Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

"Desarrollo y Pilotaje de un Examen de Español para la Educación Primaria en Baja California"

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS EDUCATIVAS

Presenta

Luis Ángel Contreras Niño

Ensenada B. C. Julio, 2000.

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Maestría en Ciencias Educativas

"Desarrollo y Pilotaje de un Examen de Español para la Educación Primaria en Baja California"

T E S I S

que para obtener el grado de

MAESTRO EN CIENCIAS EDUCATIVAS

Presenta

Luis Ángel Contreras Niño

APROBADO POR:

M. Ed. Eduardo Backhoff Escudero

Director de Tesis

Dr. Guillermo Solano Flores
Sinodal

M.Ed. Norma Larrazolo Reyna

Sinodal

M.C. Virginia Velasco Ariza
Sinodal

Ensenada B.C. Julio, 2000.

Agradecimientos

Con profundo amor, a Estrella,
mi compañera de toda la vida
y a mis hijos Ángel, Luis y Sofía,
por su apoyo incondicional y paciencia.

Con admiración, reconocimiento y respeto,
a Eduardo Backhoff Escudero, por su
multifacética participación en mi formación,
como director de Tesis, maestro y amigo.

Con profundo agradecimiento a mis maestros
y compañeros de la maestría, por compartir conmigo
sus conocimientos y experiencias durante el viaje que emprendimos
juntos para aproximarnos al complejo y fascinante fenómeno educativo

Agradecimientos

Al Lic. Felipe Martínez Rizo, autor de la idea del estudio y gestor permanente de su desarrollo.

*A los miembros del **Comité de tesis**, por su profesionalismo y por sus valiosos comentarios y sugerencias durante el desarrollo del trabajo de tesis.*

M.Ed. Eduardo Backhoff Escudero, director
Dr. Guillermo Solano Flores
M.Ed. Norma Larrazolo Reyna
M.C. Virginia Velasco Ariza

y a la Dra. Araceli Ruiz Primo, por sus valiosos comentarios al documento, realizados como un gesto de amistad

*A los miembros del **Comité Coordinador** del examen, quienes dieron orientación y sentido a las acciones del proyecto, desarrollaron instrumentos y participaron en la capacitación de los grupos de trabajo.*

M.Ed. Eduardo Backhoff escudero
M.Ed. Norma Larrazolo Reyna y
Lic. Guadalupe Tinajero Villavicencio

*A los miembros del **Comité Aplicador** del examen, quienes realizaron la aplicación de los modelos de examen bajo condiciones estandarizadas y apoyaron la captura de datos derivados de las hojas de respuesta:*

Lic. Guadalupe Tinajero Villavicencio
Psic. Estrella Lucía Roldán Berny
Ing. Ernesto Arroyo Acosta
Srita. Elalt Aguirre Lugo

*A los miembros del **Comité Diseñador** del examen, quienes efectuaron el análisis curricular del área español, el diseño de especificaciones de ítems y el análisis de la congruencia ítems-especificaciones.*

Especialmente, a la Profesora María de Lourdes Bayardo Morales y al Maestro Domingo Mendoza Villanueva, quienes fueron los pilares indiscutibles del complejo trabajo que realizó el Comité

A las Profesoras Rosa Eaton Guerrero, Sara Hernández Soto, Dora Gallardo y Lourdes Eda, por su valioso apoyo en el análisis curricular del área de español.

A la Mtra. María Elena González Robles, por el valioso apoyo que dio para elaborar los ítems y aplicar el examen en Ensenada

*A los miembros del **Comité Elaborador** de ítems, quienes desarrollaron los reactivos de los cuatro modelos de examen:*

Andrea Guadalupe Silva Ovando
Emma Sepúlveda Rivera
Beatriz Elena Tirado Montero
Julia Gloria Godoy Aldrete
Blanca Alicia Valenzuela Coronado
Irma González Ramírez
Beatriz Tinoco Mangas
María Margarita Banda Hernández
Martha Rosa Montaño Barrios
Rosa Eaton Guerrero
Yolanda Aburto Márquez
Eva Adriana Ramírez Jiménez
Consuelo Sánchez Osuna
José Manuel Alba López
Baldomero Acevedo Aguilar

A los 253 niños de las 12 escuelas primarias de Ensenada, quienes aportaron con sus respuestas la información objeto del presente estudio evaluativo

CONTENIDO

Índice de tablas

Índice de figuras

Relación de anexos

Anexo Nº 1. Manual para elaboradores de ítems

Anexo Nº 2. Criterios de calidad para un test de gran escala

Anexo Nº 3. Modelo para desarrollar exámenes nacionales de referencia normativa y criterial, orientados por el curriculum, de Anthony Nitko

Anexo Nº 4. Modelo para diseñar y pilotear una prueba de español para la educación primaria

Anexo Nº 5. Documentos seleccionados del Manual para el Comité Diseñador del examen de español para la educación primaria

Anexo Nº 6. Especificaciones de ítems para la prueba de español

Anexo Nº 7. Manual para el Comité Aplicador del examen

Anexo Nº 8. Modelo de examen Nº 4

Anexo Nº 9. Cuestionario de opinión sobre la realización de actividades de aprendizaje previstas en el eje de Lengua Hablada

Resumen

El presente estudio evaluativo tuvo como propósitos generales construir y pilotear un examen de español destinado a monitorear la calidad de los aprendizajes que logran los egresados de las escuelas primarias en Baja California. En esencia se trata de un examen criterial de gran escala, alineado con el curriculum del área de español de la educación primaria. Así, el curriculum fue la base sobre la que se construyó el examen y el criterio que determinó las decisiones respecto a qué evaluar y cómo hacerlo. El modelo psicométrico que se empleó se basa en el que propuso Anthony Nitko (1994), y establece un proceso para desarrollar exámenes que consiste en seis etapas:

1. Definir el dominio de resultados que pretende el curriculum
2. Analizar el curriculum para detectar el contenido importante a evaluar
3. Desarrollar un plan de evaluación
4. Producir y validar ítems
5. Efectuar un análisis primario de los resultados
6. Efectuar un análisis secundario de los resultados para elaborar reportes sobre la ejecución de los examinados

Cabe señalar que el desarrollo y pilotaje de la prueba, a los que se refiere este trabajo, corresponden solo a las primeras cuatro etapas del proceso. Así, en esas etapas se realizaron los siguientes procedimientos:

- Selección y capacitación de cuatro comités de trabajo: el coordinador del examen, el diseñador del instrumento, el elaborador de los ítems y el aplicador de los modelos de examen
- Análisis del contenido curricular del área de español
- Análisis complementario del contenido
- Elaboración de una retícula o modelo gráfico del contenido a evaluar
- Muestreo de resultados de aprendizaje a evaluar
- Diseño de especificaciones de ítems
- Elaboración de ítems según las especificaciones

- Revisión de la congruencia ítem-especificación
- Ensayo empírico y revisión de la dificultad y la discriminación de los ítems, así como de la confiabilidad de los modelos
- Revisión de la prueba y estructuración de una muestra de ítems representativa del dominio curricular

Un asunto de gran relevancia, fue la valoración y aseguramiento de la calidad de los productos obtenidos, de la operación de los procedimientos y de otros resultados de la aplicación del modelo. Para ello, se adoptaron los estándares de calidad comúnmente empleados para la construcción y evaluación de las pruebas de gran escala, mismos que se agruparon en tres áreas de control: calidad del contenido de los ítems del test, calidad técnica de cada ítem y calidad de las calificaciones de la prueba.

Además, en el trabajo se presentan los resultados obtenidos tras el diseño y pilotaje del instrumento, entre los cuales destacan los siguientes:

- La retícula del área de español, que contiene el dominio curricular completo que pudo ser identificado y estructurado y sobre el que se desarrolló el examen
- Un documento con 30 especificaciones para producir los ítems de la prueba
- 180 ítems, depurados inicialmente mediante: **a.** operaciones de jueceo que los declararon congruentes con las especificaciones que los produjeron, y **b.** un ensayo empírico y la consiguiente revisión de la dificultad y la discriminación de los ítems, así como de la confiabilidad de los modelos
- Cuatro modelos de examen, estructurados con una muestra de ítems representativa del dominio curricular

Para concluir cabalmente el proyecto, aún faltan por realizarse tres acciones: efectuar un ensayo empírico de gran escala y proceder a revisar los ítems ya con suficientes datos empíricos; especificar los estándares de calidad definitivos para el contenido del ítem, para su calidad técnica y para la calidad integral de los modelos del examen; y estandarizar procedimientos para aplicar y calificar el examen y analizar los resultados, en su fase de monitoreo permanente.

Capítulo 1. Introducción

1.1 Justificación del estudio

Existen al menos cuatro razones que justifican la realización del estudio evaluativo de gran escala para monitorear la calidad del aprendizaje en el área de español de la educación primaria nacional, que se describe en este trabajo: **a.** En México se carece de este tipo de instrumentos; **b.** El proceso de descentralización de la educación básica requiere de dichos instrumentos para apoyar la diversificación educativa; **c.** Poco se sabe acerca de la operación del nuevo curriculum que adoptó la SEP en 1992 para la educación primaria; y **d.** Los conocimientos y habilidades relativos a la lengua nacional son fundamentales para el desarrollo del educando y para la planeación educativa en general.

Enseguida se comentan con mayor detalle los elementos de justificación mencionados.

a. El estudio se ubica en una perspectiva nacional caracterizada por la urgente necesidad de transformación productiva, respecto a la cual la educación juega sin duda un papel preponderante. En efecto, en el momento presente se coincide en que la educación debe tener prioridad entre las estrategias nacionales promotoras del desarrollo. Al respecto, la reorientación de las políticas nacionales relativas a la educación, debe estar cimentada en un diagnóstico preciso de lo que sucede actualmente en las escuelas.

En los últimos 20 años la historia de la educación en México ha estado caracterizada por una expansión de la matrícula en todos los niveles, en una atmósfera financiera reducida y casi solo pública. Tal atmósfera está conformando una conciencia sobre la necesidad de que las instituciones

educativas rindan cuentas de su gestión mediante criterios de evaluación rigurosos, concebidos como mecanismos para mejorar y no como medios de control financiero o político. La reducción del gasto público en educación conlleva también la necesidad de hacer más eficiente al sistema educativo por la vía de mejorar el desempeño académico y administrativo de los planteles, lo cual requiere de criterios y procedimientos de evaluación más elaborados y una mayor capacidad de supervisión; es decir, mediante un monitoreo permanente de la calidad del servicio que ofrecen.

En este contexto, cabe señalar que, después de un largo periodo de desinterés generalizado por los asuntos de corte evaluativo en el ámbito educativo, durante la última década se ha observado un creciente interés por la evaluación educativa y en particular por la evaluación de la calidad de la educación (Martínez, et al, 1995). Lo anterior, es resultado de un cambio del énfasis ocurrido en las décadas de los setenta y ochenta relativo a la expansión en la matrícula, que privilegió los aspectos cuantitativos de la educación, a uno que destaca la importancia de políticas explícitamente concebidas para dar prioridad a los objetivos de calidad y equidad (OCDE, 1997).

No obstante lo anterior, en el caso de la educación básica de nuestro país son aún escasos, coyunturales y de alcance limitado los estudios evaluativos orientados al diagnóstico de la calidad del servicio educativo que ofrecen los planteles. En diversos documentos que forman parte de la literatura reciente relativa a la investigación educativa nacional, se comenta la falta de una cultura de la evaluación generalizada y una escasa investigación sobre la evaluación del aprendizaje (Martínez, 1993; Jackson-Maldonado, 1993). Lo anterior resulta extraño si se considera la importancia que representa el enorme subsistema de educación primaria como insumo para el resto del sistema educativo nacional y, en general, para el desarrollo del país. En efecto, la educación primaria atiende a 15 millones de niños entre 6 y 14 años de edad. Entre los principales retos que enfrenta este ciclo educativo están el reducir aún más los índices de

deserción (el cual pasó del 6% en 1984 a 3.4% en 1994) y de reprobación (que de 9% se redujo a 7.5% en el mismo período), así como elevar la eficiencia terminal, la cual en 1984 era del 50% y que en 1994 se incrementó al 60%. Estos indicadores revisten una importancia fundamental, porque afectan a todo el sistema educativo. Respecto a esta situación, una limitante importante radica en que el conocimiento del desempeño de los alumnos depende aún exclusivamente de las calificaciones que otorgan los profesores. Si bien se han reforzado ciertos aspectos de la evaluación con el sistema de la Carrera Magisterial, todavía se carece de suficientes instrumentos para evaluar sistemáticamente el trabajo escolar (SEP, 1996).

b. Aunado a lo anterior, la necesidad de evaluar el aprendizaje en la educación básica, se pone de manifiesto también ante la perspectiva del reciente proceso de descentralización de sus servicios. Tal política educativa requiere urgentemente de procedimientos e instrumentos, válidos y confiables, que apoyen un principio de diversificación educativa el cual debe garantizar la unidad del subsistema dentro de su deseable diversidad.

Por ello, resulta necesario que las instituciones educativas perfilen con mayor nitidez su vocación regional y las ventajas comparativas en cuanto a la operación de sus programas educativos, mediante factores como el uso de mecanismos de evaluación efectivos que subrayen la necesidad de la autoevaluación del desempeño institucional, junto con la evaluación externa de profesionales independientes o de una institución gubernamental supervisora, así como la comunicación a la sociedad de los resultados y de la calidad de los procesos educativos involucrados.

c. En México, la educación primaria se ha caracterizado por tener un curriculum único. En 1992, la Secretaria de Educación Pública (SEP) realizó un profundo cambio curricular en este ciclo educativo, que tuvo como propósitos principales fortalecer la lectura, la escritura y el manejo de las matemáticas, como medios

para elevar la calidad de la educación que reciben los niños. En el caso del área de español, dicho cambio resultó muy significativo, pues se le otorga la prioridad más alta dentro del plan de estudios. En los dos primeros años se le asigna el 45% del tiempo escolar y en los demás grados el español representa, al menos, 30% de la carga horaria total. Otra modificación importante fue la eliminación del enfoque formalista que enfatizaba el estudio de aspectos lingüísticos y los principios de la gramática estructural, en favor del desarrollo de la capacidad comunicativa mediante el ejercicio sistemático de las habilidades de expresión y comprensión, tanto oral como escrita, sin olvidar el conocimiento y uso de reglas y normas gramaticales que permiten la reflexión acerca del lenguaje (SEP, 1993). No obstante lo anterior, poco se sabe acerca de la operación de tal curriculum.

De hecho, las modificaciones curriculares efectuadas en la educación primaria forman parte de un contexto más amplio. El movimiento mundial de "regreso a lo básico", el cambio de enfoque en cuanto a la enseñanza del lenguaje, de considerarlo como materia de estudio, hacia concebirlo más como medio de comunicación y otros fenómenos relacionados, ocurridos durante la década pasada y principios de la actual, son el antecedente directo del reciente interés por desarrollar las habilidades comunicativas a partir de la educación elemental. Dicho entusiasmo ha traído aparejado un interés generalizado por la evaluación de dichas habilidades. Por otra parte, procesos socioculturales tales como el avance científico y tecnológico y la globalización socioeconómica, así como el desarrollo actual y previsible de los medios de información, producen un creciente impacto en el aspecto lexicográfico de los lenguajes y en el interés por desarrollar las habilidades y actitudes para la comunicación, particularmente las relativas al análisis simbólico (Makey, 1992).

El hecho de contar en México con un curriculum nacional en el área de español de la educación primaria, presenta paradojas históricas interesantes y tiene implicaciones importantes para la evaluación del aprendizaje de gran escala.

Así, mientras que otros países desarrollados como Estados Unidos, que carecen de un curriculum nacional en la educación básica, se han dado a la enorme tarea de generar una reforma educativa basada en estándares de ejecución, a fin de alinear con ellos el diseño de materiales, la formación docente y la evaluación estandarizada, entre otros procesos de planeación y evaluación educativos, en nuestro país donde se posee un curriculum único, no existen por ejemplo pruebas estandarizadas de gran escala. ¿Qué hace diferente a México de otros países en cuanto a contenidos curriculares o en materia de medición del aprendizaje? En el primer caso, parece que nuestro país ha tenido la ventaja de contar siempre con una política curricular coherente, por lo menos en lo relativo a la planeación de los contenidos de la educación; tal es la situación actual del área de español que se orienta al desarrollo de habilidades básicas para el aprendizaje y la comunicación. En el segundo caso, México ha estado prácticamente ausente de la historia de la psicometría moderna, misma que se dio durante casi todo el siglo veinte. Tal ausencia ha limitado la capacidad de la SEP para planear congruentemente el desarrollo de este ciclo educativo y de la sociedad en general en cuanto al conocimiento de lo que sucede en las escuelas. En este sentido, el presente trabajo constituye un esfuerzo para reducir en alguna medida dicha brecha.

En cuanto a las implicaciones que tiene, para la evaluación del aprendizaje a gran escala, el hecho de tener en nuestro país un curriculum nacional, cabe señalar que si el curriculum es lo que fue diseñado para hacer racional el proceso educativo, resulta claro que también debe ser la base racional sobre la que se sustente la evaluación educativa (Nitko, 1994). Esta consideración también afecta al tipo de metodología que se puede emplear para diseñar la evaluación, pues si la evaluación está alineada con el curriculum, este último se convierte en el criterio fundamental contra el cual se contrastarán los resultados obtenidos mediante aquella; en tal caso, parece que lo más razonable es diseñar un test criterial que este alineado con el curriculum formal, pues este ya existe.

d. En un sentido mucho más amplio, el diseño de una prueba para evaluar el aprendizaje logrado en el área de español de la educación básica reviste singular importancia, debido a que los conocimientos y habilidades relativos a la lengua nacional, son fundamentales para el desarrollo intelectual y social del educando, para facilitar su acceso a la cultura y por su valor propedéutico para la educación posterior. De esta manera, el contar con un diagnóstico preciso y sistemático de la calidad del aprendizaje que logran en esta área los alumnos que egresan de la educación primaria, tendrá un valor estratégico no sólo para este ciclo educativo, sino para la planeación educativa en general.

1.2 Adscripción teórico-metódica

El estudio incluyó diseñar y pilotear una prueba de referencia normativa y referencia criterial de lenguaje para la educación primaria en México, de gran escala, fundamentalmente con base en el modelo para desarrollar exámenes nacionales orientados por el curriculum, elaborado por Anthony Nitko (1994). Sin embargo, dicho modelo fue complementado por la metodología para la construcción de tests, tanto criteriales como normativos, propuesta por James Popham (1995); por el esquema metodológico aportado por Ronald Berk y colaboradores para la construcción de tests criteriales (1984); así como en los planteamientos psicométricos propuestos por Kellaghan y Greaney para el caso de países en desarrollo (1992; 1995). Una vez diseñada y estructurada en cuatro modelos, la prueba fue aplicada a una muestra de alumnos egresados de escuelas primarias en el estado de Baja California, y se efectuaron análisis de dificultad y discriminación de los ítems, y de confiabilidad de los modelos, con objeto de calibrarlos.

En particular el modelo de Nitko, que se describe en el Capítulo 3 del presente trabajo, establece un proceso para desarrollar exámenes que consiste en nueve etapas principales: Definición del dominio de resultados de logro; análisis

curricular; desarrollo del plan de evaluación; desarrollo de la especificación de ítems; producción y validación de ítems; ensamblaje de exámenes; establecimiento de estándares de ejecución; análisis primario de resultados; y análisis secundario de resultados.

Se eligió dicho modelo en virtud de que presenta características y ventajas que lo hacen apropiado para el contexto educativo de la escuela primaria nacional. Entre ellas, podemos destacar que:

- Mediante procedimientos especiales, crea un alineamiento entre la evaluación, el curriculum y la instrucción. Para lograrlo, se requiere que el curriculum sea el centro del desarrollo del examen y que las decisiones respecto a qué evaluar y cómo hacerlo estén muy influidas por los resultados de aprendizaje establecidos por dicho curriculum.

- Se propone obtener en un solo examen los esquemas referenciales normativo y criterial, sin menoscabo de la validez. Según Nitko, los esquemas referidos a un criterio y a una norma no son mutuamente excluyentes, sino más bien complementarios; y considera que es posible obtener ambos tipos de referencia de un solo test, habida cuenta de que se sigan procedimientos especiales cuando se diseñe y se produzca la prueba.

- Cuando se alinean los exámenes, el curriculum y la instrucción pueden esperarse beneficios educativos importantes, como los que se mencionan a continuación:

    - **Mejoras en la Operación Curricular**. Retroalimentación a escuelas y profesores, enfocada en la manera en que se comportaron los estudiantes en áreas específicas del curriculum, lo cual refuerza que los profesores enseñen el curriculum.

- **Justicia a los Estudiantes**. Si el curriculum se define con claridad, si el plan del examen se conoce y entiende por parte de los profesores, si los profesores enseñan hacia dicho plan de evaluación y si el examen refleja el plan y el curriculum, entonces el examen se vuelve justo para los estudiantes, puesto que se les habrá enseñado lo que se espera de ellos en el examen.

- **Podrá ser Evaluado el Progreso Educativo Nacional.** Es posible analizar los resultados del examen, para describir lo que los estudiantes del país son capaces de realizar. Puesto que las especificaciones del examen (el plan de evaluación) permanecen constantes a lo largo de los años, se puede monitorear el progreso logrado en el aprendizaje de metas curriculares específicas, mediante la comparación, con los años, del porcentaje de niños que lograron cada meta curricular, así como mediante la comparación de la ejecución de los estudiantes ante cúmulos de objetivos curriculares.

- **Mejoras en la Evaluación Curricular**. Los datos que derivan de los exámenes referidos a un criterio y orientados por el curriculum, pueden emplearse para identificar aquellas metas curriculares que han sido aprendidas mejor que otras.

- **Apoyo a la Orientación Vocacional de Individuos**. Una parte de la orientación consiste en identificar las fortalezas y debilidades de la persona. La interpretación referida a un criterio contribuye a dicho propósito, pues describe el grado en que se dominó cada parte del curriculum. Ello proporciona un perfil más bien específico de los conocimientos y habilidades de un estudiante, mismo que puede utilizarse con propósitos de guía.

- **Mejor Diagnóstico de las Deficiencias del Estudiante**. Los profesores pueden recibir información sobre el grado en que cada estudiante aprendió las metas de aprendizaje, de tal manera que pueda proporcionar instrucción remedial cuando se requiera.

- **Actualización de Profesores mejor Enfocada**. Al proporcionar información respecto a la ejecución de los estudiantes en áreas específicas del curriculum, por escuela, estado o en el país, es posible identificar determinados patrones que indiquen áreas específicas del curriculum que requieren ser enseñadas mejor. Si con el tiempo surgen tales patrones, es posible establecer programas de formación docente que estén enfocados efectivamente.

- **Posibilidad de una Evaluación Continua**. Las evaluaciones referidas a un criterio que están alineadas con el curriculum, permiten monitorear el progreso de los estudiantes respecto a metas de aprendizaje importantes, de tal manera que los profesores pueden determinar el estado del aprendizaje de sus alumnos que está basado en las metas de aprendizaje que realmente se enseñaron a los estudiantes.

Cabe señalar que el examen fue concebido como un instrumento de bajo impacto. Ello significa que se considera que su aplicación no compromete las expectativas de los estudiantes en cuanto a la certificación de su aprendizaje o al ingreso a la educación secundaria, por lo demás garantizado por las leyes General de Educación (Poder Legislativo Federal, 1994) y Estatal de Educación (Poder Legislativo Estatal, 1995). Asimismo, se considera que los resultados que deriven de su aplicación solo afectarán a las escuelas y a los docentes que laboran en ellas retroalimentando su trabajo académico, por lo cual su

responsabilidad por los resultados observados a partir de su aplicación, no implica consecuencias para ellos. Es decir, el uso ideal previsto para el instrumento lo ubica, junto con otros instrumentos, como parte de una estrategia de investigación descriptiva que apoye el monitoreo permanente del aprendizaje en la educación básica de nuestro país y que contribuya al mejoramiento de su calidad.

1.3 Objetivos

Considerando los elementos aportados en los puntos anteriores, los objetivos del estudio evaluativo a los que se refiere el presente trabajo fueron:

- Diseñar un examen de referencia criterial y referencia normativa, de gran escala, para evaluar el aprendizaje que logran en el área de español los alumnos que egresan de la educación primaria en Baja California.

- Iniciar el proceso de validación del instrumento, empleando para ello una muestra de escuelas primarias del estado de Baja California.

- Ensayar las condiciones necesarias para, eventualmente, establecer la aplicación del instrumento a gran escala, como un mecanismo permanente para monitorear la calidad del aprendizaje en la educación primaria que se ofrece en Baja California.

1.4 Limitaciones

Cabe señalar que tales objetivos no se lograron de manera cabal. Al respecto, pueden señalarse dos limitaciones principales:

Como se verá más adelante, las condiciones en que se realizó el estudio evaluativo impusieron restricciones asociadas principalmente a los recursos

humanos involucradas en los comités técnicos que elaboraron el examen. En parte, tales limitaciones obedecieron al hecho de que la adaptación del modelo psicométrico empleado restringió el número de comités a un mínimo indispensable; en parte, se debió a la dificultad para conseguir especialistas en el contenido con disponibilidad de tiempo necesaria para involucrarse en las complejas tareas del análisis curricular y la elaboración de especificaciones de ítems. No obstante, la calidad técnica del instrumento se mantuvo en general dentro de los límites deseables, a la luz de los principales indicadores psicométricos empleados.

