poco acerca de la selección o muestra efectuada en cada caso particular. Si partimos de la idea de que tema y contenido no son términos que remiten a un significado conceptual común y, afirmamos, al mismo tiempo, que el contenido presente en un ítem de evaluación, precisamente por no ser meramente un sinónimo de "tema", incluye siempre un particular tratamiento de éste, un recorte de ciertos aspectos o dimensiones de la temática, puede resultar de interés efectuar un segundo tipo de análisis que permite, a su vez, un mayor grado de discriminación del proceso comparativo.

La tarea consiste, en este caso, en describir para cada uno de los ítems de prueba el aspecto o faceta del tema en cuestión, el contenido central evaluado. El análisis en este segundo nivel de complejidad permite construir una tabla en la cual el contenido de cada ítem se expresa en un grado mayor de especificidad que en el caso anterior, ya que no sólo está presente -para cada uno de los ítems de prueba- el tema matemático en cuestión, sino también el tipo de capacidad puesta en juego en función de la actuación/ ejecución esperada.

Al intentar hacer esta tarea, encontramos que un punto especialmente crítico era justamente la determinación del aspecto que era objeto de evaluación en cada caso particular. Si bien en muchos ítems este dato aparecía con claridad, en una proporción importante de ellos, parecen estar en juego más de un contenido -si se atiende al tipo particular de ejecución esperada-. En este caso, la decisión metodológica fue enunciar el conjunto de los contenidos a los que hacía referencia la consigna y, en segundo lugar, seleccionar -para la comparación- aquel contenido que parecía ser el objeto central de evaluación. En otros casos, el contenido aparentemente evaluado era uno pero, por la naturaleza propia del ítem, por el tipo de alternativas que planteaba (escasa plausibilidad de una o más de ellas, por ejemplo) o por su redacción, era altamente improbable que pudiera medir ese aspecto. En este punto fue posible hacer converger los datos disponibles acerca de la validez de construcción de estos ítems.

Otro tema de interés, en este nivel de análisis, es el de las competencias o capacidades que el ítem evalúa. Al definir el contenido evaluado en cada ítem de prueba tomando en consideración no sólo el concepto, propiedad o principio en él presente sino también el tipo de actuación que la realización de la consigna requiere del alumno, el análisis se complejiza en la medida en que se acerca por momentos al tema de las competencias a las que esa actuación/ejecución nos remite. Por ejemplo, en el análisis de ítems referidos a operaciones con números naturales se hace imprescindible distinguir si se trata de un elemento de prueba que apunta a evaluar la noción de la operación, la resolución de la operación, el manejo de algoritmos o el empleo de las operaciones en la resolución de situaciones problemáticas. Aparecen allí entonces algunas cuestiones que son de interés teórico -que aquí sólo enunciamos- en la medida en que se hace evidente la importancia que pueden tener en la configuración de la tarea los aspectos técnicos referidos a la selección del ítem de prueba y su formulación[16]. En efecto, tenemos el caso de ítems que por su forma parecen apuntar a evaluar capacidades ligadas al planteo y resolución de problemas (por ejemplo, aparece un texto que remite a situaciones de la vida real) ; sin embargo por las características mismas de las variables escogidas (los datos presentes en la consigna, las alternativas presentadas, etc.) este propósito queda en gran medida desdibujado.

La inclusión de este segundo nivel aporta una nueva dimensión al análisis porque permite observar claramente cómo, en algunos casos, si bien el tema básico es el mismo en ambos operativos, existen diferencias en el/los aspectos particular/es que se priorizan en cada situación y en la cantidad de ítems destinados a la evaluación del contenido dentro del conjunto de la prueba.

