MINISTERIO DE EDUCACIÓN
Vice Ministerio de Gestión Institucional
Unidad de Medición de Calidad Educativa

EVALUACIÓN PSICOMÉTRICA DE LAS PREGUNTAS Y PRUEBAS CRECER 96

Jorge Bazán Guzmán

Lima, Noviembre 2000

En el presente documento se va a reportar la evaluación psicométrica de las preguntas y pruebas de la Aplicación CRECER 96. Se revisarán los criterios utilizados para determinar si las preguntas de las pruebas son útiles en el sentido psicométrico para calcular un puntaje total de la prueba. De otro lado proporcionar indicadores psicométricos sobre las bondades del puntaje total de la prueba para el empleo en el análisis de rendimiento escolar bajo el enfoque de normas que posteriormente se precisa. Son las pruebas nacionales aplicadas en 1996: las de Lenguaje y Matemática.

La evaluación psicométrica de las preguntas se presenta en la sección 1 y la evaluación de las pruebas en la sección 2. La evaluación de las preguntas se basa en la metodología descrita en la sección 1.1 Allí se expone el conjunto de criterios que nos permitirán presentar los resultados de la sección 1.2. En esta sección se muestra las preguntas que arrojaron mayor problema, ilustrando en cada caso los criterios y las decisiones tomadas. La evaluación de las pruebas se basa en el marco de análisis descrito en la sección 2.1, es decir, el conjunto de criterios para la evaluación psicométrica de las pruebas. La sección 2.2 contiene los constructos y contenidos de las pruebas. La sección 2.3 resume las conclusiones derivadas de los resultados encontrados presentando los indicadores estimados para las pruebas. Finalmente la sección 3 resume las conclusiones de este documento.
El anexo 1 agrega más detalle sobre la evaluación de las preguntas observadas.

1. EVALUACIÓN PSICOMÉTRICA DE LAS PREGUNTAS DE LAS PRUEBAS CRECER 96

1.1 Metodología para el análisis de las preguntas

Uno de los objetivos en la construcción de pruebas es, obtener una prueba de longitud mínima (de menor número de preguntas) que presente las mejores propiedades para la conformación del puntaje total de la prueba. Esto usualmente se logra a través del análisis de preguntas. En esta sección se exponen los criterios que han sido base para la evaluación psicométrica las preguntas. Pese a que el análisis de preguntas es una etapa de la construcción de la pruebas asociada al momento posterior de la aplicación piloto, también se da posterior a la aplicación definitiva. El objetivo en el análisis de las preguntas después de la aplicación definitiva es eliminar las preguntas que no satisfacen un conjunto de propiedades que hacen de la prueba un instrumento apropiado e insesgado en la estimación del rendimiento.

Estas propiedades están supeditadas al marco referencial teórico (modelo) que se adopte para el análisis de las pruebas. Como se sabe los modelos más usados son los de la Teoría Clásica de los tests y los modelos de la familia de la Teoría de Respuestas a Items o TRI (para una revisión de estos conceptos ver por ejemplo Muñiz, 1993; 1990). En el análisis de las pruebas CRECER 96 se ha aplicado los criterios que han sido definidos bajo el enfoque clásico tales como comparabilidad, variabilidad de la prueba y que determinan un conjunto de indicadores referenciales que a continuación se presenta. Los criterio de TRI no fueron empleados para establecer las bondades de las preguntas.

La evaluación de las preguntas de CRECER 96 se ha realizado con posterioridad a la aplicación tomando como referencia la evaluación correspondiente en CRECER 98. Algunos aspectos de más detalle de la metodología seguida en CRECER 98 puede consultarse en dichos informe.

Criterios para el análisis de las preguntas

El análisis de las preguntas se hace desde dos dimensiones, cualitativo y cuantitativo. Ambos son complementarios.

Dimensión cualitativa

Desde el punto de vista cualitativo el análisis de las preguntas incluyen su revisión en su forma (enunciado y alternativas de respuesta), atendiendo a la característica que presentan los distractores de las preguntas, la calidad de la clave o respuesta, es decir, si ésta está bien o mal especificada. Se incluye también el análisis de la distribución de las opciones en términos de porcentajes ocurridos tanto para la respuesta como para los distractores.

Este tipo de análisis es parte del proceso final de construcción de la prueba y está a cargo de los especialistas o grupo de expertos responsables del desarrollo de las pruebas desde la etapa piloto. Al momento de la definitiva lo que se hace es certificar estas propiedades pues la prueba definitiva constituye un ensamblaje de las mejores preguntas de las versiones piloteadas que no necesariamente fueron juntas o inclusive recibieron modificaciones.

Dimensión cuantitativa

Desde el punto de vista cuantitativo, el análisis de las preguntas incluye el cálculo de los índices que corresponden a las propiedades psicométricas de las preguntas. Estas propiedades se definen de acuerdo al modelo de análisis adoptado. Los criterios aquí presentados se refieren a la teoría clásica de los Tests: Validez (correlación pregunta - prueba), Dificultad, Discriminación, Índice de no respuesta.

El examen de estos indicadores, con los niveles referenciales que se presentan a continuación, nos permiten concluir si las preguntas han sido elaboradas en forma óptima, en cuyo caso el cálculo de los puntajes obtenidos por los evaluados serán válidos, confiables y diferenciables para el uso que se les desea dar. En lo que sigue se presentan los indicadores de la dimensión cuantitativa del análisis de las preguntas a las preguntas.

Es importante notar que las categorías presentadas para cada criterio (ver filas de los cuadros 1 al 4) son niveles referenciales propuestas por la Unidad de Medición de la Calidad, basadas en las apreciaciones encontradas en la literatura (por ejemplo Lazarte, 1995), así como en las decisiones tomadas por la propia unidad. En cada caso se definen y comentan estas categorías discutiendo sus implicancias para el análisis de las preguntas.

En lo que sigue las respuestas a las preguntas se consideran dicotómicas en cuanto que aciertan o no la respuesta correcta.

Validez de las preguntas

La validez de una pregunta en una prueba nos permite determinar si la pregunta mide lo que la prueba se propone medir. En este sentido la correlación pregunta-prueba (punto biserial) es el indicador que nos mide el grado de validez de la pregunta. Es un índice de validez interna pues la correlación se calcula con el puntaje total de la prueba, a falta del verdadero criterio o constructo.

El siguiente cuadro presenta las categorías que se usaron para evaluar la validez de las preguntas.