La segunda limitación deriva del hecho de que el modelo adoptado no se aplicó en su totalidad. Si bien se diseño el instrumento, se efectuó una calibración de los ítems que lo forman, y se realizó un primer ensayo de las condiciones necesarias para que las autoridades educativas estatales puedan monitorear de manera permanente la calidad del aprendizaje del área de español, faltó aplicar el instrumento a una muestra de gran escala a fin de contar con suficientes datos empíricos para observar tanto la calidad técnica de los ítems, y la de los modelos de examen, como para efectuar el análisis secundario de los datos que permita elaborar informes de resultados a los usuarios de la información, mismos que en su estado actual solo tienen un carácter ilustrativo. Estas y otras limitaciones, que se comentan posteriormente, plantean la necesidad de continuar el estudio evaluativo hasta alcanzar los estándares de calidad mínimos que se requieren para proceder a la aplicación generalizada del instrumento en las escuelas.

Capítulo 2. Evaluación del aprendizaje a gran escala y experiencias sobre la evaluación del lenguaje en México y en el mundo

Con el objeto de establecer una base conceptual mínima para el proyecto, en este apartado se presentan los aspectos psicométricos más relevantes que están involucrados en la evaluación del aprendizaje a gran escala. Así, se caracterizan brevemente los principales modelos evaluativos del aprendizaje que existen en la actualidad, así como los indicadores psicométricos y sociales que definen la calidad de un test. Además, se presentan algunos elementos conceptuales y experiencias relativos a la evaluación del aprendizaje del lenguaje, tanto en México como en el contexto internacional.

2.1. Aspectos Psicométricos Relativos a la Evaluación del Aprendizaje

Hasta el siglo pasado la construcción de pruebas fue básicamente una tarea intuitiva y subjetiva. Después de la prolongada era intuitiva, la elaboración de pruebas entró en una etapa científica durante la cual han ocurrido numerosos cambios tecnológicos que han impactado enormemente la vida escolar (Madsen, 1983). Los avances logrados durante los últimos 25 años en lo relativo a modelos psicométricos y en los métodos y procedimientos evaluativos han sido sustanciales y el interés por mediciones justas y precisas a dejado de ser una preocupación exclusiva de los sistemas educativos, para trasladarse a los gobiernos, los industriales, los políticos y a otros sectores sociales (Jones y Hambleton, 1992).

- **Clasificación de las pruebas para evaluar el aprendizaje**

Existen varios criterios para clasificar las pruebas que se emplean para evaluar el aprendizaje. Por su importancia en el contexto de la evaluación a gran escala es posible identificar, al menos, tres sistemas de clasificación de los tests: El primero de ellos

distingue a las pruebas por la *referencia* que tiene la ejecución ante ellas; así tenemos a los referidos a una norma y los referidos a un criterio. El segundo, las distingue por el *tipo de respuesta* que demandan del examinado, por lo cual pueden ser de respuesta construida o de respuesta seleccionada. Finalmente, tenemos a los que se distinguen por su *escala de aplicación*; a saber, los de pequeña escala y los de gran escala. Enseguida se comentan brevemente los tres tipos de clasificación:

Tests de Referencia Normativa y Referencia Criterial. La principal distinción entre ambos tipos de prueba depende de la manera en que interpretamos la ejecución de un examinado ante el test (Popham, 1995). En la tabla que se presenta a continuación se contrastan ambos tipos de examen con base en sus características generales y su uso básico:

Tabla Nº 1. Características generales de los tests normativos y criteriales

Tipo de test	Interpretación de la ejecución del examinado	Juicio típico que resume la interpretación	Criterio de validez de la interpretación	Característica distintiva	Propósitos educativos típicos que cumple
Normativo	En relación con la ejecución de otros que también respondieron el examen	La ejecución del examinado se encuentra arriba, abajo o dentro de la norma o promedio	La medida debe ser tomada de toda la población de estudiantes o de una muestra representativa de ellos	Proporciona una medida de la habilidad relativa del estudiante	- Selección de estudiantes - Asignación de recursos a gran escala
Criterial	En relación con el estatus del individuo respecto a un criterio o dominio evaluativo bien definido.	Comunica qué puede hacer o no el examinado en un campo del conocimiento (conocimientos y habilidades)	La medida debe ser tomada de todo el dominio de contenido o de una muestra representativa de tareas sacadas de ese dominio	Claridad con que describe lo que mide	- Evaluación de programas - Certificación de competencias académicas - Diagnóstico y diseño instruccional

Cabe señalar que, hasta hace relativamente poco tiempo, la mayor parte de los exámenes de gran escala que se elaboraban eran de tipo normativo. Por ello, presentan un nivel de desarrollo psicométrico superior al que tienen los criteriales, mismos que empezaron a desarrollarse en los años setenta.

Como ya se indicó en el capítulo 1, la referencia a un criterio y la referencia a una norma,

no son esquemas que se excluyen necesariamente, sino que pueden ser complementarios. Ambos pueden ser empleados dentro de un mismo test para lograr un conocimiento más cabal de un examinado, siempre que se sigan procedimientos especiales para su elaboración, particularmente la satisfacción del criterio de validez de la interpretación de los dos esquemas; es decir, por la vía del muestreo representativo tanto de alumnos, como de contenidos (Nitko, 1984; 1994).

Tests de respuesta construida y de respuesta seleccionada. Este criterio de clasificación se refiere al tipo de ítems que forman la prueba. Aunque en un examen pueden ser incluidos diferentes tipos de preguntas, en el contexto de la evaluación a gran escala se presentan limitaciones en este aspecto debidas principalmente al proceso de estandarización que conllevan. A continuación se presenta una tabla que describe los principales tipos de preguntas, características, ventajas, desventajas y contexto de uso.

Tabla Nº 2. Tests de respuesta construida y de respuesta seleccionada

Tipo de test	Tipo de ítem	Características	Principales Ventajas	Principales Desventajas	Contexto de uso
Respuesta construida	Ensayo	Pide al estudiante integrar libremente y por escrito, lo que sabe acerca de un tema	Permite medir aprendizajes complejos y la expresión personal	Permite un muestreo de contenido muy pobre y es difícil de calificar	Salón casi siempre
	Respuesta breve	El estudiante responde con una palabra, símbolo o frase corta	Permite buen muestreo de contenido y adivinar no conduce al éxito	Solo es útil para medir aprendizajes más bien simples	Salón casi siempre
	Ejecución	Enfatiza lo que el estudiante puede hacer, no lo que sabe, en un contexto auténtico	Permite poner a prueba conocimientos y habilidades clave, en situaciones reales y complejas	Permite un muestreo de contenido muy pobre, es difícil de calificar y consume tiempo y recursos	Salón casi siempre
Respuesta seleccionada	Opción múltiple	Pide al estudiante elegir la respuesta correcta o la mejor opción entre las que se le ofrecen	Es muy flexible, permite medir muchos tipos de aprendizaje y obtener una buena muestra de contenido	Difícil redactarlo al medir aprendizajes complejos y el alumno selecciona, no produce la respuesta	Salón y gran escala
	Respuesta alterna	Pide al estudiante juzgar la verdad o falsedad de proposiciones	Permite observar el dominio en una área y buen muestreo de contenido	Usualmente solo miden aprendizajes simples y propician copiar y adivinar	Salón y gran escala
	Asociación	Pide al estudiante relacionar conceptos u ordenar fases o eventos	Permite un buen muestreo de contenido y medir la habilidad para discriminar eventos relacionados	Solo mide información factual, es difícil encontrar material homogéneo y resulta fácil copiar	Salón y gran escala

Los aspectos considerados en la tabla, y otros como la facilidad para su administración y calificación, han orientado a los elaboradores de pruebas de gran escala a emplear casi exclusivamente los reactivos de respuesta seleccionada, particularmente los de opción múltiple, a pesar de las limitaciones y críticas encontradas. Para una descripción más detallada de los tipos de ítem, véase la nota sobre la evaluación del aprendizaje que aparece en el Manual para Elaboradores de Ítems, identificado como anexo Nº 1.

Por su parte, se ha considerado que las pruebas de respuesta construida resultan más apropiadas para el contexto de la instrucción, donde se requiere rastrear errores en la comprensión, proporcionar retroalimentación significativa al aprendizaje de los alumnos y evaluar habilidades de producción como la redacción y la expresión oral. Al respecto, cabe destacar los esfuerzos recientes por diseñar pruebas de ejecución estandarizadas y de gran escala, agrupados bajo el rubro de *tests auténticos* y que lucen muy prometedores. Sin embargo, en su estado actual de desarrollo, aún están por demostrar sus bondades a la luz de los criterios psicométricos fundamentales (Hogan, 1992) y su viabilidad para utilizarlos en contextos más amplios que el salón de clases.

Tests de pequeña escala y de gran escala. Otro criterio que se utiliza para clasificar los tests se refiere a la escala en que se emplean. Así, estos pueden ser de pequeña escala o de gran escala. Esta distinción resulta relevante debido al gran impacto social que pueden tener los exámenes. A continuación se describen ambos tipos de instrumento.

Tabla Nº 3. Exámenes de pequeña escala y de gran escala

Tipo de examen	Contexto de uso	Propósitos de la evaluación	Consecuencias de su aplicación	Requisitos técnicos	Recursos para la implementación
Pequeña escala	Salón de clases	Ubicación, monitoreo del aprendizaje	Usualmente mínimas o moderadas	Mínimos, usualmente básicos	Mínimos o moderados en cuanto a personal, costos y tiempo
Gran escala	Mós de un plantel, usualmente en un estado, región o país	Seleccionar, rendir cuentas de la gestión, certificar el logro educativo	Usualmente muy poderosas, en particular para alumnos y profesores	Altos, con procedimientos relativamente sofisticados	Altos, requieren de especialistas, tiempo, dinero e información considerables

Como puede observarse, la diferencia básica entre ambos tipos de examen radica en los propósitos de la evaluación y en el contexto de su uso.

- **Estándares de calidad para el diseño, aplicación y evaluación de pruebas de gran escala**

Por su dimensión y por el poderoso impacto social que tienen sobre las vidas de alumnos, profesores, padres, directivos escolares, autoridades educativas y la sociedad en general, los tests de gran escala plantean condiciones especiales que determinan que tanto su elaboración, como su aplicación y evaluación, deban ajustarse a rigurosos estándares de calidad tales como la definición de su uso y cobertura, la exhibición de evidencias de validez y confiabilidad, el uso de procedimientos estandarizados para la administración, calificación e interpretación de resultados, entre otros (Rudner, 1993; Joint Committee on Testing Practices, 1994; Popham, 1995).

Por ejemplo, el National Center for Research on Evaluation, Standards and Student Testing (CRESST, 1994-b), en Estados Unidos, desarrolló criterios para revisar la calidad técnica de las evaluaciones que apelan a la complejidad cognitiva (pensamiento crítico, solución de problemas y razonamiento), la calidad del contenido (contenido que represente un reto y sea importante), significatividad (que las tareas evaluativas valgan la pena y que el estudiante entienda su valor), propiedad del lenguaje (que resulte claro y al nivel del estudiante), transferencia y generabilidad (que permite generalizaciones válidas respecto a la habilidad para realizar otras tareas), justicia (que no da cabida a factores irrelevantes para el aprendizaje o no pretendidos y califica con equidad), confiabilidad (se considera que las respuestas a las preguntas representan consistentemente lo que el alumno sabe) y consecuencias pretendidas (que tiene los efectos deseados).

Así, el énfasis reciente en la calidad educativa, en la noción de asumir la responsabilidad y

el incremento del impacto social de la evaluación del aprendizaje, obligan cada vez más a quienes elaboran las pruebas y, en general a los educadores, a estar interesados e informados en los elementos técnicos que definen la calidad de una prueba educativa.

En la tabla Nº 4, que se presenta en la siguiente página, se describen algunos de los criterios, especialmente importantes, que deben satisfacer los exámenes de gran escala, ya sean normativos o criteriales.

En general la descripción de los criterios señalados en la tabla, aunque esquemática y resumida, da una idea muy clara de lo que significa desarrollar un test de gran escala que tenga calidad. Palabras como seguridad, estándares, jueceo, evidencias, validez, impacto, estadístico, ofensa y otras más que aparecen en la tabla, implican procesos sociales, técnicos, políticos y académicos que casi siempre resultan complejos, tardados y costosos; mismos que a su vez requieren del concurso de especialistas en contenido, curriculum y psicometría, autoridades educativas, profesores y estudiantes, entre muchas otras personas. Así, el desarrollo de este tipo de instrumentos usualmente queda reservado a las autoridades educativas, centros de investigación educativa y organismos especializados, que tienen condiciones, interés en hacerlo o están obligados a ello.

Una descripción más detallada de los criterios que definen la calidad de un test de gran escala se presenta en el anexo Nº 2.

Tabla Nº 4. Descripción de los principales criterios que definen la calidad de un test de gran escala

Criterio de calidad	Definición resumida	Principales tipos, formas o aspectos	Observaciones
Estandarización	Consiste en emplear procedimientos uniformes para administrar y calificar el test, así como para interpretar las calificaciones de manera que resulten comparables los resultados de los diferentes examinados	- Administración (locales, resguardo, instrucciones) - Calificación (automática, corrección por adivinar, estándares y punto de corte) - Interpretación de datos	Se trata de lograr eficiencia y seguridad para recolectar datos comparables acerca del logro académico de una gran cantidad de alumnos
Validez	Grado en que un test mide lo que dice. Técnicamente, se refiere a la obtención de evidencias que soportan las inferencias basadas en los puntajes obtenidos en el test (Es el indicador más importante de la calidad de un test)	- Contenido (¿la muestra de ítems representa al universo de contenido?)	- Basada en el juicio humano - Esencial en test criterial
		- Criterio (¿los puntajes permiten inferir la ejecución en una variable criterio? (por ejemplo, el promedio de calificaciones)	- Depende del tipo de variables criterio que se emplean - Puede ser predictiva o concurrente - Esencial en test normativo
		- Constructo (¿los puntajes son una medida del atributo psicológico de interés?	En un solo estudio no es posible obtenerla, Deben acumularse evidencias
Confiabilidad	Se refiere a la consistencia o reproducibilidad de los puntajes del test. Puede ser de ocasión a ocasión, de prueba a prueba, de ítem a ítem, de juez a juez, etc.	- Indice de estabilidad (test-retest)	Observar la consistencia en el tiempo de puntajes
		- Correlación de puntajes en formas paralelas o alternas	Los modelos deben ser equivalentes (contenido, etc.)
		- Consistencia interna (homogeneidad de los ítems)	Observar si los ítems funcionan de forma similar
Descripción de la conducta medida	Formulación explícita de los conocimientos, habilidades, aptitudes y actitudes del examinado que se pretende medir con el test	- Descripción breve (objetivo conductual usualmente)	Característica en los tests normativos
		- Descripción detallada (especificaciones de ítems)	Característica en los tests criteriales
Extensión de la prueba	Especificación del número de ítems por conducta medida en el examen. Depende de la importancia de la decisión involucrada	- Tipo de estimación de la ejecución del examinado	General: pocos ítems
			Específica:: bastantes ítems
		- Tipo de impacto (grado de afectación a los sujetos)	Bajo: pocos ítems
			Alto: bastantes ítems
Alcance de la medida	Se refiere a la amplitud del atributo que mide el test.	- Restringido (más precisa la descripción y menos ítems)	Su determinación depende de una operación de juicios de expertos
		- Amplio (menos precisa la descripción y más ítems)	
Datos comparativos	Se refiere a la cantidad y calidad de los datos normativos que permiten interpretar apropiadamente la ejecución de los examinados	. Marco referencial clave de los tests normativos	Análisis estadísticos abundantes y sofisticados
		- Marco referencial básico de los tests criteriales	Análisis estadísticos básicos
Ausencia de sesgo	Se refiere a detectar y eliminar un funcionamiento diferencial de los ítems, ante grupos diferentes de examinados, que no depende del grado de conocimiento o habilidad que se mide	- Ofensa (el ítem retrata a un grupo estereotipadamente)	Su detección requiere de pruebas empíricas y de juicios de personas que representen a los grupos potencialmente afectados
		- Penalización (un grupo falla el ítem aunque posee la misma habilidad que otro)	

2.2 Evaluación del Aprendizaje en el Area de Lenguaje

- **Experiencias en Otros Países**

Independientemente de la concepción chomskiana del desarrollo genético del lenguaje, en el ámbito de la lingüística se considera que existen tres componentes del lenguaje que son universales: la fonología, la sintaxis y la semántica. En cuanto a la fonología, todos los lenguajes comparten los mismos recursos articulatorios y todos los sonidos pueden ser clasificados en 12 categorías de acuerdo con la posición de las estructuras responsables de la articulación. Respecto a la sintaxis, se comparten todas las principales categorías sintácticas tales como sujeto, predicado, objeto, atributo, adverbio y complemento y solo difieren las reglas para ensamblar las partes. En relación con la semántica, la mayoría de las características semánticas de la mayor parte de los lenguajes son similares; por ejemplo, todos los lenguajes poseen solo tres tipos de verbo: de estado, de proceso y de acción (Davies, 1990; Weiping, 1993).

Por otra parte, procesos socioculturales tales como el avance científico y tecnológico y la globalización socioeconómica, así como el desarrollo actual y previsible de los medios de información, producen un creciente impacto en el aspecto lexicográfico de los lenguajes y en el interés por desarrollar las habilidades y actitudes para la comunicación, particularmente las relativas al análisis simbólico (Makey, 1992).

El movimiento mundial de "regreso a lo básico", el cambio de enfoque en cuanto a la enseñanza del lenguaje, de considerarlo como materia de estudio, a concebirlo más como medio de comunicación y otros fenómenos relacionados, ocurridos durante la década pasada y principios de la actual, son el antecedente directo del reciente interés por desarrollar las habilidades comunicativas a partir de la educación elemental. Dicho

entusiasmo ha traído aparejado un interés generalizado por la evaluación de dichas habilidades.

En un ámbito internacional caracterizado por el cambio de una era industrial, en la cual una persona podía salir adelante mediante habilidades elementales de lectura y aritmética, a una era de la información, que requiere de habilidad para accesar, interpretar, analizar y emplear abundantes y variados tipos de información, requeridos en el mercado laboral y en la vida cotidiana, los gobiernos y las instituciones educativas están desarrollando estándares de contenido y de ejecución ambiciosos, así como estándares evaluativos congruentes con ellos, con el propósito de mejorar el aprendizaje de los estudiantes y mantener consistentes en el tiempo dichos parámetros evaluativos (Congress of the US, 1992; Linn, 1993; Bond, 1994).

En general, los estándares están centrados en el aprendiz. Constituyen una guía para preparar al estudiante a fin de que pueda cumplir con los requerimientos de lenguaje y otros que les planteará el futuro; también propician articular una visión compartida por educadores, investigadores, padres, etc., de lo que esperan que el estudiante logre y cómo puede lograrlo, así como para promover altas expectativas para todos los estudiantes y reducir las disparidades en cuanto a oportunidades educativas (CRESST, 1994; NCTE/IRA, 1996).

Existen diversas organizaciones que han establecido estándares evaluativos relativos al aprendizaje del lenguaje en la educación primaria, entre los que destacan por la dimensión y el alcance que tienen, así como por el ámbito de su influencia, los preparados por la International Reading Association (IRA) y las norteamericanas National Council of Teachers of English (NCTE), la National Assessment of Educational Progress (NAEP), el Center for Research on Evaluation, Standards and Student Testing (CRESST) y el National Assessment Governing Board (NAGB), entre otras.

Una de las áreas del conocimiento que ha recibido más atención por parte de los diseñadores de instrumentos de evaluación del aprendizaje es, sin duda, el lenguaje. Solamente la colección de cerca de 10,000 tests del *Educational Testing Service* (ETS), incluye más de 250 pruebas de lenguaje, la mayor parte de las cuales está destinada a la educación básica. Algunas pruebas como el *California Achivement Test* (CAT/5, s/f), el *Iowa Test of Basic Skills* (ITBS) y el *Comprehensive Testing Program* (CTP, 1993) del ETS, se han constituido en verdaderos parámetros evaluativos en el ámbito internacional. En la tabla siguiente se comparan estas pruebas en cuanto a las áreas de contenido que consideran y se incluye también el test de evaluación de la educación primaria elaborado por el INCE del Ministerio de Educación y Ciencia de España (INCE, 1995).

Tabla Nº 5. Comparación entre cuatro tests de lenguaje destacados internacionalmente

Áreas de contenido incluidas	Examen			
	CAT	ITBS	CTP	INCE
Análisis de la palabra	X	X		
Vocabulario	X	X	X	X
Comprensión oral	X		X	X
Comprensión escrita	X	X	X	X
Expresión oral	X			
Expresión escrita	X	X	X	X
Ortografía	X	X		X
Gramática	X	X		X
Habilidades de estudio	X			
Habilidad verbal		X	X	

Al respecto, cabe señalar que una revisión de los descriptores, términos de contenido y resúmenes de 30 tests de lenguaje para la educación básica de la colección del ETS, reveló que los aspectos que son considerados con más frecuencia para la evaluación del lenguaje en este nivel, coinciden con los que contempla el CAT (ETS, 1996).

- **Experiencias en México**

En el caso de nuestro país, hasta muy recientemente, no existían antecedentes relacionados con el establecimiento de estándares nacionales o estatales de evaluación o de sistemas de captación de información como podrían ser los exámenes nacionales

(Martínez, 1993). En parte, dicha ausencia se explica por el hecho de que tradicionalmente el curriculum de la educación primaria en México ha sido único y no se ha considerado necesario establecer estándares relativos al lenguaje, como en otros países. Sin embargo, ello no explica porqué no han surgido exámenes nacionales o estatales de lenguaje. Tal vez dicha ausencia se explica más bien por la compleja relación que la SEP mantiene con los profesores a través de su representación sindical.

Así, la historia de los trabajos sistemáticos relativos a la adquisición, evaluación y desarrollo de la lengua en las escuelas de nuestro país, es muy corta, escasa y ha estado caracterizada por problemas como la poca información publicada que, en su mayor parte, es de circulación interna y confidencial en las instituciones educativas que imparten el servicio. Por ejemplo, los trabajos pioneros de Hurtado y colaboradores en 1982 y 1984, en la Dirección General de Educación Especial de la SEP, acerca del modelo de gramática universal a partir de supuestos chomskianos y sobre estructuras sintácticas en niños entre siete y once años de edad, son de circulación interna de dicha dependencia, al igual que la Batería de Evaluación de la Lengua Española, elaborada por Bárbara Merino y adaptada a México por Donna Jackson-Maldonado (Jackson-Maldonado, 1993).

En la página siguiente, se presenta una tabla que resume las principales experiencias en materia de evaluación del aprendizaje del lenguaje que surgieron durante la década de los noventa en nuestro país.

Tabla Nº 6. Exámenes de lenguaje elaborados o aplicados en México durante la década de los noventa

Examen	Contenidos que evalúa	Observaciones	Referencias
ECOLE	Comprensión de lectura en niños de 1o. a 6o. grado	- Estudio realizado por la OEA y la Universidad de las Américas	Gómez, *et al*., 1990
Prueba de competencia para la comunicación	Interpretación de imágenes, traducción entre lenguajes, comprensión y expresión escritas, entre otras	- Forma parte de un estudio internacional coordinado por la UNESCO - No alineada con los contenidos u objetivos programáticos	Schmelkes, 1994 Ornelas, 1994
Pruebas de Español para 3º, 4º, 5º y 6º grados	Lengua escrita (comprensión y redacción, ortografía, etc.) Recreación literaria (creación de cuentos, poemas, etc.) Reflexión sobre la lengua (uso de elementos oracionales, tiempos verbales, concordancia, etc.)	- Estudio realizado por investigadores de la Universidad Autónoma de Aguascalientes y del Instituto de Educación de Aguascalientes - Pruebas de gran escala, de tipo criterial, alineadas con el curriculum	Ruiz y Martínez, 1996 Zorrilla, *et al*., 1996
EXHCOBA (escala verbal, subescalas de conocimientos básicos de la lengua española y de habilidad verbal)	**Habilidades verbales** (comprensión, uso gramatical y razonamiento de enunciados y párrafos; vocabulario; uso de referencias) y **español básico** (gramática, sintaxis, comprensión de lectura y literatura)	- Estudio realizado por investigadores de la Universidad Autónoma de Baja California - Es un examen normativo de gran escala, diseñado para el ingreso a la universidad, pero las subescalas corresponden con el contenido de la educación básica	Backhoff, *et al*., 1996
Prueba de estándares	Se desconocen	- Estudio realizado para la Secretaría de Educación Pública - Prueba de tipo criterial alineada con estándares de ejecución	Schmelkes, *et al*., 1999

En general, la información contenida en esta sección, pero particularmente la que aparece en la tabla, permiten efectuar entre otras las siguientes observaciones:

- Resulta interesante ver que, después de una nula presencia de los educadores e investigadores educativos mexicanos en el panorama de la evaluación del aprendizaje a gran escala en la educación básica, al parecer comienza a surgir un interés visible por tales asuntos a partir de esta década. Si este incipiente movimiento es producto de corrientes globales con interés en la educación comparada; si se trata de un interés legítimo, producto de necesidades sentidas, que empieza a surgir en las instituciones educativas; o si los proyectos mencionados son de naturaleza coyuntural y solo reflejan, como moda, los vigorosos movimientos internacionales que se están dando en el ámbito de la evaluación del aprendizaje a gran escala, es prematuro saberlo. Sin embargo, es

indiscutible que durante la presente década se han elaborado, aplicado y dado a conocer más pruebas, que en toda la historia de la evaluación del aprendizaje en la educación primaria mexicana.