Detengámonos en algunos ejemplos de este tipo de situación:

• Consideremos el tema: "Sistema de numeración decimal". En primer término, puede decirse que fue efectivamente objeto de evaluación en ambos operativos. Si consideramos el total de ítems dedicados al tema, vemos que en el Operativo 93 cuenta con tres ítems, mientras que en el 94, sólo con uno. El ítem que las dos pruebas tienen en común está dedicado a tareas de "composición y descomposición de números". En los otros dos ítems que se dedican al tema en 1993, no se evalúa este mismo aspecto sino otros: ejercicios de equivalencia entre unidades de distinto orden y manejo de términos (en este caso, la unidad de mil). Ello permite concluir que no se trata solamente de una diferencia cuantitativa entre ítems dedicados a un mismo contenido en ambos operativos, sino que, además de la diferencia cuantitativa a nivel del tema general, aparecen dimensiones o aspectos del contenido considerados en un caso y no en el otro.

• Otro ejemplo interesante se refiere al tema "Operaciones con números naturales". Se trata de un contenido que, si bien es evaluado en forma independiente a través de ítems específicos, aparece también como contenido -aunque en forma secundaria o derivada- en ítems que apuntan aparentemente a la medición de otro tipo de aspectos (por ejemplo: cálculo del perímetro de un polígono). Teniendo

CUADRO 5

CONTENIDO CENTRAL EVALUADO	1993	1994
I. NÚMEROS Y OPERACIONES		
• Números naturales		
Sistema de numeración		
Denominación y reconocimiento de unidades de distinto orden	1	0
Equivalencia entre unidades de distinto orden	1	0
Composición y descomposición de números	1	1
Operaciones		
Suma, resta, multiplicación y división	1	0
Aplicación de algoritmos	3	4
Resolución de cálculos combinados y ecuaciones		
Con enunciado verbal	1	2
Expresión aritmética (sin paréntesis)	0	0
Expresión aritmética (con uso de paréntesis)	1	4
Resolución de problemas de combinatoria	1	0
• Números racionales		
Fracciones		
Concepto y formas de representación	5	2
Fracción de un número decimal	1	0
Operaciones con fracciones		
Suma,resta, multiplicación y división	0	1
Aplicación de algoritmos	2	1
Resolución de cálculos combinados con uso de paréntesis	1	0
Resolución de situaciones problemáticas	1	1
Números decimales		
Expresiones decimales: denominación y expresión numérica	1	2
Operaciones		
Suma y resta. Multiplicación y división	1	1
Aplicación de algoritmos	2	3
Resolución de cálculos combinados		1
• Proporcionalidad		
Relaciones de proporcionalidad directa		
Concepto	0	1
Resolución de situaciones problemáticas	2	1
Relaciones de proporcionalidad inversa		
Resolución de situaciones problemáticas	1	1
Porcentaje		
Cálculo del porcentaje de un N	1	1
II. MEDICIÓN		
Sistemas de unidades de medida		
Unidades de tiempo		
Conversión de unidades	1	2
Resolución de situaciones problemáticas	1	1
Unidades de longitud		
Conversión de unidades	1	3
Resolución de situaciones problemáticas	1	0
Perímetro		
Cálculo del perímetro de un polígono irregular	1	1
Área		
Cálculo del área para la resolución de situaciones problemáticas	1	0
Volumen: concepto	1	0
III. NOCIONES GEOMÉTRICAS		
• Rectas: relaciones de paralelismo y perpendicularidad.	0	1
• Ángulos		
Clasificación, propiedades	1	1
• Figuras geométricas		
Triángulo		
Concepto, elementos	1	0
Propiedades, relaciones	0	2
Cuerpos		
Cubo: elementos	0	1
IV. NOCIONES DE ESTADÍSTICA		
Lectura de gráficos de barras	1	1
Lectura de gráficos de pastel	1	0