Cuadro 1:

Clasificación de la validez según correlación pregunta prueba

Clasificación	Índice de Validez
MUY BUENA	0.20 – 1
ACEPTABLE	0 - 0.19
ELIMINAR	< 0

Dificultad de las preguntas

La dificultad de las preguntas se mide en base al porcentaje de aciertos en la pregunta. El indicador varía entre 0 y 1. Una pregunta con una dificultad de 0.3 es más difícil que una pregunta con una dificultad de 0.8. En el primer caso solo el 30% acertaron la pregunta y en el segundo caso acertaron el 80%. Es decir, y como se ha observado en la literatura, el indicador se mide en forma inversa a la dificultad de la pregunta.

Aplicada a las pruebas CRECER, el índice de dificultad nos indica el grado de dificultad de cada pregunta en las áreas seleccionadas. Si una pregunta tiene un índice de dificultad cercana a 0 ó 1, la pregunta generalmente debería de ser alterada o descartada por no estar dando información acerca de las diferencias entre las habilidades de los examinados.

Una pregunta tiene una dificultad media si fue respondida correctamente por aproximadamente el 50% de los examinados, o sea, su índice de dificultad es cercano a 0.5. Generalmente índices de dificultad entre 0.3 y 0.7 maximizan la información que el test provee sobre la diferencia entre los examinados.

El siguiente cuadro presenta las categorías para determinar el grado de dificultad de las preguntas de las pruebas CRECER.

Cuadro 2

Clasificación del nivel de dificultad de las preguntas

Clasificación	Índice de dificultad
MUY FÁCIL	0.75 – 1
FÁCIL	0.55 - 0.74
INTERMEDIO	0.45 - 0.54
DIFÍCIL	0.25 - 0.44
MUY DIFÍCIL	0.00 - 0.24

Para la selección de las preguntas al momento de elaborar las pruebas definitivas, la Unidad de Medición de la Calidad, UMC, priorizó preguntas con índices de dificultad intermedia (de 0.45 a 0.54). En menor medida se consideró las preguntas fáciles y difíciles. No se consideraron preguntas muy fáciles o muy difíciles (mayor de 0.74 o menor de 0.25 respectivamente).

Adicionalmente a la consideración del porcentaje de examinados que eligieron la opción correcta es importante también el análisis del porcentaje de examinados que eligieron cada opción en cada pregunta. Las opciones de cada pregunta que son incorrectas se llaman distractores.

Discriminación de las preguntas

La discriminación de una pregunta se mide por el grado en que la pregunta ayuda a ampliar las diferencias estimadas entre los que obtuvieron un puntaje total de la prueba relativamente alto de los que obtuvieron un puntaje relativamente bajo. El índice de este indicador, varía entre –1 y 1. Valores positivos indican que la pregunta discrimina a favor del grupo superior, negativo indica que la pregunta es discriminadora que favorece al grupo inferior.

El siguiente cuadro presenta las categorías del índice de discriminación usadas en el análisis de las preguntas.

Cuadro 3

Clasificación de la discriminación de las preguntas

Clasificación	Índice de Discriminación
MUY ALTA	0.40 – 1
ALTA	0.30 - 0.39
MODERADA	0.20 - 0.29
MUY BAJA	0 - 0.19
ELIMINAR	< 0

La discriminación de un item es muy alta si su valor está entre 0.4 y 1. La discriminación del item es muy baja si su valor está entre 0 y 0.19. Si los valores son negativos la pregunta debe eliminarse.

No Respuesta de las preguntas

El índice de no respuesta de una pregunta, se mide por la proporción de personas que no contestan la pregunta (no señalan ninguna opción como la correcta) y/o por la proporción de personas que indicando una opción como la correcta, la invalidan. El programa de lectura óptica de CRECER 96 invalida una respuesta a una opción si hay además otras opciones elegidas, o cuando esta opción tiene un trazo que no es detectado por el programa.

El índice de no-respuesta , Nr_i , se define por

$$Nr_i = 1 - p_i - q^*_i$$

donde p_i es la dificultad de la pregunta y q^*_i es la proporción de personas que no aciertan la pregunta pues marcaron opciones erróneas.

Categorías de no respuestas

El cuadro siguiente presenta una clasificación de niveles de tasa de no-respuesta de las preguntas de acuerdo a los índices de no-respuesta.

Cuadro 4

Clasificación de la no-respuesta de las preguntas

Clasificación	Índice de no-respuesta
ADECUADA	0 - 0.15
ACEPTABLE	0.16 - 0.20
TOLERABLE	0.21 - 0.29
ELIMINABLE	0.30-1

Este indicador es informativo de si el evaluado ha desarrollado o no contenidos relacionados a lo que la pregunta mide. Así, si el indicador es alto es muy probable que el evaluado no conteste la pregunta por desconocimiento del contenido evaluado en ella.

También en el caso de que relacionáramos este indicador con el orden de la pregunta puede ser indicativo del tamaño de la prueba. La posibilidad de que el alumno no haya tenido suficiente tiempo para terminar la prueba, especialmente cuando el índice es más alto en las últimas preguntas, justificaría un tamaño menor de la prueba. Si las últimas preguntas presentan una tendencia a mayores índices de no-respuesta podemos suponer que el alumno no revisó estas preguntas. En este caso la dificultad de las preguntas está subestimada porque no se sabe si los que trabajan más despacio acertarían esta pregunta.
Para la selección de preguntas al momento de elaborar las pruebas definitivas, la UMC priorizó preguntas con índices de no-respuesta que estuvieran en la categoría adecuados. En menor medida consideró las preguntas en las categorías tolerable y aceptabe. No se consideraron preguntas con índices de no-respuesta por encima de 0.30.

Finalmente debemos indicar que para los análisis de las pruebas (véase informe respectivo) la respuesta de no acierto considera además la tasa de no-respuesta. Si bien en el análisis de preguntas es necesario distinguir la no-respuesta, en los resultados finales, la no-respuesta ha sido tomada como no acierto.

Proceso del análisis de las preguntas

En el análisis de las preguntas de las pruebas CRECER 96 se empleó (corrió) el programa ITEMAN (modelo clásico). Para cada prueba se obtuvo un reporte ITEMAN de las preguntas basado en el procesamiento de aproximadamente 45,771 registros (estudiantes) por prueba.

Se formaron equipos de especialistas para revisar el output (salidas) generado por estos programas. Estos especialistas chequearon las propiedades óptimas descritas en las secciones anteriores para cada pregunta de cada prueba. Con el ITEMAN se obtuvo la información de la dimensión cuantitativa, es decir, sobre los índices de discriminación, dificultad, no respuesta y validez (correlación biserial) de las preguntas. Un programa adicional permitió clasificar los índices en las categorías señalas arriba.

