EVALUACIŒN PSICOMÉTRICA DE LAS PRUEBAS CRECER 1998

JORGE BAZ«N
ŒSCAR MILLONES

Esta sección tiene por finalidad evaluar las pruebas CRECER 1998 desde el punto de vista psicométrico, es decir, determinar si éstas son óptimas para ser usadas en el análisis del rendimiento escolar desde el enfoque de normas que posteriormente se precisa. Este documento técnico es complementario de otro[1] en el que se detallan las características psicométricas de las preguntas de las pruebas.

1 EL MARCO PARA EL AN«LISIS DE LAS PRUEBAS

Es importante aclarar los conceptos del marco de análisis para evaluar psicométricamente las pruebas, que incluye los alcances y usos que pudieran hacerse de los puntajes obtenidos en éstas.

PROPÓSITO DE LAS PRUEBAS

Una de las primeras preguntas que debe resolverse en el análisis de las pruebas es qué se espera de ellas; es decir, qué se pretende hacer con los resultados y qué tipo de conclusiones podremos obtener. Este aspecto ha sido recientemente señalado como clave para determinar una de las propiedades que debe cumplir la prueba: su validez[2].

Existen dos enfoques en la interpretación de los resultados: el de normas y el de criterios. En el caso de CRECER 1998, el propósito principal de las pruebas es reportar niveles relativos entre grupos de la población. Así, en cuanto al análisis de los resultados se adoptó un criterio de normas más que uno de criterios.

La interpretación de normas de una prueba busca estimar las distancias relativas entre grupos de interés o subpoblaciones. Estas comparaciones pueden definirse de varias formas. De esta manera, para algunos usos es interesante referirse a una norma nacional o regional. En otros casos se puede hacer referencia a subdistribuciones de grupos más específicos como podrían ser, por ejemplo, los de la gestión pública o privada. Se dice que las interpretaciones basadas en tales comparaciones son referenciadas con base en normas. Estadísticas adicionales como percentiles o cuartiles de la distribución relevante son útiles para especificar mejor las comparaciones entre estos grupos de interés.

En la interpretación de criterios no se hace referencia directa al desempeño de otros exa-

1 Bazán, J. y O. Millones: Evaluación psicométrica de las preguntas de las pruebas CRECER 1998. Lima: Ministerio de Educación-Unidad de Medición de la Calidad de la Educación, 2000. Documento de trabajo, incluido en la sección anterior de este volumen.

2 AERA, APA, NCME: *Standards for Educational and Psychological Testing*, preparado por un comité conjunto de la American Educational Research Association, American Psychological Association y el National Council on Measurement in Education. Washington: AERA, 1999.

minados. El interés se centra en determinar la probabilidad de éxito (individual) respecto de algún dominio de preguntas. Este tipo de interpretación también toma una variedad de formas. Por ejemplo, es posible referirse a la probabilidad de éxito o de respuesta correcta para un subconjunto de preguntas de la prueba (un dominio de la prueba) o para un dominio más amplio de preguntas. Otras interpretaciones pueden generarse si se consideran otros tipos de pruebas construidas para otros ámbitos fuera del área del rendimiento educativo.

En general, las pruebas —sea para su interpretación en normas o en criterios— deben cumplir con un conjunto de requisitos como validez y confiabilidad. Adicionalmente, están sujetas a principios en el desarrollo de pruebas, en las escalas generadas desde los puntajes originales y en las condiciones de su administración.

Respecto de la construcción de pruebas, éste es un proceso que incluye el planeamiento de la prueba, la selección de áreas a incluirse en ella, la proposición de un conjunto de preguntas que cubren las áreas elegidas, la administración de una prueba piloto para el ensayo de las preguntas seleccionadas, el proceso de análisis de las preguntas (lo que lleva a la selección de las mejores) y una administración final a partir de una muestra que servirá para la última versión de la prueba. El lector que quiera tener una síntesis de estos pasos puede revisar el anexo 2 de esta sección, que resume este proceso de construcción de pruebas.

CRITERIOS PARA LA EVALUACIÓN DE LAS PRUEBAS: VALIDEZ DE LAS PRUEBAS

El análisis de las pruebas CRECER 1998 se basa en criterios presentados en la literatura psicométrica tanto desde el enfoque clásico[3] como desde el moderno[4]. Un aspecto del nuevo enfoque es el concerniente al significado ampliado que tiene el concepto de validez y el término asociado de "constructo".

Validez de la prueba

Desde el enfoque clásico, la validez de una prueba es la medida en que ésta mide el constructo que pretende medir. El término "constructo" se refiere a las características que no pueden ser medidas directamente sino que se infieren a partir de un conjunto de observaciones. El enfoque moderno del concepto de validez es más amplio: validez es el grado en que la evidencia acumulada (teórica o empírica) soporta las interpretaciones derivadas de los puntajes obtenidos en las pruebas[5]. Estas interpretaciones se refieren a los constructos o los conceptos que las pruebas se proponen medir (por ejemplo, rendimiento en Matemática). En este sentido, ya no se habla de diferentes tipos de validez (por ejemplo validez de contenido, concurrente o de constructo), sino de diferentes líneas o formas de evidenciar validez.

Este documento presenta un conjunto de criterios que proveen información relevante para determinar la validez de las pruebas. Los criterios que se describen incluyen: (i) el juicio de expertos; (ii) el análisis de unidimensionalidad de las preguntas que componen las pruebas; (iii) la confiabilidad de las pruebas; (iv) otras características basadas en las propiedades psicométricas de las preguntas, como son el nivel de dificultad, el grado de discriminación y los índices de no respuesta; y, (v) las propiedades derivadas de la construcción de las escalas y las transformaciones hechas para los objetivos de las pruebas. También puede derivarse evidencia complementaria de validez del diseño muestral y de la administración de las pruebas.

(i) Juicio de expertos

La opinión de los expertos tiene como finalidad analizar la correspondencia entre el contenido de las pruebas y los constructos que éstas intentan medir. El juicio de expertos se basa

3 Lord y Novick: *Statistical Theories of Mental Test Scores*. New York: Adisson-Wesley, 1974.

4 Por ejemplo, Moss P.A.: *Concepciones cambiantes de validez en la medición educativa: Implicaciones para la medición del desempeño*. Traducido por Juan Esquivel Alfaro. Tomado de *Review of Educational Research*, otoño de 1992, (62), 3, 1992, pp. 229-258; AERA, APA, NCME: *Standards for Educational..., op. cit.*, 1999.

5 AERA, APA, NCME: *ibidem.*

en el análisis curricular y las tablas de especificaciones que generaron los especialistas responsables de las pruebas (véase ejemplos de especificaciones en el anexo 1). Estas especificaciones fueron sometidas a juicio de expertos, y participaron en su elaboración los especialistas del MED y diversos consultores nacionales e internacionales.

(ii) Análisis de unidimensionalidad

En el esquema moderno del concepto de validez se incluye la evidencia de unicidad, es decir, la propiedad de una prueba de medir únicamente un constructo (unicidad de la prueba medible).

Para establecer si el conjunto de preguntas de una prueba mide una sola cosa —es decir, para evaluar la unidimensionalidad—, se usó el Modelo de Análisis de Correspondencias[6]. Este análisis indica el grado de homogeneidad de los conceptos medidos por el conjunto de preguntas que componen la prueba. El criterio para determinar la unidimensionalidad es el porcentaje de varianza explicada por el conjunto de preguntas de la prueba. Si en la primera solución (para la primera dimensión) esta varianza explicada es de 70% o más, se concluye que esta dimensión es suficiente para explicar la varianza total; es decir, no es necesario considerar más dimensiones para explicar la varianza de la prueba.

(iii) Confiabilidad de la prueba

La confiabilidad de una prueba mide el grado en que es consistente con los puntajes que de ella se obtienen. Idealmente se determina tomando dos o más veces la misma prueba a un examinado y revisando si los puntajes obtenidos son consistentes (idénticos o similares). En la práctica, la consistencia se determina de diversas formas alternativas, una de las cuales se basa en la consistencia interna de la prueba; es decir, por ejemplo, cuán consistente es la mitad de una prueba respecto de su otra mitad. Este criterio de consistencia interna de la prueba puede ser calculado por el coeficiente "alfa" de Cronbach[7]. Otra alternativa para un indicador de confiabilidad son los coeficientes derivados de formas paralelas de la prueba[8].

El criterio usado en CRECER 1998 es el de consistencia interna, es decir, coeficientes basados en la relación entre las preguntas de la prueba o entre un grupo de éstas. El supuesto del análisis de confiabilidad de las pruebas es que otros factores que afectan la distribución de puntajes son constantes o estables. Entre estos otros factores se encuentran las condiciones de administración de la prueba, la influencia del examinador y la inestabilidad de los estados del examinado, que son ajenas a los objetivos de la medición[9].

(iv) Criterios basados en índices psicométricos de las preguntas

Algunos criterios usados para las pruebas se sustentaron en los promedios de los índices psicométricos de sus preguntas. Los índices psicométricos de las preguntas incluyen la correlación pregunta-prueba, la discriminación, el nivel de dificultad y los índices de no respuesta. Así, para un índice agregado de correlación pregunta-prueba se ha tomado el promedio de las correlaciones pregunta-prueba de las preguntas que componen la prueba. Para un índice agregado de discriminación de las preguntas de las pruebas se ha tomado el promedio de los coeficientes de discriminación de las preguntas que componen la prueba. El nivel de dificultad de las pruebas se ha estimado con el promedio de los niveles de dificultad de cada pregunta. El nivel de no respuesta es el promedio de no

---

6 Véase Nishisato, S.: *Dual Scaling*. Toronto: University of Toronto Press, 1994. Para el análisis de homogeneidad, Visauta, V.B.: *Análisis estadístico con SPSS para Windows*, vol. II: *Estadística multivariante*. Madrid: McGraw-Hill, 1998.