- Con excepción de las pruebas de Aguascalientes, es significativa la falta de alineamiento de esos instrumentos con el curriculum pues, como se comentó en la introducción de este trabajo, en nuestro país siempre ha existido un currículum único para la educación primaria y casi nada se ha sabido de los resultados de su operación. En cambio, el interés es de naturaleza diversa: reflejar el dominio del conocimiento básico, reflejar el logro respecto a estándares de ejecución definidos, observar el desarrollo de habilidades de lectoescritura u observar el impacto de factores socioeconómicos sobre el logro educativo, entre otros.
- Lo que sí han tenido en común las pruebas mencionadas en la tabla, es la preocupación por conocer los resultados de aprendizaje de la expresión y la comprensión escritas.
- Cabe destacar dos puntos que aparecen en la tabla: la presencia de una investigadora, Silvia Schmelkes, en dos de los proyectos evaluativos; y el censo que se hizo de las pruebas elaboradas en Aguascalientes en las escuelas primarias de ese estado.

Capítulo 3. Modelo para evaluar el aprendizaje del área de español en la educación primaria

Como ya se indicó, para elaborar la prueba de español fue utilizado como base el modelo psicométrico propuesto por Anthony Nitko para crear exámenes alineados con el curriculum. Así, el curriculum es la base sobre la que se construye el examen y las decisiones respecto a qué evaluar y cómo hacerlo están determinadas por los resultados de aprendizaje que establece dicho curriculum (Nitko, 1994). El modelo propone un proceso de desarrollo que consiste en nueve etapas, mismas se describen brevemente a continuación:

Tabla Nº 7. Modelo de Anthony Nitko para crear exámenes nacionales alineados con el curriculum

Etapas	Procedimientos	Propósitos
1. Definir el dominio de resultados que pretende el curriculum	- Análisis del contenido curricular	- Revisar y sintetizar fuentes de la planeación curricular y guías de la operación - Detectar los resultados importantes pretendidos por el curriculum - Prefigurar las tareas de evaluación
2. Analizar el curriculum	- Elaboración de un mapa curricular del dominio de contenido a evaluar	- Estructurar los resultados de aprendizaje importantes que pretende el curriculum, para crear el sistema de evaluación
	- Establecimiento de concordancia	- Clarificar y consensar las partes del curriculum sobre las cuales los alumnos serán evaluados
3. Desarrollar un plan de evaluación	- Muestreo de resultados de aprendizaje a evaluar	- Reducir el curriculum operacional que aparecerá en el examen
	- Diseño de prototipos de especificaciones de ítems	- Ilustrar el proceso de elaboración de las especificaciones de ítems
	- Establecimiento de comités de examen	- Crear una organización con funciones de diseño, monitoreo, gestión, evaluación y control de calidad
4. Desarrollar la especificación de ítems	- Diseño de especificaciones de ítems	- Refinar las tareas prototipo para que sean válidas y aclaren a los elaboradores de ítems cuales son válidas y cuales no
5. Producir y validar ítems	- Crear ítems según las especificaciones	- Elaborar los ítems del examen - Asegurar la calidad de la relación **ítem**-representa **curriculum**
	- Revisar la congruencia ítem-especificación	
	- Ensayo empírico y revisión de ítems	
6. Ensamblar los exámenes	- Estructuración de una muestra de ítems representativa del dominio curricular	- Asegurar que los modelos de examen representen el dominio curricular
7. Establecer estándares	- Definición de estándares de ejecución, justos, mínimos y que representen ejecuciones comparables en el tiempo	- Garantizar justicia para todos - Garantizar comparabilidad de resultados con el tiempo (monitoreo de la calidad)
8. Análisis primario	- Especificación de estándares de calidad para el contenido del ítem, para la calidad técnica del ítem y para la calidad integral de los puntajes del examen - Diseño y aplicación de procedimientos para calificar el examen y analizar los resultados	- Crear y mantener un programa de control de calidad y relevancia del examen - Definir las cualidades que el examen debe exhibir antes de que se utilice oficialmente para monitorear la calidad de los aprendizajes
9. Análisis secundario	- Análisis de resultados para elaborar los reportes de resultados por escuela y municipio, por eje temático del plan de estudios y por destinatario	- Contar con reportes de resultados del examen apropiados a las necesidades de información de los destinatarios

En el proceso descrito en la tabla están implícitos dos elementos cruciales que se comentan enseguida: el asunto de la validez y la relación entre el curriculum, la instrucción y la evaluación.

- En el capítulo anterior, se dijo que la validez es el criterio más importante que define la calidad de un test. Asimismo se indicó que, para el caso de los tests criteriales, el tipo fundamental de evidencias que soportan la validez de las inferencias que hacemos acerca de la ejecución del examinado, es el que tiene que ver con el contenido. Con el propósito de atender dicha exigencia Nitko, por así decirlo, interconstruye en su modelo el proceso de validación. Al hacerlo propone, en primer lugar, alinear el examen con el curriculum; lo cual significa que el curriculum es el criterio o la base sobre la que se construye la prueba y que las decisiones respecto a qué evaluar y cómo hacerlo están determinadas por los resultados de aprendizaje que establece dicho curriculum. En segundo lugar, establece la especificación del dominio de tales resultados mediante tres cortes: el análisis curricular (que hace explícito el universo de contenido), la estructuración del dominio del contenido importante a evaluar (que define el universo de medida) y la obtención de una muestra del mismo (es decir, el examen). Finalmente, establece los procesos técnicos (especificación, producción y validación de ítems) y sociales (concordancia, jueceo, estándares) que permitirán el desarrollo del instrumento. De esta manera, el proceso de validación de la prueba recorre una trayectoria que va, desde su estado final al inicial, de: ítem probado empíricamente ← que empata con la especificación acordada que lo produjo ← misma que fue obtenida de la muestra representativa de contenido ← que a su vez es parte de la estructura de contenido que se juzgó importante ← que representa al dominio curricular a evaluar. En resumen, si podemos garantizar la relación **ítem-representa-curriculum** (o dicho con más propiedad: si el conjunto de ítems que llamamos examen, representa al universo de contenido que llamamos curriculum), entonces podemos tener seguridad de que la ejecución del examinado ante el test, es

una evidencia que nos permite inferir válidamente como sería su ejecución en el resto del dominio que llamamos curriculum. Desde luego, la clave para garantizarlo es el juicio humano, el cual está presente a lo largo del proceso.

- El alineamiento de la evaluación con el curriculum (o, en su defecto, con estándares de ejecución) tiene también una función político-educativa. Entre los especialistas de la evaluación del aprendizaje a gran escala está arraigada la idea de que es posible mejorar la educación por la sola vía del examen. Tal idea descansa en el efecto observado de la aplicación de pruebas de gran escala, especialmente las de alto impacto (*high stakes*), sobre los profesores, quienes se ven obligados a enseñar el examen y no el curriculum en respuesta a las presiones sociales que producen los resultados de su aplicación.

  La estrategia que propone Nitko en su modelo para controlar las fuerzas del alto impacto, consiste en hacer que el curriculum operacional corresponda muy cercanamente con el curriculum formal, mediante la elaboración de exámenes que estén muy alineados con el curriculum. Como resultado, se crea un alineamiento instrucción → examen → curriculum; es decir, la fuerza que motiva a los profesores a enseñar el examen se domina: enseñar el examen es lo mismo que enseñar el curriculum.

Una descripción más detallada del modelo de Nitko aparece en el anexo Nº 3.

3.1 Modelo para desarrollar la prueba de español para la educación primaria

Para diseñar, elaborar y pilotear la prueba de español para la educación primaria en el estado de Baja California, se efectuó una adaptación del modelo de Nitko misma que se ilustra en la siguiente tabla:

Tabla Nº 8. Modelo para diseñar y pilotear una prueba de español para la educación primaria

Etapas	Procedimientos
1. Definir el dominio de resultados que pretende el curriculum	- Selección y capacitación del Comité Diseñador del examen
	- Análisis del contenido curricular del área de español
	- Análisis complementario
2. Analizar el curriculum	- Elaboración de la retícula del contenido a evaluar
3. Desarrollar un plan de evaluación	- Muestreo de resultados de aprendizaje a evaluar
	- Diseño de especificaciones de ítems
	- Capacitación del Comité Elaborador de ítems
4. Producir y validar ítems	- Elaboración de ítems según las especificaciones
	- Revisión de la congruencia ítem-especificación
	- Ensayo empírico y revisión de ítems
	- Revisión de la prueba y estructuración de una muestra de ítems representativa del dominio curricular
	- Ensayo empírico de gran escala y revisión de ítems
5. Análisis primario	- Especificación de estándares de calidad para el contenido del ítem, para la calidad técnica del ítem y para la calidad integral de los puntajes del examen - Diseño y aplicación de procedimientos para calificar el examen y analizar los resultados
6. Análisis secundario	- Análisis de resultados para elaborar los reportes de resultados por escuela y municipio, por eje temático del plan de estudios y por destinatario

En la adaptación realizada se redujo el número de etapas del modelo, de las nueve originales a seis y se integraron a ellas dos procedimientos. Dichos cambios obedecieron principalmente a la necesidad de ajustar el modelo al alcance especificado para el presente estudio evaluativo, que considera el desarrollo de la prueba solo hasta su pilotaje. De esta manera, las etapas ***Desarrollo de la Especificación de Ítems*** y ***Ensamblaje de Exámenes***, fueron eliminadas y, por razones de secuencia lógica y cronológica, sus procedimientos *Diseño de Especificaciones de Ítems y Estructuración de una Muestra de Ítems Representativa del Dominio Curricular* fueron integrados a las etapas ***Desarrollar un Plan de Evaluación*** y ***Producir y Validar Ítems***, respectivamente. Además, la etapa ***Establecer Estándares*** de ejecución, se omitió por ser una actividad que es preferible realizar tras la operación regular de la prueba, como señala el propio Nitko en su modelo. Ello coincide también con la idea que se apuntó en la introducción, en el sentido de concebir al examen como un instrumento de bajo impacto y un carácter eminentemente descriptivo (la versión completa del modelo adaptado se describe en el anexo Nº 4).

Un asunto de gran relevancia, fue la valoración y aseguramiento de la calidad de los productos obtenidos, de la operación de los procedimientos y de otros resultados de la adaptación y aplicación del modelo. Para ello se adoptaron básicamente los mismos estándares que propone Nitko en su modelo, mismos que quedaron agrupados en tres áreas de control: calidad del contenido de los ítems del test, calidad técnica de cada ítem y calidad de las calificaciones de la prueba. En la tabla que se presenta enseguida se describen los estándares, las medidas asociadas a ellos y el criterio que debe satisfacerse en cada caso:

Tabla Nº 9. Áreas de control de calidad, estándares, medidas y criterios, para la prueba de español

Área de control de calidad	Estándar	Medida	Criterio que debe satisfacer
Calidad del contenido de los ítems del test	1. Corrección del contenido 2. Corrección de la respuesta correcta 3. Relevancia e importancia de la tarea a ejecutar 4. Congruencia del ítem del test con el contenido 5. Correspondencia del ítem del test con su especificación	1. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos) 2. Juicio de expertos en contenido (si-no) 3. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos) 4. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos) 5. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos)	1. Promedio de clasificación de 3.5 por ítem 2. Todos confirman que la respuesta correcta lo es, o es la mejor opción 3. Promedio de clasificación de 3.5 por ítem 4. Promedio de clasificación de 3.5 por ítem 5. Promedio de clasificación de 3.5 por ítem
Calidad técnica de cada ítem del test	1. Escritura de ítems sin defecto 2. Vocabulario apropiado 3. Dificultad apropiada 4. Discriminación apropiada 5. Evitación de estereotipos étnicos y de género (ofensa) 6. Evitación de sesgo	1. Revisión del ítem por un profesional de la escritura de ítems 2. Todas las palabras en el ítem deben estar incluidas en los materiales usados en la instrucción (juicio del Comité Diseñador: si-no) 3. Valor "p" del ítem, obtenido de la muestra de ensayo 4. Índice de discriminación 5. Juicios del comité Diseñador 6. Juicios del Comité Diseñador	1. Cada ítem debe estar exento de cualquier defecto de escritura 2. Cada ítem contiene, exclusivamente, palabras incluidas en los materiales empleados en clase 3. $.05 < p < .95$ 4. $r_{bis} > .2$ 5. Ningún ítem juzgado contiene estereotipos 6. Ningún ítem juzgado pone en ventaja a algún grupo
Calidad de las calificaciones del test	1. Alta confiabilidad 2. Alto indicador de confiabilidad para ensayos	1. Coeficiente alfa o Kuder-Richardson 20 2. Porcentaje de acuerdo	1. Coeficiente mayor o igual que .85 en cada modelo 2. Porcentaje de acuerdo de .90 ó mayor en cada modelo

Nitko propone también otros estándares que no fueron incorporados. En el área de ***Calidad de contenido de los ítems del test***, no se consideró el estándar *Correspondencia del ítem del test con la categoría de pensamiento*, pues

requería del juicio de especialistas en los constructos incluidos en el área de español, que no estaban disponibles. Otro tanto sucedió con cuatro estándares correspondientes al área de ***Calidad de las calificaciones del test***: *Distribución de índices de dificultad y de discriminación de los ítems de cada modelo, en el año, para cada sujeto*; *Alta consistencia de decisión*; y *Alta validez convergente*, mismos que requieren de estudios longitudinales que estaban más allá de las posibilidades de este estudio.

Dado que en la adaptación que se hizo del modelo, las etapas y procedimientos especificados están descritos en términos generales, se requiere de una especificación mayor en el contexto particular donde se aplique. Así, en los capítulos 4 al 7 del presente trabajo, se describen con mayor detalle las etapas, procedimientos, instrumentos, productos y otros eventos relacionados con el modelo, y también se presentan las experiencias y resultados obtenidos al aplicarlo para diseñar, elaborar y pilotear la prueba de español para la educación primaria en el estado de Baja California.

Cabe señalar que el estilo de presentación en los capítulos que siguen obedece a la intención de recoger las experiencias obtenidas durante el desarrollo del examen, de tal manera que permitan dar cuenta del proceso de construcción y, además, puedan servir como marco de referencia para otras personas que proyecten desarrollar un instrumento similar.

Capítulo 4. Definición del dominio de resultados pretendidos por el curriculum del área de español de la educación primaria

Una vez definida la metodología general para construir y pilotear la prueba, se procedió a constituir un **Comité Coordinador** del Examen, con funciones de diseño general, capacitación, piloteo de instrumentos, análisis de datos y control de calidad, así como elaboración de materiales e informes, integrado por investigadores del Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California, quienes poseían experiencia en el desarrollo de pruebas de gran escala. Dicho comité procedió a diseñar y seleccionar los procedimientos y materiales que se mencionan más adelante, así como a operar esta primera etapa, que contempla los siguientes procedimientos:

4.1 Selección y capacitación del Comité Diseñador del examen.

El primer procedimiento para construir la prueba fue seleccionar y capacitar a un **Comité Diseñador** del examen, integrado por especialistas en las áreas de lenguaje, diseño curricular y evaluación, en aspectos relativos al análisis curricular, en la elaboración de redes de contenido (retículas) y en el diseño de especificaciones de items, a fin de contar con un grupo de especialistas bien entrenado que fuera la base para contruir el instrumento.

La integración del comité fue una tarea mucho más difícil de lo que se esperaba inicialmente. Se trataba de formar un grupo de especialistas en el campo del lenguaje, particularmente en el área de español de la educación primaria, que analizara el curriculum de dicha área, detectara y estructurara el contenido importante a evaluar, elaborara las especificaciones de ítems y finalmente que juzgara la congruencia ítem - especificación. Al parecer, la diversidad y complejidad de tales tareas, junto con la limitada disponibilidad de tiempo para realizarlas por parte de quienes colaboraron en el comité, ocasionaron serios

problemas de rotación de los miembros y un retraso considerable en las actividades. Por ejemplo, el análisis del contenido curricular del área de español, procedimiento clave en esta etapa, se prolongó casi un año, y la estructuración del contenido importante a evaluar, que corresponde a la etapa siguiente, tuvo una duración similar. En total, participaron en las actividades del Comité Diseñador nueve personas: dos linguístas, un psicólogo educativo, cuatro profesores de escuela normal con especialidad en español y dos directivos escolares con experiencia vigente en la operación de los nuevos programas de estudio. Cabe señalar que solo tres de ellos participaron en los trabajos hasta concluir las actividades programadas para el comité.

Otra razón que explica las dificultades encontradas, tiene que ver con la manera en que se adaptó la metodología propuesta por Nitko para trabajar en esta y otras etapas del modelo. El autor propone constituir grupos de especialistas independientes para analizar el curriculum y estructurar el contenido importante a evaluar, para establecer concordancia, para diseñar las especificaciones de ítems y para juzgar la congruencia entre los ítems y las especificaciones que los producen. La idea es que unos comités juzguen, de manera independiente, las decisiones adoptadas por otros a fin de propiciar la objetividad, sumar consensos y mejorar con ello la validez. Sin embargo, debido a las condiciones modestas en que se realizó el estudio, en el modelo adaptado todas esas actividades fueron adscritas al Comité Diseñador.

Por las razones expuestas, la capacitación no pudo efectuarse de manera formal como estaba previsto, sino que se trabajó en sesiones individuales o en pequeños grupos, según las condiciones, lugar de residencia y disponibilidad de tiempo de los miembros del comité. Para apoyar la capacitación, fue elaborado un manual que incluyó los siguientes documentos:

- *Materiales curriculares de la educación primaria* (plan y programas de estudios, libros de texto, guías para el profesor y el alumno, etc.), necesarios para identificar el dominio

del contenido curricular sobre el cual versa el examen.

- *Documento de reticulación y modelo de retícula***,** para ilustrar a los especialistas en contenido acerca de la estructuración de conocimientos y habilidades, de tal manera que estuvieran en condiciones de estructurar una retícula del contenido importante a evaluar del área de español.

- *Prototipos de especificaciones de items,* para ilustrar a los especialistas en contenido acerca de las normas para desarrollar cada item, de tal manera que pudieran elaborar las especificaciones que se entregaron a quienes posteriormente elaboraron los reactivos.

- *Formatos para el registro de información*, necesarios para elaborar especificaciones, consignar propuestas, analizar datos, elaborar reportes y otras.

Los principales documentos que integran el manual se presentan en el anexo Nº 5, entre ellos la descripción de fases y procedimientos del modelo, el documento de criterios para la estructuración de los contenidos del área de español, la descripción de la estrategia de reticulación y el modelo de retícula.

4.2 Análisis del contenido curricular del área de español.

Como en la práctica ningún documento contiene todo lo que se debe enseñar o lo que es importante, en esta primera etapa se efectuó un análisis de contenido de diversas fuentes, tanto formales como informales, que definen el curriculum del área de español de la educación primaria, tales como el plan y los programas de estudios, materiales instruccionales y prácticas educativas de maestros experimentados, así como de aspectos particulares de las teorías cognitiva y curricular, a fin de hacer explícito el dominio de resultados de logro pretendidos por el curriculum en esta área y determinar su alcance.

Puesto que en un test criterial la calidad de los ítems es juzgada constantemente contra los resultados pretendidos por el curriculum, la validez del examen depende críticamente de que tan bien estén definidas las metas de aprendizaje del curriculum. Las principales acciones que permitieron dicha definición fueron:

- Revisar y sintetizar las fuentes de la planeación curricular y las guías de operación que elaboró la SEP para apoyar el trabajo de directivos escolares y profesores, así como otras disponibles. Aquí, el trabajo principal consistió en identificar los presupuestos, concepciones pedagógicas, intenciones educativas y estrategias, que están implícitos en la documentación generada por la instancia planeadora, para posteriormente efectuar una síntesis de dicha información.

- Detectar los resultados importantes pretendidos por el curriculum del área. En este punto no se hizo ningún esfuerzo por establecer criterios *ad oc* que definieran la importancia relativa de los contenidos analizados, puesto que la intención primaria era dejar claro lo que es importante para quienes planearon el curriculum del área.

- Simultáneamente, considerar de manera preliminar las posibles acciones de evaluación asociadas con los contenidos identificados como importantes. Para evitar una autolimitación prematura, se enfatizó el considerar las formas de evaluación más apropiadas, según la naturaleza del contenido importante, independientemente de que fueran o no factibles de realizar en el contexto de la evaluación a gran escala que se estaba desarrollando.

Cabe señalar que, para el caso específico de estas actividades, así como para las demás correspondientes a otras etapas que requirieron del juicio de los miembros del Comité Diseñador para definir sobre algún asunto, se adoptó el consenso como criterio general para acordar. En parte, este fue otro de los factores que ocasionaron retraso en los

trabajos, sobre todo al inicio de las interacciones; pero después, a medida que se fueron definiendo con claridad los liderazgos académicos, se formó un espíritu de concertación que permitió avanzar.

El producto de tales acciones fue el dominio curricular completo que pudo ser identificado y sobre el cual se desarrolló el examen. El dominio quedó registrado en una enorme tabla de doble entrada que presenta en las columnas los seis grados escolares de la educación primaria y en los renglones los ejes y sub-ejes de contenido que aparecen en los programas de estudio. La organización del contenido curricular que fue hecha explícita se muestra en la retícula del área de español que aparece en la página siguiente. También en la retícula se representa el dominio completo del contenido del área.

De esta manera, en cada celda de la retícula aparecen los contenidos que corresponden a un grado escolar y a un sub-eje, que a su vez pertenece a uno de los cuatro ejes de contenido que constituyen el área de español: *Lengua Hablada; Lengua Escrita; Recreación Literaria* y *Reflexión sobre la Lengua*. Además de los contenidos identificados, dentro de cada celda aparecen marcados con color aquellos que, en principio, se juzgaron como resultados de aprendizaje importantes que establece el curriculum, según los propios diseñadores.

4.3 Análisis complementario

Con el propósito de complementar el análisis curricular efectuado, los miembros del comité diseñador solicitaron la colaboración de profesores de primaria en servicio, de diferentes grados escolares y que tuvieran experiencia en la operación de los nuevos programas de estudio, para que identificaran los que a su juicio son los contenidos específicos más importantes del área de español, en cuanto a que promueven la adquisición y el ejercicio de conocimientos, habilidades y actitudes que todos los egresados de la educación primaria deberán ser capaces de manifestar en su educación

Insertar aquí una liga a la retícula

posterior y a lo largo de su vida. Este procedimiento operó mediante reuniones grupales con docentes en las que se pidió su opinión por grado educativo y por eje curricular.

Inicialmente se consideró solicitar la opinión de los profesores mediante un cuestionario formal con preguntas abiertas, pero la cronología de los eventos no lo permitió. Así no se cuenta con un registro que permita, por ejemplo, comparar en que medida coinciden las opiniones de los profesores con las de los especialistas del comité, un cuanto a lo que se considera importante dentro del área de español. Sin embargo, en opinión de los miembros del comité, la participación de los profesores en servicio fue de mucha utilidad pues no solo reflejó el impacto que tiene la planeación curricular en las aulas, sino que además permitió dar cuenta de cambios curriculares que ha introducido la SEP con el tiempo y que fueron considerados para el análisis.

Capítulo 5. Análisis del curriculum del área de español

El análisis curricular efectuado en la etapa anterior permitió hacer explícito el dominio de resultados de aprendizaje que establece el curriculum del área de español; es decir, el universo de contenido sobre el que versó la prueba. En esta etapa, el análisis estuvo orientado a definir el universo de medida sobre el que se construyó el instrumento, mediante la estructuración del contenido importante a evaluar. Para ello, se siguieron los procedimientos que se describen a continuación:

5.1 Elaboración de la retícula del contenido a evaluar.

Con base en el dominio definido previamente y con el apoyo de los materiales incluidos en el manual correspondiente, el Comité Diseñador efectuó un análisis curricular del mismo a fin de estructurar los resultados de aprendizaje importantes que pretende el curriculum del área. El análisis se efectuó con base en la estrategia de diseño y evaluación curriculares denominada *reticulación (*Robredo *et al*, 1983*)*, la cual permite mostrar los contenidos y las relaciones de servicio entre ellos y cuyo producto notable es una retícula o modelo gráfico que identifica:

- Contenidos *fuente*, que prestan servicios a otros contenidos.
- Contenidos *sintéticos*, los cuales reciben servicios de otros contenidos.
- Contenidos *rama* o de enlace, que dan y reciben servicios simultáneamente.
- Contenidos aislados, que no presentan relaciones con otros contenidos.

De esta manera, el análisis consistió en representar gráficamente los contenidos del área de español, por eje curricular y grado educativo, haciendo explícitas las relaciones de servicio entre los contenidos. Para efectuar la estructuración se utilizó como base la tabla previamente elaborada para representar el dominio de resultados de aprendizaje (misma que se menciona en la página Nº 33), pues en ella los contenidos presentaban

ya un cierto nivel de organización.

Puesto que en la reticulación las relaciones entre los contenidos pueden ser de naturaleza epistemológica, pedagógica, disciplinaria o de algún otra clase, el número y tipo de enlaces que establece un contenido, son elementos estratégicos para definir su importancia relativa respecto a otros. Así, el análisis de la importancia de los contenidos del área de español que fue iniciada en la etapa anterior fue completado en ésta, con base en los criterios de relevancia y estrategia que se describen a continuación:

- Un contenido es relevante si proporciona numerosos servicios de contenido. En este caso, su relevancia radica en ser requisito de otros; es decir, si no se logra su aprendizaje, el aprendizaje de los que dependen de él se verá afectado.