En la dimensión cualitativa se analizó información sobre distractores competitivos, respuestas correctas mal especificadas. Todos estos aspectos generaron las perspectivas de los especialistas respecto a la inclusión o eliminación de las preguntas para el puntaje final. El anexo 1 detalla el conjunto de problemas presentados en las preguntas observadas de las distintas pruebas.

Adicionalmente se utilizó el programa RASCAL (modelo Rasch de un parámetro). Con este programa se evaluó la posibilidad de extender el análisis para la consideración de estimar las habilidades de los alumnos en futuras aplicaciones (para la distinción teórica entre puntaje y habilidad véase Muñiz, 1990). Los resultados generados por ambos programas se encuentran disponibles en los documentos "Evaluación del rendimiento estudiantil CRECER 96: Análisis de itemes" (ver serie MED, 2000).

1.2 Resultados de la evaluación de las preguntas

Cuadro 5: Número de preguntas que caen dentro de la clasificación de validez, dificultad, discriminación y no respuesta

		Lenguaje	Matemática
Validez	Muy Buena	30	34
Dificultad	Muy difícil	3	3
	difícil	11	14
	intermedio	8	8
	Fácil	4	8
	Muy fácil	4	1
Discriminación	muy baja	1	
	moderada	2	2
	alta	5	3
	muy alta	22	29
No respuesta	eliminable		11
	aceptable		4
	aceptable		6
	adecuada	30	13

Como se observa en el cuadro 5 la validez de las preguntas medidas por el coeficiente de correlación pregunta- prueba (correlación biserial) es ‘Muy Buena’ en todas las preguntas de ambas pruebas.

Como se menciona en la metodología, no se consideraron preguntas que fueron “muy fáciles” o “muy difíciles” en las pruebas pilotos y que son los rangos extremos de dificultad (índice de dificultad mayor de 0.74 o menor de 0.25 respectivamente). Con las pruebas finales se observó un número relativamente bajo de preguntas que caen en estos rangos extremos. Como puede verse en el cuadro 5 el porcentaje varía de 23.3 % (7 preguntas sobre un total de 30) en Lenguaje a 12 % (4 preguntas de 34) en Matemática.

La discriminación de una pregunta es muy alta si su valor está entre 0.4 y 1; y es alta si el índice está entre 0.30 y 0.40 (ver metodología). Como puede apreciarse en el cuadro 5, en estas dos categorías caen la mayoría de las preguntas. El porcentaje de preguntas que caen en el rango de Alta y Muy Alta varía de 84.4% en Lenguaje a 94.2 % para Matemática.

Como se observa en el cuadro 5, los índices de No Respuesta son adecuados para el caso de Lenguaje. En Matemática sin embargo se ha encontrado 11 preguntas eliminables pues sus porcentajes de no-respuesta son superiores al 30 %, hay también 4 preguntas en el rango de tolerable (entre el 20 y 30 % de no –respuesta).

En el siguiente cuadro se identifican cuáles han sido las preguntas, para cada prueba , que han sido observadas por las razones arriba mencionadas. En el Anexo 1, se detallan los problemas presentados para las preguntas observadas, así como las decisiones tomadas en cada caso.

Cuadro 6

Identificación de las preguntas que presentaron problemas

Problema:	Lenguaje	Matemática
Validez:		
-Muy buena	Todas	Todas
Dificultad:		
-Preg. Muy Fácil	1,2,5,9,15	16
-Preg. Muy Difícil	8,18	24,26,27
Discriminación:		
-Muy Baja	8	
No respuesta		
-Eliminable		25,26,27,28,29,30,31,32,33,34,35
Distractores Competitivos	4,6,14,23,25	1,2,6,14,22,23,31
Mal especificación	7,8,18,19	24,26,27
Conclusión:		
-Preg. Observadas	4,6,7,8,9,15,18,19	18,24,26,27
-Preg. Eliminadas	19	18

* Ver Anexo 1

2. EVALUACIÓN PSICOMÉTRICA DE LAS PRUEBAS DE LAS PRUEBAS CRECER 96

2.1 El Marco de análisis de las pruebas

Es importante aclarar los criterios y conceptos que conforman el marco de análisis para evaluar psicométricamente las pruebas. Esta sección también incluye los alcances y usos que pudieran hacerse de los puntajes de las pruebas así como las limitaciones implicadas por ella.

Las subsecciones siguientes se refieren a: el propósito de las pruebas, criterios para la evaluación de las pruebas, el diseño muestral, la aplicación de las pruebas y otros enfoques alternativos o modelos de análisis. Al final se resume los aspectos considerados en el análisis de las pruebas.

Propósito de las pruebas

Una de las primeras preguntas que debe resolverse en el análisis de las pruebas es acerca de lo que se espera del conjunto de las pruebas. Es decir, qué se pretende hacer con los resultados, qué tipo de conclusiones podremos generar y también qué tipo de conclusiones no será posible realizar. Esta discusión es base para la validez de las pruebas.

El propósito de las pruebas está ligado al grado de especificidad de la medición del instrumento y también al tipo de enfoque con que se interpretarán los resultados.

Las pruebas CRECER 96 son pruebas basadas en normas. Las pruebas de normas se oponen a la de criterios los cuales requieren de una completa descripción de las conductas medidas. En los tests basado en Normas hay más flexibilidad o generalidad en definir las conductas a medir. Esta diferencia puede ilustrarse con el siguiente ejemplo.

Consideremos el currículo escolar de cuarto grado de primaria en Matemática. Distingamos dos niveles de análisis: i) las áreas y ii) las capacidades. En el primer nivel encontramos áreas como “Números naturales”, “Conjuntos”. Con estas definiciones se observa que el nivel de áreas es más general que el nivel de capacidades. Una prueba, por lo tanto, puede desarrollarse con el objetivo de medir aspectos solo a nivel de áreas o alternativamente de medir solo a nivel de capacidades.

En una prueba de Normas incorporaremos preguntas representativas del primer nivel (i) de las áreas, pues solo se desea es tener representatividad general del currículo. Un aspecto adicional a considerar es si es que se miden todas las competencias (el universo) o solo una selección de ellas (una muestra). Pero este aspecto no se discutirá en este reporte.

En las pruebas basadas en Normas principalmente se está interesado en los contrastes relativos entre los examinados. Es por ello que las pruebas basadas en Normas sirven para estimar las “distancias” relativas entre los estudiantes según diferentes criterios de desagregación. Las Pruebas CRECER 96, al seguir los criterios de Normas, están diseñadas de tal manera que constituyen un instrumento eficaz que nos permite estimar los contrastes o las diferencias de interés entre sub-poblaciones de examinados. Se trata además de las primeras pruebas nacionales.

Definida la orientación en el enfoque de Normas se inició la tarea de construir las pruebas. La construcción de pruebas es un proceso que incluye el planeamiento de la prueba, la selección de áreas a incluirse 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 en base a una muestra que servirá para la versión final de la prueba.