7 Véase el anexo 2 (Glosario de términos psicométricos) de la sección anterior, p. 164.

8 Véase, por ejemplo, Nunnally, J. e I. Berstein: *Teoría psicométrica*. México: McGraw-Hill, 1995.

9 Véase Bazán y Millones: Evaluación psicométrica..., *op. cit.*

respuesta de las preguntas que componen la prueba[10].

(v) Construcción de escalas, transformaciones y normalidad. Comparabilidad de puntajes

Típicamente, los puntajes son las sumas de las respuestas correctas de la prueba. Así, los puntajes altos denotan mayor rendimiento en la prueba. Sin embargo, es necesario aclarar que los puntajes están determinados en parte por el número de preguntas, el tiempo que dura la prueba y las dificultades que presentan las preguntas. Estas características hacen que diferentes puntajes sean a veces difíciles de interpretar en ausencia de mayor información. En el caso de las pruebas CRECER, el puntaje está definido por el número de aciertos.

Construcción de escalas y transformaciones. La interpretación de los puntajes y su análisis estadístico pueden facilitarse convirtiendo los puntajes en un conjunto diferente de valores llamados puntajes derivados o puntajes de escala. La literatura presenta diversas escalas; las más populares son la escala de Rasch y la escala porcentual. La UMC emplea ambas.

La *escala de Rasch*[11] requiere el cumplimiento de ciertos principios en aspectos como: a) el comportamiento de los alumnos durante las pruebas; b) las características de las pruebas; y, c) la aplicación misma. Se puede comentar estos aspectos en las pruebas CRECER 1998.

Respecto de lo primero, se ha estimado que los alumnos que sabían las respuestas tuvieron más oportunidad de responder correctamente que los que no las sabían. Igualmente, no hubo evidencia de que los alumnos no resolvieran la prueba de manera independiente.

Con respecto a las características de las pruebas, se verificó que éstas evaluaron el rendimiento del alumno de manera unidimensional. Esto sugiere que una sola habilidad sería suficiente para explicar la ejecución de la prueba. Adicionalmente, por construcción, las preguntas miden sólo una variable (evidencia recogida de los índices de correlación pregunta-prueba). También se sostiene que la respuesta a una pregunta no es afectada por las respuestas a otras, resultado que incluye el caso de las pruebas de Lenguaje que corresponden a un mismo estímulo, sea éste un texto o una imagen.

En lo que concierne a la aplicación de las pruebas, los tiempos asignados para su resolución fueron suficientes. Una evidencia de esto son las bajas tasas de no respuesta encontradas. Sólo en los casos de las pruebas de Matemática para cuarto y quinto de secundaria la no respuesta es más alta, aun cuando no significativamente (véase los indicadores del acápite 3: Resultados de los indicadores en la evaluación de las pruebas).

El modelo de Rasch postula que la relación entre el rendimiento y la dificultad de una pregunta sigue una función determinada que permite obtener la probabilidad de acertar una pregunta determinada para un rendimiento específico. La escala de Rasch es una estimación de las habilidades de los alumnos a partir del modelo de Rasch. Sin embargo, es importante anotar que en las pruebas CRECER 1998 sólo se ha empleado la escala como una transformación no lineal de la escala porcentual. No se han usado las otras características e información generada por este modelo. La transformación realizada toma valores de 50 a 550, y corresponde a una transformación lineal estandarizada de la escala *logit* del modelo de acuerdo con ciertas ponderaciones. Tiene media 300 y varianza 50, y la correlación entre esta escala transformada y la escala porcentual está por encima de 0,98 en todas las pruebas.

La *escala porcentual* corresponde al porcentaje de acierto de la prueba. Esta escala va de 0 a 100. Con ésta se consigue uniformar la presentación de los resultados, independientemente del número de preguntas de las pruebas.

Se han señalado limitaciones en el uso de escalas porcentuales para la presentación

10 Véase, por ejemplo, M. Martin y D. Kelly, editores: *Technical Report*, vol. III: *Implementation and Analysis.* Final Test of Secondary School (Population 3). Third International Mathematics and Science Study, 1998.

11 Véase Muñiz, J.: *Teoría de respuesta a los ítemes: Un nuevo enfoque en la evolución psicológica y educativa*. Madrid: Ediciones Pirámide S.A., 1990.

de los resultados[12]; entre las principales están: 1) no existe, *a priori*, ningún valor que pueda considerarse como rendimiento insatisfactorio; 2) no indican qué es lo que saben o lo que ignoran los alumnos; 3) no tienen en cuenta la dificultad de las preguntas; 4) no pueden referirse de ninguna manera a los contenidos; 5) no indican la importancia de las preguntas no contestadas correctamente, ni cuántos son los sujetos que no las contestaron; 6) no permiten hacer comparaciones entre pruebas distintas: por ejemplo, sería erróneo interpretar que el resultado de Matemática en tercero de secundaria (51% medio de aciertos) es inferior a los resultados del quinto grado de primaria (53% medio de aciertos).

Aunque las escalas no porcentuales como la de Rasch pueden superar estas limitaciones, la ventaja principal de tener una escala no porcentual es que los usuarios no intentarán determinar cuántos alumnos han sido aprobados (que no es el sentido de la prueba), sino que buscarán saber qué grupos de alumnos han salido mejor que otros.

Normalidad

Es deseable que las escalas sigan una distribución normal para el uso de la inferencia paramétrica y para la eventualidad de formar grupos de rendimiento utilizando las medias y desviaciones estándar.

Esta propiedad ha sido difícil de obtener con las distribuciones de las pruebas CRECER 1998.

Sin embargo, en términos estadísticos esta exigencia no es necesaria. Las características de los puntajes de las pruebas (el número de preguntas de la mayoría es 30) determinan que cualquiera de las escalas presente sólo 30 valores diferentes, pues éstas son transformaciones biyectivas de aquéllos. Además, el tamaño efectivo de las muestras (aproximadamente 17 000) hace que las escalas presenten un número grande de "empates" (valores repetidos). De esta manera, la escala porcentual no necesariamente sigue una distribución normal. Sin embargo, las escalas de Rasch, por construcción, siguen una distribución normal.

Uso de las escalas para la comparabilidad

Se ha mencionado que uno de los objetivos de la evaluación basada en normas es sostener la comparabilidad entre grupos. Tanto la escala porcentual como la de Rasch garantizan las comparaciones de una misma prueba entre los estratos de la muestra. Adicionalmente, la escala de Rasch refuerza este objetivo cuando consigue que todas las pruebas presenten la misma media o valor central para la muestra nacional.

Los resultados de ambas escalas servirán para presentar los reportes globales (o nacional) y por los estratos de interés (gestión, región, departamento, etcétera). La presentación de los resultados por estratos puede incluir los reportes de los promedios, errores estándares, cuartiles de distribución y porcentaje de alumnos dentro del estrato.

Los resultados deben ser presentados para cada curso y para cada grado, para evitar la crítica 6) ya expuesta. Se deben considerar además los factores de ponderación, que toman en cuenta qué porcentaje tiene el estrato elegido en la población y en la muestra.

Homogeneización de pruebas

La homogeneización de las pruebas es un conjunto de procedimientos que hacen posible convertir el sistema de puntajes de una prueba al sistema de puntajes de otra prueba para que la interpretación sea equivalente. Para que esto suceda, la prueba debe mapear y mostrar que ambas miden la misma variable.

Dos pruebas son equivalentes si el desempeño o los puntajes pueden ser directamente trasladados de una a otra para que la elección de la prueba sea independiente del desempeño. Si la misma variable subyace en ambas pruebas, es indiferente cuál se use para

12 De la Orden Hoz, A.; R. Bisquerra; J. Gaviria; G. Gil; J. Jornet; F. López; J. Sánchez; M. Sánchez; J. Sierra y F. Tourón: *Los resultados escolares. Diagnóstico del sistema educativo 1997*. Madrid: Ministerio de Educación y Cultura, Instituto Nacional de Calidad y Evaluación, 1998.

obtener la medida de la posición de una persona en la variable. Conjuntos diferentes de ítemes pueden ser usados para evaluar grupos de personas diferentes.

En CRECER 1998 se utilizaron dos formatos de prueba de Matemática para el cuarto y quinto de secundaria. En cuarto, ambos formatos presentan tres preguntas iguales y dos similares (varían en el orden de las alternativas y, por ende, en la clave de respuesta). En quinto, ambos formatos presentan tres preguntas iguales y tres similares (varían en el orden de las alternativas y, por tanto, en la clave de respuesta). En ninguno de los dos grados coincide el lugar que ocupan los ítemes iguales o similares.

La utilización de dos formatos fue sugerida a raíz de las altas tasas de no respuesta detectadas en el piloto (prueba única de 40 preguntas y 90 minutos de duración). De cada objetivo o área de la prueba se tomó la mitad de los ítemes para determinar las formas A y B. Adicionalmente, se incorporaron cinco ítemes nuevos a cada prueba. Las pruebas fueron administradas a muestras equivalentes: la mitad del aula por cada escuela seleccionada del país. Cada muestra equivalente es para aproximadamente 8000 sujetos.

En principio, se trata de un diseño de anclaje con muestras equivalentes, algo que no debía ser difícil de equiparar. Un proceso de equiparación busca homogeneizar los puntajes (escalas) de las formas para elaborar un solo reporte de ambas.