- Un contenido es relevante si recibe numerosos servicios de otros contenidos. En este caso, la relevancia estriba en la función sintética que desempeña el contenido en el contexto del programa; es decir, se trata de un contenido sintético, probablemente difícil, sobre el cual confluyen varios servicios de contenido que deben ser integrados por él antes de que el programa educativo pueda continuar.

- Un contenido es importante porque, independientemente de los servicios que da o recibe de otros, su relevancia es disciplinaria; es decir, lingüística, psicológica, pedagógica u otra involucrada.

- Un contenido es importante por razones de estrategia evaluativa, como sería el caso de evaluar el aprendizaje de un contenido poco relevante en sí mismo, pero para el cual es más fácil redactar preguntas de examen de tipo objetivo y, así, observar si se logró o no otro más relevante, que lo implica en algún sentido. Por ejemplo, la elaboración de resúmenes de textos, que está incluida en los programas

de tercero y cuarto grados, siendo una habilidad muy importante en el contexto del plan, podría evaluarse explorando otra habilidad menos importante pero que la supone, como sería la redacción de un telegrama (que aparece en el programa de quinto grado).

También en este caso, las decisiones adoptadas para considerar vinculados de algún modo los contenidos o para determinar su importancia relativa, fueron tomadas de manera consensual por los miembros del Comité Diseñador. En todo caso, la idea fue contar con un mapa reticulado del contenido importante a evaluar, que permitiera crear el sistema de evaluación. La retícula que se produjo mediante tales acciones se muestra en la Figura Nº 1, que aparece en la página 38.

En la retícula pueden observarse ciertos aspectos que son importantes para su adecuada interpretación:

- Todos los contenidos están enmarcados por un rectángulo y las relaciones entre ellos están representadas mediante flechas de diversos tipos. La punta de la flecha indica la dirección de los servicios de contenido; normalmente de izquierda a derecha, en el sentido de la secuencia pedagógica que va de antes a después. Así, el eje horizontal de la retícula representa al tiempo. Lo anterior significa que este tipo de relación es de naturaleza antecedente-consecuente. Al respecto cabe señalar que, aunque de hecho se presenta la interacción entre contenidos o causalidad recíproca, este tipo de relación implica una administración de los procesos académicos que no está exenta de la variable temporal; es decir, aunque dos contenidos interactúen, necesariamente uno de ellos tiene que ser presentado para su aprendizaje antes que el otro.

- Todos los contenidos que aparecen en la retícula representan el dominio o universo de contenido del área de español (producto de la etapa anterior) y, entre ellos, figuran los contenidos que están conectados mediante flechas, los cuales

representan el universo de medida; es decir, el contenido estructurado que se juzgó importante evaluar, de conformidad con los criterios antes expuestos.

- En consecuencia, los contenidos que no están conectados, siendo parte del universo de contenido, no fueron considerados para el diseño de la prueba; ya sea porque no pueden ser evaluados en un examen de gran escala con reactivos de respuesta seleccionada, no son tan importantes como otros o porque siéndolo ya fueron considerados de algún otro modo dentro de la estrategia evaluativa. Por ejemplo, el uso de las conjunciones y preposiciones no fue incluido pues se consideró que la función de la educación primaria es introducir dichas nociones, pero que su uso más significativo se da en la educación secundaria.

- En general, la mayor parte de los contenidos considerados como importantes forman parte de cadenas de contenido que empiezan en primero o segundo grado y terminan en quinto o sexto. Como se verá más adelante, cuando se describan las especificaciones de ítems, en tales casos usualmente se seleccionó el contenido más integrador; es decir, el que usualmente estaba al final de la cadena.

- En las cadenas, los contenidos y las líneas que los unen tienen un color que los diferencia de otros, según una taxonomía que se elaboró para clasificar los contenidos. Fue necesario elaborar dicha taxonomía a fin de que resultara significativa la organización de los resultados de aprendizaje importantes, pues la organización en ejes y subejes que se había utilizado hasta el momento resultó ya demasiado general para ser útil. De esta manera, el nuevo nivel de organización de los contenidos, al que se refiere la taxonomía elaborada, fue denominado *línea de formación* y se muestra en la segunda columna de la Tabla Nº 10. Además, la taxonomía aparece, como la sección de código, en la parte inferior derecha de la versión final de la retícula del área de español que se presenta en la Figura Nº 1.

- En la retícula se pueden observar con claridad varios casos que ilustran el uso de los criterios de relevancia antes expuestos. Por ejemplo, en la celda que corresponde a la intersección del primer grado, con los conocimientos, habilidades y actitudes de la lengua hablada, aparece un contenido tipo *fuente* que proporciona cuatro servicios de contenido, y otro que está en ese mismo subeje, pero en el segundo grado, que da tres servicios a otros tantos contenidos; Por otro lado, se observan varios contenidos *sintéticos;* por ejemplo en el tercer grado se encuentran, uno de ellos en el subeje de conocimientos, habilidades y actitudes de la lengua escrita, mismo que recibe tres servicios de contenido, y otros dos en el subeje de situaciones comunicativas que reciben cinco servicios cada uno de ellos.

En las secciones correspondientes del anexo Nº 5, se describen con mayor detalle tanto la estrategia de programación reticular, como los criterios que permitieron definir la importancia relativa de los contenidos.

Tabla 10. Versión final de la tabla de especificaciones para el examen

Eje curricular	Línea de formación (contenidos estructurados)	Nº de contenidos	Relevancia*	Nº de especif.	Nº de ítems	Tipo de ítem	Número del Ítem	Tipo de evaluación
Lengua hablada				**6 / 3**	**13 / 3**			
- Conocimientos y habilidades	Exposición y entrevistas	19	esencial	1 / 2	2 / 2	Opc. mult.	2 y 3 / 43 y 44	Opinión / Logro
	Seguimiento y registro de noticias	1	no esencial	1	2	Opc. mult.	4 y 5	Opinión
	Uso de vocabulario apropiado según el contexto	2	esencial	1	1	Opc. mult.	21	Logro
- Situaciones comunicativas	Narración y descripción orales	8	-	-	-	-	-	-
	Exposición y entrevista	10	esencial	2	4	Opc. mult.	11 y 12; 13 y 14	Opinión
	Argumentación, discusión e intervención mediada	6	esencial	1	2	Opc. mult.	9 y10	Opinión
	Seguimiento y exposición de noticias	1	no esencial	1	3	Opc. mult.	6, 7 y 8	Opinión
Lengua escrita				**1 / 10**	**1 / 18**			
- Conocimientos y habilidades	Manejo de letras y sílabas: ortografía	10	esencial	4	9	Opc. Mult.	8, 9, 10, 11, 12, 13, 38, 39, 40	Logro
	Comprensión lectura y redacción textos breves	10	-	-	-	-	-	-
	Redacción y elaboración de resúmenes	10	esencial	2	3	Opc. Mult.	14, 15, 32	Logro
	Uso de signos de interrogación y exclamación	3	-	-	-	-		-
	Comprensión de textos (instrucciones, normas, fichas)	11	esencial	1	2	Opc. Mult.	41 y 42	Logro
	Manejo de materiales de consulta	7	esencial	1	1	Opc. Mult.	18	Logro
	Uso de signos de puntuación	3	-	-	-	-	-	-
- Situaciones comunicativas	Comprensión de ilustraciones y textos	6	esencial	1	2	Opc. Mult.	19 y 20	Logro
	Comprensión textos (instrucciones, noticias, anuncios)	8	-	-	-	-	-	-
	Redacción de preguntas, cartas, anuncios y solicitudes	8	esencial	1	1	Ensayo	45	Logro
	Uso de técnicas para resumir y tomar apuntes	4	-	-	-	-	-	-
	Organización de bibliotecas	4	no esencial	1	1	Opc. Mult.	1	Opinión
Recreación literaria				**0 / 9**	**0 / 12**			
- Conocimientos y habil.	Comprensión y creación de géneros populares	5	esencial, no esencial	2	5	Opc. Mult.	2, 3, 4, 5; 36	Logro
- Situaciones comunicativas	Comunicación lúdica	1	-	-	-	-	-	-
	Representación literaria	5	esencial	1	1	Opc. Mult.	33	Logro
	Creación oral y escrita de géneros populares	10	esencial	6	6	Opc. Mult.	1, 34,7,6, 35,37	Logro
Reflexión sobre Lengua				**0 / 8**	**0 / 12**			
- Conocimientos y habilidades	Uso de oraciones	4	-	-	-	-	-	-
	Uso de palabras en la oración: Sujeto, verbo y predicado	7	esencial	5	9	Opc. Mult.	22;24,25,26,27, 28;29,30,31	Logro
	Desarrollo vocabulario mediante campos semánticos	6	esencial	1	1	Opc. Mult.	17	Logro
	Tiempos verbales: matices de significado copretérito y pospretérito	1	no esencial	1	1	Opc. Mult.	16	Logro
	Aporte de otras lenguas al español: galicismos y anglicismos	1	no esencial	1	1	Opc. Mult.	23	Logro
- Situaciones comunicativas	Corrección de textos	8	-	-	-	-	-	-
	Elaboración de campos semánticos	3	-	-	-	-	-	-
Totales	**30**	**182**		**7 / 30**	**14 / 45**			

***** Los contenidos juzgados como esenciales fueron censados. Los no esenciales son una muestra al azar de los importantes que se evaluaron

Capítulo 6. Desarrollo de un plan de evaluación

Una vez definido el dominio o universo de contenido del área de español, y habiendo sido identificados y estructurados en una retícula los resultados importantes a evaluar, en esta etapa el reto principal fue elaborar un plan de evaluación para construir la prueba. El plan incluyó tres tipos de acciones: reducir el universo de medida al nivel de un examen de gran escala, diseñar especificaciones para las preguntas del examen y capacitar a las personas que elaboraron los ítems. Los procedimientos y resultados obtenidos mediante esas acciones se describen a continuación.

6.1 Muestreo de resultados de aprendizaje a evaluar.

En última instancia todo plan evaluativo conduce a estrechar el curriculum operacional. Ello es así, debido a que existen muchos más resultados de aprendizaje de los que es posible evaluar en una sola ocasión. Por ello, como ya se comentó en el capítulo 3, en este procedimiento se considera que la validez de un test criterial solo es posible cuando se evalúa a un estudiante en el dominio completo de las metas de aprendizaje definidas por el curriculum o cuando se han seguido procedimientos especiales para obtener una muestra representativa de ellos.

De hecho, uno de tales procedimientos fue desarrollado en la etapa anterior. Ante la imposibilidad de evaluar todo el dominio de resultados que pretende el curriculum del área de español, se seleccionó y estructuró la parte que fue considerada más importante. Ahora ante una dificultad similar, pues el universo de medida es aún demasiado grande, resultaba necesario obtener una muestra que fuera representativa tanto del dominio curricular estructurado, como del dominio completo del área. Para lograrlo, el Comité Diseñador efectuó un muestreo intencional mediante el siguiente procedimiento:

- Se tomaron en consideración todos los aspectos importantes del curriculum que fueron definidos en la retícula del área, en el procedimiento anterior.

- A continuación, se hizo explícito qué partes de curriculum siempre serían evaluadas en el examen, por tratarse de los resultados de aprendizaje de primer nivel de importancia o esenciales y cuales otras serían incluidas sobre la base de un muestreo al azar, ya que por razones prácticas no es posible evaluarlas en su totalidad.

- Se determinó el peso relativo que tendría cada parte del examen, de conformidad con la organización del contenido del área que había sido definida con anterioridad; es decir, por eje y subeje curriculares, así como por línea de formación.

Tales decisiones quedaron registradas en una tabla de especificaciones del examen que, a partir de este momento, continuó desarrollándose hasta que se concluyó el plan de evaluación. La versión final de la tabla de especificaciones se presenta a continuación.

Insertar aquí la tabla de especificaciones completa Tabla No. 10

Para interpretar adecuadamente la información contenida en la tabla, resultan necesarias las siguientes aclaraciones:

- Como puede observarse, la primera columna incluye a los ejes y subejes propios de la organización curricular del área de español de la educación primaria.

- En la segunda columna se muestra el nivel taxonómico creado con el propósito de que el agrupamiento de los contenidos fuera más informativo.

- La tercera columna registra el número de contenidos incluidos en cada línea de formación que fueron estructurados; es decir, los considerados importantes.

- Por su parte, la cuarta columna presenta el dictamen del Comité Diseñador respecto a la importancia relativa de los contenidos estructurados. Así, la categoría ***esencial*** incluye los contenidos que se consideraron de primera importancia y que por ello fueron censados. En cambio, la categoría ***no esencial*** incluye los contenidos estructurados que no fueron considerados de la más alta relevancia, pero que finalmente quedaron incorporados al examen debido a que fueron obtenidos al azar tras un proceso de insaculación de todos aquellos contenidos que no fueron considerados esenciales, pero para los cuales se reservó un espacio en el examen, de conformidad con el criterio de obtener una muestra representativa del contenido del área de español. De esta manera, los contenidos que son evaluados en el examen incluyen los considerados esenciales y los importantes que fueron seleccionados al azar.

- En las dos columnas siguientes aparecen, respectivamente, el número de especificaciones de ítems y el número de reactivos que se consideraron necesarios para evaluar los contenidos de cada línea de formación. Como se verá más adelante, en el siguiente procedimiento, el haber llegado a tales

definiciones fue un proceso iterativo largo y difícil. En ambas columnas, aparecen dos números divididos por una diagonal en los subtotales correspondientes a los ejes de lengua hablada, lengua escrita, recreación literaria y reflexión sobre la lengua, así como en los respectivos totales y en la línea de formación ***Exposición y entrevistas***. En tales casos, los números que están a la izquierda indican que se trata de especificaciones o ítems que corresponden a un cuestionario de opinión que se aplicó a los niños y que se comentará más adelante. En cambio, los números que aparecen a la derecha de la diagonal identifican a las especificaciones o los ítems de logro que integraron el examen. Lo anterior se indica también en la última columna de la tabla.

- En la columna que hace referencia al tipo de ítem se observa que, con excepción de un contenido correspondiente a la línea de formación ***Redacción de preguntas, cartas, anuncios, y solicitudes***, para el cual se estableció un ítem de ensayo, para los demás contenidos solo se especificaron reactivos de opción múltiple.
- La columna restante, identifica el número de ítem que correspondió finalmente a cada reactivo tanto del cuestionario de opinión, como del examen.

Una vez creado el plan, las demás etapas de la construcción de la prueba fueron más técnicas.

6.2 Diseño de especificaciones de ítems.

A partir de los productos elaborados previamente, en particular la retícula del área de español y la tabla de especificaciones, y con el apoyo del manual correspondiente, el Comité Diseñador elaboró especificaciones técnicas, tanto desde el punto de vista del contenido como del psicométrico, para la construcción de cada uno de los ítems de la prueba.

Las especificaciones de ítems tuvieron como propósito principal proporcionar a los elaboradores de ítems el contenido específico, derivado de cada línea de formación, y los detalles técnicos necesarios para generar reactivos efectivos. Sin embargo, otros propósitos asociados fueron el comunicar a los usuarios del examen qué es lo que cada ítem mide, así como la cobertura y alcance de las competencias evaluadas, además de proporcionar un marco contextual que ayude a interpretar la ejecución de los estudiantes. La estructura general de cada especificación de ítems fue la siguiente:

- Una descripción general de la tarea de evaluación. La descripción incluyó el enunciado del contenido y su ubicación en el grado educativo correspondiente. Además, casi siempre se hizo una interpretación del sentido del contenido, un comentario acerca de su importancia, una delimitación del segmento del contenido que debería cubrir o alguna otra nota que dejara claro cuál era el concepto o la habilidad que se quería evaluar y cómo debería evaluarse.

- Una descripción de los atributos de los estímulos y de las respuestas que debería presentar el ítem, según fueran las necesidades más o menos específicas en cada caso.

- La redacción de un ítem muestra que ilustrara la manera en que se aplicaron los demás elementos de la especificación. Cabe señalar que el conjunto de ítems que sirvieron de muestra para las especificaciones, finalmente fue estructurado para formar el modelo 4 del examen.

Para ejemplificar a los miembros del Comité Diseñador la aplicación de estos elementos, se adaptaron dos prototipos de especificaciones de ítems, uno muy detallado y otro de nivel medio, mismos que forman parte del manual que se elaboró para ellos. En general, se prefirió el prototipo de nivel medio porque a probado ser más funcional y menos restrictivo (Nitko, 1994; Popham, 1990).

La idea fue contar con un marco normativo compacto, claro y significativo, que permitiera a los elaboradores de los reactivos producir ítems válidos y saber cuando no lo eran, así como al propio Comité Diseñador tener estándares contra los cuales contrastar posteriormente el valor de los ítems elaborados.

Para definir la cantidad de especificaciones y de ítems que sería necesario elaborar para la prueba, se siguieron los criterios relativos al tipo de estimación de la ejecución del examinado y el tipo de impacto de la prueba, señalados para determinar la extensión del test, los cuales aparecen en la tabla Nº 4 del capítulo 2, y otros de tipo operativo. Al respecto, se efectuaron las siguientes consideraciones:

- En consonancia con el enfoque básicamente descriptivo que se pretendía dar al examen, se requería una descripción de la ejecución de los examinados que fuera más bien específica, a fin de poder dar cuenta con cierto detalle de la ejecución que tendrían en los ejes, subejes y líneas de formación del área de español. Por ello se requerían bastantes reactivos.

- En cuanto al impacto esperado del examen, se pensó que implicaba una afectación mínima a los examinados o a sus profesores, por lo cual se necesitaban más bien pocos ítems.

- Dadas la contradicción emanada de estas consideraciones y para tener una medida de referencia, se revisó la extensión de otros instrumentos de gran escala diseñados para la educación elemental en México y en otros lugares, tales como el CAT, el del INCE y la prueba de Aguascalientes, que ya se mencionaron previamente.
- Además, se efectuó una pequeña prueba empírica para determinar la cantidad de tiempo que les llevaría a niños de 5º y 6º de primaria y de 1º de secundaria, responder a los 40 ítems muestra del primer conjunto de especificaciones que fue

elaborado. El rango de tiempo osciló entre 1.5 y 2 horas y en general los niños no mostraron cansancio o falta de atención.

Una vez definidos estos aspectos, se decidió elaborar un total de 30 especificaciones para producir 45 ítems. En el anexo Nº 6, se presenta el conjunto final de las especificaciones para producir los ítems de la prueba.

A manera de ilustración, enseguida se presenta una especificación de ítems completa:

- **Eje*: Reflexión sobre la lengua;* Subeje*: Conocimientos, habilidades y actitudes;* Línea de formación*: Uso de las palabras en la oración: sujeto, verbo y predicado***

Identificación del sujeto y del predicado en las oraciones (tercer grado)

El propósito es conocer el grado en que el niño de este nivel domina los conceptos de sujeto y predicado, cuando estos elementos gramaticales se encuentran en el contexto de la oración. Para ello, se requieren cinco ejercicios que contengan, cada uno de ellos, un enunciado que presente los dos elementos oracionales. Se trata de determinar si el niño aprendió dichos elementos por su conceptualización o simplemente por el orden en que se ubican comúnmente en la oración; es decir, en forma mecánica. En cada enunciado que se presente, se deberá variar la ubicación tanto del sujeto como del predicado. Por ejemplo:

- Los niños llegaron hasta la meta.
- Hasta la meta llegaron los niños.
- Llegaron los niños hasta la meta

En las instrucciones para responder se establecerá la forma en que los examinados deberán localizar ambos elementos de la oración. En todo caso, se les pedirá que identifiquen en cada enunciado solo uno de los dos elementos oracionales. Ejemplo de ítem:

Instrucciones. Elige la opción que presenta el sujeto en el siguiente enunciado:

Llegaron los niños hasta la meta después de mucho esfuerzo

❑ meta
❑ después
❑ esfuerzo
❑ Niños*****

Figura Nº 1 *bis*. Ejemplo de especificación de ítems para el examen de español

Haciendo un balance general de las especificaciones de ítems que fueron elaboradas podríamos decir que, aunque todas corresponden a contenidos que pertenecen a los

cuatro ejes del área de español, su distribución dista mucho de ser equitativa. Así, para el eje de Lengua Hablada se elaboraron tres especificaciones; para Lengua Escrita diez; para Recreación Literaria nueve; y para Reflexión sobre la Lengua ocho. Entre las razones que explican este hecho, podemos mencionar las siguientes:

- Tanto el dominio de resultados de aprendizaje del área de español que fue identificado, como la estructura del contenido importante a evaluar que se determinó a partir de dicho dominio, dejan ver con claridad el gran énfasis que se otorga en el curriculum a la lengua hablada y a la escrita. Lo anterior puede ser constatado en la versión final de la tabla de especificaciones del examen que se muestra en la tabla Nº 10. En ella, se observa que 47 contenidos de un total de 182 estructurados (el 25.8% de los contenidos del área) corresponden al eje de lengua hablada; Por su parte, 84 contenidos (el 46.1 % del total del área) pertenecen al eje de lengua escrita. Otro tanto sucede con el contenido total del área (al respecto, véase la retícula que aparece en la página 38).

- Por tratarse de una prueba de gran escala, las especificaciones se orientan a producir ítems de respuesta seleccionada, de opción múltiple. Sin embargo, como puede observarse en la retícula que aparece en la citada página 38, el enfoque del nuevo curriculum del área de español enfatiza en gran medida el desarrollo de habilidades comunicativas de producción, principalmente en los ejes de lengua hablada, lengua escrita y recreación literaria, mismas que para su evaluación apropiada requieren de ítems de respuesta construida, como los de ejecución. Con el propósito de salvar esta situación, los miembros del comité hicieron un esfuerzo considerable cuando diseñaron las especificaciones de ítems, de tal manera que al evaluar una habilidad pudiera seleccionarse alguna dimensión de ella cuya medición aportara información relevante sobre su dominio y que, a la vez, fuera susceptible de ser evaluada con un ítem de opción múltiple. En los casos en que lo

anterior no fue posible, las especificaciones tienden a ignorar esas partes del curriculum, a pesar de ser importantes.

- Una excepción a lo anterior, fue la decisión de incorporar al examen una especificación de ítems para producir un reactivo de ejecución orientado a capturar una muestra de la habilidad para redactar de los niños, correspondiente al eje de Lengua Escrita y que se consideró demasiado importante en el contexto del currículo como para ser ignorada.

- Estrechamente relacionado con los dos puntos anteriores, para el caso de los contenidos correspondientes al eje de Lengua Hablada, que representan casi un tercio del curriculum del área de español (y el 25.8% del contenido que se juzgó importante evaluar), y que en gran medida se refieren a habilidades eminentemente de producción, se tomó la decisión de elaborar especificaciones de ítems para producir preguntas que permitieran indagar la opinión de los examinados respecto a la realización de actividades de aprendizaje que prevé el curriculum y que tienen como propósito desarrollar habilidades como hacer presentaciones orales de temas, realizar entrevistas, participar en debates y otras que, por su naturaleza evanescente, no pueden ser evaluadas mediante un test de gran escala con ítems de opción múltiple.

  La alternativa era no evaluar esos repertorios de ejecución. Sin embargo, con ello se desaprovecharía el costo de oportunidad que representa aplicar un examen de esta naturaleza para conocer, aunque sea en una medida muy limitada e inadecuada, si se han realizado esta clase de actividades que marcan los programas, cómo se han realizado y el impacto que han tenido en el aprendizaje de los niños. No obstante, en los casos en que fue posible obtener una medida más directa del aprendizaje logrado por los alumnos en el eje de la lengua hablada, las especificaciones de ítems marcan reactivos que miden el logro educativo. Dicho

cuestionario de opinión quedó integrado con 14 preguntas y aparece en el anexo Nº 9. Con el propósito de ilustrar el tipo de preguntas que contiene, enseguida se presenta una especificación de ítems que contiene una pregunta del cuestionario de opinión que fue aplicado.

- **Práctica del debate (6º grado).**

Porque los niños deben desarrollar su capacidad para expresarse oralmente con claridad, coherencia y sencillez y puesto que la práctica del debate facilita esta finalidad, es necesario verificar la frecuencia con que el alumno participó en la práctica del debate; para esto habrá que elaborar un ítem, como en el siguiente ejemplo:

Instrucciones. Con la siguiente pregunta, se quiere saber qué tan seguido practicaste el debate en tu salón. Para contestar, marca la opción que describe mejor tu caso.

¿En cuántos debates participaste durante este año escolar?

❑ En ninguno.
❑ Participé en uno.
❑ Participé en dos o más debates.
❑ Solo preparamos el debate, pero no lo realizamos.

Figura Nº 2. Ejemplo de especificación de ítems para el cuestionario de opinión

6.3 Capacitación del Comité Elaborador de ítems.

Como resultado de un convenio informal con las autoridades educativas del municipio de Ensenada, quienes tuvieron la generosidad de comisionar a un grupo de trabajo, fue constituido un comité integrado por 17 personas, entre quienes se encontraban cinco asesores técnicos, cuatro profesores de primaria y seis directivos escolares, todos ellos especialistas en el área de español y con experiencia en la redacción de reactivos de tipo objetivo, quienes elaboraron los ítems de la prueba. Para apoyar la capacitación de este grupo, fue elaborado un manual que incluyó, entre otros, los siguientes documentos (ver Anexo No. 1): retícula del área de español y documento descriptivo, especificaciones de ítems, documento para la redacción técnica de ítems, tabla de

especificaciones de la prueba y formatos de registro de información. El Comité Diseñador fue responsable tanto de la selección y diseño de los materiales, como del entrenamiento correspondiente, el cual se dio primero mediante cinco sesiones en un curso formal de 20 horas, y posteriormente continuó por medio de asesorías individuales, según las requirieron los elaboradores de los reactivos.