Criterios para la evaluación de las pruebas: Validez de las pruebas

El análisis de las pruebas CRECER 1996 se basa en criterios presentados en la literatura psicométrica tanto bajo el enfoque clásico (Lord & Novick, 1974) como el moderno (por ejemplo Moss, 1992, AERA, APA, NCME, 1999). Uno de los nuevos enfoques es el concerniente al significado ampliado que tiene el concepto de validez y el término asociado de "constructo".

Validez de la prueba
En el enfoque clásico validez de una prueba es la medida en que la prueba mide el constructo que pretende medir. El término "constructo" se refiere a las características que no pueden ser medidas directamente sino que pueden ser inferidas desde 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 (AERA; APA, NCME, 1999). Estas interpretaciones se refieren a los constructos o los conceptos que las pruebas se proponen medir, como por ejemplo, rendimiento en matemáticas. 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 en su totalidad proveen de información que son relevantes 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. Evidencia complementaria de validez puede también derivarse del diseño muestral y de la administración de las pruebas.

(i) Juicio de Expertos
La opinión de los expertos tiene por finalidad analizar la correspondencia entre el contenido de las pruebas y los constructos que las pruebas intentan de medir. El juicio de expertos se basa en el análisis curricular y las tablas de especificaciones que generaron los especialistas responsables de las pruebas (ver especificaciones en el anexo 1). Estas especificaciones fueron sometidas a juicio de expertos y en lo que participaron los especialistas del Ministerio de Educación 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 pruebde medir únicamente un constructo (unicidad de la prueba medible).

Para establecer si el conjunto de preguntas dentro de una prueba mide una sola cosa, es decir para evaluar la unidimensionalidad, se usó el modelo de Análisis de Correspondencias (ver Nishisato 1994, ver también análisis de homogeneidad, HOMALS, Visauta 1998). Este análisis nos indica el grado de homogeneidad en 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 nos mide el grado en que una prueba es consistente en 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 idénticos o similares. En la práctica la consistencia se determina de formas alternativas, una de las cuales se basa en la consistencia interna de la prueba, es decir, por ejemplo, cuan consistente mide la mitad de una prueba respecto a su otra mitad. Este criterio de consistencia interna de la prueba es calculado por el coeficiente "alfa" de Cronbach

(iv) Criterios basados en índices psicométricos de las preguntas
Algunos criterios usados para las pruebas se basaron 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 -pruebas, 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 que tienen cada pregunta componente de la prueba. El nivel de no respuesta es el promedio de no respuesta de las preguntas que componen la prueba.

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

Típicamente los puntajes son las sumas aditivas de las repuestas 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 de 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, siendo la más popular la escala porcentual.

Las escala que se ha considerado en las pruebas CRECER 96 es la escala porcentual, que corresponde al porcentajes de acierto de la prueba. Esta escala va de 0 a 100. Con esta escala se consigue uniformizar la presentación de los resultados independientemente del número de preguntas de las pruebas.

Se han señalado (de la Orden Hoz et al.1998) limitaciones en el uso de escalas porcentuales para la presentación de los resultados, entre las principales están: 1) no existe *a priori* ningún valor que pueda considerarse como rendimiento insatisfactorio, 2) falta de indicación de qué es lo que saben o lo que ignoran los alumnos, 3) no tiene

en cuenta la dificultad de las preguntas, 4) no puede referirse de ninguna manera a los contenidos, 5) no indica la importancia de las preguntas no contestadas correctamente, ni cuántos son los sujetos que no las han contestado, 6) no permite hacer comparaciones entre pruebas distintas. Por ejemplo sería erróneo interpretar que el resultado de Matemática en el 3er grado de secundaria, 51 % medio de aciertos, es inferior a los resultados en el 5to grado de primaria, 53 % medio de aciertos.

Aunque las escalas no porcentuales tienen la ventaja de superar las limitaciones expuestas, sin embargo la ventaja principal de tener una escala 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.

Por lo expuesto, las críticas mencionadas son superables. Así en los puntos 1), 2) y 4), las Pruebas CRECER 96 siguen el enfoque de Normas y por lo tanto buscan estimar las comparaciones entre grupos antes de determinar los niveles mismos de rendimiento. Por otro lado una de las escalas propuestas en la literatura es la escala de Rasch (Hambleton, R. K., Swaminathan, H. y Rogers, H. J. , 1991) que sí toma en cuenta los puntos 3), 4) y 5).

La escala de Rasch 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 (Muñiz, 1990). El cumplimiento de estos principios son satisfactorios en las pruebas CRECER 96.

Respecto a lo primero, los alumnos que sabían las respuestas tuvieron más oportunidad de responder correctamente las preguntas, por el contrario los que no sabían, tuvieron menos oportunidades. Igualmente los alumnos resolvieron la prueba de manera independiente.

Con respecto a lo segundo, hemos verificado que las pruebas evalúan el rendimiento del alumno de manera unidimensional y por tanto una sola habilidad es suficiente para explicar la ejecución de la prueba. Adicionalmente, por construcción, las preguntas miden una y una sola variable (ver los índices de correlación pregunta-prueba). También sostenemos que la respuesta a una pregunta no se afecta por las respuestas a otras preguntas de la prueba, incluso en el caso de las pruebas de Lenguaje que corresponden a un mismo estímulo, sea este un texto o una imagen.

Finalmente, con respecto a lo tercero, consideramos que los tiempos asignados a las pruebas fueron suficientes para su ejecución por los alumnos, una evidencia de esto está en las tasas de no respuesta bajas encontradas en las pruebas. Sólo en los casos de las pruebas de Matemática en el cuarto y quinto grados de secundaria la no-respuesta es más alta sin ser esto significativo (ver los indicadores de la sección 3).

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 bajo el modelo de Rasch. Sin embargo es importante anotar que en las pruebas CRECER 96, sólo hemos empleado la escala como una transformación no-lineal de la escala porcentual. No hemos 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 logits del modelo bajo ponderaciones.

Tiene media 300 y varianza 50 y la correlación entre esta escala transformada y la escala porcentual está por encima de .98 en todas las pruebas.

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 obtenerla con las distribuciones de las pruebas CRECER 98.

Sin embargo, en términos estadísticos, esta exigencia no es necesaria. Las características de los puntajes de las pruebas (el número de preguntas es 29 y 34) determina que cualquiera de las escalas presente sólo 29 o 34 valores diferentes, pues estas son transformaciones biyectivas de ellos. Además el tamaño efectivo de las muestras (aproximadamente 45 771) determina que las escalas presenten un número grande de "empates"(valores repetidos). De esta manera, no necesariamente la escala porcentual 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 en 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 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) así como los resultados por los estratos de interés (gestión, región, departamento, etc.). La presentación de los resultados por estratos pueden 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) arriba expuesta. Se debe considerar además los factores de ponderación, que toman en cuenta qué porcentaje tiene el estrato elegido en la población y en la muestra.