presente el carácter "instrumental" del contenido "Operaciones con números naturales" y procurando aislar aquellos ítems en los que constituye el aspecto central a evaluar, puede verse que en 1993 hay siete ítems (entre cuarenta) destinados a este tema, mientras que en 1994 este valor asciende a diez ítems. Discriminando más el análisis podemos decir que en 1993 esos siete ítems apuntan a la evaluación en los siguientes aspectos en particular: interpretación y aplicación de algoritmos -tres ítems-; resolución de operaciones -un ítem- ; resolución de cálculos -tres ítems-, y un ítem destinados a problemas de combinatoria. Con respecto al Operativo 1994, puede señalarse que, de un total de diez ítems dedicados al tema, de los cuales cuatro se refieren a la aplicación de algoritmos y seis que apuntan a la resolución de cálculos aritméticos combinados. Nuevamente puede concluirse entonces que, si bien el tema en cuestión es el mismo en ambos Operativos, hay diferencias en los aspectos considerados en uno y otro caso, como así también en la cantidad de ítems destinados a medirlo en el total de la prueba.

• En el caso del tema "Números fraccionarios", habíamos visto que mientras que en 1993 el tema aparece evaluado en diez ítems, en 1994 sólo se mide en cinco ítems. Si se analiza el contenido específico presente en la consigna puede verse que, entre los diez ítems de 1993 hay cinco sobre el concepto de fracción y formas de representación, dos referidos a la aplicación de algoritmos; uno sobre resolución de cálculo (con empleo de paréntesis); uno sobre fracciones decimales y otro que apunta a la resolución de situaciones problemáticas. En 1994, de los cinco ítems referidos al tema, dos se refieren al concepto de fracción; uno a resolución de operaciones; uno sobre aplicación de algoritmos y uno que apunta a la resolución de situaciones problemáticas.

• Por último, consideremos el caso del tema "Proporcionalidad". En el cuadro puede observarse que mientras que tanto en 1993 como en 1994 hay tres ítems en la prueba referidos al tema. Sin embargo, encontramos diferencias en cuanto al contenido evaluado en cada caso. En efecto, en la prueba 1993 esos tres ítems se refieren a resolución de situaciones problemáticas de casos de proporcionalidad directa (dos de ellos) y proporcionalidad inversa (un ítem). En 1994, en cambio, hay un ítem -que no tiene equivalente en el 93- referido al concepto mismo de relaciones de proporcionalidad directa y los otros dos ítems presentes se refieren a resolución de situaciones problemáticas -uno de proporcionalidad directa y otro de inversa-.

El tipo de información que aporta este análisis sobre el proceso de evaluación establece un marco más definido para la comparación entre ambos Operativos, permitiendo un mayor nivel de precisión. Esto es importante si se tiene en cuenta que en la mayoría de los documentos sobre presentación de resultados las categorías utilizadas para comparar resultados obtenidos en un año y otro son de carácter general -específicamente, del grado de generalidad propio de nuestro primer nivel de análisis- y, en ese sentido, la contrastación no permite poner en evidencia estos matices que son fundamentales a la hora de interpretar la información estadística disponible.

Es posible aún situarse en un tercer nivel de análisis que permite la comparación entre ambos operativos con un grado de discriminación aún mayor. En este sentido nos propusimos describir cada ítem de prueba en función de los siguientes aspectos :

• Contenido/s evaluado/s : en este punto se detallan, con el mayor grado de especificidad posible, el aspecto de contenido matemático (concepto, principio, procedimiento, etc.) que efectivamente el ítem permite evaluar -considerando no sólo el tipo de pregunta que formula sino también la forma y el contenido de las alternativas de respuesta-.

• Ejecución central requerida: en este punto se describe, a partir del análisis de la enunciación de cada una de las consignas incluidas en la prueba, qué es lo que se espera que el alumno haga en términos de ejecuciones (actuaciones) para resolver el ítem. Se describe, entonces, el tipo de demanda que el ítem propone al alumno, incluyendo -en los casos que se considera pertinente- la explicitación de los aspectos de la propia consigna que operan como condiciones -restrictivas o facilitadoras- en términos de la resolución .