Para realizar este procedimiento se tiene que partir del supuesto de que las poblaciones o que las pruebas son equivalentes. Es importante verificar ambos y decidir cuál de ellos será la base para la construcción de la escala (a partir de las capacidades de los alumnos o la dificultad de los ítemes).

Lo que se hizo fue un análisis de Rasch de los datos para comparar los indicadores de los ítemes en común. En principio, los indicadores deberían arrojar resultados muy similares, lo que sugeriría que, como se planificó al construir las pruebas, la escala de habilidades subyacentes es común. Se encontró que las dificultades de los ítemes de anclaje no son las mismas en las dos pruebas. Los ítemes que aparecieron primero en cualquiera de las dos formas fueron menos difíciles. El efecto de orden parece importante.

Tomando esto en cuenta, hemos asumido que las dos formas de pruebas tienen ítemes diferentes, pues las dificultades de los ítemes de anclaje así lo indican. Para certificar que es posible sobreponer las estimaciones RASCAL (desde el enfoque de Rasch) basadas en las muestras equivalentes, nos hemos preguntado si las pruebas son realmente diferentes en la medición del contructo. Hemos encontrado que no: las pruebas tienen idéntica distribución. Esto nos permite suponer que las formas midieron lo mismo y son "idénticas" en distribución de habilidades: 2 *logits* de habilidad en una "significa" 2 *logits* en la otra; es decir, las pruebas son equiparables.

DISEÑO MUESTRAL

Un requerimiento importante para el efectivo uso de las pruebas es obtener muestras y tamaños de muestras representativos y apropiados. Las pruebas de 1998 se aplicaron a una muestra representativa de centros educativos (CE) polidocentes urbanos en el ámbito nacional. En el nivel de educación primaria incluye a 17 370 estudiantes del cuarto grado y a igual número de alumnos de sexto, y en el nivel de educación secundaria a 17 400 estudiantes del cuarto grado y a igual cantidad de alumnos de quinto.

Estratos considerados

El enfoque de normas seguido sugirió la selección de un conjunto de criterios de utilidad para las comparaciones y para los futuros usos de los resultados de las pruebas. Estos criterios sirvieron de base para la estratificación de la población y de la muestra. Los estratos de la muestra sólo se refieren a la zona urbana, y en este ámbito los criterios de clasificación fueron: gestión del CE, región y departamento a los que pertenecen los examinados.

De acuerdo con el criterio de regiones se definieron tres regiones longitudinales (costa, sierra y selva) y tres transversales (norte, centro y sur), que determinaron las nueve regiones de comparación: costa norte, centro y sur; sierra norte, centro y sur; y selva norte, centro

y sur. Los departamentos fueron veinticinco en total (incluyendo Callao), y la gestión se refirió a escuela pública (estatal) y escuela privada (particular).

Selección de la muestra

La muestra se seleccionó con base en un sistema de muestreo bietápico, estratificado y por conglomerados, con las escuelas como unidades de primera etapa y los estudiantes como unidades de segunda etapa. Es decir, en cada estrato se seleccionó una muestra probabilística de CE, y, luego, en cada CE se escogió, mediante sorteo aleatorio simple, a los estudiantes que integrarían la muestra.

El procedimiento fue el siguiente: dados los estratos considerados, en cada uno de ellos se determinó la cuota de colegios por elegir (dividiendo entre 30 la cantidad de estudiantes por seleccionar en el estrato y redondeando). Esta operación da más probabilidad de que una escuela sea seleccionada en función del número de alumnos que estudian en ella. Además, se fijó un mínimo de dos colegios por estrato de sorteo, para poder estimar el error de muestreo en todos los casos.

La segunda etapa consistió en la selección de estudiantes. Los 30 estudiantes de cada sección fueron elegidos por sorteo aleatorio simple. En el caso de CE con más de una sección se realizó un sorteo simple de secciones, para evitar complicaciones excesivas de logística. Esta etapa adicional del sorteo no sesga la muestra (sigue siendo representativa del colegio), pero tampoco permite retirar del error de estimación el posible efecto de diferencias entre secciones del colegio, si las hubiera. Pero es un riesgo calculado por el que se optó para controlar los errores "no de muestreo".

Aunque para cada nivel primario y secundario son dos las poblaciones objetivo (cuarto y sexto grado en primaria, por ejemplo), por tratarse de CE polidocentes urbanos (que tienen tanto cuarto como quinto grado) se optó por un diseño que contemplaba el muestreo simultáneo en ambas poblaciones; esto es, las muestras de cuarto y sexto se tomaron en los mismos CE. Las consideraciones anteriores, aplicadas a los estratos, dieron una lista de 579 CE.

Se trata, pues, de una muestra con probabilidad de selección proporcional al tamaño en la primera etapa, pero que asegura una igual oportunidad de formar parte de la muestra a todos los estudiantes en la segunda etapa. Por esto, las estimaciones que se hagan requieren el uso de ponderaciones. Esto quiere decir que en una muestra planificada por estratos de muestreo (por departamentos, regiones y gestión) es necesario calcular los pesos o ponderaciones de los estudiantes para garantizar la representatividad total de la muestra y restituir la proporcionalidad[13].

Factores de ponderación

Para los efectos de los cálculos agregados y cálculos de los promedios de los puntajes, se usaron las ponderaciones correspondientes para corregir la no proporcionalidad respecto de los tamaños de los estratos del universo.

Errores de estimación

La estimación de las varianzas y el error estándar para el cálculo de los promedios y otros estadísticos debe ser tomada en consideración para cualquier inferencia paramétrica. Debido al diseño de muestreo utilizado, las estimaciones del error estándar pueden ser empleadas usando las fórmulas correspondientes del muestreo estratificado o, alternativamente, pueden ser estimadas por otros métodos como Jackknife o Bootstrap. Es decir, no es recomendable el uso de las fórmulas de las varianzas del muestreo simple aleatorio[14].

13 Para mayores detalles, véase Calderón, Arturo; Cholly Farro y Jorge Bazán: "Diseño muestral en la aplicación CRECER 98". Lima: Ministerio de Educación-Unidad de Medición de la Calidad de la Educación, 2000. Documento de trabajo.

14 Para mayor información sobre la estimación de dichos errores véase la sección Estimación del error estándar en las pruebas CRECER 1998 en este mismo documento.

APLICACIÓN DE LAS PRUEBAS

La versión final de las pruebas se aplicó en diciembre de 1998. La aplicación fue supervisada por especialistas y coordinadores en cada uno de los CE seleccionados. Ningún profesor del CE donde se realizaba la evaluación participó en la aplicación de las pruebas: la tarea recayó sobre profesores de otros ámbitos especialmente entrenados para esta aplicación.

Las pruebas fueron elaboradas en los meses previos y tomadas al final de un proceso metodológico que se inició en 1997 a través de la aplicación piloto de cinco a diez formas de prueba para cada una de las doce pruebas nacionales. El anexo 2 resume el proceso de construcción de las pruebas.

OTROS ENFOQUES ALTERNATIVOS

Las propiedades psicométricas mencionadas están relacionadas con el marco referencial teórico (modelo) que se adopta en el análisis de las pruebas. Como se sabe, los modelos más usados son: (i) los de la Teoría Clásica de los Tests; y, (ii) los modelos de la familia de la Teoría de Respuestas a Ítemes o TRI. El modelo de Rasch ya mencionado es un caso especial de la familia de los TRI.

La Teoría Clásica de los Tests

La Teoría Clásica de los Tests es un enfoque según el cual el resultado de la medición de una variable depende de la prueba utilizada y de los sujetos evaluados. El énfasis que pone esta teoría en las pruebas usadas ha motivado críticas, pues en esta estrategia una variable es inseparable del instrumento utilizado para medirla. Esto constituye una seria limitación, pues inevitablemente se acabaría definiendo de manera operativa la variable por el instrumento con que se mide.

La Teoría Clásica de los Tests, denominada también Teoría del Puntaje Verdadero, se apoya en un modelo lineal con error de medición formulado por Spearman en 1904. El puntaje obtenido en la prueba tiene dos componentes: su verdadero valor y un error de medición. A partir de una axiomática simple y con base en la noción de pruebas paralelas se definen las propiedades mencionadas: confiabilidad, validez y discriminación.

Las propiedades de las preguntas que se incluyen en las pruebas son las definidas por el modelo clásico: validez, dificultad, índice de discriminación e índice de no respuesta[15].

TRI *versus* teoría clásica

Otras opciones metodológicas respecto del análisis de las pruebas podrían ser seguidas con el uso de la TRI. La ventaja de tomar en cuenta otros enfoques es la oportunidad de estimar mediciones psicológicas adicionales que no pueden ser proporcionadas por la teoría clásica. Es importante anotar, sin embargo, que el enfoque TRI no contradice ni los supuestos ni las conclusiones fundamentales de la teoría clásica. Son sólo enfoques que nos dan información adicional sobre si la metodología empleada y los requisitos adicionales se cumplen. Por ello, el carácter de estos modelos TRI es complementario al de la teoría clásica.

Siguiendo una tendencia reciente en reportes especializados sobre evaluaciones de sistemas educativos nacionales e internacionales (NAEP-National Assessment Educational Progress, IAEP-International Assessment Educational Progress, TIMMS-Third International Mathematics and Science Study y LLECE-Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación), las pruebas nacionales CRECER 1998 fueron sometidas a un análisis con el modelo de medición de Rasch, uno de los modelos de Teoría de Respuesta al Ítem. Los modelos de Rasch son utilizados en sistemas de evaluación educativa de países como Australia, Inglaterra, Alemania, Estados Unidos, Colombia, Holanda y Dinamarca.