Diseño Muestral.-

Un requerimiento importante para el efectivo uso de las pruebas es obtener muestras y tamaños de muestras que sean representativos y apropiados. Las pruebas de 1996 se aplicaron a una muestra representativa de centros educativos polidocentes urbanos y rurales en el ámbito nacional, la misma que contempló, en el nivel de educación primaria, a 45 771 estudiantes del cuarto grado de primaria.

Estratos considerados

El enfoque de normas seguido sugirió la selección de un conjunto de criterios que sean 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, solo se refieren a la zona urbana y rural y a al gestión pública y privada dentro de cada subpoblación.

Selección de la muestra
La muestra se seleccionó dentro de un sistema de Muestreo Bietápico, por Conglomerados y estratificado. Cada Departamento fue tomado como una subpoblación independiente. Con las escuelas como unidades de primera etapa y los estudiantes como unidades de segunda etapa. Es decir, dentro de cada departamento se formó conglomerados al interior de los cuales se seleccionó una muestra probabilística de centros educativos proporcional al tamaño dentro de cada estrato.

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. La muestra efectiva difiere de la muestra planificada debido a las dificultades de aplicación en algunas sedes. Estas dificultades se aunaron después a la decisión de limitar la muestra a los evaluados en el último día de aplicación. Por lo tanto es necesario calcular los pesos o ponderaciones de los estudiantes para garantizar la representatividad total de la muestra y restituir la proporcionalidad. El marco muestral corresponde a 1993.

Aplicación de la pruebas.-

La versión final de las pruebas se aplicó en noviembre de 1996. La aplicación fue supervisada por especialistas y coordinadores en cada uno de los centros educativos seleccionados. Ningún profesor del centro educativo 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 aplicadas al final de un proceso metodológico que se inició en 1995 a través de la aplicación piloto de formas de prueba para cada una de las pruebas nacionales.

Otros enfoques Alternativos
Las propiedades psicométricas mencionadas están relacionadas al 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 Items o TRI. El modelo de Rasch mencionado arriba es un caso especial de la familia de los TRI.

La teoría Clásica de los Test
La teoría Clasica 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 la teoría clásica de los test en las pruebas utilizadas ha sido causa de críticas pues en esta estrategia, una variable es inseparable del instrumento utilizado para medirla y ello constituye una seria limitación, pues inevitablemente se acabaría definiendo operativamente 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 basándose en la noción de pruebas paralelas se definen las propiedades mencionadas como son: confiabilidad, validez y discriminación.

Las propiedades de las preguntas que se consideran en las pruebas son las definidas bajo el modelo clásico: Validez, Dificultad, Índice de Discriminación e Índice de No Respuesta. Se podrá encontrar más detalle en el Glosario de Términos.

TRI vs. Teoría Clásica.
Otras opciones metodológicas respecto al análisis de las pruebas podrían ser seguidas con el uso de la teoría de las respuestas a items o TRI. La ventaja de considerar otros enfoques es la oportunidad de estimar mediciones psicológicas adicionales que no pueden ser proporcionados 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 solo enfoques que nos dan información adicional, si es que la metodología empleada y los requisitos adicionales se cumplen. Por ello el carácter de estos modelos TRI es complementario a los de la teoría clásica.

Siguiendo una tendencia reciente en reportes especializados sobre evaluaciones de sistema educativos nacionales e internacionales: NAEP (National Assessment Educational Progress), IAEP (International Assessment Educational Progress), TIMMS (Third International Mathematics an Science Study), LLECE (Laboratorio Latinoamericano de Evaluación de la calidad de la educación), las pruebas Nacionales CRECER 96 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. El uso de este modelo en las pruebas de 1996 corresponden a una decisión posterior de reanálisis de las pruebas.

Programas computacionales
Respecto al software computacional, en el análisis de las preguntas de las pruebas CRECER 98 se usaron tanto el ITEMAN (modelo clásico) como el RASCAL con un parámetro (modelo de Rasch). (Para información véase http://www.assess.com/softmenu.html). Cada equipo de especialistas revisó las preguntas para chequear las propiedades generadas por las corridas.

Con el ITEMAN se obtuvo la información sobre nivel de discriminación, dificultad de items, e información sobre distractores. Con el RASCAL se evaluó la posibilidad de extender el análisis para la consideración de estimar las habilidades de los examinados. Este análisis está supeditado al chequeo del supuesto de unidimensionalidad que se desprende del análisis de las pruebas. Sin embargo el RASCAL fue usado solo para la parte de la transformación de las variables, que se ha explicado anteriormente.

Usos de la prueba: Elección de la escala
Dado que el objetivo de las pruebas es el de comparar los resultados entre grupos relevantes y no de medir el nivel mismo del rendimiento de los examinados, cualquier transformación sobre los puntajes obtenidos es apropiada para los fines de comparación.

En esta nueva escala lo que se busca es determinar qué grupo de alumnos han salido mejor que otros. Los resultados de la escala servirán para presentar los reportes globales (o nacional) así como los resultados por los estratos de interés (gestión, región, departamento, etc.). La presentación de los resultados por estratos pueden incluir, los reportes de los promedios (previa ponderación por no proporcionalidad en el tamaño de los de los estratos en la población), desviaciones estándares, 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 puntaje total.

2.2 Constructos y contenidos de las pruebas

A fines de Noviembre de 1996 se realizó la **Aplicación Nacional CRECER 96**, a los estudiantes de cuarto grado de primaria. Esta aplicación de pruebas fue complementada con encuestas a Padres de Familia o Tutores, y Directores de Centros Educativos donde estudiaban los estudiantes.

La Aplicación Nacional incluyó 1 pruebas de selección múltiple de Matemática, una prueba de selección múltiple de Lenguaje con una subprueba de respuesta abierta o de producción de textos, 1 encuesta a directores, 1 encuesta de padres o apoderados, 1 encuesta de profesor.