• Posibles procedimientos de resolución: en este punto se detallan algunas de la/s alternativas posibles de resolución del ítem. Esto significa que, para la realización de la ejecución principal requerida en la consigna, existen distintas vías de resolución por las que puede optar el alumno para resolverlo apropiadamente. Cada una de estas alternativas -en los casos que existen varias- pueden presentar niveles de complejidad diferentes y poner en juego facetas del contenido matemático también distintas. La elección de la alternativa está en gran medida condicionado por la propia formulación del ítem -tanto en el núcleo como en las opciones- . Una cuestión importante a destacar aquí es que el análisis, en este punto, permite incluir las cuatro opciones presentadas para cada ítem que pueden, como dijimos, condicionar las estrategias posibles incorporando o excluyendo la resolución por vía del descarte o la estimación.

Por razones de extensión del trabajo no es posible desarrollar aquí los resultados obtenidos en este último nivel de análisis[17]. Sin embargo, es importante decir que la descripción de cada ítem en términos de estos tres aspectos posibilitó efectuar comparaciones más precisas entre ítems correspondientes a distintos operativos. Ello es de particular interés en aquellos casos en que, de acuerdo con los análisis precedentes, el contenido evaluado es el mismo en ambos ítems. De este modo, encontramos que aún en las situaciones en las que cantidad de ítems referido a un mismo contenido fuera equivalente en los dos operativos, había diferencias entre ellos en términos de las ejecuciones requeridas y las estrategias posibles de resolución del ítem.

Como se enfatizó anteriormente, la cuestión más importante aquí está centrada en la comparabilidad de las pruebas correspondientes a cada Operativo. Es evidente que para poder efectuar estudios comparativos año tras

año, los instrumentos de medición -y por consiguiente los datos de referencia- deben ser comparables. Por tal motivo, la literatura sobre el tema señala en forma recurrente que para cumplir tal propósito las pruebas deben ser equivalentes.

Llegados a este punto, entonces, puede afirmarse que, de acuerdo con la información disponible, el trabajo realizado pone de manifiesto que las características de las pruebas de un año y otro no cumplen satisfactoriamente el requisito de comparabilidad. No estamos haciendo aseveraciones acerca del grado de dificultad de cada una de las pruebas y de este aspecto en términos comparativos, sino que la idea es que las pruebas correspondientes a distintos años no son equivalentes en términos del objeto evaluado en cada caso y esta limitación de consistencia interna restringe las posibilidades de efectuar comparaciones válidas entre Operativos.

A modo de cierre

Los resultados alcanzados a través del análisis técnico cobran sentido si se los ubica en el marco de los propósitos del SNEC. Recordemos que el Sistema responde a una necesidad de generar información básica para la formulación de políticas de mejoramiento de la calidad.

Los propósitos planteados para el SNEC expresan claramente el carácter regulador -en distintos sentidos- del dispositivo y su estrecha vinculación con el nivel político. Se trata, como dijimos, de un proyecto de evaluación de la calidad a escala nacional que procura proveer información sistemática a los organismos de planeamiento y gobierno del sistema con el fin de derivar políticas referidas al campo del curriculum, de la investigación y de la formación docente, entre las más importantes.

Puede estimarse aún más la importancia y el impacto de un dispositivo de este tipo si consideramos además la presencia que la implementación de los operativos y sus resultados han tenido en los medios de comunicación gráficos y televisivos.

Se hace evidente en este punto que la magnitud del proyecto y la importancia de las decisiones que están en juego genera, necesariamente, un imperativo de rigurosidad técnico-metodológica, dado que sólo cuando estos aspectos están adecuadamente resueltos, puede darse valor a los resultados hallados y emplearlos para el diseño de políticas.