Programas computacionales

Respecto del *software* computacional, en el análisis de las preguntas de las pruebas CRECER

15 Para mayores detalles, véase el anexo 2 de la sección anterior, p. 164.

1998 se usaron tanto el ITEMAN (modelo clásico) como el RASCAL con un parámetro (modelo de Rasch)[16]. Cada equipo de especialistas revisó las preguntas para verificar las propiedades generadas por las corridas.

Con el ITEMAN se obtuvo la información sobre nivel de discriminación, dificultad de ítemes y aquella sobre distractores. Con el RASCAL se evaluó la posibilidad de extender el análisis para estimar las habilidades de los examinados. Este análisis está supeditado a la verificación del supuesto de unidimensionalidad que se desprende del análisis de las pruebas. Sin embargo, el RASCAL fue usado sólo para la parte de la transformación de las variables, que se ha explicado anteriormente.

ALCANCES

El objetivo de CRECER 1998 de comparar estratos relevantes constituye un primer paso para la evaluación del rendimiento escolar en el sistema educativo peruano. Permite conocer la estructura del rendimiento educativo para los grados elegidos en 1998 e identificar sectores críticos relativos que necesiten mayor atención y prioridad en la implementación de políticas de mejoramiento de la educación. En este sentido, los resultados de las pruebas CRECER 1998 se colocan en una secuencia que provee de información para delinear los siguientes pasos en futuras aplicaciones.

Con las propiedades presentadas y las escalas elaboradas se puede obtener una idea de cómo (en términos relativos) han salido los estudiantes de diferentes grupos. El propósito de generar estos resultados es la comparación entre los grupos que corresponden a los criterios seleccionados y que son presentados en el tercer acápite de este documento.

Con respecto a las pruebas desarrolladas para la fase de estudio piloto y las versiones finales de las pruebas, es importante notar un aspecto cualitativo del muestreo. Los valores que se obtienen para las propiedades estadísticas analizadas pueden variar de muestra a muestra. A este tipo de variación hay que añadir la variación debida a la modificación (mejoramiento) de las preguntas. Por tanto, es necesario volver a revisar las propiedades psicométricas por cambios de muestra y por modificaciones como resultado del desarrollo o construcción de *tests*. El poder inferencial de los resultados sobre la población muestreada está supeditado tanto a las posibles variaciones y modificaciones que ocurren como efecto del análisis de las pruebas en sí, cuanto a las muestras utilizadas entre la fase piloto y definitiva.

Este aspecto está vinculado a una de las críticas dirigidas a la Teoría Clásica de los Tests por los seguidores de la Teoría de Respuesta a Ítemes (TRI), en el sentido de que los resultados de los *tests* no sólo dependen de los *tests* mismos sino también de los examinados. De ahí la necesidad, según el enfoque TRI, de obtener mediciones que no varíen en función del instrumento utilizado y la necesidad de disponer de instrumentos de medida cuyas propiedades no dependan de los examinados.

Elección de la escala

Dado que las pruebas son usadas sobre todo para comparar los resultados entre grupos relevantes, cualquier transformación de los puntajes obtenidos a una nueva escala es apropiada para los fines de comparación.

En esta nueva escala lo que se busca es determinar qué grupo de alumnos ha salido mejor. Los resultados de la escala servirán para presentar los reportes globales (o nacionales), así como los resultados por los estratos de interés (gestión, región, departamento, etcétera). La presentación de los resultados por estratos puede incluir los reportes de los promedios (previa ponderación por no proporcionalidad en el tamaño de los estratos en la población), desviaciones estándar, cuartiles de distribución para cada materia estudiada y para cada grado por separado.

Para el análisis de las preguntas siguiendo el enfoque clásico, la revisión de las propiedades de las pruebas y de las preguntas de cada prueba se hará con la escala original de

16 Para mayor información, véase http: //www.assess.com/softmenu.html

puntaje total. Para el análisis de los resultados, sin embargo, se empleará la escala de Rasch y la escala porcentual, dependiendo de los análisis por implementar.

2 CONSTRUCTOS Y CONTENIDOS DE LAS PRUEBAS

A fines de noviembre y los primeros días de diciembre de 1998 se realizó la Aplicación Nacional CRECER 1998 a los estudiantes de cuarto y sexto grado de primaria y de cuarto y quinto de secundaria. Esta aplicación de pruebas fue complementada con encuestas a padres de familia o tutores y directores de CE en los que estudiaban los evaluados.

La Aplicación Nacional incluye 18 pruebas (12 de selección múltiple y 6 de respuesta abierta o de desempeño), 2 a directores, 4 a padres o apoderados, 4 a estudiantes y 10 a profesores (que incluyen 4 denominadas de oportunidades para aprender).

CONTENIDO DE LAS PRUEBAS

En 1998 los estudiantes del cuarto grado de primaria estaban desarrollando sus aprendizajes con la nueva estructura curricular organizada en áreas de desarrollo. Así, las pruebas fueron aplicadas en las áreas de Comunicación Integral, Lógico-Matemática, Personal Social y Ciencia y Ambiente. Los estudiantes del sexto grado de primaria, en cambio, desarrollaban la estructura curricular organizada en líneas de acción educativa, por lo que las pruebas fueron aplicadas en las asignaturas de Lenguaje, Matemática, Ciencias Histórico-Sociales y Ciencias Naturales. Los estudiantes de cuarto y quinto de secundaria desarrollaban también la estructura curricular organizada en asignaturas, por lo que las pruebas fueron aplicadas en Lenguaje y Matemática.

LAS ÁREAS O ASPECTOS EVALUADOS

4° grado de primaria

Personal Social. Incluye cuidado de la salud personal y colectiva, convivencia democrática, sentimiento de pertenencia y conocimiento de su medio sociohistórico y natural.

Ciencia y Ambiente. Cubre las áreas de conservación de la salud, conservación del medio ambiente y la intervención humana en el medio.

Comunicación Integral. Incluye las áreas de comunicación escrita-lectura, reflexión sobre la lengua y lectura de imágenes.

Lógico-Matemática. Revisa el conocimiento de los números y la numeración; la habilidad operativa y el cálculo; la medición y la organización del espacio.

Cuadro 1
Pruebas consideradas en CRECER 1998

Primaria	N° de ítemes	Tiempo de prueba	Secundaria	N° de ítemes	Tiempo de prueba
4° de primaria			**4° de secundaria**		
Comunicación Integral	30	60	Lenguaje y Literatura	38	60
Lógico-Matemática	30	75	Matemática Forma 1	25	60
Ciencia y Ambiente	30	60	Matemática Forma 2	25	60
Personal Social	27	60			
6° de primaria			**5° de secundaria**		
Lenguaje	31	60	Lenguaje y Literatura	38	60
Matemática	30	75	Matemática Forma 1	25	60
Ciencias Naturales	30	60	Matemática Forma 2	25	60
Ciencias Histórico- Sociales					

6° grado de primaria

Ciencias Histórico-Sociales. Incluye historia y cultura peruana, forjadores de la peruanidad (personajes modelos de patriotismo), el Estado peruano, el reconocimiento del universo, el medio geográfico y la convivencia en sociedad.

Ciencias Naturales. Abarca las áreas de transformación de la materia, funciones del cuerpo humano, interacciones materia-energía y conservación del medio ambiente.

Lenguaje. Las áreas temáticas son comprensión de lectura, nociones gramaticales, vocabulario y análisis de imágenes (identificación de textos gráficos, sus usos, análisis e intencionalidad).

Matemática. Incluye números naturales, fracciones, números decimales, medición y geometría.

4° de secundaria

Lenguaje y Literatura. Presenta áreas de comprensión lectora, nociones y reglas gramaticales, análisis de imágenes y razonamiento verbal.

Matemática. Incluye Aritmética, Álgebra, Estadística y Geometría.

5° de secundaria

Lenguaje y Literatura. Abarca temas de comprensión lectora, nociones y reglas gramaticales, análisis de imágenes y razonamiento verbal.

Matemática. Comprende conjuntos, Aritmética, Álgebra, Estadística, Geometría y Trigonometría.

REPRESENTATIVIDAD DE CONTENIDOS

Uno de los aspectos fundamentales de la validez de un test se refiere al grado y extensión con que los constructos son representados por los contenidos de la prueba. Al respecto, es importante determinar si el conjunto de contenidos desarrollados en una prueba representa o no los constructos teóricos por medir. Puede manifestarse carencia de validez por una subrepresentación o representación irrelevante del constructo. En presencia de subrepresentación de constructos, los puntajes de las pruebas pierden la capacidad de representar el verdadero rendimiento del área evaluada.

Este riesgo de no representatividad se presenta en diversos grados de acuerdo con el objetivo del uso de la prueba, es decir, según el enfoque sea de normas o de criterios. Al construir una prueba con el enfoque de normas hay menos especificidad en los contenidos correspondientes al área a ser evaluada. No obstante, las áreas representadas en la prueba deben ser una muestra representativa del dominio del constructo.

La diferencia entre los análisis del enfoque de normas y de criterios es que el análisis de criterios requiere de una completa descripción de la(s) conducta(s) medida(s), mientras que en el análisis de normas hay más flexibilidad o generalidad para definir las conductas por medir. Esta diferencia puede ilustrarse con el siguiente ejemplo. Si consideramos el currículo escolar de cuarto grado de primaria en el área Lógico-Matemática, distinguimos dos niveles de análisis: (i) las competencias; y, ii) las capacidades y actividades. En el primer nivel encontramos competencias como “conocimiento de los números y la numeración”, “conocimiento de las operaciones con números naturales”, etcétera. En el segundo nivel hallamos capacidades como “elaborar sucesiones numéricas crecientes y decrecientes”, o “establecer y graficar relaciones numéricas y no numéricas”.