Contenido de las pruebas

En 1996 los estudiantes del cuarto grado de primaria estaban desarrollando sus aprendizajes con la nueva estructura curricular organizada en líneas de acción educativa, en este sentido, las pruebas aplicadas fueron en Lenguaje y Matemática. La prueba de Lenguaje contenía 30 preguntas y tuvo una duración 45 minutos. La prueba de Matemática contenía 35[1] preguntas y tuvo una duración de 60 minutos

Comunicación Integral.- las áreas que se incluyen son la comprensión de lectura, comprensión de lenguaje oral, nociones gramaticales y vocabulario
Lógico Matemática.- revisa conjuntos, fracciones, geometría, números decimales y naturales, y sistema internacional de unidades y sistema monetario

2.3 Resultados

Con respecto a los indicadores considerados los siguientes fueron los resultados cuantitativos más saltantes en las pruebas CRECER (ver tablas 6 y 8).

Confiabilidad (consistencia interna)
A nivel de pruebas los resultados se presentan en base al indicador de consistencia interna (alfa de Cronbach).

La prueba de Lenguaje presenta una confiabilidad de 0.83 y la de Matemática 0.89.

Unidimensionalidad

Todas las pruebas arrojaron un grado de unidimensionalidad (93 % y 91 % de la variancia explicada de la 1ra dimensión para Matemática y Lenguaje respectivamente.

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

[1] La prueba aplicada tenía 35 preguntas, pero la pregunta 18 no se consideró en los análisis definitivos por problemas en la impresión que dificultan la interpretación de las alternativas de respuesta

truncado, es decir, se limita a un rango que va de 0 a aproximadamente 34 (que es el máximo puntaje que se puede obtener). Considerando que en cada prueba hay 45,771 alumnos y muchos empates la distribución de esta frecuencia es limitada al rango mencionado. Un punto importante de anotar es que en la práctica muchas de las comparaciones se hacen a nivel más agregado, como por ejemplo, la escuela que toma los promedios en las aulas. En este caso hay dos efectos favorables hacia la normalidad que son una disminución de casos (aproximadamente hay 1421 valores correspondientes a las escuelas) y una suavización de la distribución de casos (los promedios se distribuyen más continuamente) mejorando la densidad en frecuencia del intervalo de la escala.

La consecuencia de esto es que el uso de instrumentos paramétricos para el uso de las pruebas debe tomarse con cuidado. Es recomendable usar métodos complementarios basados en el análisis de datos categóricos o cualitativos donde la asociación y correlación entre variables es estimada a base de supuestos más flexibles. En el caso del análisis paramétrico 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).

Otros análisis basados en las preguntas

Los siguientes resultados a nivel de prueba se basan en los resultados del análisis de las preguntas (ver documento técnico sobre Análisis psicométrico de las preguntas de las pruebas CRECER). Es decir los resultados de dificultad media, discriminación media, correlación pregunta -prueba media e índice de respuestas medios, son los promedios de los indicadores respectivos de las preguntas que componen cada prueba.

-Correlación pregunta prueba
Los coeficientes de Validez fueron 41.45 % y 45.61 % para Lenguaje y Matemática respectivamente. El rango está sobre el nivel del 20% de validez, que califica una buena validez.

-Dificultad
El rango de dificultad medio para Lenguaje fue de 50.34 % y para Matemática de 45.96%.

-Discriminación
La prueba que más discriminó fue la prueba de Matemática con un coeficiente de discriminación medio de 53.55%, mientras que la prueba que menos discriminó fue la prueba de Lenguaje.

-No-Respuesta
El rango de no-respuesta se encuentra entre 2.4% para la prueba de Lenguaje y 20.74% para la prueba de Matemáticas. La prueba de Matemática de 4to grado, arroja tasas de no-respuesta

Cuadro 7
Criterios de validez cuantitativa de las preguntas de las Pruebas CRECER 96

ESTADÍSTICAS FINALES	COMUNICA. INTEGRAL	LÓGICO MATEMÁTICA
CONFIABILIDAD		

Alfa de Cronbach	0.83	0.89
Índices Psicométricos		
Índ. Dificultad Medio	50.34%	45.96%
Índ. Discriminación Medio	44.81%	53.55%
Índ. Validez Medio	41.45%	45.61%
Índ. No-Respuesta Medio	2.41%	20.74%
Total de Preguntas	29	34
Total de Evaluados	45771	45771

Cuadro 8
Criterios de Validez basado en la Unidimensionalidad de las Pruebas CRECER 96

PRUEBAS	%Varianza explicada (1ra dimensión)	Ratio de 1ra sobre 2da dimensión	Coef. Confiabilidad Theta
Lógico Matemática	92.92	3.62	0.89
Comunicación Integral	90.81	3.14	0.85

3. CONCLUSIONES

- En general, el análisis de las preguntas basados en la metodología presentada arrojó resultados moderados tanto desde la perspectiva de los indicadores cualitativos como de los cuantitativos.
- En Matemática deben excluirse una pregunta. En la prueba de Lenguaje debe excluirse una pregunta. Los motivos de exclusión de las preguntas mencionadas son expuestas en el Anexo.
- Respecto a la validez de las pruebas, y dado que el objetivo trazado es el de 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 cuantitatvos en el análisis de las preguntas) sugieren que las pruebas finales presentan propiedades psicométricas óptimas para su uso y 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 Rasch, teniendo en cuenta las ventajas y desventajas de cada cual expresadas en este documento. Cuando se quieran estimar totales u otros estadísticos agregados, el sistema de ponderaciones debe ser usado para recuperar la proporcionalidad del universo.
- La estimación de las varianzas y el error estándar para el cálculo de los promedios y otros estadísticos, deben ser tomados en consideración para cualquier inferencia paramétrica. Debido al diseño de muestreo usado, las estimaciones del error estándar pueden ser usadas usando las fórmulas correspondientes del muestreo estratificado o alternativamente pueden ser estimadas por otros métodos como “jack nife” o “bootstrap”. Es decir, el uso de las fórmulas de las varianzas del muestreo simple aleatorio no son recomendables.
- Las pruebas son unidimensionales y presentan alta confiabilidad. Los índices de validez, basados en las preguntas son aceptables, lo que nos hace concluir que a nivel global de prueba, éstas poseen buenas características desde el punto de vista psicométrico.

- Las características adicionales presentadas nos reflejan un conjunto de pruebas con buena discriminación, están 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 matemáticas han resultado poco significativas.