Desde esta perspectiva, la validez constituye una cuestión vital. En el caso en estudio, puede observarse que se presentan diferentes niveles en los que deben tomarse decisiones de carácter teórico y metodológico que atañen directamente a la validez del dispositivo. Una de las primeras y más importantes es la elección de las dimensiones y los indicadores del objeto que se está procurando medir. En el marco del SNEC ello significa, concretamente, la respuesta a la pregunta acerca de cómo se concibe, se define y operacionaliza la calidad de la educación. El análisis de la información provista por los documentos permite sostener con claridad que, en este punto, la estrategia planteada sobredimensionó uno de los indicadores posibles: el rendimiento académico de los alumnos.

Pero, además, luego de haber optado este indicador para estimar la calidad, el rendimiento académico se constituye en la variable que debe, a su vez, ser operacionalizada. Ello implica la necesidad de buscar indicadores que la reemplacen en forma válida . Este nuevo problema de "representación" -como le denomina Barbier- fue resuelto , en este caso, a través de la administración de pruebas de aprovechamiento en las distintas áreas, siendo el porcentaje de respuestas correctas en cada prueba el indicador último de los logros.

Consideramos que este estudio ha permitido identificar algunos de los factores que restringen la posibilidad de garantizar la validez de los resultados de estas pruebas en, por lo menos, dos sentidos. En primer término, nos referimos a los problemas de validez de cada prueba considerada individualmente. Esto es, los problemas de construcción de cada instrumento de prueba empleado en los Operativos 93 y 94 van en detrimento de las posibilidades de considerar a esos puntajes de prueba -en este caso indicadores- como mediciones válidas del aprovechamiento de los alumnos en la medida en que, en un número importante de casos, el ítem presenta problemas relativos a su construcción.

En segundo lugar, y en estrecha relación con esto, si los resultados de cada Operativo no cuentan con el grado de validez suficiente para ser representativos de los logros de los estudiantes, tampoco puede ser empleada esa información de modo válido para describir la evolución de la calidad educativa a través de los años, aspecto que constituía una de las metas específicas del Programa.

El análisis de las potencialidades y debilidades de los programas de este tipo como estrategia para mejorar la calidad de un sistema educativo y generar procesos de cambio sin duda excede los límites de un análisis como el que aquí se ha desarrollado. Merece consideraciones acerca de cuestiones tales como: el tipo de tradición evaluativa de los docentes, las condiciones institucionales y áulicas en el marco de las cuales se desarrollan los procesos de enseñanza y aprendizaje y las relativas al trabajo docente en general, el tipo de interpretación que se hace de los resultados y las acciones que se toman en función de ello, por mencionar algunas de las más relevantes.

De ningún modo, sin embargo -y aún en el marco de análisis políticos sobre el tema- puede eludirse el debate técnico sobre el tema, en la medida en que este aspecto condiciona todo el conjunto de decisiones de índole más sustantivo (político, institucional, pedagógico) que sobre el asunto puedan tomarse.

Notas

[1] Se trata del Proyecto UBACyT F215 "El Sistema Nacional de Evaluación: evaluación de su calidad". Tiene sede de trabajo en el Instituto de Investigaciones en Ciencias de la Educación, Facultad de Filosofía y Letras, UBA.

[2] El SNEC es un programa de evaluación de la calidad educativa a escala nacional que fue implementado por el Ministerio de Educación de la Nación (Secretaría de Programación y Evaluación Educativa) a partir de 1993. Está constituído, básicamente, por pruebas objetivas de rendimiento administradas a estudiantes de los últimos años del nivel primario y medio en las distintas áreas curriculares. Además, hay instrumentos de carácter complementario tendientes a relevar información acerca de las variables que pueden estar asociadas al rendimiento. Esta información es recogida a través de una encuesta administrada a distintos tipos de unidades informantes: padres, directivos, docentes y estudiantes.

Hasta el presente se han realizado cinco Operativos Nacionales ampliando en forma progresiva el número de pruebas administradas.

[3] El trabajo de Lundgren (Lungren, 1996, *op.cit.*) es un buen ejemplo de las relaciones que pueden establecerse entre la implementación de programas de evaluación de los sistemas educativos y los procesos de descentralización/centralización de los sistemas educativos.