El análisis de normas se dirige a preguntas representativas del primer nivel de las competencias (seleccionadas del total de competencias y de acuerdo con los criterios de la Teoría Clásica de los Tests). Aunque muchas de estas competencias corresponden a capacidades y actitudes del segundo nivel, lo que se desea es tener representatividad general del nivel de competencias.

En un análisis de criterios se seleccionan capacidades (del nivel ii) que son relevantes por sí mismas; por ejemplo, “clasificar números naturales de acuerdo con diversos criterios”. Sobre la base de una correcta especificación de esta capacidad (nivel ii) se analiza

un conjunto de preguntas y se incorpora un criterio de logro (por ejemplo, 75%).

Como se verá luego, el análisis de las pruebas CRECER 1998 sigue la orientación de normas y, por tanto, su uso se restringe a los objetivos específicos que persiguen este tipo de pruebas.

Este punto es importante cuando se considera otra diferencia entre los análisis de los enfoques de normas y criterios. En el enfoque de normas importan principalmente los contrastes relativos entre los examinados, mientras que en el enfoque de criterios se busca estimar específicamente lo que los examinados pueden o no hacer (los logros). Adicionalmente, el grado de rigor requerido en las pruebas basadas en normas para definir las conductas medidas puede ser más global o general, aunque representativo de los dominios del análisis.

La construcción de pruebas es un proceso que incluye el planeamiento de la prueba, la selección de áreas a incluir en la prueba, la proposición de un conjunto de preguntas que cubren las áreas elegidas, la administración de una prueba piloto para el ensayo de las preguntas seleccionadas, el proceso de análisis de las preguntas (lo que lleva a la selección de las mejores preguntas) y una administración final a partir de una muestra que servirá para la versión final de la prueba.

3 RESULTADOS DE LOS INDICADORES EN LA EVALUACIŒN DE LAS PRUEBAS

Con respecto a los indicadores considerados, los siguientes fueron los resultados cuantitativos más relevantes de las pruebas CRECER 1998 (véase los cuadros 2, 3 y 4).

CONFIABILIDAD (CONSISTENCIA INTERNA)

En lo que a las pruebas se refiere, los resultados se presentan a partir del indicador de consistencia interna (alfa de Cronbach).

Tanto las pruebas de primaria como las de secundaria arrojaron altos índices de consistencia interna. El rango se sitúa entre 0,73 en Matemática Forma 1 en cuarto de secundaria y 0,85 en Lenguaje de sexto grado de primaria. Los coeficientes de consistencia interna resultaron relativamente menores en el nivel secundario respecto del primario, aunque esta diferencia no es significativa.

Es de notar también que en el área de Matemática de secundaria se incluyen temas como Aritmética, Álgebra, Geometría/Trigonometría y Estadística, mientras que en primaria los temas de Matemática son más homogé-

Cuadro 2
Criterios de validez cuantitativa de las preguntas de las pruebas CRECER 1998: Educación primaria

Estadísticas finales	4° grado				6° grado			
	Comunicación Integral	Lógico-Matemática	Personal Social	Ciencia y Medio Ambiente	Lenguaje	Matemática	Ciencias Histórico-Sociales	Ciencias Naturales
Confiabilidad								
Alfa de Cronbach	0,816	0,773	0,826	0,764	0,852	0,806	0,773	0,774
Índices psicométricos								
Índice de dificultad medio (%)	53,96	54,53	58,44	52,78	56,06	48,68	58,17	61,90
Índice de discriminación medio (%)	45,69	36,42	48,69	40,65	49,57	45,61	41,01	39,92
Índice de validez medio (%)	39,75	35,56	42,69	35,67	42,90	39,26	36,37	36,61
Índice de no respuesta medio (%)	2,75	3,58	1,14	1,17	3,69	4,93	1,15	0,78
Total de preguntas	30	30	27	30	31	30	30	30
Total de evaluados	16 997	16 827	16 819	16 744	16 833	16 716	16 759	16 650

Cuadro 3
Criterios de validez cuantitativa de las preguntas de las pruebas CRECER 1998: Educación secundaria

Estadísticas finales	4° grado			5° grado		
	Lenguaje y Literatura	Matemática Forma 1	Matemática Forma 2	Lenguaje y Literatura	Matemática Forma 1	Matemática Forma 2
Confiabilidad						
Alfa de Cronbach	0,774	0,733	0,755	0,79	0,772	0,755
Indicadores psicométricos						
Índice de dificultad medio (%)	44,12	46,44	45,65	54,94	47,18	46,27
Índice de discriminación medio (%)	37,26	39,44	43,20	39,78	40,87	44,09
Índice de validez medio (%)	32,32	36,82	38,21	33,97	38,90%	39,83
Índice de no respuesta medio (%)	6,58	8,73	7,65	5,22	8,80	10,82
Total de preguntas	38	25	25	38	25	25
Total de evaluados	16 939	8294	8274	16 710	8206	8034

Cuadro 4
Criterios de validez basados en la unidimensionalidad de las pruebas CRECER 1998: Educación primaria y secundaria

Pruebas	% varianza explicada (1ª dimensión)	*Ratio* de 1ª sobre 2ª dimensión	*Ratio* de 2ª sobre 3ª dimensión	Coeficiente de confiabilidad Theta
4° de primaria				
Lógico-Matemática	73,52	2,69	1,25	0,89
Comunicación Integral	81,85	3,80	1,11	0,87
Personal Social	84,71	3,89	1,28	0,84
Ciencia y Ambiente	77,33	3,23	1,17	0,89
6° de primaria				
Matemática	81,15	3,59	1,13	0,86
Lenguaje	86,45	4,42	1,08	0,84
Ciencias Histórico-Sociales	80,98	3,92	1,05	0,88
Ciencias Naturales	80,64	3,51	1,21	0,88
4° de secundaria				
Matemática Forma 1	70,23	2,54	1,23	0,89
Matemática Forma 2	74,77	2,89	1,25	0,88
Lenguaje y Literatura	76,73	3,26	1,10	0,91
5° de secundaria				
Matemática Forma 1	76,56	3,26	1,06	0,88
Matemática Forma 2	78,01	3,25	1,19	0,87
Lenguaje y Literatura	79,47	3,37	1,19	0,90

neos. Por otro lado, en Lenguaje de sexto de primaria las áreas temáticas son similares a las de secundaria.

UNIDIMENSIONALIDAD

Todas las pruebas arrojaron un grado de unidimensionalidad (porcentaje de la varianza explicada de la primera dimensión) de 70% o más, lo que nos permite concluir que el conjunto de preguntas mide un solo aspecto. Las pruebas de Matemática arrojaron unidimensionalidad relativamente baja (entre 70 y 80%). Otras áreas como Lenguaje, Personal Social, Comunicación Integral, Ciencias Histórico-Sociales y Ciencias Naturales arrojaron una unidimensionalidad mayor de 80%.

Como en el índice anterior, es posible que las pruebas de Matemática midan áreas relativamente más heterogéneas que las otras pruebas. Sin embargo, como la consistencia interna de las pruebas es alta, se puede concluir que la principal fuente de variación no es el contenido de las pruebas sino la variación proveniente de los estudiantes. En el primer acápite se comentó, además, que las variaciones provenientes de otros factores situacionales del muestreo no fueron importantes. Un paso siguiente en el análisis de unidimensionalidad es aplicar el esquema de la Teoría de Respuesta al Ítem en el sentido de analizar la varianza proveniente de los sujetos (habilidades y actitudes).

OTROS ANÁLISIS BASADOS EN LAS PREGUNTAS

Los siguientes datos en el nivel de prueba se basan en los resultados del análisis de las preguntas[17]; esto quiere decir que los resultados de dificultad media, discriminación media, correlación pregunta-prueba media e índice de respuestas medio son los promedios de los indicadores respectivos de las preguntas que componen cada prueba.

Correlación pregunta-prueba

Los coeficientes de validez se encontraron en el rango 32,32%-42,90%, y corresponden a las pruebas de Lenguaje y Literatura de cuarto de secundaria y Lenguaje de sexto de primaria respectivamente. El rango está sobre el nivel del 20%, que califica una buena validez.

Dificultad[18]

El rango de dificultad medio varió desde 44,12% para Lenguaje y Literatura de cuarto grado de secundaria a 61,90% para Ciencias Naturales de sexto de primaria. Este rango no se aparta mucho del nivel de dificultad promedio recomendable de 50%. Merece notarse que las pruebas de primaria, salvo la de Matemática de sexto grado, estuvieron por encima de una dificultad del 50%. Por otro lado, y contrariamente, las pruebas de secundaria, excepto la de Lenguaje y Literatura de quinto de secundaria, estuvieron por debajo del valor medio de dificultad.

Discriminacion[19]

La prueba que más discriminó fue la de Lenguaje de sexto de primaria, con un coeficiente de discriminación medio de 49,57%, mientras que la que menos discriminó fue la prueba de Lógico-Matemática de cuarto de primaria. El promedio general de discriminación fue de 36,42%.