Sugerencias

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

- Los cambios sustantivos de las versiones originalmente planteadas y piloteadas tienen efectos importantes en los valores de las propiedades de las preguntas. En este sentido es necesario que cuando se haga una modificación post-piloto, las versiones corregidas deben estar igualmente sujetas a pruebas de campo para su validación final.
- La misma recomendación se sugiere para otros cambios “menores” en la estructura de la pregunta (enunciado y distractores). Estas modificaciones no garantizan la idoneidad de la pregunta en una versión definitiva. Se trata de una pregunta prácticamente nueva
- En el trabajo de las pruebas piloto los equipos de revisión de las preguntas observadas deben considerar tanto los indicadores psicométricos como los propios criterios pedagógicos para decidir la conveniencia de la eliminación de una pregunta. En ese sentido el equipo debe ser integral y estar conformado por el especialista de psicometría, el constructor de la prueba y especialistas en el área evaluada.
- Debe procurarse una correspondencia entre el enfoque de evaluación (normas y criterios) y la metodología de construcción de pruebas (Teoría clásica, Teoría de Respuesta al ítem) de manera que esta última proporcione las condiciones para el uso final de la información
- Debe procurase la supervisión contínua de los equipos responsables de las pruebas para garantizar el seguimiento similar (uniforme) y correcto de la metodología de construcción de pruebas de manera que no se observe diferencias en los procedimientos seguidos.
- Debe procurarse tanto los análisis desde una dimensión cuantitativa como cualitativa de manera que las decisiones a tomar durante el seguimiento de la metodología de construcción de pruebas sean las más adecuadas.
- Debe implementarse 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.

Referencias

AERA, APA, NCME (1999). Standars 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.

De la Orden Hoz, Arturo, Bisquerra, R., Gaviria, J., Gil, G., Jornet, J., López, F. Sánchez, J., Sánchez, M. Siierra, J. Tourón, F. (1996). Los resultados escolares. Diagnóstico del Sistema Educativo. 1997. Madrid: Ministerio de Educación y Cultura, Instituto Nacional de Calidad y Evaluación.

Hambleton, R. K., Swaminathan, H. y Rogers, H. J. (1991*). Fundamental of item response theory*. Beverly Hills, CA: Sage

Lazarte, A (1995). Análisis de Preguntas. Separata del curso PSB234. PUCP. Facultad de Psicología 3p.

Moss P.A. (1992). 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, Fall 1992, (62), 3, pp: 229-258.

Muñiz, J. (1990) . Teoría de Respuesta a los Ítems. Un nuevo enfoque en la evolución psicológica y educativa. Madrid: Ediciones Pirámide, S.A.

Muñiz Fernández J. (1993). Teoría clásica de los tests.

Nishisato S. (1994). Dual Scaling. Toronto: University of Toronto Press.

Lord and Novick (1974). Statisitical Theories of Mental Test Scores. New York: Adisson-Wesley

Visauta V.B. (1996). Análisis estadístico con SPSS para Windows. Volumen II, Estadística multivariante. Madrid: McGraw Hill.

ANEXO

Análisis de las preguntas observadas

En Lenguaje:
A continuación se muestran los resultados de las preguntas 4,6,7,8,9,15,18,19, que permitieron la decisión de no ser tomadas en cuenta..

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
4	0-4	.37	.41	.36	A	.35	.36	.28	-.07	
					B	.21	.31	.11	-.20	
					C	.37	.18	.60	.36	*
					D	.06	.14	.01	-.22	
					Other	.01	.00	.00	-.07	

Dimensión cuantitativa: No tiene problemas en ningún indicador
Dimensión cualitativa: La pregunta 4 presenta dos respuestas competitivas con la respuesta correcta. La respuesta C es más abstracta e implica las respuestas A y B. De esta manera se trata de una pregunta mal especificada. Son respuestas valederas sin embargo de acuerdo al especialista la alternativa C expresa el fin último o justificación de propósito final en relación a la organización de los vecinos.
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
6	0-6	.34	.52	.45	A	.28	.37	.14	-.20	
					B	.34	.43	.19	-.21	
					C	.02	.04	.00	-.13	
					D	.34	.13	.65	.45	*
					Other	.02	.00	.00	-.05	

Dimensión cuantitativa: No tiene problemas en ningún indicador
Dimensión cualitativa: La pregunta 6 presenta dos respuestas competitivas con la respuesta correcta. La respuesta D es más abstracta e implica las respuestas A y B. De esta manera se trata de una pregunta mal especificada. Son respuestas valederas sin embargo de acuerdo al especialista la alternativa D expresa engloba ambas respuestas
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
7	0-7	.36	.37	.33	A	.54	.59	.40	-.16	
					B	.36	.20	.57	.33	*
					C	.04	.11	.01	-.19	
					D	.05	.08	.02	-.14	
					Other	.01	.00	.00	-.08	

Dimensión cuantitativa: No tiene problemas en ningún indicador
Dimensión cualitativa: La pregunta 7 está mal especificada. Hay dos alternativas que acaparan las respuestas, y no se tiene buenos distractores. Hay ambiguedad en la

identificación de la lectura: poema o cuento. Sin embargo expresa cómo los niños no tienen noción sobre el tipo de texto que se les presenta.
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
8	0-8	.18	.19	.22	A	.16	.26	.04	-.24	
					B	.64	.55	.62	.04	
					C	.18	.14	.33	.22	*
					D	.02	.03	.01	-.09	
					Other	.00	.00	.00	-.08	

Dimensión cuantitativa: La pregunta es muy difícil, el índice de discriminación es muy bajo .
Dimensión cualitativa: La pregunta 8 está mal especificada. Hay tres alternativas que acaparan las respuestas. Los niños expresan mayoritariamente que habla un mago (alternativa A) sin embargo esta respuesta no es correcta. Probablemente esto refleja dificultad para abstraerse del relato fantástico quien habla. A juicio del especialista esta es una habilidad que se desea medir.
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
9	0-9	.82	.39	.41	A	.09	.21	.02	-.26	
					B	.03	.05	.01	-.12	
					C	.82	.58	.97	.41	*
					D	.06	.14	.01	-.24	
					Other	.01	.00	.00	-.09	

Dimensión cuantitativa: La pregunta es muy fácil.
Dimensión cualitativa: La pregunta 9 está correctamente especificada y los distractores son muy pobres de allí que la respuesta es obvia.
Sin embargo el ítem discrimina adecuadamente entre el grupo que puntúa alto y bajo y resulta válido con el puntaje total.
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
15	0-15	.80	.37	.39	A	.08	.15	.02	-.19	
					B	.06	.14	.01	-.22	
					C	.04	.09	.01	-.17	
					D	.80	.59	.96	.39	*
					Other	.01	.00	.00	-.12	

Dimensión cuantitativa: La pregunta es muy fácil.
Dimensión cualitativa: La pregunta 15 está correctamente especificada y los distractores son muy pobres de allí que la respuesta es obvia.
Sin embargo el ítem discrimina adecuadamente entre el grupo que puntúa alto y bajo y

resulta válido con el puntaje total.
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
18	0-18	.24	.42	.42	A	.48	.48	.39	-.09	
					B	.22	.30	.08	-.21	
					C	.06	.09	.02	-.12	
					D	.24	.09	.51	.42	*
					Other	.02	.00	.00	-.14	