[4] Sobre las instituciones como sistemas culturales de regulación pueden verse los desarrollos de Fernández, L. *Instituciones educativas.* Bs. As., Paidós, 1994.

[5] Los resultados de esta línea de análisis pueden encontrarse en Diker,G. y Feeney, S., *op.cit.*

[6] El proyecto global está integrado por cuatro líneas de trabajo de desarrollo simultáneo:

a. Descripción del Sistema Nacional de Evaluación.

b. Descripción y análisis de los instrumentos utilizados en los distintos Operativos.

c. Estudio del procesamiento de datos y presentación de resultados empleados en los distintos Operativos.

d. Descripción del estado actual de instrumentos y técnicas de evaluación utilizados en una muestra de escuelas primarias .

[7] House efectúa un profundo análisis del concepto desde una perspectiva filosófica y epistemológica y explicita los supuestos y definiciones de validez que caracteriza a cada uno de los distintos enfoques vigentes de evaluación de programas.

[8] Desarrollo completos sobre estos aspectos pueden encontrarse en Thorndike y Hagen, *op.cit.*,Ebel,R. *op.cit.*, Camilloni, A., *op.cit.*, Gronlound, N.,*op.cit.*, de Ketele, J.M., *op.cit.*

[9] En el trabajo de análisis de la calidad de construcción de las pruebas participaron Laura Basabe y Verónica Chetman.

[10] Un desarrollo de los criterios teóricos y metodológicos que orientaron el trabajo, puede encontrarse en Cols, E., Basabe, L. y Chetman, V.: Informe Final UBACyT, 1998. Allí también se presentan el conjunto de las matrices de base para el análisis y la descripción de las el observaciones para cada uno de los ítems analizados.

[11] Dos tipos de aclaraciones son necesarias para la correcta lectura del cuadro. En primer lugar, los números de ítems no indican equivalencia del ítem entre un año y otro en términos del contenido evaluado. Esto significa que, por ejemplo, el ítem 2 de 1993 no equivale al N°2 del 94 sino que o bien aparece en forma similar con otro número o bien desaparece como ítem.

En segundo lugar, como ya se mencionó, un ítem puede tener más de una observación en su construcción. Ello puede deberse a que no esté de acuerdo con más de un criterio -en cuyo caso tendrá una marca en cada uno de ellos-; o bien porque tenga más de una marca en un mismo criterio debido a que hay más de una razón por la que desatiende esa pauta.

[12] Algunos de estas opiniones han adoptado carácter público, presentándose en los medios gráficos de divulgación general y especializada, la comparación de los resultados de ambos Operativos como evidencia de un incremento en los niveles de calidad del sistema.

[13] En el Informe de Avance correspondiente a esta área de trabajo, se detallan los aspectos metodológicos y se presentan los cuadros de base del análisis de contenido de cada uno de los ítems analizados.

[14] Este trabajo no hubiera sido posible sin la valiosa colaboración de la Lic. Claudia Broitman, como especialista en Didáctica del área.

[15] Se excluyó del análisis uno de los ítems correspondientes a la prueba de Matemática 1993, por suponer que se trataba de un ítem con un error en el tipeo o la impresión de la prueba. Por eso, el total de ítems en el operativo 93 es igual a 39 y no a 40.

[16] Nos referimos aquí a la noción de tarea en términos cognitivos. (Ver, por ejemplo, Doyle, W.: "Content representation in teachers definition of academic work." En Curriculum Studies, Vol.18, N°4, 1986; y Newman, D., Griffin, D. y Cole, M. : *La zona de construcción del conocimiento*. Madrid, Morata, 1992.)

[17] El análisis completo y el detalle de los aspectos metodológicos, pueden encontrarse en el Informe Final, UBACyT, 1998.

Bibliografía

Barbier,J.M. (1993) *La evaluación de los procesos de formación*, Barcelona, Paidós, MEC.