Capítulo 7. Producción y validación de ítems

Una vez elaboradas las especificaciones de ítems y habiendo sido entrenado el grupo de elaboradores de reactivos, en esta etapa se procedió a desarrollar los reactivos, analizarlos a la luz de las especificaciones elaboradas, probarlos empíricamente ante una muestra de alumnos que estaban por egresar de escuelas primarias, efectuar un análisis de las respuestas a los ítems obtenidas mediante su aplicación, así como revisar los ítems y estructurar la versión final la prueba. Estos cinco procedimientos se describen a continuación, así como los resultados obtenidos al operarlos.

7.1 Elaboración de ítems.

Con base en el manual de especificaciones y en el entrenamiento recibido, el **Comité Elaborador** desarrolló un conjunto de 200 ítems para la prueba de conformidad con las normas, a fin de propiciar su validez. La distribución de los ítems entre los elaboradores no fue equitativa; se dejó que ellos se dividieran el trabajo en función de su especialidad, su interés y la disponibilidad de tiempo para la tarea. Por los resultados obtenidos tal mecánica fue adecuada, con alguna excepción.

En general, los elaboradores diseñaron los ítems que les correspondió trabajar de manera independiente o en pequeños grupos y posteriormente los aplicaron a sus alumnos o a niños de otra escuela, a fin de efectuar una calibración inicial de las preguntas y retroalimentar así su trabajo. Lo anterior no pudo realizarse en todos los casos, particularmente no se realizó con los ítems correspondientes al eje de Recreación Literaria, debido a que en su mayoría contienen dibujos y ello retrasó su elaboración, la cual concluyó poco antes de la aplicación.

Durante este procedimiento se presentaron diferentes tipos de interacciones entre los miembros de los comités elaborador y coordinador, siendo las más comunes las que

tenían por objeto hacer aclaraciones y retroalimentar las actividades. Al finalizar la elaboración, los responsables entregaron los ítems que diseñaron, junto con los resultados de la pequeña prueba empírica que realizaron.

Algunos elaboradores entregaron más ítems de los que les correspondió elaborar, exceso que fue producto del uso del método de generación de ítems que emplearon, principalmente el de transformaciones lingüísticas (mismo que se describe en la parte final del anexo Nº 1).

En general, la elaboración de los ítems duró unos 15 días y en los casos en que los reactivos contenían dibujos, se prolongó una semana más.

Gracias al apoyo financiero que otorgó al proyecto el Programa Interinstitucional de Investigaciones sobre la Educación Superior (PIIES) de la SEP, fue posible compensar parcialmente el esfuerzo que realizaron los elaboradores, a quienes se pagó una pequeña cantidad por ítem limpio; es decir, por cada reactivo elaborado de conformidad con las normas, dictaminado favorablemente por el Comité Diseñador que analizó su congruencia con la especificación correspondiente, analizado empíricamente y, en su caso, corregido de las fallas detectadas mediante dichos procedimientos.

7.2 Revisión formal de la congruencia ítem-especificación.

Una vez elaborados, los reactivos fueron sometidos a una detallada revisión de contenido, psicométrica y lógica, contra las especificaciones de ítems correspondientes (las que aparecen en el anexo Nº 6). Además, para la revisión se consideraron los estándares, medidas y criterios correspondientes a la calidad del contenido de los ítems del test y a la calidad técnica de cada ítem del test, que fueron definidos previamente en la tabla Nº 9 que se presentó en el capítulo 3. Esta tarea fue realizada por el **Comité Diseñador,** y tuvo como propósito garantizar la calidad de la relación:

ítem *representa* **curriculum** y, con ello, la validez de los ítems. El procedimiento de revisión operó de la manera que se ilustra en el siguiente diagrama de flujo:

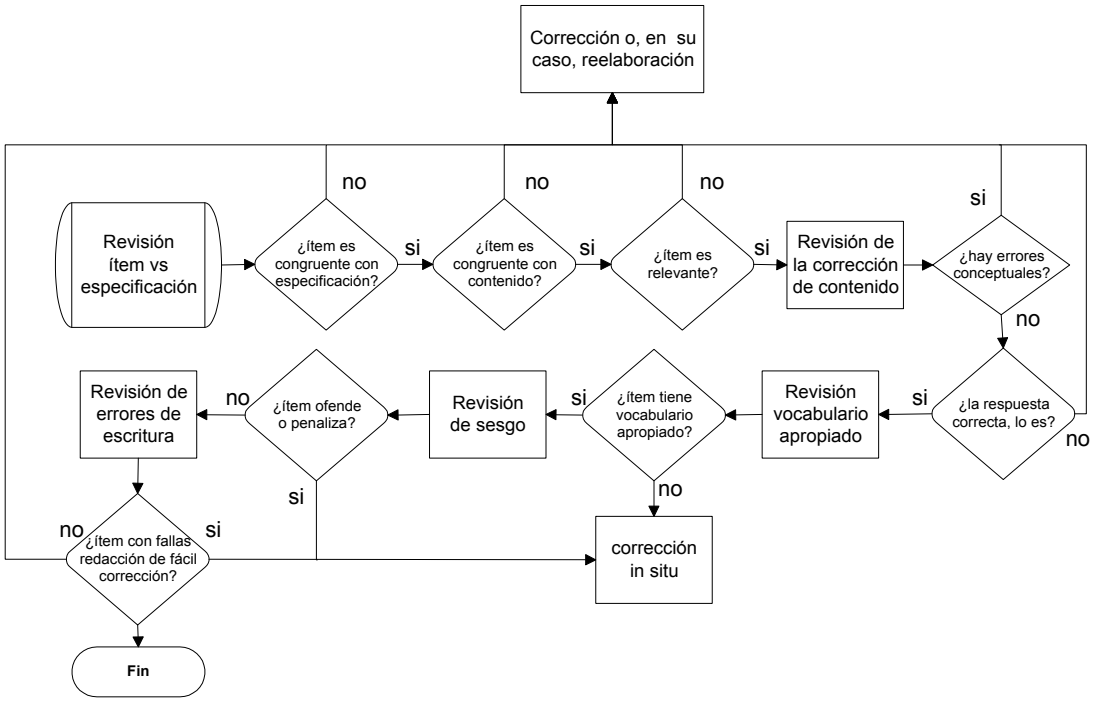

Figura Nº 3 Diagrama de flujo del proceso de revisión formal de los ítems de la prueba

A continuación se describe el proceso ilustrado en el diagrama:

- Inicialmente, cada ítem fue contrastado con la especificación que lo produjo. Para ello el Comité Diseñador, constituido en un panel de expertos, dictaminó si el ítem era congruente con la especificación y con el contenido correspondiente y si además resultaba relevante. También en esta ocasión las decisiones del comité se adoptaron de manera consensual. Cuando el dictamen resultó favorable en los tres casos, se continuó con el análisis; cuando no lo fue, se envió el ítem a corrección o reelaboración según fuera la naturaleza de las observaciones.

- Posteriormente, se revisó la corrección del contenido del ítem; es decir, que

estuviera exento de errores conceptuales y que la respuesta correcta lo fuera. Cuando el dictamen fue favorable en ambos casos, se continuó el proceso y en los casos en que no lo fue se indicó lo procedente.

- Enseguida, se revisó que el ítem presentara un vocabulario apropiado al nivel de los examinados. Cuando no lo tuvo, se hicieron las correcciones necesarias ahí mismo.

- Además, se revisó que el ítem no presentara estereotipos étnicos o de género y que no pusiera en ventaja a un grupo sobre otros. Lo mismo se hizo posteriormente con el conjunto de ítems que formaron cada modelo. En ningún caso se juzgó que estuviera presente algún tipo de sesgo.

- Finalmente, se revisó que el ítem estuviera exento de errores de escritura, tanto técnica como de tipo mecanográfico. En los casos en que se presentaron determinantes específicos, opciones poco plausibles y otras fallas similares, se efectuaron las correcciones en el momento. Cuando no pudo hacerse de inmediato y según la naturaleza de la falla detectada, la corrección se dejó a cargo de una subcomisión y su dictamen quedó pendiente para la siguiente sesión.

En general, el procedimiento de revisión resultó bastante fluido. En parte, debido a que la mecánica de trabajo consideró como un requisito para poder efectuar una reunión la revisión previa, por parte de cada miembro del comité, de todos los reactivos a evaluar en cada sesión. Parcialmente también, la fluidez de los trabajos se debió al reducido número de miembros del comité que continuaron y concluyeron estos trabajos, mismos que al final fueron tres. Esto último, explica además porqué se decidió adoptar al consenso como criterio para decidir y porqué el procedimiento se prolongó considerablemente.

Al igual que los elaboradores de ítems, por esta y las demás actividades que realizaron durante el complejo diseño de la prueba, los miembros del Comité Diseñador recibieron

una modesta compensación económica, que fue posible otorgarles gracias al apoyo del PIIES.

7.3 Ensayo empírico de los ítems.

Además de la aplicación del examen, este procedimiento incluyó tres acciones previas: estructurar los modelos de examen, seleccionar la muestra de examinados y capacitar a los aplicadores del examen, mismas que se describen a continuación:

Estructuración de los modelos de examen.

Con el propósito de efectuar una primera calibración formal de la calidad técnica de los ítems y modelos, en cuanto a su nivel de legibilidad (redacción clara para los sujetos, vocabulario que corresponde a su nivel y conceptualizaciones acordes con su desarrollo), dificultad, discriminación, funcionamiento diferencial ante los sujetos y confiabilidad, el **Comité Coordinador** estructuró los reactivos revisados, de conformidad con la tabla de especificaciones, en cuatro modelos de examen con 45 ítems cada uno de ellos.

En términos generales, las siguientes características están presentes en cada modelo:

- Al inicio del examen se presenta una página que describe a los niños el propósito del examen, seguido de instrucciones generales para responder y de un ejercicio de práctica que supervisa el aplicador.

- En la segunda página se presenta el reactivo Nº 1 el cual requiere, para ser respondido por los examinados, que el aplicador lea previamente en voz alta un pequeño cuento. En cada una de las 21 páginas que tiene cada modelo, aparecen impresos el mismo número y tipo de reactivos que en esa misma

página tienen los demás modelos. La idea es que, a partir de los modelos originales, puedan estructurarse y ponerse a prueba psicométrica diferentes formas paralelas del examen.

- Al inicio de cada ítem se incluyen las indicaciones necesarias para responderlo.

Con el propósito de ilustrar la configuración final que se dio a cada modelo, en el anexo Nº 8 se presenta el modelo de examen 4 completo; el mismo que se estructuró con los reactivos que sirvieron de muestra en las especificaciones de ítems. Los demás modelos no se incluyen, debido a la necesidad de preservar la integridad del instrumento y con ello su validez para ser empleado posteriormente de manera efectiva.

Como ya se explicó al final del capítulo anterior, además de los cuatro modelos de examen constituidos por reactivos que miden el logro académico de los niños, se estructuró un cuestionario para explorar la opinión de los examinados sobre la realización de ciertas actividades curriculares que establecen los programas de estudio para los ejes de Lengua Hablada y Lengua Escrita. Para estructurar el cuestionario no se elaboraron nuevos ítems, sino que se integró el instrumento con los ítems muestra de las especificaciones correspondientes a dichos ejes curriculares. Tales especificaciones se muestran al final del anexo Nº 6, y el cuestionario de opinión que se aplicó a los niños se presenta en el anexo Nº 9.

Los cuatro modelos de examen fueron aplicados a una muestra inicial de 253 niños, de 12 escuelas, que en el momento de la aplicación estaban por egresar de la educación primaria. Por su parte, el cuestionario de opinión fue aplicado a 90 niños, de 6 escuelas, entre los que se encontraban buena parte de los que respondieron el examen y quienes lo contestaron al terminar de ser examinados; además, se aplicó a otros niños que no fueron examinados y que solo respondieron las preguntas del cuestionario de opinión.

Obtención de la muestra de sujetos a quienes se aplicó el examen.

Previa a la aplicación, fue seleccionada una muestra intencional de niños para ser examinados, la cual presentó las siguientes características:

- Alumnos que se encontraban cursando el 6º grado de primaria en el momento de aplicarse la prueba.

- Con el propósito de explorar el comportamiento del examen ante poblaciones diferentes, así como la ejecución de ellas en el test, se trató de que en conjunto los niños presentaran una variedad de condiciones en cuanto a:

    - Tipo de escuela (pública, privada)
    - Diversidad delegacional (en Ensenada)
    - Tipo de población (urbana, urbana marginal, rural)
    - Nivel sociocultural (alto, medio, bajo), según la valoración del director de la escuela, del inspector escolar y del aplicador del examen
    - Turno (matutino, vespertino)
    - Lugar de origen (BC, estado fronterizo, centro del país, sureste del país)
    - Sexo (masculino, femenino)
    - Edad (hasta 12 años, 13-14 años, 15 o más años)

En general, estas condiciones quedaron representadas de manera apropiada en la muestra, de tal manera que los examinados presentaron variedad en cuanto a las características requeridas para pilotear los modelos de examen y calibrar los ítems. Así, participaron en el examen escuelas de cuatro delegaciones del municipio de Ensenada; siete fueron urbanas, tres urbanas marginales y dos rurales; diez de ellas son públicas y dos privadas; diez ofrecen el servicio en el turno matutino y 2 de ellas en el vespertino; y una tiene alumnos que típicamente pertenecen al nivel socioeconómico

alto, otra más al medio alto, siete al nivel medio y 3 de ellas al bajo.

Cabe señalar que el examen solo pudo ser aplicado en el municipio de Ensenada debido a que las pláticas que se habían sostenido hasta el momento con las autoridades educativas del estado de Baja California, no culminaron oportunamente en el convenio necesario para una aplicación a escala estatal, a pesar del interés que habían manifestado. No obstante, con el generoso apoyo que otorgaron en esta y otras etapas del proceso las autoridades educativas del municipio de Ensenada, fue posible efectuar la aplicación del examen en escuelas de dos inspecciones escolares, casi al término del ciclo escolar 1998-1999.

Como resultado de lo anterior, tras aplicar el examen fue posible obtener una caracterización de la población que fue examinada, misma que se presenta en la tabla Nº 11, que aparece en la página siguiente.

Entre los principales aspectos que destacan en la tabla, podemos mencionar que:

- En general, las condiciones que parecen caracterizar a los examinados son: que se trata de estudiantes del turno matutino de escuelas oficiales de la zona urbana de Ensenada; que pertenecen a los niveles socioeconómicos medio y bajo; que tienen alrededor de 12 años de edad; de ambos sexos por igual; nacidos principalmente en Baja California; cuyas madres estudiaron hasta la secundaria y sus padres son empleados; que aspiran a estudiar en la universidad y que viven en casa propia, lo cual no es extraño en esta región del país.

- Al parecer, los datos de la tabla coinciden con los reportados en muchos otros trabajos que exploran los factores sociales y económicos que intervienen en la educación de los niños: qué quienes pertenecen a niveles socioeconómicos bajos tienen madres con baja escolaridad, menos aspiración por estudios

Tabla Nº 11. Descripción de la población estudiada en el municipio de Ensenada.

Condición		Nº de alumnos por escuela													Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	
Delegación	Ciudad	23	36	16	-	-	-	24	30	-	21	30	25	15	220
	San Antonio	-	-	-	3	-	-	-	-	-	-	-	-	-	3
	Otra	-	-	-	-	13	17	-	-	-	-	-	-	-	30
Tipo de población	Urbana	23	-	16	-	-	-	24	-	-	21	30	25	15	154
	Urbana marginal	-	36	-	-	13	-	-	30	-	-	-	-	-	79
	Rural	-	-	-	3	-	17	-	-	-	-	-	-	-	20
Tipo de escuela	Pública	23	36	-	3	13	17	24	30	-	21	30	25	-	222
	Privada	-	-	16	-	-	-	-	-	-	-	-	-	15	31
Turno	Matutino	23	36	16	3	-	17	24	30	-	-	30	25	15	219
	Vespertino	-	-	-	-	13	-	-	-	-	21	-	-	-	34
Nivel socioeconómico	Alto	-	-	-	-	-	-	-	-	-	-	-	-	15	15
	Medio alto	-	-	16	-	-	-	-	-	-	-	-	-	-	16
	Medio	23	-	-	-	13	-	24	30	-	21	30	25	-	166
	Bajo	-	36	-	3	-	17	-	-	-	-	-	-	-	56
Edad	10 a 12	19	32	14	2	7	13	23	25	-	10	30	13	14	202
	13 a 14	4	4	2	-	4	3	1	4	-	10	-	12	1	45
	15 a 20	-	-	-	1	2	1	-	1	-	1	-	-	-	6
Sexo	Femenino	10	19	7	2	7	8	9	20	-	13	13	14	8	130
	Masculino	13	17	9	1	6	9	15	10	-	8	17	11	7	123
Ciudad de origen	Sureste del país	3	-	-	-	1	-	-	2	-	-	-	-	-	6
	Centro del país	1	4	4	2	2	6	3	11	-	4	4	4	2	47
	Fronteriza	1	1	1	-	2	1	2	-	-	5	3	2	3	21
	Baja California	18	31	11	1	8	10	19	17	-	12	23	19	10	179
Años de vivir en Ensenada	Menos de 5	1	2	2	-	5	-	1	5	-	2	4	-	3	25
	Entre 5 y 10	2	8	3	1	2	3	4	9	-	3	10	7	3	55
	Más de 10	20	26	11	2	6	14	19	16	-	16	16	18	9	173
Escolaridad máxima de la madre	Primaria	8	3	-	2	8	8	2	4	-	3	3	-	-	41
	Secundaria	4	8	-	-	3	4	4	6	-	10	5	6	-	50
	Bachillerato	3	6	2	-	2	2	3	4	-	3	1	7	-	33
	Técnica	-	3	-	-	-	-	1	1	-	2	-	-	-	7
	Normalista	-	5	-	-	-	-	1	-	-	-	3	2	2	13
	Universidad	3	1	13	-	-	-	3	1	-	-	13	4	2	40
	No sabe	5	10	1	1	-	3	10	14	-	3	5	6	11	69
Profesión del padre	Obrero o campesino	1	-	-	1	-	3	3	3	-	3	1	-	-	15
	Empleado de oficina	5	31	1	1	1	9	4	13	-	6	7	9	6	93
	Comerciante	4	3	2	-	-	1	1	2	-	1	2	2	4	22
	Técnico	-	2	-	-	3	-	1	3	-	-	4	7	-	20
	Otra	10	-	13	-	7	1	11	6	-	7	13	4	5	77
	No sabe	3	-	-	1	2	3	4	3	-	4	3	3	-	26
Aspiración máxima de estudios	Secundaria	1	-	-	1	-	5	3	-	-	1	-	-	-	11
	Bachillerato	2	-	-	-	1	2	1	5	-	1	-	1	-	13
	Carrera corta	2	3	-	1	2	3	2	4	-	5	-	1	-	23
	Universidad	18	33	16	1	10	7	17	21	-	14	30	23	15	205
Vivienda	Propia	21	34	15	2	9	16	20	27	-	15	23	23	12	217
	Rentada	2	1	1	-	4	1	3	1	-	4	6	1	2	26
	Prestada	-	-	-	-	-	-	1	2	-	2	1	-	1	7
	Otra	-	-	-	1	-	-	-	-	-	-	-	-	-	1
	No sabe	-	1	-	-	-	-	-	-	-	-	-	1	-	2

superiores, tienen padres que son obreros, campesinos o empleados, y viven en zonas rurales o urbanas marginales. Por lo contrario, los estudiantes de las dos escuelas privadas, que fueron clasificadas como pertenecientes a los niveles socioeconómicos medio alto y alto, tienen madres con alta escolaridad y padres comerciantes o empleados de oficina y todos aspiran a estudiar en la universidad.

- Cuando se contrastan estas condiciones con los aciertos que obtuvieron los niños en el examen, los resultados son similares. Lo anterior puede observarse en la tabla que se muestra enseguida, la cual presenta únicamente aquellas correlaciones entre las condiciones de la muestra y el número de aciertos en el examen, que resultaron significativas estadísticamente:

Tabla Nº 12. Correlaciones entre aciertos en el examen y las condiciones de la muestra de examinados, que son significativas estadísticamente.

Condición	Nivel de significancia de la relación	
	al 0.05	al 0.01
Delegación		.253
Tipo de escuela		.342
Turno		.220
Nivel socioeconómico		.226
Escolaridad de la madre		.209
Aspiración máxima de estudios	.158	

Así, en términos de correlación, presentan más aciertos de manera significativa (en los niveles de 0.05 y 0.01, respectivamente) quienes estudiaron en la zona típicamente urbana de Ensenada, en escuelas privadas, en el turno matutino, pertenecen a los niveles socioeconómicos alto y medio alto, tienen madres con mejor escolaridad y aspiran a estudios más altos.

- Por razones que se comentan más adelante, en la escuela Nº 9 no fue posible aplicar el examen. Así, no aparecen los datos correspondientes a esos estudiantes en la tabla Nº 13 y en las demás tablas que se presentarán posteriormente.

Selección y capacitación del Comité Aplicador del examen.

Para aplicar el examen, fue seleccionado un Comité Aplicador integrado por cuatro personas, quienes fueron entrenadas para operar los modelos bajo condiciones estandarizadas, particularmente las relativas a la organización de las actividades, el manejo confidencial de los ejemplares de examen y hojas de respuesta, así como la impartición de instrucciones a los niños, el seguimiento de su ejecución en el ensayo de práctica para responder y la solución a sus dudas sobre cómo contestar la prueba. La capacitación se proporcionó en dos sesiones de trabajo y como material de apoyo fue elaborado un manual, mismo que aparece en el anexo Nº 7.

En general, el instructivo para la aplicación y los materiales contenidos en el manual resultaron funcionales, y la participación de los aplicadores fue la adecuada. Sin embargo, ciertos detalles de coordinación causaron algunos problemas. Por ejemplo, uno de los mapas con la ubicación de las escuelas causó dificultades para llegar a un plantel rural y se retrasó la aplicación; en otro caso, la prueba no pudo aplicarse en la mencionada escuela Nº 9 debido a que previamente había sido programada una salida de los alumnos, la cual coincidió con la fecha prevista para aplicación del examen en el plantel.

Como puede observarse, además de dar comienzo a la fase empírica del proceso de calibración de la calidad técnica de los ítems y modelos de examen, este procedimiento tuvo como propósito adicional explorar inicialmente las condiciones necesarias para estandarizar su aplicación.

7.4 Análisis estadístico de los ítems.

Una vez aplicados los modelos de examen, el **Comité Coordinador** procedió a capturar los resultados y a efectuar un análisis de ítems y de confiabilidad de los

modelos, para estimar su calidad técnica. En el primer caso, las hojas de respuesta fueron leídas mediante un *scanner* de alta velocidad y la información se registró en una base de datos. Para el segundo caso, se obtuvieron índices de dificultad apropiada para cada ítem y modelo mediante tres procedimientos:

- Se calculó el *índice de dificultad* o valor ***p*** del reactivo. Es decir, la proporción de examinados que contestaron correctamente el ítem. Para ello se empleó la fórmula:

$$p = c / t,$$

Donde,
c = número de examinados que tuvieron correcto el ítem
t = total de examinados que respondieron el ítem

El valor **p** fue obtenido de la muestra de ensayo y, para ser apropiado, se consideró que debería ser mayor que 0.05 y menor que 0.95. Tanto en este caso, como en los dos siguientes, el análisis se orientó por los criterios, medidas y estándares que fueron adoptados para controlar la calidad para la prueba y que aparecen en la tabla Nº 9, del capítulo 3.

- Se obtuvo el *índice de discriminación* o valor ***D*** del ítem. Es decir, la diferencia entre ***p*** para el grupo alto (27%) y ***p*** para el grupo bajo (27%). Para calcularlo se utilizó la fórmula:

$$D = pa - pb$$

Donde,
pa = dificultad del ítem para el grupo alto; es decir, el grupo de examinados que obtuvieron las calificaciones más altas en el examen (el 27% del total de examinados).

Pb = dificultad del ítem para el grupo bajo; es decir, el grupo de examinados que obtuvieron las calificaciones más bajas en el examen (el 27% del total de examinados).

El valor discriminativo del ítem se consideró apropiado si ***D*** >= 0.2. Además, se obtuvo el *coeficiente de discriminación*, mediante la correlación del punto biserial; es decir, la relación entre las respuestas al reactivo (0 o 1) y las calificaciones en el test de todos los examinados y no solo del 54% de ellos.

- Para estimar la calidad de los puntajes del test, o sea su calidad integral como instrumento de medida, fue estimada la *confiabilidad* mediante el *coeficiente de consistencia interna* de cada modelo de examen; es decir, el coeficiente *alfa de Cronbach* o *kuder-Richardson 20*, que puede ser considerado como la correlación promedio obtenida de todas las posibles estimaciones de confiabilidad, mediante división por mitades, de los reactivos de un modelo de examen (Popham, 1990, p. 133), mismo que para este caso requería ser mayor o igual que 0.85 en cada caso. Para determinar la confiabilidad de los modelos se calculó el coeficiente alfa mediante la fórmula:

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma\sigma_i^2}{\sigma_t^2}\right)$$

donde α = estimador de la confiabilidad
n = número de reactivos de la prueba
σ_t^2 = varianza de la prueba
$\Sigma\sigma_i^2$ = sumatoria de la varianza de los reactivos

Los resultados del análisis de ítems se muestran en la tabla compuesta que se presenta en la página siguiente. La idea de juntar en una sola, las tablas correspondientes al índice de dificultad de los reactivos de los cuatro modelos, al índice

de discriminación de sus ítems, al coeficiente de discriminación de los mismos y al coeficiente de confiabilidad de los modelos, es presentar una visión panorámica que permita observar con facilidad y tener una referencia rápida de los detalles y las relaciones entre sus elementos.