No respuesta[20]

El rango de no respuesta se encuentra entre 0,78% para la prueba de Ciencias Naturales de sexto de primaria y 10,82% para la prueba de Matemática Forma 2 de quinto de secundaria. Sin embargo, la mediana para primaria es de sólo 1,96%, y para secundaria de 7,11%, lo que indica que la no respuesta fue un pro-

17 Véase la sección precedente, Evaluación psicométrica de las preguntas de las pruebas CRECER 1998.
18 Véase el anexo 2 (glosario) de la sección anterior, p. 164.
19 *Idem* nota 18.
20 *Idem* nota 18.

blema sólo en las pruebas de secundaria. Se observó que la Forma 2 de Matemática de quinto de secundaria, junto con las pruebas de Matemática de cuarto grado, arrojaron tasas de no respuesta relativamente altas comparadas con el resto de pruebas. Estos resultados nos llevan a concluir que son las pruebas de Matemática de secundaria las que arrojaron las más altas tasas de no respuesta.

4 CONCLUSIONES Y SUGERENCIAS

CONCLUSIONES

Respecto de la validez de las pruebas, y dado que el objetivo trazado es comparar grupos de interés, los resultados obtenidos del análisis de las preguntas a través de una variedad de tipos de evidencias (que van desde aspectos cualitativos como la opinión de expertos sobre los contenidos de las pruebas hasta indicadores más cuantitativos en el análisis de las preguntas) sugieren que las pruebas finales presentan propiedades psicométricas óptimas para su empleo en el análisis de resultados.

En cuanto a las escalas y el uso de los puntajes, los resultados de las pruebas pueden ser reportados con el uso de los porcentajes y la transformación a la escala de Rasch, teniendo en cuenta las ventajas y desventajas de cada cual ya expresadas en este documento. Cuando se quiera estimar totales u otros estadísticos agregados debe usarse el sistema de ponderaciones para recuperar la proporcionalidad del universo.

La estimación de las varianzas y el error estándar para el cálculo de los promedios y otros estadísticos debe tomar en consideración el diseño de muestreo usado. Pueden utilizarse las fórmulas correspondientes del muestreo estratificado o, alternativamente, pueden ser estimados por otros métodos como el de Jackknife[21]. Dicho de otro modo, no es recomendable el uso de las fórmulas de las varianzas del muestreo simple aleatorio.

Otro aspecto en la inferencia es la distribución subyacente en la muestra. La ausencia de normalidad en las distribuciones de los porcentajes de la muestra tiene algunas explicaciones. En primer lugar, el rango posible de valores para los puntajes es truncado, es decir, se limita a uno que va de 0 a aproximadamente 30 (que es el máximo puntaje que se puede obtener). Considerando que en cada prueba hay 16 000 alumnos y muchos empates, la distribución de esta frecuencia es limitada al rango mencionado. Es importante anotar que, en la práctica, muchas de las comparaciones se hacen en un nivel más agregado, como por ejemplo la escuela que toma los promedios en las aulas. En este caso se presentan dos efectos favorables hacia la normalidad: una disminución de casos (existen aproximadamente 580 valores correspondientes a las escuelas) y una suavización de la distribución de casos (los promedios se distribuyen más continuamente), lo que mejora la densidad de frecuencia del intervalo de la escala.

La consecuencia de esto es que el uso de estadísticas paramétricas en las pruebas debe tomarse con cuidado. Es recomendable utilizar métodos complementarios sustentados en el análisis de datos categóricos o cualitativos en los que la asociación y correlación entre variables es estimada a partir de supuestos más flexibles. En el caso de los análisis paramétricos se sugiere el análisis parcial referido a grupos pequeños de categorías relevantes o modelos integrados que estiman efectos parciales en forma escalonada (modelos jerárquicos de análisis).

En síntesis, de los cuadros 2, 3 y 4 podemos concluir que:

- Las pruebas son unidimensionales y presentan alta confiabilidad. Los índices de validez basados en las preguntas son aceptables, lo que permite concluir que en el nivel global de prueba éstas poseen buenas características desde el punto de vista psicométrico.
- Las características adicionales presentadas reflejan un conjunto de pruebas con buena discriminación, centradas en una dificultad intermedia que garantiza las comparaciones por estratos de la muestra. Adicionalmente, las tasas de no respuesta, con excepción de las pruebas de Mate-

21 Véase Calderón, Farro y Bazán: "Diseño muestral...", *op. cit.*

mática de secundaria, han resultado poco significativas.
- El número de preguntas de las pruebas CRECER 1998 resultó apropiado, pues se minimizó la tasa de no respuesta de la etapa piloto, especialmente en el área de Matemática (en primaria el problema ha desaparecido y en secundaria se ha minimizado). La estrategia de utilizar dos formas de las pruebas permite obtener mayor número de preguntas.

SUGERENCIAS

En el aspecto metodológico fue útil recoger una serie de recomendaciones que nos servirán para futuras aplicaciones. Las principales sugerencias son:

- Procurar una correspondencia entre el enfoque de evaluación (normas y criterios) y la metodología de construcción de pruebas (Teoría Clásica de los Tests, Teoría de Respuesta al Ítem), de manera que esta última proporcione las condiciones para el uso final de la información.
- Realizar una supervisión continua de los equipos responsables de las pruebas para garantizar el seguimiento similar (uniforme) y correcto de la metodología de construcción de pruebas y así eliminar diferencias en los procedimientos seguidos.
- Realizar análisis cuantitativos y cualitativos, de manera que las decisiones a tomar durante el seguimiento de la metodología de construcción de pruebas sean las más adecuadas.
- Implementar una etapa de verificación de cambios entre la etapa piloto y la aplicación definitiva, de manera que se anticipe el comportamiento final de las pruebas en términos de validez, confiabilidad e indicadores psicométricos agregados.
- Incorporar en la etapa piloto el análisis de sesgos (validez transcultural) y estudios complementarios de diferenciabilidad de las preguntas y prueba.

Desde una perspectiva más amplia, las bondades que nos ofrecen las estimaciones de los indicadores del modelo clásico pueden ser complementadas por otros modelos como el de Teoría de Respuesta a los Ítemes. Para ello será necesario evaluar también las preguntas desde el punto de vista de si los supuestos implícitos de estos modelos se cumplen o no. Sin embargo, dado que se ha planteado el enfoque de normas, que conduce a una metodología útil para las comparaciones de grupos, el uso de las propiedades que se desprenden de la Teoría Clásica de los Tests es suficiente para la evaluación de estas pruebas.

ANEXO 1

TABLAS DE ESPECIFICACIONES DE LAS PRUEBAS CRECER 1998

Prueba de Lógico Matemática de 4° grado de educación primaria					
Competencias		**Contenidos**	**Ítemes**	**Total**	**%**
1. Conocimiento de los números y la numeración	101	Relación de orden en los números naturales	2		
	102	Relación de orden con los números fraccionarios	2		
	103	Relación de orden con los números decimales	1	5	16,67
2. Habilidad operativa y cálculo	201	Resolución de ejercicios	2		
	202	Resolución de problemas utilizando las cuatro operaciones fundamentales con los números naturales	4		
	203	Resolución de problemas utilizando las operaciones fundamentales con fracciones homogéneas	2		
	204	Conversión de fracciones a números decimales	2		
	205	Resolución de problemas utilizando las operaciones fundamentales con números decimales	2	12	40,00
3. Medición	301	Estima longitud y masa en unidades convencionales y no convencionales	3		
	302	Resuelve problemas de compra y venta relacionados con las unidades de masa	2		
	303	Resuelve problemas de compra y venta relacionados con las unidades de longitud	2		
	304	Resuelve problemas y ejercicios de compra y venta relacionados con las unidades de tiempo	2	9	30,00
4. Geometría	401	Clasificación de ángulos y rectas paralelas y perpendiculares	2		
	402	Identifica figuras geométricas en el plano	2	4	13,33
			Total	30	100,00

Prueba de Comunicación Integral de 4° grado de educación primaria					
Competencias		**Contenidos**	**Ítemes**	**Total**	**%**
1. Comunicación escrita: lectura	101	Identifica tipos de textos	2		
	102	Reconoce información específica de un texto	5		
	103	Extrae la idea principal	2		
	104	Reconoce e infiere la causa y efecto de los sucesos	2		
	105	Encuentra el significado de palabras y frases por el contexto	2	13	43,30
2. Reflexión sobre la lengua	201	Construye oraciones con sentido	2		
	202	Identifica el sujeto y el predicado de la oración	2		
	203	Reconoce sustantivos y adjetivos	2		
	204	Identifica y usa los tiempos fundamentales de los verbos	2		
	205	Reconoce y utiliza los sinónimos y antónimos	2		
	206	Ordena alfabéticamente palabras en un contexto dado	3	13	43,30
3. Lectura de imágenes	301	Identifica diversos textos gráficos	1		
	302	Obtiene información a partir de ilustraciones	1		
	303	Analiza los elementos de las ilustraciones y gráficos	1		
	304	Descubre la intencionalidad del autor	1	4	13,30
			Total	30	100,00

Prueba de Personal Social de 4° grado de educación primaria					
Competencias		**Contenidos**	**Ítemes**	**Total**	**%**
1. Cuidado de su salud personal y colectiva	101	Riesgos para su salud física y comportamientos que previenen enfermedades	3		
	102	Identificación y respeto de las normas de seguridad vial	2		
	103	Cumplimiento de las indicaciones de Defensa Civil	2	7	25,90
2. Convivencia democrática	201	Comprensión de los deberes y derechos al interior de la escuela, en la familia y comunidad	4		
	202	Respeto de las diferentes costumbres, opiniones y gustos	2	6	22,20
3. Sentimiento de pertenencia	301	Identificación de los principales acontecimientos del proceso histórico peruano y ubicación cronológica	3		
	302	Identificación de los símbolos patrios y sus características	2		
	303	Análisis de las funciones de las principales instituciones de la localidad	2		
	304	Reconocimiento de la importancia de la conservación del patrimonio cultural del Perú	2	9	33,30
4. Conocimiento de su medio sociohistórico y natural	401	Ubicación de los países y departamentos limítrofes del Perú	2		
	402	Reconocimiento de las actividades económicas de las diversas regiones del Perú	3	5	18,50
			Total	27	100,00