Camilloni, A. (1998) "La calidad de los programas de evaluación y de los instrumentos que los integran", en: Camilloni, A., Celman, S., Litwin, E. y Palou de Maté, M.C. *La evaluación de los aprendizajes en el debate didáctico contemporáneo,* Buenos Aires, Paidós.

Ebel, R. (1977) *Fundamentos de la medición educacional,* Buenos Aires,Guadalupe.

Diker, G. Y Feeney, S. "La evaluación de la calidad en Argentina: un análisis del discurso oficial", en: *Revista del IICE*, Buenos Aires (en prensa).

De Ketele, J.M. (1984) *Observar para educar*, Madrid, Aprendizaje Visor.

Gronlund, N. (1973) *Medición y evaluación en la enseñanza*, México, Pax.

House, E. (1994) *Evaluación, ética y poder*, Madrid, Morata.

Lundgren, U. (1996) "Formulación de la política educativa, descentralización y evaluación", en: Pereyra, M., García Mínguez, J., Gómez, A. y Beas, M. (comps.) *Globalización y descentralización de los sistemas educativos*, Barcelona, Pomares.

Sachs Adams, G. (1970) *Medición y evaluación*, Barcelona, Herder.

Schwab, J. (1973) *Un enfoque práctico como lenguaje para el currículum*, Buenos Aires, El Ateneo.

Thorndike, R. y Hagen, E. (1970) *Medición y evaluación en psicología y en educación*, Barcelona, Herder.

[2] El SNEC es un programa de evaluación de la calidad educativa a escala nacional que fue implementado por el Ministerio de Educación de la Nación (Secretaría de Programación y Evaluación Educativa) a partir de 1993. Está constituído, básicamente, por pruebas objetivas de rendimiento administradas a estudiantes de los últimos años del nivel primario y medio en las distintas áreas curriculares. Además, hay instrumentos de carácter complementario tendientes a relevar información acerca de las variables que pueden estar asociadas al rendimiento. Esta información es recogida a través de una encuesta administrada a distintos tipos de unidades informantes: padres, directivos, docentes y estudiantes.

Hasta el presente se han realizado cinco Operativos Nacionales ampliando en forma progresiva el número de pruebas administradas.

[3] El trabajo de Lundgren (Lungren, 1996, *op.cit.*) es un buen ejemplo de las relaciones que pueden establecerse entre la implementación de programas de evaluación de los sistemas educativos y los procesos de descentralización/centralización de los sistemas educativos.

[4] Sobre las instituciones como sistemas culturales de regulación pueden verse los desarrollos de Fernández, L. *Instituciones educativas.* Bs. As., Paidós, 1994.

[5] Los resultados de esta línea de análisis pueden encontrarse en Diker,G. y Feeney, S., *op.cit.*

[6] El proyecto global está integrado por cuatro líneas de trabajo de desarrollo simultáneo:

a. Descripción del Sistema Nacional de Evaluación.

b. Descripción y análisis de los instrumentos utilizados en los distintos Operativos.

c. Estudio del procesamiento de datos y presentación de resultados empleados en los distintos Operativos.

d. Descripción del estado actual de instrumentos y técnicas de evaluación utilizados en una muestra de escuelas primarias .

[7] House efectúa un profundo análisis del concepto desde una perspectiva filosófica y epistemológica y explicita los supuestos y definiciones de validez que caracteriza a cada uno de los distintos enfoques vigentes de evaluación de programas.

[8] Desarrollo completos sobre estos aspectos pueden encontrarse en Thorndike y Hagen, *op.cit.*,Ebel,R. *op.cit.*, Camilloni, A., *op.cit.*, Gronlound, N.,*op.cit.*, de Ketele, J.M., *op.cit.*

[9] En el trabajo de análisis de la calidad de construcción de las pruebas participaron Laura Basabe y Verónica Chetman.