Dimensión cuantitativa: La pregunta es muy difícil.
Dimensión cualitativa: La pregunta 18 está mal especificada pues presenta un distractor que tiene mayor porcentaje de respuesta y otro que compite con la respuesta correcta. Sin embargo el ítem discrimina adecuadamente entre el grupo que puntúa alto y bajo y resulta válido con el puntaje total. Los alumnos no reconocen el significado de la palabra ansiedad en relación al sentimiento expresado en el texto.
Decisión: **conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
19	0-19	.32	.23	.21	A	.10	.12	.08	-.06	
					B	.17	.25	.08	-.18	
					C	.39	.36	.39	.02	
					D	.32	.22	.45	.21	*
					Other	.02	.00	.00	-.14	

Dimensión cuantitativa: La pregunta discrimina moderadamente y la correlación pregunta prueba es apenas conveniente.
Dimensión cualitativa: La pregunta 19 está mal especificada pues presenta un distractor que tiene mayor porcentaje de respuesta y otro que compite con la respuesta correcta. El ítem no discrimina adecuadamente entre el grupo que puntúa alto y bajo y resulta medianamente válido con el puntaje total. La respuesta c puede considerarse igualmente válida por lo que se tiene dos respuestas correctas. De esta manera no resulta apropiado para el puntaje total.
Decisión: **No conservar la pregunta para el puntaje total de la Prueba de Lenguaje**

En Matemática:
A continuación se muestran los resultados de las preguntas 24,26,27,31 que permitieron la decisión de no ser tomadas en cuenta..

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
24	0-24	.18	.27	.32	A	.19	.12	.26	.14	
					B	.18	.07	.34	.32	*
					C	.30	.21	.31	.07	
					D	.08	.09	.04	-.07	
					Other	.26	.00	.00	-.43	

Dimensión cuantitativa: La pregunta es muy difícil y discrimina moderadamente. La tasa de no-respuesta es alta.
Dimensión cualitativa: La pregunta 24 sugiere una mala especificación pues presenta un

distractor que tiene mayor porcentaje de respuesta y otro que compite con la respuesta correcta. Sin embargo la respuesta correcta está bien definida. Los resultados denotan que los alumnos no saben resolver situaciones problemáticas donde hay que aplicar operaciones combinadas de multiplicación y división. A juicio del especialista este es el tipo de cosas que se desea evaluar.

Decisión: **Conservar la pregunta para el puntaje total de la Prueba de Matemática**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
26	0-26	.17	.30	.36	A	.06	.07	.04	-.04	
					B	.17	.05	.35	.36	*
					C	.09	.08	.08	-.00	
					D	.33	.20	.40	.15	
					Other	.35	.00	.00	-.41	

Dimensión cuantitativa: La pregunta es muy difícil. La tasa de no-respuesta es alta.

Dimensión cualitativa: La pregunta 26 sugiere una mala especificación pues presenta un distractor que tiene mayor porcentaje de respuesta. Sin embargo la respuesta correcta está bien definida. Los resultados denotan que los alumnos no saben resolver situaciones problemáticas donde hay que aplicar operaciones combinadas de multiplicación y división. A juicio del especialista este es el tipo de cosas que se desea evaluar. Por otro lado hay que anotar que la tasas de no –respuesta puede verse influida por la posición de la pregunta.

Decisión: **Conservar la pregunta para el puntaje total de la Prueba de Matemática**

No.	-Item	Correct	Index	Biser.	Alt.	Total	Low	High	Biser.	Key
27	0-27	.12	.22	.31	A	.12	.04	.26	.31	*
					B	.09	.07	.11	.06	
					C	.22	.15	.27	.10	
					D	.17	.11	.21	.11	
					Other	.39	.00	.00	-.41	

Dimensión cuantitativa: La pregunta es muy difícil. El índice de discriminación es moderado y la tasa de no-respuesta es alta.

Dimensión cualitativa: La pregunta 27 sugiere una mala especificación pues presenta dos distractores que tienen mayor porcentaje de respuesta. Sin embargo la respuesta correcta está bien definida. Los resultados denotan que los alumnos o no entendieron la situación problemática presentada o no manejan correctamente los algoritmos de la división y multiplicación de números decimales. A juicio del especialista este es el tipo de cosas que se desea evaluar. Por otro lado hay que anotar que la alta tasa de no – respuesta puede verse influida por la posición de la pregunta.

Decisión: **Conservar la pregunta para el puntaje total de la Prueba de Matemática**

Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	High	Point Biser.	Key
31	0-31	.28	.37	.35	A	.10	.07	.13	.07	
					B	.20	.13	.24	.10	
					C	.10	.10	.08	-.04	
					D	.28	.11	.48	.35	*
					Other	.32	.00	.00	-.44	

Dimensión cuantitativa: No tiene problemas en ningún indicador.
Dimensión cualitativa: La pregunta 31 presenta un distractor competitivo con la respuesta. Sin embargo la respuesta correcta está bien definida. Los resultados denotan que los alumnos no tienen clara la noción de paralelismo y de proporcionalidad. A juicio del especialista este es el tipo de cosas que se desea evaluar. Por otro lado hay que anotar que la alta tasa de no –respuesta puede verse influida por la posición de la pregunta, esto es más claro cuando se observa que los alumnos que alcanzaron a "mirar" la pregunta la contestaron correctamente.
Decisión: **Conservar la pregunta para el puntaje total de la Prueba de Matemática**

Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
18	0-18	.27	.40	.38	A	.12	.12	.09	-.04	
					B	.17	.14	.15	.00	
					C	.27	.09	.50	.38	*
					D	.19	.15	.21	.05	
					Other	.26	.00	.00	-.41	

Dimensión cuantitativa: La tasa de no-respuesta es alta.
Dimensión cualitativa: La pregunta 18 no presenta distractores competitivos con la respuesta.
Sin embargo la versión impresa de la pregunta que se muestra más adelante no corresponde a la versión piloteada ni deseada. EN la versión impresa el símbolo "-" aparece en vez del símbolo "=". De esta manera, por este error involuntario las respuestas a estas preguntas no permiten medir adecuadamente lo deseado. Por lo que no resulta conveniente conservar esta pregunta.
Decisión: **No conservar la pregunta para el puntaje total de la Prueba de Matemática**

La pregunta 18 fue presentada como sigue

18) **¿Cuántas** de las siguientes expresiones **son falsas?**

I) $0{,}75 - \frac{75}{100}$

II) $0{,}30 - \frac{3}{10}$

III) $0{,}08 - \frac{8}{10}$

IV) $0{,}01 - \frac{1}{10}$

RESPUESTA :

A) 0
B) 1
C) 2
D) 3