Prueba de Ciencia y Ambiente de 4° grado de educación primaria					
Competencias		**Contenidos**	**Ítemes**	**Total**	**%**
1. Conservación de su salud	101	Recuperación de energías mediante la alimentación	4		
	102	Normas y códigos para conservar la salud	2	6	20,00
2. Conservación del medio ambiente	201	Plantas y animales más representativos de diferentes medios	3		
	202	Plantas y animales como recursos fundamentales para la supervivencia humana. Sus cuidados y uso racional	6		
	203	Contaminantes más frecuentes del aire, agua y suelo, y su influencia en los seres vivos	3		
	204	La importancia del agua para la vida y sus cambios de estado físico por acción del calor	5	17	56,67
3. Intervención humana en el medio	301	Recursos naturales de la región, recursos renovables y no renovables	3		
	302	Las transformaciones del paisaje y los recursos naturales producidas por los seres humanos para satisfacer sus necesidades	4	7	23,33
			Total	30	100,00

Prueba de Matemática de 6° grado de educación primaria					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Números naturales	101	Relación de orden en los números naturales	2		
	102	Ejercicios de multiplicación y división	2		
	103	Operaciones combinadas de adición, sustracción, multiplicación y división (hasta dos operaciones)	2		
	104	Resolución de problemas que requieren de dos operaciones fundamentales	1		
	105	Uso del algoritmo de mínimo común múltiplo	1	8	26,67
2. Fracciones	201	Relación de orden y simplificación de fracciones	2		
	202	Ejercicios de operaciones con fracciones	2		
	203	Resolución de problemas de la vida cotidiana utilizando operaciones con fracciones	2	6	20,00
3. Números decimales	301	Relación de orden de decimales	1		
	302	Operaciones combinadas de decimales	1		
	303	Resolución de problemas de la vida cotidiana utilizando operaciones con decimales	2	4	13,33
4. Proporcionalidad	401	Resolución de problemas de la vida cotidiana utilizando regla de tres simple	1		
	402	Resolución de problemas de la vida cotidiana utilizando porcentajes	1	2	6,67
5. Medición	501	Resolución de problemas de la vida cotidiana utilizando unidades de medición	3		
	502	Resolución de problemas de compra usando la conversión de unidades de medición	1	4	13,33
6. Geometría	601	Clasificación de ángulos y rectas	2		
	602	Cálculo y resolución de problemas con base en el perímetro de un cuadrilátero	2		
	603	Concepto de sólidos geométricos	1		
	604	Cálculo del volumen conjunto de cubos	1	6	20,00
			Total	30	100,00

Prueba de Lenguaje de 6° grado de educación primaria					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Comprensión de lectura	101	Identificar los hechos o ideas principales de textos leídos	4		
	102	Extraer información específica de un texto	4		
	103	Descubrir la intencionalidad del autor	2		
	104	Reconocer las marcas de cohesión textual	1	11	35,48
2. Nociones gramaticales	201	Reconocer la oración gramatical como unidad de sentido	2		
	202	Reconocer los núcleos del sujeto y del predicado	2		
	203	Reconocer las clases de palabras y sus accidentes	2		
	204	Conocer el empleo de los signos de puntuación	2		
	205	Usar correctamente las grafías: g-j, x, s-c-z, h, b-v, ll-y	2	10	32,26
3. Vocabulario	301	Encontrar el significado de palabras por el contexto	2		
	302	Identificar las familias de palabras	2		
	303	Reconocer las relaciones semánticas entre las palabras	2	6	19,35
4. Análisis de imágenes	401	Identificar diversos textos gráficos	1		
	402	Obtener información a partir de ilustraciones	1		
	403	Analizar los elementos de las ilustraciones	1		
	404	Descubrir la intencionalidad del autor	1	4	12,90
			Total	31	100,00

Prueba de Ciencias Naturales de 6° grado de educación primaria					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Transformaciones de la materia; mezcla y combinación	101	Transformaciones de la materia. Cambios de estado	3	3	10
2. Funciones del cuerpo humano	201	Sistema digestivo: órganos, funciones, relaciones con otros sistemas	4		
	202	Sistema respiratorio: órganos, relaciones: respiración, aire y fotosíntesis	2		
	203	Sistema circulatorio: órganos y funciones	2		
	204	Función excretora, sistema urinario	2		
	205	Función de relación: irritabilidad (sentidos)	2		
	206	Función de reproducción, órganos, paternidad responsable	2	14	47
3. Interacciones entre materia y energía	301	Fotosíntesis, luz, propiedades	4		
	302	El calor y climas. Buenos y malos conductores del calor	2		
	303	La energía eléctrica, fuentes y canales de transmisión	2		
	304	Imanes, propiedades. Magnetismo, sus aplicación en los aparatos inventados por el hombre	2	10	33
4. Conservación del medio ambiente	401	Contaminación ambiental, problemas y alternativas de solución	3	3	10
			Total	30	100

Prueba de Ciencias Histórico-Sociales de 6° grado de educación primaria					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Historia y cultura peruana	101	Identificación del origen de la cultura peruana y el proceso histórico nacional	3		
	102	Explicación de los factores ideológicos, políticos, económicos y sociales en la independencia	2		
	103	Análisis del proceso de formación de la conciencia nacional	2		
	104	Identificación de los símbolos patrios	1	8	26,70
2. Forjadores	201	Comprensión de las acciones de los personajes modelos de patriotismo	3	3	10,00
3. Estado peruano	301	Análisis de los principales tratados limítrofes firmados por el Estado peruano	2		
	302	Reconocimiento de la labor de los principales poderes del Estado	2	4	13,30
4. Reconocimiento del universo	401	Reconocimiento de la estructura y características del sistema planetario solar	2		
	402	Identificación de los elementos y movimiento de la tierra	1		
	403	Reconocimiento de las líneas imaginarias de la tierra y su utilización para la ubicación de puntos geográficos	1	4	13,30
5. Medio geográfico	501	Ubicación de las regiones donde se encuentran los distintos recursos naturales y actividades económicas	4		
	502	Análisis de las causas y efectos de la migración y el crecimiento urbano	2	6	20,00
6. Convivencia en sociedad	601	Comprensión de los deberes y derechos de la persona al interior de la escuela, en la familia y la comunidad	3		
	602	Reconocimiento de las normas de Defensa Civil	1		
	603	Comprensión de las principales normas de educación y seguridad vial	1	5	16,70
			Total	30	100,00

Prueba de Lenguaje y Literatura de 4° grado de educación secundaria					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Comprensión lectora	101	Extraer, deducir e interpretar información del texto	3		
	102	Identificar y deducir características de los personajes	4		
	103	Identificar la intencionalidad del autor	2		
	104	Identificar la idea central y derivar conclusiones de un texto	1		
	105	Reconocer géneros, estructuras, técnicas y estilos literarios	7	17	44,74
2. Nociones y reglas gramaticales	201	Usar los modos y tiempos verbales	1		
	202	Establecer y aplicar reglas de concordancia	2		
	203	Identificar, clasificar y usar correctamente las categorías gramaticales	4	7	18,42
3. Análisis de imágenes	301	Interpretar y analizar imágenes fijas, deduciendo su intencionalidad	4	4	10,53
4. Razonamiento verbal	401	Identificar el significado de palabras y expresiones según el contexto	2		
	402	Relacionar palabras a través de analogías	2		
	403	Diferenciar la formación de palabras	2		
	404	Establecer relaciones semánticas entre palabras	2		
	405	Reconocer las familias de palabras	2	10	26,32
			Total	38	100,00

Prueba de Lenguaje y Literatura de 5° grado de educación secundaria					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Comprensión lectora	101	Extraer, deducir e interpretar información del texto	5		
	102	Identificar y deducir características de los personajes	3		
	103	Identificar la intencionalidad del autor	2		
	104	Identificar la idea central y derivar conclusiones de un texto	5		
	105	Reconocer géneros, estructuras, técnicas y estilos literarios	3	18	47,37
2. Nociones y reglas gramaticales	201	Usar los modos y tiempos verbales correctamente	2		
	203	Reconocer las oraciones compuestas	2		
	203	Utilizar las preposiciones y adverbios correctamente	2		
	204	Reconocer la función sintáctica de la oración subordinada o proposición	2	8	21,05
3. Análisis de imágenes	301	Interpretar imágenes fijas y reconocer el contexto en que se realizan	1		
	302	Identificar la intencionalidad del autor o emisor	1	2	5,26
4. Razonamiento verbal	401	Deducir el significado de palabras y expresiones según el contexto	2		
	402	Relacionar palabras a través de analogías	2		
	403	Reconocer elementos comunes en la formación de palabras	2		
	404	Establecer relaciones semánticas entre palabras	2		
	405	Reconocer las familias de palabras	2	10	26,32
			Total	38	100,00