Tabla Nº 13. Resultados del análisis de respuestas a los ítems y a los modelos de examen: Índices de dificultad y discriminación, y coeficientes de discriminación y de confiabilidad, de los ítems y modelos.

Dificultad de los ítems del examen

Item	Modelo 1	Modelo 2	Modelo 3	Modelo 4
1	0.04	0.61	0.56	0.75
2	0.61	1.00	0.90	0.81
3	0.96	0.78	0.76	0.79
4	0.96	0.83	0.85	0.82
5	0.79	0.94	0.74	0.88
6	0.43	0.47	0.54	0.86
7	0.18	0.14	0.13	0.32
8	0.54	0.56	0.76	0.67
9	0.75	0.61	0.03	0.45
10	0.64	0.47	0.76	0.60
11	0.68	0.28	0.51	0.41
12	0.57	0.69	0.92	0.58
13	0.71	0.67	0.86	0.73
14	0.11	0.69	0.22	0.16
15	0.32	0.28	0.26	0.44
16	0.00	0.36	0.31	0.53
17	0.46	0.56	0.22	0.46
18	0.79	0.53	0.18	0.50
19	0.57	0.69	0.50	0.41
20	0.96	0.78	0.83	0.88
21	0.43	0.36	0.67	0.59
22	0.89	0.78	0.79	0.68
23	0.68	0.50	0.60	0.44
24	0.79	0.61	0.68	0.85
25	0.79	0.42	0.65	0.41
26	0.71	0.67	0.72	0.85
27	0.64	0.50	0.53	0.65
28	0.61	0.50	0.69	0.54
29	0.86	0.81	0.82	0.91
30	0.36	0.69	0.44	0.50
31	0.39	0.56	0.39	0.32
32	0.36	0.44	0.39	0.44
33	0.57	0.36	0.43	0.50
34	0.71	0.58	0.68	0.52
35	0.79	0.67	0.74	0.65
36	0.82	0.42	0.24	0.56
37	0.18	0.44	0.35	0.22
38	0.68	0.50	0.36	0.45
39	0.57	0.28	0.68	0.21
40	0.64	0.47	0.19	0.26
41	0.43	0.31	0.22	0.50
42	0.25	0.33	0.24	0.26
43	0.29	0.25	0.40	0.22
44	0.43	0.44	0.42	0.15
Dif.Prom.	0.57	0.54	0.53	0.54

Discriminación de los ítems del examen (rbis)

Item No.	Modelo 1	Modelo 2	Modelo 3	Modelo 4
1	-0.31	0.10	-0.03	-0.06
2	0.21	0.00	0.12	0.16
3	0.28	0.17	0.12	-0.06
4	0.25	-0.10	0.11	0.26
5	-0.36	0.34	0.29	0.20
6	-0.16	0.25	0.13	0.01
7	0.28	0.07	0.06	0.12
8	0.65	0.27	0.07	0.22
9	0.48	0.42	-0.01	0.20
10	0.50	-0.02	0.37	0.34
11	0.30	0.28	0.16	0.27
12	0.55	0.35	0.34	0.24
13	0.31	0.49	0.26	0.11
14	0.10	0.11	0.32	-0.05
15	0.08	0.06	0.18	0.07
16	0.00	0.48	0.35	0.33
17	0.12	0.35	0.14	-0.01
18	0.36	0.50	0.27	0.42
19	0.53	0.50	0.25	0.35
20	0.25	0.15	0.39	0.13
21	0.34	-0.15	0.44	0.34
22	0.35	0.32	0.40	0.30
23	0.40	0.23	0.24	0.14
24	0.62	0.54	0.40	0.32
25	0.36	0.08	0.50	0.12
26	0.62	0.41	0.51	0.34
27	-0.07	0.31	0.31	0.25
28	0.58	0.28	0.36	0.16
29	0.38	0.14	0.27	0.30
30	0.59	0.44	0.46	0.18
31	0.34	0.43	0.20	0.28
32	-0.01	0.30	0.39	0.22
33	0.37	0.03	0.39	0.27
34	-0.12	0.34	0.15	0.25
35	0.42	0.47	0.20	0.28
36	0.30	0.20	0.20	0.41
37	0.03	0.16	0.33	0.04
38	0.58	0.19	0.27	0.06
39	0.41	0.13	0.36	0.21
40	0.16	0.60	0.19	0.24
41	0.05	0.12	0.09	0.06
42	-0.08	-0.03	0.04	-0.10
43	0.17	0.34	0.35	0.23
44	0.14	0.35	0.47	-0.09
Disc Prom	**0.26**	**0.25**	**0.26**	**0.18**
No casos	28	36	65	105
Alfa	**0.80**	**0.79**	**0.80**	**0.70**

Discriminación de los ítems del examen (altos vs bajos)

Item No.	Modelo 1	Modelo 2	Modelo 3	Modelo 4
1	-0.10	0.31	0.08	0.00
2	0.30	0.00	0.12	0.21
3	0.10	0.15	0.12	0.02
4	0.10	-0.08	0.12	0.26
5	-0.40	0.15	0.27	0.16
6	-0.10	0.31	0.04	0.07
7	0.20	0.00	0.15	0.23
8	0.90	0.38	0.08	0.33
9	0.50	0.46	0.04	0.37
10	0.70	0.00	0.35	0.40
11	0.30	0.31	0.12	0.47
12	0.70	0.46	0.19	0.35
13	0.40	0.62	0.27	0.21
14	0.10	0.08	0.27	0.07
15	0.20	0.15	0.23	0.23
16	0.00	0.54	0.46	0.37
17	0.20	0.46	0.08	0.14
18	0.40	0.54	0.15	0.49
19	0.60	0.54	0.31	0.53
20	0.10	0.15	0.23	0.12
21	0.40	0.00	0.50	0.37
22	0.30	0.31	0.35	0.30
23	0.50	0.46	0.46	0.16
24	0.60	0.69	0.46	0.28
25	0.40	0.08	0.62	0.23
26	0.70	0.46	0.58	0.33
27	0.10	0.31	0.35	0.33
28	0.80	0.38	0.38	0.28
29	0.30	0.23	0.38	0.16
30	0.70	0.46	0.54	0.30
31	0.40	0.62	0.35	0.35
32	0.20	0.46	0.54	0.37
33	0.50	0.08	0.50	0.35
34	-0.10	0.31	0.23	0.28
35	0.50	0.62	0.31	0.44
36	0.40	0.31	0.12	0.49
37	0.00	0.31	0.46	0.07
38	0.70	0.23	0.35	0.16
39	0.50	0.15	0.50	0.26
40	0.20	0.77	0.08	0.33
41	0.00	0.08	0.15	0.23
42	-0.10	0.08	0.12	0.00
43	0.20	0.31	0.42	0.23
44	0.20	0.46	0.54	-0.02
Disc Prom	**0.31**	**0.31**	**0.29**	**0.26**

Estándares: Dificultad entre 0.05 - 0.95; Discriminación r >=0.20; Confiabilidad en escala: Alfa >= 0.85

No satisface el estándar -->

Los resultados de los análisis efectuados, que aparecen en la tabla compuesta, permiten hacer varios comentarios evaluativos sobre la información obtenida, mismos que se agruparon en tres categorías: dificultad, discriminación y confiabilidad.

Respecto a la dificultad de los ítems:

- En la parte izquierda de la tabla se muestra el análisis de dificultad de los 44 ítems de cada modelo de examen. Como puede observarse al pie de la tabla, los cuatro modelos presentaron una dificultad media.

- Prácticamente, solo el modelo 1 tuvo ítems que no satisfacen el estándar de calidad especificado para el índice de dificultad (entre 0.05 y 0.95). Los ítems Nº 1 y 16 resultaron muy difíciles y los ítems Nº 3, 4 y 20 demasiado fáciles. En el Modelo 2, se salió del rango especificado el ítem Nº 2, el cual fue tan fácil que lo respondieron correctamente todos los niños. Por su parte, el reactivo Nº 9 del modelo 3 fue excesivamente difícil. Solo el modelo 4 parece no tener problemas con la dificultad, pues todos sus reactivos cayeron dentro del rango apropiado.

- Para los reactivos que no satisficieron el estándar para la dificultad, el dictamen fue claro: pasar de nueva cuenta a una detallada revisión, pero ahora con el valioso apoyo de los datos derivados de la aplicación. Los resultados de dicha tarea se comentan en el siguiente procedimiento.

- Por otra parte, para observar como se distribuyó la dificultad de los ítems en los 4 modelos, se presenta la siguiente figura:

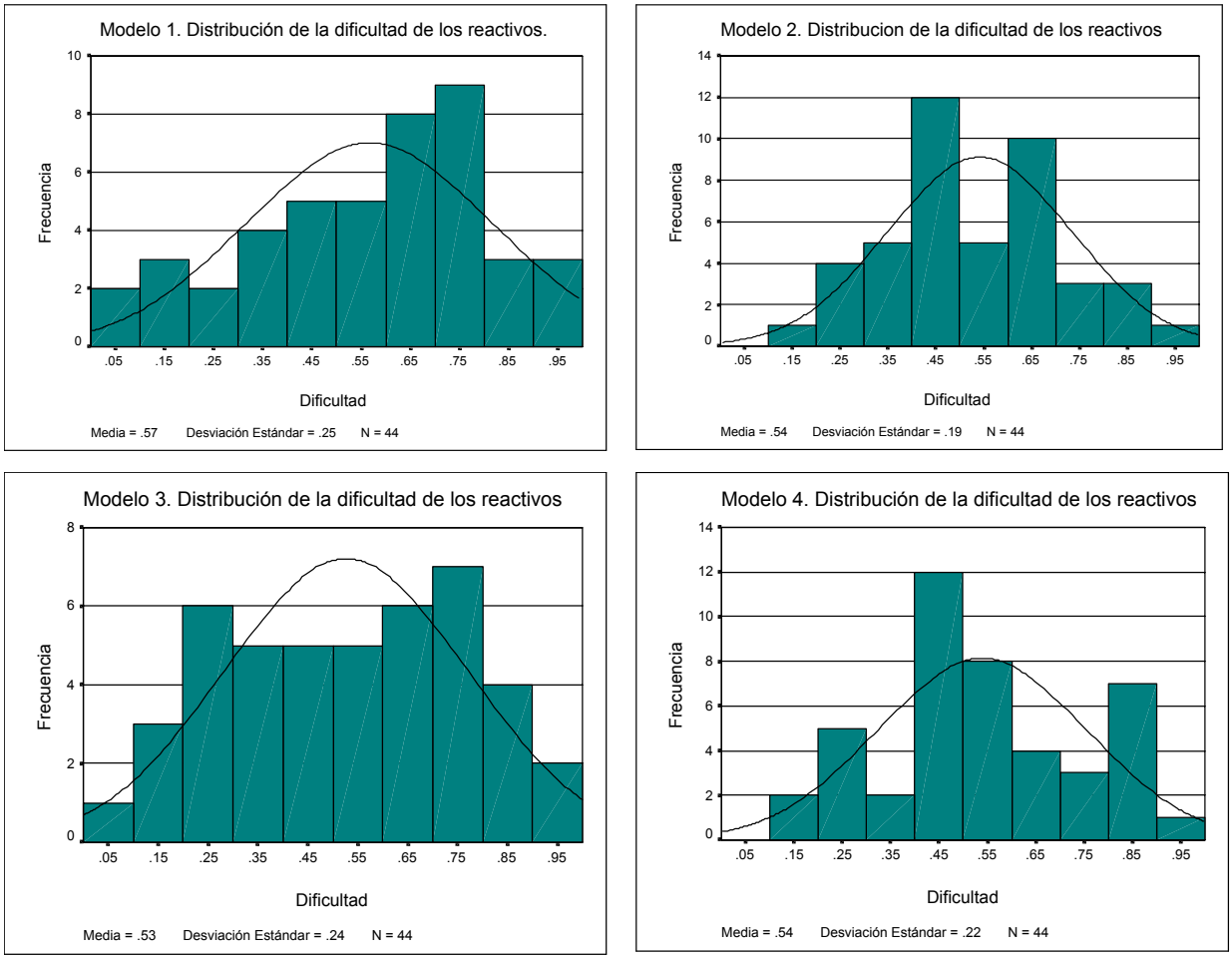

Figura Nº 4. Distribución de la dificultad de los ítems en los 4 modelos de examen

Con excepción del modelo 4, que presentó irregularidades más pronunciadas en su distribución, en los demás modelos la distribución de la dificultad de los ítems se aproximó más a la normal, aunque con diferentes sesgos y kurtosis. Lo anterior no tiene mayor significado en sí mismo, pues un modelo de examen alineado con el curriculum puede ser más o menos difícil que otro y sus ítems presentar una distribución de la dificultad diferente. Sin embargo, para pretender la equivalencia de los modelos de examen se requiere que los indicadores psicométricos, entre ellos la dificultad, tengan un comportamiento equiparable. En ese sentido, los modelos 1 y 3 presentan más similitud en lo que atañe a la distribución de la dificultad de sus ítems.

En conclusión, podemos decir que no se observaron fallas graves en cuanto a la dificultad de los ítems en los modelos, ni en cuanto a la forma en que esta se distribuye en cada modelo. Consecuentemente, la solución de los problemas detectados en este sentido fue bastante ágil, como se verá más adelante en el inciso 7.5.

Respecto a la discriminación de los ítems:

- En el caso de la discriminación se observa un panorama diferente. En la parte central de la tabla compuesta, se muestran los resultados del análisis de discriminación de los reactivos correspondientes a los 4 modelos de examen, efectuado mediante el cálculo de la correlación del punto biserial ($\mathbf{r_{bis}}$) de los reactivos de cada modelo. El análisis muestra que:

  - En el modelo 1, 16 de sus ítems tuvieron un coeficiente de discriminación menor que 0.2; es decir, el 36% de sus reactivos caen por debajo del estándar de calidad mínimo establecido.

  - Por su parte, en el modelo 2, 18 ítems tuvieron un coeficiente menor que 0.2, los cuales representan un 40% de los ítems.

  - Para el caso del modelo 3, el número de ítems cuyo coeficiente de discriminación fue menor que el mínimo aceptable fue 15, lo que significa el 34% de sus reactivos.

  - En cuanto al modelo 4, se presentaron 19 reactivos por debajo del 0.2; es decir, el 43% del total de ítems.

  - El número de ítems, en cada modelo, que presentaron una discriminación negativa fue el siguiente: siete reactivos con valor negativo en el modelo 1; cuatro ítems en el modelo 2; dos reactivos en el modelo 3; y en el caso del

modelo 4, observamos que seis ítems discriminan negativamente. Además, únicamente los modelos 1 y 2 presentaron un caso con 0.0 discriminación, cada uno de ellos.

- Finalmente, ocho ítems en el modelo 1 y 13 ítems en cada uno de los demás modelos, tienen valores que se encuentran entre 0.01 y 0.19; es decir, que discriminan positivamente pero por debajo del estándar especificado.

- Por su parte, a la derecha de la tabla compuesta aparecen los datos del análisis de discriminación de los reactivos correspondientes a los cuatro modelos, realizado mediante el cálculo del índice de discriminación de los reactivos de cada modelo; es decir, la diferencia entre la dificultad del ítem para el subgrupo de examinados que obtuvo las calificaciones altas y para el subgrupo con las calificaciones más bajas.

  El análisis de la información obtenida muestra que:

  - En el modelo 1, 13 de sus ítems tuvieron un índice de discriminación menor que 0.2; es decir, casi el 30% de sus reactivos caen por debajo del estándar de calidad mínimo establecido.

  - El modelo 2, tuvo 15 ítems con un índice menor que 0.2, mismos que representan un 34% de los ítems.

  - En el modelo 3, el número de ítems cuyo índice de discriminación fue menor que el mínimo aceptable fue 16, lo que significa el 36% de sus reactivos.

  - En cuanto al modelo 4, se presentaron 13 reactivos con un índice por debajo del 0.2; es decir, casi el 30% del total de ítems.

- El número de ítems, en cada modelo, que presentaron una discriminación negativa fue el siguiente: cinco reactivos con valor negativo en el modelo 1; un ítem en el modelo 2; ningún reactivo en el modelo 3; y en el modelo 4, observamos que un ítem discrimina negativamente. Además, el número de ítems que tienen un índice de discriminación de 0.0 en los modelos 1, 2, 3 y 4 es de tres, cuatro, cero y dos, respectivamente.

- Finalmente, cinco ítems en el modelo 1; 10 ítems en el modelo 2; 16 ítems en el modelo 3; y 10 ítems en el modelo 4, tuvieron índices de discriminación que se encuentran entre 0.01 y 0.19; es decir, que discriminan positivamente pero que no alcanzaron el estándar especificado.

- Al comparar los resultados del análisis discriminativo de los ítems, en cada modelo del examen, obtenidos mediante ambos métodos (r_{bis} vs altos-bajos), puede decirse que la correlación del punto biserial es en general un procedimiento más "duro", lo cual es producto de comparar la respuesta al ítem de todos los examinados y no solo la de los grupos extremos contrastados. Lo anterior puede observarse en la tabla siguiente:

Tabla Nº 14. Comparación de los resultados obtenidos mediante los métodos r_{bis} y alto-bajo

Modelo	Porcentaje de ítems que satisfacen el estándar		Número de ítems que no satisface el estándar, por categoría discriminativa					
			Discriminación = 0.0		Discriminación negativa		Discriminación entre 0.01 y 0.19	
	r_{bis}	Alto -bajo	r_{bis}	Alto -bajo	r_{bis}	Alto -bajo	r_{bis}	Alto -bajo
1	64	70	1	3	7	5	8	5
2	60	66	1	4	4	1	13	10
3	66	64	0	0	2	0	13	16
4	57	70	0	2	6	1	13	10

Sin embargo, la diferencia no parece ser tan determinante cuando se observa el perfil discriminativo que producen todos los ítems en cada modelo. Para ilustrar este efecto, obsérvese la siguiente figura que muestra la discriminación de los ítems del modelo 1, obtenida mediante ambos métodos:

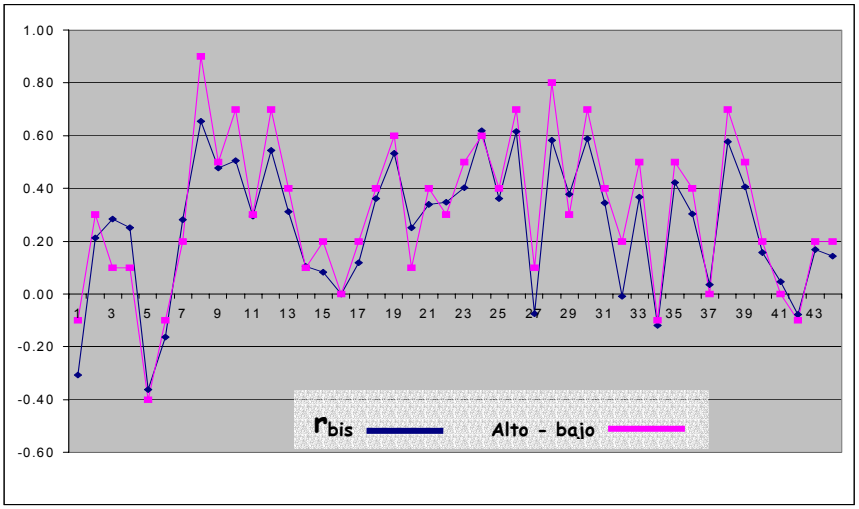

Figura Nº 5. Perfil de discriminación de los ítems del modelo 1, obtenido mediante los Métodos de correlación biserial (r_{bis}) y el de grupos extremos contrastados (alto y bajo)

Con las diferencias ya comentadas, en ambos casos el perfil general producido es similar; aunque se ve que el índice de discriminación tiene una ligera tendencia a beneficiar con valores más altos a los ítems del modelo.

- Para observar también como se distribuyó la discriminación de los ítems en los 4 modelos, en la página siguiente se presenta la correspondiente gráfica compuesta:

En este caso las diferencias son menos pronunciadas y, en general, en los 4 modelos la distribución de la discriminación de los ítems se aproximó más a la normal, con similares sesgos pero con diferente kurtosis. Considerando la posible equivalencia entre los modelos de examen, podemos decir que según este indicador psicométrico las distribuciones de los modelos 3 y 4 lucen más parecidas entre sí que con los otros modelos.

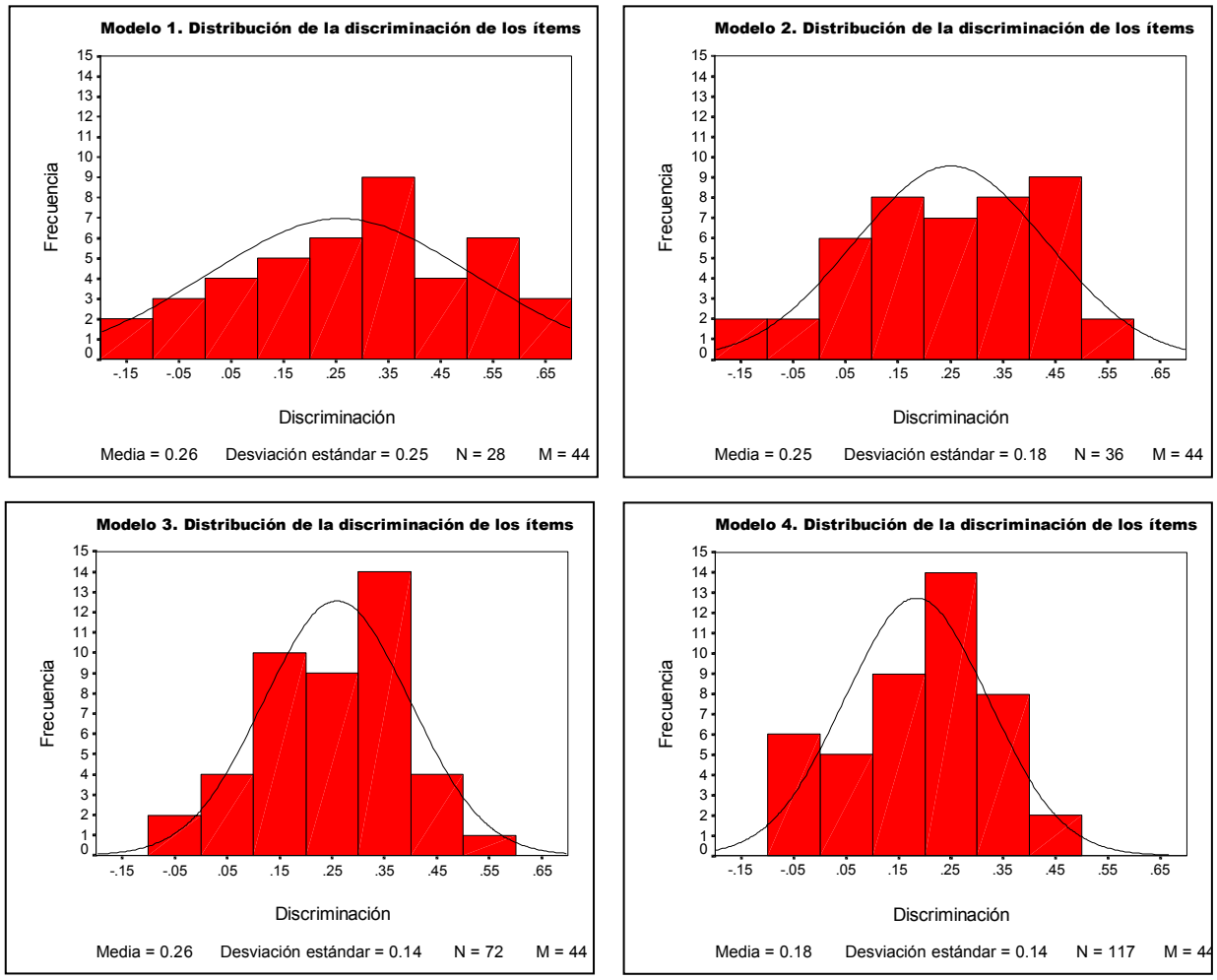

Figura Nº 6. Distribución del índice de discriminación de los ítems en los 4 modelos de examen

De hecho, puede decirse que un valor psicométrico clave de la discriminación radica en ponerle una bandera al ítem defectuoso. En particular, el coeficiente respectivo (o índice, en su caso) proporciona una buena orientación sobre el origen o el sentido de la falla; por ello, su principal ventaja se reveló durante el procedimiento de revisión de los ítems.

- No obstante, lo dicho hasta el momento en esta sección no oculta el hecho de que en todos los modelos de examen alrededor de un tercio de los ítems, un número considerable, presentaron fallas en cuanto a su poder discriminativo y requirieron de

una cuidadosa revisión para ser modificados o eliminados, como se verá más adelante.

- **Respecto a la confiabilidad de los modelos**

  En la parte inferior de la tabla compuesta (tabla Nº 13), se presenta también el valor de alfa que obtuvieron los modelos de examen; es decir, su coeficiente de confiabilidad. Como puede apreciarse, los 4 modelos quedaron cortos respecto al estándar de calidad definido para este indicador psicométrico, el cual establecía un mínimo de 0.85. Los modelos 1 y 3 son los que más se acercan al criterio con 0.80 de confiabilidad, cada uno de ellos; el modelo 2 obtuvo un coeficiente de 0.79 y el más alejado del estándar es el modelo 4, con 0.70 de confiabilidad.