[10] Un desarrollo de los criterios teóricos y metodológicos que orientaron el trabajo, puede encontrarse en Cols, E., Basabe, L. y Chetman, V.: Informe Final UBACyT, 1998. Allí también se presentan el conjunto de las matrices de base para el análisis y la descripción de las el observaciones para cada uno de los ítems analizados.

[11] Dos tipos de aclaraciones son necesarias para la correcta lectura del cuadro. En primer lugar, los números de ítems no indican equivalencia del ítem entre un año y otro en términos del contenido evaluado. Esto significa que, por ejemplo, el ítem 2 de 1993 no equivale al N°2 del 94 sino que o bien aparece en forma similar con otro número o bien desaparece como ítem.

En segundo lugar, como ya se mencionó, un ítem puede tener más de una observación en su construcción. Ello puede deberse a que no esté de acuerdo con más de un criterio -en cuyo caso tendrá una marca en cada uno de ellos-; o bien porque tenga más de una marca en un mismo criterio debido a que hay más de una razón por la que desatiende esa pauta.

[12] Algunos de estas opiniones han adoptado carácter público, presentándose en los medios gráficos de divulgación general y especializada, la comparación de los resultados de ambos Operativos como evidencia de un incremento en los niveles de calidad del sistema.

[13] En el Informe de Avance correspondiente a esta área de trabajo, se detallan los aspectos metodológicos y se presentan los cuadros de base del análisis de contenido de cada uno de los ítems analizados.

[14] Este trabajo no hubiera sido posible sin la valiosa colaboración de la Lic. Claudia Broitman, como especialista en Didáctica del área.

[15] Se excluyó del análisis uno de los ítems correspondientes a la prueba de Matemática 1993, por suponer que se trataba de un ítem con un error en el tipeo o la impresión de la prueba. Por eso, el total de ítems en el operativo 93 es igual a 39 y no a 40.

[16] Nos referimos aquí a la noción de tarea en términos cognitivos. (Ver, por ejemplo, Doyle, W.: "Content representation in teachers definition of academic work." En Curriculum Studies, Vol.18, N°4, 1986; y Newman, D., Griffin, D. y Cole, M. : *La zona de construcción del conocimiento*. Madrid, Morata, 1992.)

[17] El análisis completo y el detalle de los aspectos metodológicos, pueden encontrarse en el Informe Final, UBACyT, 1998.

Bibliografía

Barbier,J.M. (1993) *La evaluación de los procesos de formación*, Barcelona, Paidós, MEC.

Camilloni, A. (1998) "La calidad de los programas de evaluación y de los instrumentos que los integran", en: Camilloni, A., Celman, S., Litwin, E. y Palou de Maté, M.C. *La evaluación de los aprendizajes en el debate didáctico contemporáneo,* Buenos Aires, Paidós.

Ebel, R. (1977) *Fundamentos de la medición educacional,* Buenos Aires,Guadalupe.

Diker, G. Y Feeney, S. "La evaluación de la calidad en Argentina: un análisis del discurso oficial", en: *Revista del IICE*, Buenos Aires (en prensa).

De Ketele, J.M. (1984) *Observar para educar*, Madrid, Aprendizaje Visor.

Gronlund, N. (1973) *Medición y evaluación en la enseñanza*, México, Pax.

House, E. (1994) *Evaluación, ética y poder*, Madrid, Morata.

Lundgren, U. (1996) "Formulación de la política educativa, descentralización y evaluación", en: Pereyra, M., García Mínguez, J., Gómez, A. y Beas, M. (comps.) *Globalización y descentralización de los sistemas educativos*, Barcelona, Pomares.

Sachs Adams, G. (1970) *Medición y evaluación*, Barcelona, Herder.

Schwab, J. (1973) *Un enfoque práctico como lenguaje para el currículum*, Buenos Aires, El Ateneo.

Thorndike, R. y Hagen, E. (1970) *Medición y evaluación en psicología y en educación*, Barcelona, Herder.