Prueba de Matemática de 4° grado de educación secundaria F1 y F2					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Arimética	101	Resolver problemas de la vida real usando las operaciones fundamentales con números racionales (Q)	1		
	102	Resolver problemas de la vida real usando las operaciones fundamentales con números decimales	1		
	103	Resolver situaciones problemáticas usando las operaciones fundamentales con números reales (R)	1		
	104	Resolver problemas de la vida real que impliquen el uso de los conceptos de proporcionalidad directa e inversa	1		
	105	Resolver problemas de la vida real usando el sistema internacional de unidades y sistema monetario peruano	2		
	106	Resolver ejercicios de racionalización	1	7	28,00
2. Álgebra	201	Factorizar polinomios	1		
	202	Operar con fracciones algebraicas	1		
	203	Resolver ecuaciones de segundo grado	1		
	204	Resolver problemas de la vida cotidiana utilizando ecuaciones de primer grado	1		
	205	Resolver situaciones problemáticas de la vida real utilizando sistemas de ecuaciones lineales con dos incógnitas	1		
	206	Resolver ecuaciones lineales e inecuaciones de primer grado con o sin valor absoluto	2	7	28,00
3. Estadística	301	Interpretar gráficas y cuadros estadísticos	2		
	302	Resolver problemas de la vida real con medidas de tendencia central	1	3	12,00
4. Geometría	401	Resolver problemas utilizando conceptos geométricos y propiedades de ángulos y rectas	3		
	402	Resolver situaciones problemáticas aplicando propiedades de triángulos	1		
	403	Resolver problemas aplicando propiedades de la circunferencia	1		
	404	Resolver situaciones problemáticas que impliquen el uso de propiedades de proporcionalidad entre figuras geométricas	1		
	405	Resolver situaciones problemáticas que impliquen el cálculo de perímetros y áreas de cuadriláteros y hexágonos	1		
	406	Resolver situaciones problemáticas que impliquen el cálculo de áreas y volúmenes de prismas y cilindros	1	8	32,00
			Total	25	100,00

Prueba de Matemática de 5° grado de educación secundaria F1 y F2					
Objetivos		**Contenidos**	**Ítemes**	**Total**	**%**
1. Conjuntos	101	Resolver situaciones problemáticas utilizando las operaciones fundamentales de conjuntos	1	1	4,00
2. Arimética	201	Desarrollar ejercicios sobre progresión aritmética y geométrica	1		
	202	Resolver problemas de la vida cotidiana usando las operaciones fundamentales con números reales	1		
	203	Resolver situaciones problemáticas de la vida real utilizando las unidades del sistema internacional	2		
	204	Resolver situaciones problemáticas que impliquen la aplicación de conceptos y propiedades de proporcionalidad	1		
	205	Resolver problemas de la vida real usando regla de interés	1	6	24,00
3. Álgebra	301	Realizar operaciones combinadas de adición, sustracción, multiplicación y división de polinomios	1		
	302	Resolver problemas de la vida cotidiana con expresiones algebraicas	2		
	303	Resolver ecuaciones de segundo grado	1		
	304	Resolver problemas de la vida real utilizando ecuaciones y sistema de ecuaciones lineales con dos incógnitas	1	5	20,00
4. Estadística	401	Resolver problemas de la vida real con medidas de tendencia central e interpretación de cuadros y gráficos	3		
	402	Resolver problemas de la vida real utilizando propiedades fundamentales del análisis combinatorio	1	4	16,00
5. Geometría	501	Resolver problemas sobre triángulos y cuadriláteros aplicando ángulos formados por dos paralelas y una secante	2		
	502	Resolver problemas utilizando la relación de proporcionalidad entre figuras geométricas	2		
	503	Resolver situaciones problemáticas que impliquen calcular el área de un polígono cualquiera	1		
	504	Resolver problemas sobre áreas de polígono regular inscrito y circunscrito en una circunferencia	1	6	24,00
6. Trigonometría	601	Resolver problemas sobre longitud de arco y sector circular usando el concepto y las propiedades de ángulo trigonométrico	1		
	602	Resolver problemas de la vida real sobre ángulos de elevación y depresión usando propiedades de funciones trigonométricas	1		
	603	Resolver problemas que impliquen la reducción de ángulos al primer cuadrante y el uso de ángulos notables	1	3	12,00
			Total	25	100,00

ANEXO 2

DESARROLLO Y CONSTRUCCIŒN DE LAS PRUEBAS

En la elaboración de las pruebas participaron prestigiados expertos nacionales e internacionales en aspectos de evaluación, sistemas nacionales de evaluación, política educativa y análisis de datos, quienes trabajaron en coordinación con especialistas de diferentes dependencias del MED. También se contó con la participación de profesores en diversos talleres de sensibilización y capacitación en los que se generó parte de las preguntas de las pruebas.

1. SELECCIŒN DE LAS «REAS ELEGIDAS

ANÁLISIS CURRICULAR

La primera etapa consistió en revisar la estructura curricular (los programas curriculares) correspondiente a las áreas de desarrollo evaluadas. Este proceso derivó en la selección de aquellos aspectos que fueron contemplados en las pruebas. La tarea estuvo a cargo de especialistas de la UMC, en coordinación con los expertos en currículo de la Dirección de Educación Primaria y con profesores especialistas de cada una de las áreas de desarrollo evaluadas.

La selección curricular se expresa en los cuadros de especificaciones elaborados a manera de matrices para guiar la generación o redacción de las preguntas. De manera específica, en el segundo acápite y en el anexo 1 podemos ver los temas seleccionados para las pruebas en cada área escogida.

REDACCIÓN DE LAS PREGUNTAS DE LAS PRUEBAS

Una vez definidos los aspectos por evaluar, sintetizados en los cuadros de especificaciones, se prepararon pautas para la elaboración de las preguntas. En esta parte del proceso, profesores especialistas de diferentes lugares del país participaron en talleres de sensibilización. Se trata de una selección de maestros de aula de las 16 USE de Lima y de las ciudades de Huaraz, Tumbes, Iquitos, Huánuco, Pucallpa, Tacna, Puno, Ayacucho, Junín, Cajamarca, Madre de Dios y Piura.

Asistieron 432 profesores que fueron capacitados para la tarea encomendada y elaboraron preguntas que, después de ser revisadas, pasaron a formar parte del banco de preguntas para cada área o asignatura. Por su parte, además de revisar las preguntas generadas en los talleres, los diferentes equipos de la UMC elaboraron sus propias preguntas para someterlas a prueba durante la aplicación piloto.

En las 12 pruebas aplicadas las preguntas generadas son del tipo de selección de respuesta (o de opción múltiple). Este tipo de pregunta consta de un enunciado que interroga o plantea una situación problemática y de cuatro alternativas de respuesta, de las cuales sólo una es correcta. La calificación de las preguntas asigna 1 punto a la respuesta correcta y 0 puntos a la respuesta incorrecta o a la pregunta sin respuesta.

En cuanto a la redacción o formulación de este tipo de preguntas, se tomaron en consideración diversas pautas para formular el enunciado y para estructurar las alternativas. Asimismo, y gracias a la tabla de especificaciones, se pudieron redactar preguntas que exploraran distintos niveles del dominio cognoscitivo.

2. LAS PRUEBAS PILOTO

Una vez construidas y revisadas las preguntas se elaboraron 10 formatos de prueba para cada área de desarrollo evaluada, con un promedio de 35 preguntas cada una. Estas pruebas preliminares se aplicaron en el mes de noviembre de 1997 a una muestra nacional de 3240 estudiantes (324 por formato) distribuidos en 156 escuelas del país.

3. EL AN«LISIS DE LAS PREGUNTAS

Las respuestas dadas por los sujetos a las pruebas piloto fueron sometidas a análisis estadísticos (análisis psicométricos) con el fin de conocer el nivel de dificultad del ítem, su poder discriminatorio (que posibilite la comparación entre grupos de estudiantes), el porcentaje de no respuesta, entre otros. Esto permitió eliminar parte importante de las preguntas de la prueba y mantener, por consiguiente, las más pertinentes. Los resultados de esta etapa se presentan en la sección Evaluación psicométrica de las preguntas de las pruebas CRECER 1998.

Las preguntas también fueron evaluadas por especialistas que vieron el grado de correspondencia de cada una con la tabla de especificaciones. Este proceso permitió seleccionar las preguntas más apropiadas y satisfacer los requerimientos de validez interna y de contenido de aquellas que pasaron a formar parte de la prueba final.

4. SELECCIŒN DE TEMES DEFINITIVOS

Las mejores preguntas (a partir de criterios presentados en el acápite Diseño muestral) sirvieron de base para la confección de las pruebas finales. Esta última versión cuenta con criterios óptimos de validez, pertinencia pedagógica y la confiabilidad del caso.

La versión final de las pruebas se aplicó en diciembre de 1998. La aplicación fue supervisada por especialistas y coordinadores en cada uno de los CE seleccionados. Ningún profesor del CE donde se realizaba la evaluación participó en la aplicación de las pruebas; la tarea recayó en profesores de otros ámbitos especialmente entrenados para esta aplicación. GRADE participó y apoyó logísticamente este proceso.

5. PROCESAMIENTO Y AN«LISIS

Una vez aplicadas todas las pruebas de rendimiento, los cuadernillos se remitieron a la UMC para su procesamiento computarizado. En vista de que los estudiantes marcaron sus respuestas en formatos de lectora óptica, éstos fueron leídos y luego ingresados automáticamente en la base de datos.

La información incluida se encuentra codificada de tal forma que posibilita el desarrollo de los análisis planificados. Las bases de datos han sido depuradas con el fin de identificar adecuadamente las respuestas a las pruebas.

La información ingresada a la base de datos permitirá generar múltiples reportes y comunicaciones oficiales acerca de los resultados de las pruebas CRECER 1998. Se trata de reportes diversos que pretenden ofrecer insumos relevantes para la toma de decisiones a los diferentes actores del entorno educacional y, asimismo, motivar la reflexión abierta sobre nuestras principales debilidades y fortalezas.