  Se sabe que la confiabilidad se mejora incrementando el número de ítems de un examen. Sin embargo, esa no parece ser la solución en este caso. Sin ser muy grave la situación, más bien los problemas de confiabilidad están relacionados con las fallas del poder discriminativo de los ítems anotadas en el punto anterior, las cuales tendrían que ser resueltas primero.

Para tener una visión global de los resultados obtenidos, a continuación se presenta un resumen de la información derivada del análisis de ítems y modelos que se presentó en los puntos anteriores.

Tabla Nº 15. Resumen de la información derivada del ítem análisis

Modelo	Porcentaje de ítems con dificultad apropiada	Porcentaje de ítems con discriminación apropiada (r_{bis})	Porcentaje de ítems con discriminación apropiada (altos vs bajos)	Confiabilidad (alfa)
1	89	64	70	0.80
2	98	60	66	0.79
3	98	66	64	0.80
4	100	57	70	0.70

Como puede observarse, tras la primera aplicación del examen el análisis mostró que tanto los reactivos, como los modelos el examen, aún no alcanzan los niveles de calidad requeridos para su uso extensivo como instrumento para monitorear el aprendizaje que logran los niños en el área de español de la educación primaria. El punto fuerte del examen es que, con pocas excepciones, los ítems tuvieron el nivel de dificultad apropiado para los examinados. En cambio, sus puntos débiles consisten en que cerca de un tercio de las preguntas, en cada modelo, no discriminaron apropiadamente entre los examinados que dominan los contenidos respectivos y quienes no lo hacen; y en que los modelos mismos no obtienen los puntajes del test de manera suficientemente consistente. Desde luego, hasta este punto, los datos mostraron que no se está lejos de lograr los estándares especificados en esos rubros. Justamente, en esa dirección estuvieron encaminados los esfuerzos en el siguiente procedimiento.

7.5 Revisión de Ítems y estructuración de la prueba.

Con base en los resultados del análisis estadístico de los ítems, el **Comité Diseñador** y el **Comité Elaborador** revisaron cuidadosamente los ítems que no cumplieron con los estándares de calidad establecidos, a fin de determinar sus fallas y decidir cuáles de ellos podían ser corregidos y cuáles deberían ser reelaborados.

Para la revisión de cada ítem se tomó en consideración, además de los índices de dificultad y discriminación y del coeficiente de discriminación respectivos, el análisis de los distractores. Es decir, de manera complementaria se observó la ejecución de los niños ante cada una de las opciones de respuesta. De esta manera, pudieron ser identificadas varias clases de fallas en los ítems, respecto a las cuales se adoptaron las siguientes decisiones: su conservación, modificación o eliminación. Los tipos de fallas más comunes que fueron detectadas y las decisiones adoptadas ante ellas, se describen en la tabla siguiente:

Tabla Nº 16. Taxonomía de fallas más comunes y decisiones adoptadas para el mejoramiento de los ítems

Tipo de falla	Ejemplos	Tipo de decisión adoptada para el mejoramiento
Complejidad cognitiva	Muy difícil	Cuidadoso análisis del contenido del ítem y, en su caso, redacción más categórica, hacer más o menos atractivas las opciones o sustituirlas
	Muy fácil	
Discriminación errónea	Discriminación negativa	Cambiar la opción que atrae a los que saben, cambiar distractores poco elegidos
	Discriminación baja	Cambiar la opción que parece esta creando el problema
Edición	Escritura confusa	Corregir errores mecanográficos, hacer dibujos más claros
Redacción	Conceptos complejos	Simplificar las conceptualizaciones, cambiar la opción confusa
Mixta	Respuesta al azar	Hacer más categóricas la base y la respuesta correcta, o sustituir distractores

De hecho, para revisar cada ítem y con ello detectar las posibles fallas que tenía a fin de proceder posteriormente a su correspondiente modificación, se efectuó una especie de análisis estructural cuyos elementos fueron las evidencias disponibles a partir del análisis de reactivos y del análisis de los distractores. A continuación se presenta un diagrama que ilustra el proceso general que se siguió para analizar y modificar el ítem Nº 20 del modelo 4:

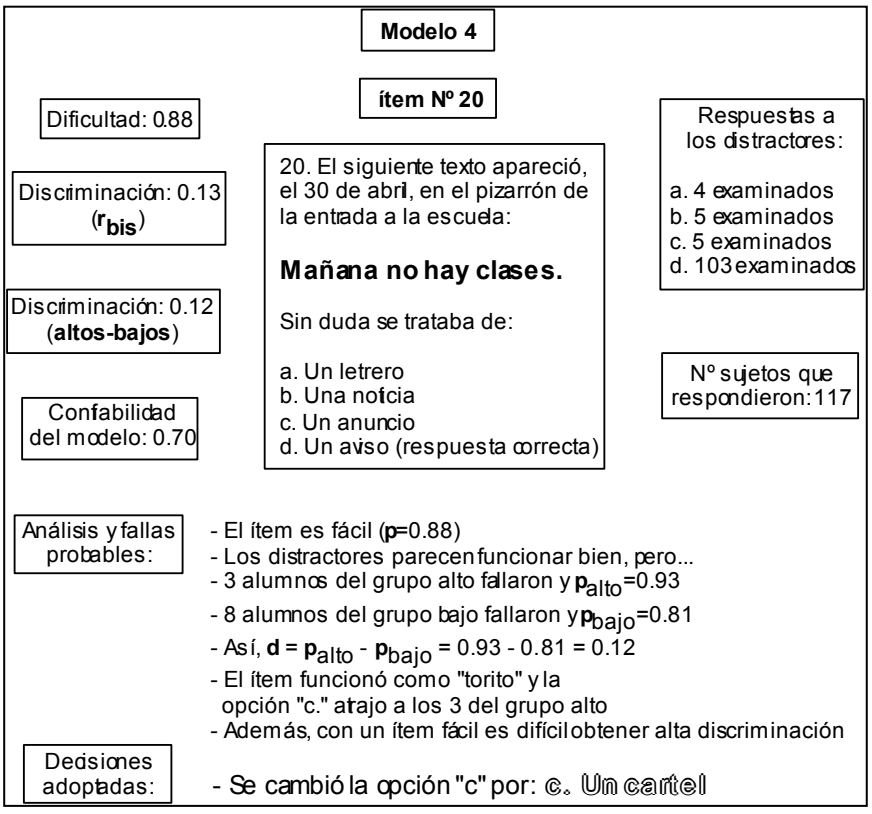

Figura Nº 7. Ilustración del proceso para analizar y modificar un ítem

La idea principal al presentar el diagrama es mostrar el poder que tiene el uso conjunto de los indicadores psicométricos para detectar y corregir fallas en los ítems y modelos de examen.

Una vez corregidos o reelaborados los ítems que presentaron fallas en cada modelo, el **Comité Coordinador** procedió a estructurar una muestra de ítems que fuera representativa del dominio curricular del área de español; es decir, conformó de nueva cuenta los modelos de examen de tal manera que, mediante una especificación y un muestreo adecuados, mostraran los resultados del curriculum con un peso y una proporción apropiados.

Con esta última acción, se dieron por concluidos el diseño, la elaboración y el pilotaje de la prueba y, con ello, se cumplen los propósitos del presente trabajo. Sin embargo, de conformidad con la metodología propuesta en el capítulo 3, aún faltan por desarrollarse fases y procedimientos que resultan necesarios para pasar el instrumento al nivel de gran escala y probarlo en esa dimensión a fin de garantizar su calidad integral para monitorear permanentemente el aprendizaje de los niños. Tales acciones se describen más adelante en el capítulo de conclusiones.

Capítulo 7. Producción y validación de ítems

Una vez elaboradas las especificaciones de ítems y habiendo sido entrenado el grupo de elaboradores de reactivos, en esta etapa se procedió a desarrollar los reactivos, analizarlos a la luz de las especificaciones elaboradas, probarlos empíricamente ante una muestra de alumnos que estaban por egresar de escuelas primarias, efectuar un análisis de las respuestas a los ítems obtenidas mediante su aplicación, así como revisar los ítems y estructurar la versión final la prueba. Estos cinco procedimientos se describen a continuación, así como los resultados obtenidos al operarlos.

7.1 Elaboración de ítems.

Con base en el manual de especificaciones y en el entrenamiento recibido, el **Comité Elaborador** desarrolló un conjunto de 200 ítems para la prueba de conformidad con las normas, a fin de propiciar su validez. La distribución de los ítems entre los elaboradores no fue equitativa; se dejó que ellos se dividieran el trabajo en función de su especialidad, su interés y la disponibilidad de tiempo para la tarea. Por los resultados obtenidos tal mecánica fue adecuada, con alguna excepción.

En general, los elaboradores diseñaron los ítems que les correspondió trabajar de manera independiente o en pequeños grupos y posteriormente los aplicaron a sus alumnos o a niños de otra escuela, a fin de efectuar una calibración inicial de las preguntas y retroalimentar así su trabajo. Lo anterior no pudo realizarse en todos los casos, particularmente no se realizó con los ítems correspondientes al eje de Recreación Literaria, debido a que en su mayoría contienen dibujos y ello retrasó su elaboración, la cual concluyó poco antes de la aplicación.

Durante este procedimiento se presentaron diferentes tipos de interacciones entre los miembros de los comités elaborador y coordinador, siendo las más comunes las que

tenían por objeto hacer aclaraciones y retroalimentar las actividades. Al finalizar la elaboración, los responsables entregaron los ítems que diseñaron, junto con los resultados de la pequeña prueba empírica que realizaron.

Algunos elaboradores entregaron más ítems de los que les correspondió elaborar, exceso que fue producto del uso del método de generación de ítems que emplearon, principalmente el de transformaciones lingüísticas (mismo que se describe en la parte final del anexo Nº 1).

En general, la elaboración de los ítems duró unos 15 días y en los casos en que los reactivos contenían dibujos, se prolongó una semana más.

Gracias al apoyo financiero que otorgó al proyecto el Programa Interinstitucional de Investigaciones sobre la Educación Superior (PIIES) de la SEP, fue posible compensar parcialmente el esfuerzo que realizaron los elaboradores, a quienes se pagó una pequeña cantidad por ítem limpio; es decir, por cada reactivo elaborado de conformidad con las normas, dictaminado favorablemente por el Comité Diseñador que analizó su congruencia con la especificación correspondiente, analizado empíricamente y, en su caso, corregido de las fallas detectadas mediante dichos procedimientos.

7.2 Revisión formal de la congruencia ítem-especificación.

Una vez elaborados, los reactivos fueron sometidos a una detallada revisión de contenido, psicométrica y lógica, contra las especificaciones de ítems correspondientes (las que aparecen en el anexo Nº 6). Además, para la revisión se consideraron los estándares, medidas y criterios correspondientes a la calidad del contenido de los ítems del test y a la calidad técnica de cada ítem del test, que fueron definidos previamente en la tabla Nº 9 que se presentó en el capítulo 3. Esta tarea fue realizada por el **Comité Diseñador,** y tuvo como propósito garantizar la calidad de la relación:

ítem *representa* **curriculum** y, con ello, la validez de los ítems. El procedimiento de revisión operó de la manera que se ilustra en el siguiente diagrama de flujo:

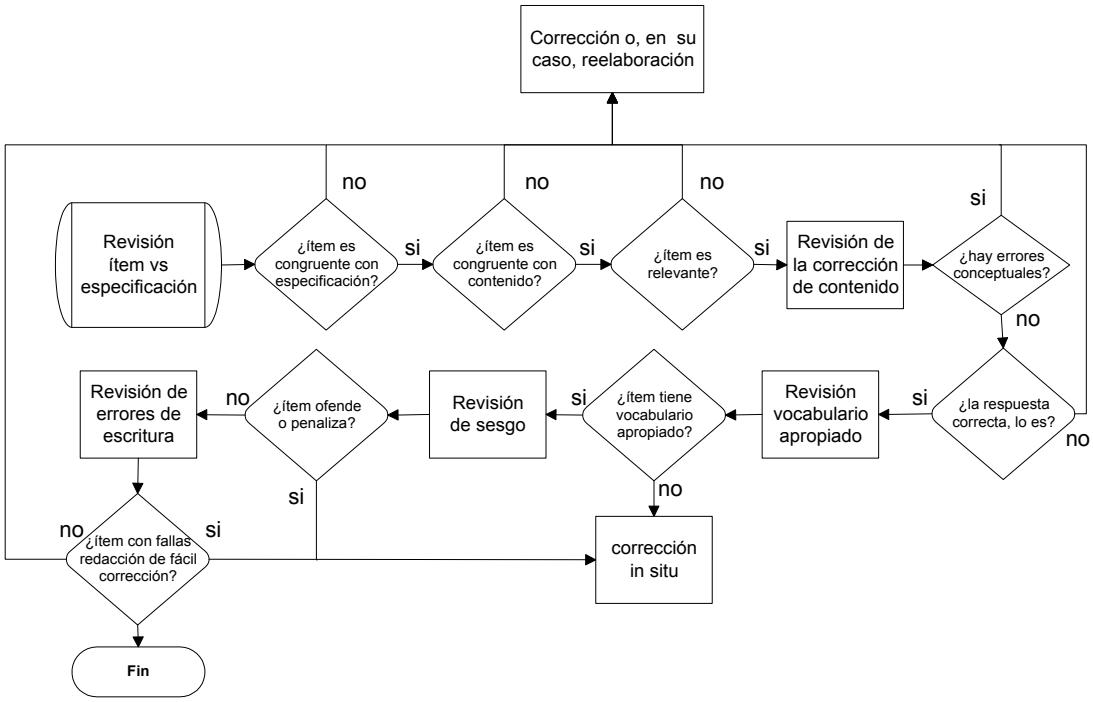

Figura Nº 3 Diagrama de flujo del proceso de revisión formal de los ítems de la prueba

A continuación se describe el proceso ilustrado en el diagrama:

- Inicialmente, cada ítem fue contrastado con la especificación que lo produjo. Para ello el Comité Diseñador, constituido en un panel de expertos, dictaminó si el ítem era congruente con la especificación y con el contenido correspondiente y si además resultaba relevante. También en esta ocasión las decisiones del comité se adoptaron de manera consensual. Cuando el dictamen resultó favorable en los tres casos, se continuó con el análisis; cuando no lo fue, se envió el ítem a corrección o reelaboración según fuera la naturaleza de las observaciones.

- Posteriormente, se revisó la corrección del contenido del ítem; es decir, que

estuviera exento de errores conceptuales y que la respuesta correcta lo fuera. Cuando el dictamen fue favorable en ambos casos, se continuó el proceso y en los casos en que no lo fue se indicó lo procedente.

- Enseguida, se revisó que el ítem presentara un vocabulario apropiado al nivel de los examinados. Cuando no lo tuvo, se hicieron las correcciones necesarias ahí mismo.

- Además, se revisó que el ítem no presentara estereotipos étnicos o de género y que no pusiera en ventaja a un grupo sobre otros. Lo mismo se hizo posteriormente con el conjunto de ítems que formaron cada modelo. En ningún caso se juzgó que estuviera presente algún tipo de sesgo.

- Finalmente, se revisó que el ítem estuviera exento de errores de escritura, tanto técnica como de tipo mecanográfico. En los casos en que se presentaron determinantes específicos, opciones poco plausibles y otras fallas similares, se efectuaron las correcciones en el momento. Cuando no pudo hacerse de inmediato y según la naturaleza de la falla detectada, la corrección se dejó a cargo de una subcomisión y su dictamen quedó pendiente para la siguiente sesión.

En general, el procedimiento de revisión resultó bastante fluido. En parte, debido a que la mecánica de trabajo consideró como un requisito para poder efectuar una reunión la revisión previa, por parte de cada miembro del comité, de todos los reactivos a evaluar en cada sesión. Parcialmente también, la fluidez de los trabajos se debió al reducido número de miembros del comité que continuaron y concluyeron estos trabajos, mismos que al final fueron tres. Esto último, explica además porqué se decidió adoptar al consenso como criterio para decidir y porqué el procedimiento se prolongó considerablemente.

Al igual que los elaboradores de ítems, por esta y las demás actividades que realizaron durante el complejo diseño de la prueba, los miembros del Comité Diseñador recibieron

una modesta compensación económica, que fue posible otorgarles gracias al apoyo del PIIES.

7.3 Ensayo empírico de los ítems.

Además de la aplicación del examen, este procedimiento incluyó tres acciones previas: estructurar los modelos de examen, seleccionar la muestra de examinados y capacitar a los aplicadores del examen, mismas que se describen a continuación:

Estructuración de los modelos de examen.

Con el propósito de efectuar una primera calibración formal de la calidad técnica de los ítems y modelos, en cuanto a su nivel de legibilidad (redacción clara para los sujetos, vocabulario que corresponde a su nivel y conceptualizaciones acordes con su desarrollo), dificultad, discriminación, funcionamiento diferencial ante los sujetos y confiabilidad, el **Comité Coordinador** estructuró los reactivos revisados, de conformidad con la tabla de especificaciones, en cuatro modelos de examen con 45 ítems cada uno de ellos.

En términos generales, las siguientes características están presentes en cada modelo:

- Al inicio del examen se presenta una página que describe a los niños el propósito del examen, seguido de instrucciones generales para responder y de un ejercicio de práctica que supervisa el aplicador.

- En la segunda página se presenta el reactivo Nº 1 el cual requiere, para ser respondido por los examinados, que el aplicador lea previamente en voz alta un pequeño cuento. En cada una de las 21 páginas que tiene cada modelo, aparecen impresos el mismo número y tipo de reactivos que en esa misma

página tienen los demás modelos. La idea es que, a partir de los modelos originales, puedan estructurarse y ponerse a prueba psicométrica diferentes formas paralelas del examen.

- Al inicio de cada ítem se incluyen las indicaciones necesarias para responderlo.

Con el propósito de ilustrar la configuración final que se dio a cada modelo, en el anexo Nº 8 se presenta el modelo de examen 4 completo; el mismo que se estructuró con los reactivos que sirvieron de muestra en las especificaciones de ítems. Los demás modelos no se incluyen, debido a la necesidad de preservar la integridad del instrumento y con ello su validez para ser empleado posteriormente de manera efectiva.

Como ya se explicó al final del capítulo anterior, además de los cuatro modelos de examen constituidos por reactivos que miden el logro académico de los niños, se estructuró un cuestionario para explorar la opinión de los examinados sobre la realización de ciertas actividades curriculares que establecen los programas de estudio para los ejes de Lengua Hablada y Lengua Escrita. Para estructurar el cuestionario no se elaboraron nuevos ítems, sino que se integró el instrumento con los ítems muestra de las especificaciones correspondientes a dichos ejes curriculares. Tales especificaciones se muestran al final del anexo Nº 6, y el cuestionario de opinión que se aplicó a los niños se presenta en el anexo Nº 9.

Los cuatro modelos de examen fueron aplicados a una muestra inicial de 253 niños, de 12 escuelas, que en el momento de la aplicación estaban por egresar de la educación primaria. Por su parte, el cuestionario de opinión fue aplicado a 90 niños, de 6 escuelas, entre los que se encontraban buena parte de los que respondieron el examen y quienes lo contestaron al terminar de ser examinados; además, se aplicó a otros niños que no fueron examinados y que solo respondieron las preguntas del cuestionario de opinión.

Obtención de la muestra de sujetos a quienes se aplicó el examen.

Previa a la aplicación, fue seleccionada una muestra intencional de niños para ser examinados, la cual presentó las siguientes características:

- Alumnos que se encontraban cursando el 6º grado de primaria en el momento de aplicarse la prueba.

- Con el propósito de explorar el comportamiento del examen ante poblaciones diferentes, así como la ejecución de ellas en el test, se trató de que en conjunto los niños presentaran una variedad de condiciones en cuanto a:

    - Tipo de escuela (pública, privada)
    - Diversidad delegacional (en Ensenada)
    - Tipo de población (urbana, urbana marginal, rural)
    - Nivel sociocultural (alto, medio, bajo), según la valoración del director de la escuela, del inspector escolar y del aplicador del examen
    - Turno (matutino, vespertino)
    - Lugar de origen (BC, estado fronterizo, centro del país, sureste del país)
    - Sexo (masculino, femenino)
    - Edad (hasta 12 años, 13-14 años, 15 o más años)

En general, estas condiciones quedaron representadas de manera apropiada en la muestra, de tal manera que los examinados presentaron variedad en cuanto a las características requeridas para pilotear los modelos de examen y calibrar los ítems. Así, participaron en el examen escuelas de cuatro delegaciones del municipio de Ensenada; siete fueron urbanas, tres urbanas marginales y dos rurales; diez de ellas son públicas y dos privadas; diez ofrecen el servicio en el turno matutino y 2 de ellas en el vespertino; y una tiene alumnos que típicamente pertenecen al nivel socioeconómico

alto, otra más al medio alto, siete al nivel medio y 3 de ellas al bajo.

Cabe señalar que el examen solo pudo ser aplicado en el municipio de Ensenada debido a que las pláticas que se habían sostenido hasta el momento con las autoridades educativas del estado de Baja California, no culminaron oportunamente en el convenio necesario para una aplicación a escala estatal, a pesar del interés que habían manifestado. No obstante, con el generoso apoyo que otorgaron en esta y otras etapas del proceso las autoridades educativas del municipio de Ensenada, fue posible efectuar la aplicación del examen en escuelas de dos inspecciones escolares, casi al término del ciclo escolar 1998-1999.

Como resultado de lo anterior, tras aplicar el examen fue posible obtener una caracterización de la población que fue examinada, misma que se presenta en la tabla Nº 11, que aparece en la página siguiente.

Entre los principales aspectos que destacan en la tabla, podemos mencionar que:

- En general, las condiciones que parecen caracterizar a los examinados son: que se trata de estudiantes del turno matutino de escuelas oficiales de la zona urbana de Ensenada; que pertenecen a los niveles socioeconómicos medio y bajo; que tienen alrededor de 12 años de edad; de ambos sexos por igual; nacidos principalmente en Baja California; cuyas madres estudiaron hasta la secundaria y sus padres son empleados; que aspiran a estudiar en la universidad y que viven en casa propia, lo cual no es extraño en esta región del país.

- Al parecer, los datos de la tabla coinciden con los reportados en muchos otros trabajos que exploran los factores sociales y económicos que intervienen en la educación de los niños: qué quienes pertenecen a niveles socioeconómicos bajos tienen madres con baja escolaridad, menos aspiración por estudios

Tabla Nº 11. Descripción de la población estudiada en el municipio de Ensenada.

Condición		Nº de alumnos por escuela													Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	
Delegación	Ciudad	23	36	16	-	-	-	24	30	-	21	30	25	15	220
	San Antonio	-	-	-	3	-	-	-	-	-	-	-	-	-	3
	Otra	-	-	-	-	13	17	-	-	-	-	-	-	-	30
Tipo de población	Urbana	23	-	16	-	-	-	24	-	-	21	30	25	15	154
	Urbana marginal	-	36	-	-	13	-	-	30	-	-	-	-	-	79
	Rural	-	-	-	3	-	17	-	-	-	-	-	-	-	20
Tipo de escuela	Pública	23	36	-	3	13	17	24	30	-	21	30	25	-	222
	Privada	-	-	16	-	-	-	-	-	-	-	-	-	15	31
Turno	Matutino	23	36	16	3	-	17	24	30	-	-	30	25	15	219
	Vespertino	-	-	-	-	13	-	-	-	-	21	-	-	-	34
Nivel socioeconómico	Alto	-	-	-	-	-	-	-	-	-	-	-	-	15	15
	Medio alto	-	-	16	-	-	-	-	-	-	-	-	-	-	16
	Medio	23	-	-	-	13	-	24	30	-	21	30	25	-	166
	Bajo	-	36	-	3	-	17	-	-	-	-	-	-	-	56
Edad	10 a 12	19	32	14	2	7	13	23	25	-	10	30	13	14	202
	13 a 14	4	4	2	-	4	3	1	4	-	10	-	12	1	45
	15 a 20	-	-	-	1	2	1	-	1	-	1	-	-	-	6
Sexo	Femenino	10	19	7	2	7	8	9	20	-	13	13	14	8	130
	Masculino	13	17	9	1	6	9	15	10	-	8	17	11	7	123
Ciudad de origen	Sureste del país	3	-	-	-	1	-	-	2	-	-	-	-	-	6
	Centro del país	1	4	4	2	2	6	3	11	-	4	4	4	2	47
	Fronteriza	1	1	1	-	2	1	2	-	-	5	3	2	3	21
	Baja California	18	31	11	1	8	10	19	17	-	12	23	19	10	179
Años de vivir en Ensenada	Menos de 5	1	2	2	-	5	-	1	5	-	2	4	-	3	25
	Entre 5 y 10	2	8	3	1	2	3	4	9	-	3	10	7	3	55
	Más de 10	20	26	11	2	6	14	19	16	-	16	16	18	9	173
Escolaridad máxima de la madre	Primaria	8	3	-	2	8	8	2	4	-	3	3	-	-	41
	Secundaria	4	8	-	-	3	4	4	6	-	10	5	6	-	50
	Bachillerato	3	6	2	-	2	2	3	4	-	3	1	7	-	33
	Técnica	-	3	-	-	-	-	1	1	-	2	-	-	-	7
	Normalista	-	5	-	-	-	-	1	-	-	-	3	2	2	13
	Universidad	3	1	13	-	-	-	3	1	-	-	13	4	2	40
	No sabe	5	10	1	1	-	3	10	14	-	3	5	6	11	69
Profesión del padre	Obrero o campesino	1	-	-	1	-	3	3	3	-	3	1	-	-	15
	Empleado de oficina	5	31	1	1	1	9	4	13	-	6	7	9	6	93
	Comerciante	4	3	2	-	-	1	1	2	-	1	2	2	4	22
	Técnico	-	2	-	-	3	-	1	3	-	-	4	7	-	20
	Otra	10	-	13	-	7	1	11	6	-	7	13	4	5	77
	No sabe	3	-	-	1	2	3	4	3	-	4	3	3	-	26
Aspiración máxima de estudios	Secundaria	1	-	-	1	-	5	3	-	-	1	-	-	-	11
	Bachillerato	2	-	-	-	1	2	1	5	-	1	-	1	-	13
	Carrera corta	2	3	-	1	2	3	2	4	-	5	-	1	-	23
	Universidad	18	33	16	1	10	7	17	21	-	14	30	23	15	205
Vivienda	Propia	21	34	15	2	9	16	20	27	-	15	23	23	12	217
	Rentada	2	1	1	-	4	1	3	1	-	4	6	1	2	26
	Prestada	-	-	-	-	-	-	1	2	-	2	1	-	1	7
	Otra	-	-	-	1	-	-	-	-	-	-	-	-	-	1
	No sabe	-	1	-	-	-	-	-	-	-	-	-	1	-	2

superiores, tienen padres que son obreros, campesinos o empleados, y viven en zonas rurales o urbanas marginales. Por lo contrario, los estudiantes de las dos escuelas privadas, que fueron clasificadas como pertenecientes a los niveles socioeconómicos medio alto y alto, tienen madres con alta escolaridad y padres comerciantes o empleados de oficina y todos aspiran a estudiar en la universidad.

- Cuando se contrastan estas condiciones con los aciertos que obtuvieron los niños en el examen, los resultados son similares. Lo anterior puede observarse en la tabla que se muestra enseguida, la cual presenta únicamente aquellas correlaciones entre las condiciones de la muestra y el número de aciertos en el examen, que resultaron significativas estadísticamente:

Tabla Nº 12. Correlaciones entre aciertos en el examen y las condiciones de la muestra de examinados, que son significativas estadísticamente.

Condición	Nivel de significancia de la relación	
	al 0.05	al 0.01
Delegación		.253
Tipo de escuela		.342
Turno		.220
Nivel socioeconómico		.226
Escolaridad de la madre		.209
Aspiración máxima de estudios	.158	

Así, en términos de correlación, presentan más aciertos de manera significativa (en los niveles de 0.05 y 0.01, respectivamente) quienes estudiaron en la zona típicamente urbana de Ensenada, en escuelas privadas, en el turno matutino, pertenecen a los niveles socioeconómicos alto y medio alto, tienen madres con mejor escolaridad y aspiran a estudios más altos.

- Por razones que se comentan más adelante, en la escuela Nº 9 no fue posible aplicar el examen. Así, no aparecen los datos correspondientes a esos estudiantes en la tabla Nº 13 y en las demás tablas que se presentarán posteriormente.

Selección y capacitación del Comité Aplicador del examen.

Para aplicar el examen, fue seleccionado un Comité Aplicador integrado por cuatro personas, quienes fueron entrenadas para operar los modelos bajo condiciones estandarizadas, particularmente las relativas a la organización de las actividades, el manejo confidencial de los ejemplares de examen y hojas de respuesta, así como la impartición de instrucciones a los niños, el seguimiento de su ejecución en el ensayo de práctica para responder y la solución a sus dudas sobre cómo contestar la prueba. La capacitación se proporcionó en dos sesiones de trabajo y como material de apoyo fue elaborado un manual, mismo que aparece en el anexo Nº 7.

En general, el instructivo para la aplicación y los materiales contenidos en el manual resultaron funcionales, y la participación de los aplicadores fue la adecuada. Sin embargo, ciertos detalles de coordinación causaron algunos problemas. Por ejemplo, uno de los mapas con la ubicación de las escuelas causó dificultades para llegar a un plantel rural y se retrasó la aplicación; en otro caso, la prueba no pudo aplicarse en la mencionada escuela Nº 9 debido a que previamente había sido programada una salida de los alumnos, la cual coincidió con la fecha prevista para aplicación del examen en el plantel.

Como puede observarse, además de dar comienzo a la fase empírica del proceso de calibración de la calidad técnica de los ítems y modelos de examen, este procedimiento tuvo como propósito adicional explorar inicialmente las condiciones necesarias para estandarizar su aplicación.

7.4 Análisis estadístico de los ítems.

Una vez aplicados los modelos de examen, el **Comité Coordinador** procedió a capturar los resultados y a efectuar un análisis de ítems y de confiabilidad de los

modelos, para estimar su calidad técnica. En el primer caso, las hojas de respuesta fueron leídas mediante un *scanner* de alta velocidad y la información se registró en una base de datos. Para el segundo caso, se obtuvieron índices de dificultad apropiada para cada ítem y modelo mediante tres procedimientos:

- Se calculó el *índice de dificultad* o valor ***p*** del reactivo. Es decir, la proporción de examinados que contestaron correctamente el ítem. Para ello se empleó la fórmula:

$$p = c / t,$$

Donde,
c = número de examinados que tuvieron correcto el ítem
t = total de examinados que respondieron el ítem

El valor **p** fue obtenido de la muestra de ensayo y, para ser apropiado, se consideró que debería ser mayor que 0.05 y menor que 0.95. Tanto en este caso, como en los dos siguientes, el análisis se orientó por los criterios, medidas y estándares que fueron adoptados para controlar la calidad para la prueba y que aparecen en la tabla Nº 9, del capítulo 3.

- Se obtuvo el *índice de discriminación* o valor ***D*** del ítem. Es decir, la diferencia entre ***p*** para el grupo alto (27%) y ***p*** para el grupo bajo (27%). Para calcularlo se utilizó la fórmula:

$$D = pa - pb$$

Donde,
pa = dificultad del ítem para el grupo alto; es decir, el grupo de examinados que obtuvieron las calificaciones más altas en el examen (el 27% del total de examinados).

Pb = dificultad del ítem para el grupo bajo; es decir, el grupo de examinados que obtuvieron las calificaciones más bajas en el examen (el 27% del total de examinados).

El valor discriminativo del ítem se consideró apropiado si ***D*** >= 0.2. Además, se obtuvo el *coeficiente de discriminación*, mediante la correlación del punto biserial; es decir, la relación entre las respuestas al reactivo (0 o 1) y las calificaciones en el test de todos los examinados y no solo del 54% de ellos.

- Para estimar la calidad de los puntajes del test, o sea su calidad integral como instrumento de medida, fue estimada la *confiabilidad* mediante el *coeficiente de consistencia interna* de cada modelo de examen; es decir, el coeficiente *alfa de Cronbach* o *kuder-Richardson 20*, que puede ser considerado como la correlación promedio obtenida de todas las posibles estimaciones de confiabilidad, mediante división por mitades, de los reactivos de un modelo de examen (Popham, 1990, p. 133), mismo que para este caso requería ser mayor o igual que 0.85 en cada caso. Para determinar la confiabilidad de los modelos se calculó el coeficiente alfa mediante la fórmula:

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma\sigma_i^2}{\sigma_t^2}\right)$$

donde α = estimador de la confiabilidad
n = número de reactivos de la prueba
σ_t^2 = varianza de la prueba
$\Sigma\sigma_i^2$ = sumatoria de la varianza de los reactivos

Los resultados del análisis de ítems se muestran en la tabla compuesta que se presenta en la página siguiente. La idea de juntar en una sola, las tablas correspondientes al índice de dificultad de los reactivos de los cuatro modelos, al índice

de discriminación de sus ítems, al coeficiente de discriminación de los mismos y al coeficiente de confiabilidad de los modelos, es presentar una visión panorámica que permita observar con facilidad y tener una referencia rápida de los detalles y las relaciones entre sus elementos.

Tabla Nº 13. Resultados del análisis de respuestas a los ítems y a los modelos de examen: Índices de dificultad y discriminación, y coeficientes de discriminación y de confiabilidad, de los ítems y modelos.

Dificultad de los ítems del examen

Item	Modelo 1	Modelo 2	Modelo 3	Modelo 4
1	0.04	0.61	0.56	0.75
2	0.61	1.00	0.90	0.81
3	0.96	0.78	0.76	0.79
4	0.96	0.83	0.85	0.82
5	0.79	0.94	0.74	0.88
6	0.43	0.47	0.54	0.86
7	0.18	0.14	0.13	0.32
8	0.54	0.56	0.76	0.67
9	0.75	0.61	0.03	0.45
10	0.64	0.47	0.76	0.60
11	0.68	0.28	0.51	0.41
12	0.57	0.69	0.92	0.58
13	0.71	0.67	0.86	0.73
14	0.11	0.69	0.22	0.16
15	0.32	0.28	0.26	0.44
16	0.00	0.36	0.31	0.53
17	0.46	0.56	0.22	0.46
18	0.79	0.53	0.18	0.50
19	0.57	0.69	0.50	0.41
20	0.96	0.78	0.83	0.88
21	0.43	0.36	0.67	0.59
22	0.89	0.78	0.79	0.68
23	0.68	0.50	0.60	0.44
24	0.79	0.61	0.68	0.85
25	0.79	0.42	0.65	0.41
26	0.71	0.67	0.72	0.85
27	0.64	0.50	0.53	0.65
28	0.61	0.50	0.69	0.54
29	0.86	0.81	0.82	0.91
30	0.36	0.69	0.44	0.50
31	0.39	0.56	0.39	0.32
32	0.36	0.44	0.39	0.44
33	0.57	0.36	0.43	0.50
34	0.71	0.58	0.68	0.52
35	0.79	0.67	0.74	0.65
36	0.82	0.42	0.24	0.56
37	0.18	0.44	0.35	0.22
38	0.68	0.50	0.36	0.45
39	0.57	0.28	0.68	0.21
40	0.64	0.47	0.19	0.26
41	0.43	0.31	0.22	0.50
42	0.25	0.33	0.24	0.26
43	0.29	0.25	0.40	0.22
44	0.43	0.44	0.42	0.15
Dif.Prom.	0.57	0.54	0.53	0.54

Discriminación de los ítems del examen (rbis)

Item No.	Modelo 1	Modelo 2	Modelo 3	Modelo 4
1	-0.31	0.10	-0.03	-0.06
2	0.21	0.00	0.12	0.16
3	0.28	0.17	0.12	-0.06
4	0.25	-0.10	0.11	0.26
5	-0.36	0.34	0.29	0.20
6	-0.16	0.25	0.13	0.01
7	0.28	0.07	0.06	0.12
8	0.65	0.27	0.07	0.22
9	0.48	0.42	-0.01	0.20
10	0.50	-0.02	0.37	0.34
11	0.30	0.28	0.16	0.27
12	0.55	0.35	0.34	0.24
13	0.31	0.49	0.26	0.11
14	0.10	0.11	0.32	-0.05
15	0.08	0.06	0.18	0.07
16	0.00	0.48	0.35	0.33
17	0.12	0.35	0.14	-0.01
18	0.36	0.50	0.27	0.42
19	0.53	0.50	0.25	0.35
20	0.25	0.15	0.39	0.13
21	0.34	-0.15	0.44	0.34
22	0.35	0.32	0.40	0.30
23	0.40	0.23	0.24	0.14
24	0.62	0.54	0.40	0.32
25	0.36	0.08	0.50	0.12
26	0.62	0.41	0.51	0.34
27	-0.07	0.31	0.31	0.25
28	0.58	0.28	0.36	0.16
29	0.38	0.14	0.27	0.30
30	0.59	0.44	0.46	0.18
31	0.34	0.43	0.20	0.28
32	-0.01	0.30	0.39	0.22
33	0.37	0.03	0.39	0.27
34	-0.12	0.34	0.15	0.25
35	0.42	0.47	0.20	0.28
36	0.30	0.20	0.20	0.41
37	0.03	0.16	0.33	0.04
38	0.58	0.19	0.27	0.06
39	0.41	0.13	0.36	0.21
40	0.16	0.60	0.19	0.24
41	0.05	0.12	0.09	0.06
42	-0.08	-0.03	0.04	-0.10
43	0.17	0.34	0.35	0.23
44	0.14	0.35	0.47	-0.09
Disc Prom	**0.26**	**0.25**	**0.26**	**0.18**
No casos	28	36	65	105
Alfa	**0.80**	**0.79**	**0.80**	**0.70**

Discriminación de los ítems del examen (altos vs bajos)

Item No.	Modelo 1	Modelo 2	Modelo 3	Modelo 4
1	-0.10	0.31	0.08	0.00
2	0.30	0.00	0.12	0.21
3	0.10	0.15	0.12	0.02
4	0.10	-0.08	0.12	0.26
5	-0.40	0.15	0.27	0.16
6	-0.10	0.31	0.04	0.07
7	0.20	0.00	0.15	0.23
8	0.90	0.38	0.08	0.33
9	0.50	0.46	0.04	0.37
10	0.70	0.00	0.35	0.40
11	0.30	0.31	0.12	0.47
12	0.70	0.46	0.19	0.35
13	0.40	0.62	0.27	0.21
14	0.10	0.08	0.27	0.07
15	0.20	0.15	0.23	0.23
16	0.00	0.54	0.46	0.37
17	0.20	0.46	0.08	0.14
18	0.40	0.54	0.15	0.49
19	0.60	0.54	0.31	0.53
20	0.10	0.15	0.23	0.12
21	0.40	0.00	0.50	0.37
22	0.30	0.31	0.35	0.30
23	0.50	0.46	0.46	0.16
24	0.60	0.69	0.46	0.28
25	0.40	0.08	0.62	0.23
26	0.70	0.46	0.58	0.33
27	0.10	0.31	0.35	0.33
28	0.80	0.38	0.38	0.28
29	0.30	0.23	0.38	0.16
30	0.70	0.46	0.54	0.30
31	0.40	0.62	0.35	0.35
32	0.20	0.46	0.54	0.37
33	0.50	0.08	0.50	0.35
34	-0.10	0.31	0.23	0.28
35	0.50	0.62	0.31	0.44
36	0.40	0.31	0.12	0.49
37	0.00	0.31	0.46	0.07
38	0.70	0.23	0.35	0.16
39	0.50	0.15	0.50	0.26
40	0.20	0.77	0.08	0.33
41	0.00	0.08	0.15	0.23
42	-0.10	0.08	0.12	0.00
43	0.20	0.31	0.42	0.23
44	0.20	0.46	0.54	-0.02
Disc Prom	**0.31**	**0.31**	**0.29**	**0.26**

Estándares: Dificultad entre 0.05 - 0.95; Discriminación r >=0.20; Confiabilidad en escala: Alfa >= 0.85

No satisface el estándar -->

Los resultados de los análisis efectuados, que aparecen en la tabla compuesta, permiten hacer varios comentarios evaluativos sobre la información obtenida, mismos que se agruparon en tres categorías: dificultad, discriminación y confiabilidad.

Respecto a la dificultad de los ítems:

- En la parte izquierda de la tabla se muestra el análisis de dificultad de los 44 ítems de cada modelo de examen. Como puede observarse al pie de la tabla, los cuatro modelos presentaron una dificultad media.

- Prácticamente, solo el modelo 1 tuvo ítems que no satisfacen el estándar de calidad especificado para el índice de dificultad (entre 0.05 y 0.95). Los ítems Nº 1 y 16 resultaron muy difíciles y los ítems Nº 3, 4 y 20 demasiado fáciles. En el Modelo 2, se salió del rango especificado el ítem Nº 2, el cual fue tan fácil que lo respondieron correctamente todos los niños. Por su parte, el reactivo Nº 9 del modelo 3 fue excesivamente difícil. Solo el modelo 4 parece no tener problemas con la dificultad, pues todos sus reactivos cayeron dentro del rango apropiado.

- Para los reactivos que no satisficieron el estándar para la dificultad, el dictamen fue claro: pasar de nueva cuenta a una detallada revisión, pero ahora con el valioso apoyo de los datos derivados de la aplicación. Los resultados de dicha tarea se comentan en el siguiente procedimiento.

- Por otra parte, para observar como se distribuyó la dificultad de los ítems en los 4 modelos, se presenta la siguiente figura:

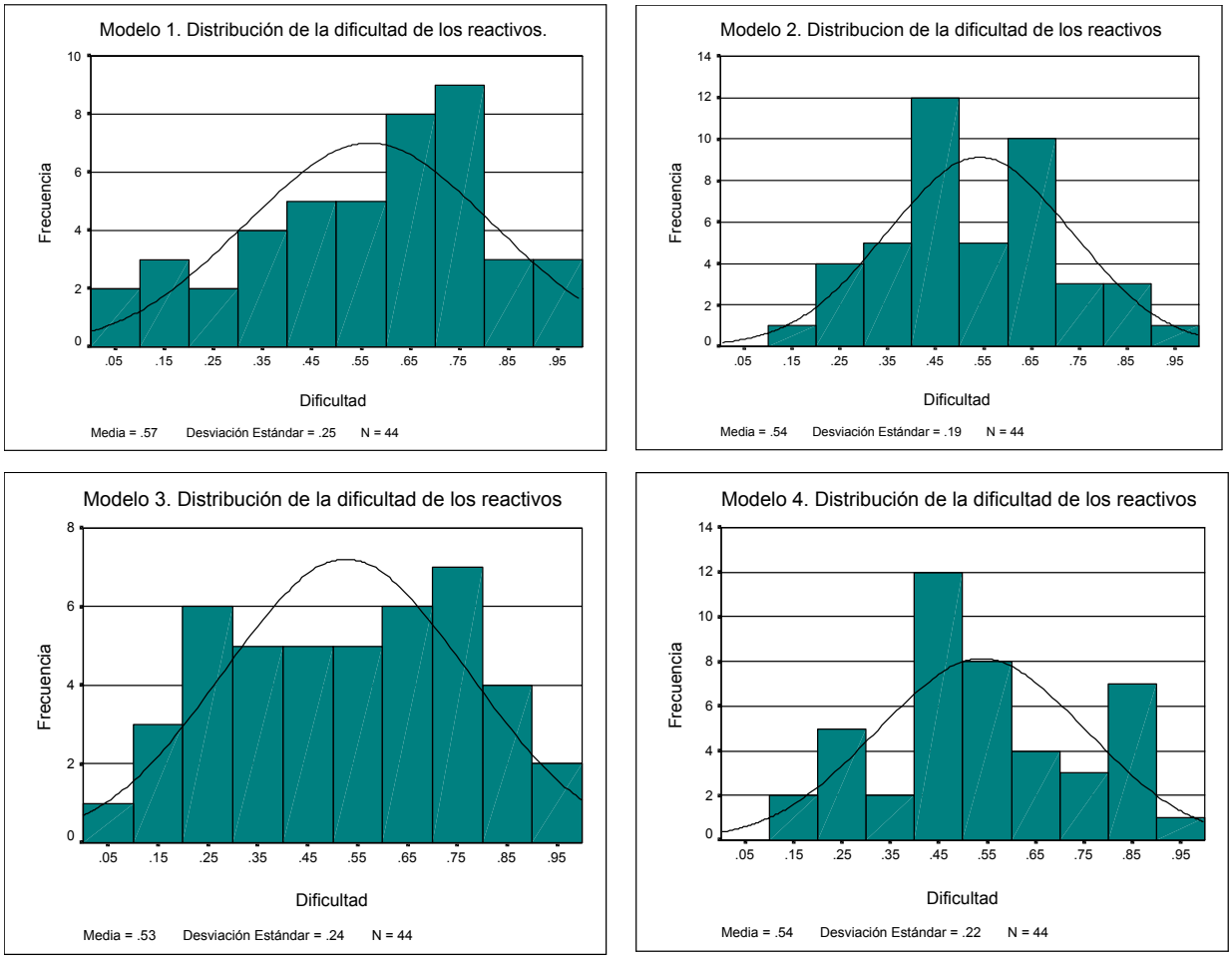

Figura Nº 4. Distribución de la dificultad de los ítems en los 4 modelos de examen

Con excepción del modelo 4, que presentó irregularidades más pronunciadas en su distribución, en los demás modelos la distribución de la dificultad de los ítems se aproximó más a la normal, aunque con diferentes sesgos y kurtosis. Lo anterior no tiene mayor significado en sí mismo, pues un modelo de examen alineado con el curriculum puede ser más o menos difícil que otro y sus ítems presentar una distribución de la dificultad diferente. Sin embargo, para pretender la equivalencia de los modelos de examen se requiere que los indicadores psicométricos, entre ellos la dificultad, tengan un comportamiento equiparable. En ese sentido, los modelos 1 y 3 presentan más similitud en lo que atañe a la distribución de la dificultad de sus ítems.

En conclusión, podemos decir que no se observaron fallas graves en cuanto a la dificultad de los ítems en los modelos, ni en cuanto a la forma en que esta se distribuye en cada modelo. Consecuentemente, la solución de los problemas detectados en este sentido fue bastante ágil, como se verá más adelante en el inciso 7.5.

Respecto a la discriminación de los ítems:

- En el caso de la discriminación se observa un panorama diferente. En la parte central de la tabla compuesta, se muestran los resultados del análisis de discriminación de los reactivos correspondientes a los 4 modelos de examen, efectuado mediante el cálculo de la correlación del punto biserial ($\mathbf{r_{bis}}$) de los reactivos de cada modelo. El análisis muestra que:

  - En el modelo 1, 16 de sus ítems tuvieron un coeficiente de discriminación menor que 0.2; es decir, el 36% de sus reactivos caen por debajo del estándar de calidad mínimo establecido.

  - Por su parte, en el modelo 2, 18 ítems tuvieron un coeficiente menor que 0.2, los cuales representan un 40% de los ítems.

  - Para el caso del modelo 3, el número de ítems cuyo coeficiente de discriminación fue menor que el mínimo aceptable fue 15, lo que significa el 34% de sus reactivos.

  - En cuanto al modelo 4, se presentaron 19 reactivos por debajo del 0.2; es decir, el 43% del total de ítems.

  - El número de ítems, en cada modelo, que presentaron una discriminación negativa fue el siguiente: siete reactivos con valor negativo en el modelo 1; cuatro ítems en el modelo 2; dos reactivos en el modelo 3; y en el caso del

modelo 4, observamos que seis ítems discriminan negativamente. Además, únicamente los modelos 1 y 2 presentaron un caso con 0.0 discriminación, cada uno de ellos.

- Finalmente, ocho ítems en el modelo 1 y 13 ítems en cada uno de los demás modelos, tienen valores que se encuentran entre 0.01 y 0.19; es decir, que discriminan positivamente pero por debajo del estándar especificado.

- Por su parte, a la derecha de la tabla compuesta aparecen los datos del análisis de discriminación de los reactivos correspondientes a los cuatro modelos, realizado mediante el cálculo del índice de discriminación de los reactivos de cada modelo; es decir, la diferencia entre la dificultad del ítem para el subgrupo de examinados que obtuvo las calificaciones altas y para el subgrupo con las calificaciones más bajas.

  El análisis de la información obtenida muestra que:

  - En el modelo 1, 13 de sus ítems tuvieron un índice de discriminación menor que 0.2; es decir, casi el 30% de sus reactivos caen por debajo del estándar de calidad mínimo establecido.

  - El modelo 2, tuvo 15 ítems con un índice menor que 0.2, mismos que representan un 34% de los ítems.

  - En el modelo 3, el número de ítems cuyo índice de discriminación fue menor que el mínimo aceptable fue 16, lo que significa el 36% de sus reactivos.

  - En cuanto al modelo 4, se presentaron 13 reactivos con un índice por debajo del 0.2; es decir, casi el 30% del total de ítems.

- El número de ítems, en cada modelo, que presentaron una discriminación negativa fue el siguiente: cinco reactivos con valor negativo en el modelo 1; un ítem en el modelo 2; ningún reactivo en el modelo 3; y en el modelo 4, observamos que un ítem discrimina negativamente. Además, el número de ítems que tienen un índice de discriminación de 0.0 en los modelos 1, 2, 3 y 4 es de tres, cuatro, cero y dos, respectivamente.

- Finalmente, cinco ítems en el modelo 1; 10 ítems en el modelo 2; 16 ítems en el modelo 3; y 10 ítems en el modelo 4, tuvieron índices de discriminación que se encuentran entre 0.01 y 0.19; es decir, que discriminan positivamente pero que no alcanzaron el estándar especificado.

- Al comparar los resultados del análisis discriminativo de los ítems, en cada modelo del examen, obtenidos mediante ambos métodos (r_{bis} vs altos-bajos), puede decirse que la correlación del punto biserial es en general un procedimiento más "duro", lo cual es producto de comparar la respuesta al ítem de todos los examinados y no solo la de los grupos extremos contrastados. Lo anterior puede observarse en la tabla siguiente:

Tabla Nº 14. Comparación de los resultados obtenidos mediante los métodos r_{bis} y alto-bajo

Modelo	Porcentaje de ítems que satisfacen el estándar		Número de ítems que no satisface el estándar, por categoría discriminativa					
			Discriminación = 0.0		Discriminación negativa		Discriminación entre 0.01 y 0.19	
	r_{bis}	Alto -bajo	r_{bis}	Alto -bajo	r_{bis}	Alto -bajo	r_{bis}	Alto -bajo
1	64	70	1	3	7	5	8	5
2	60	66	1	4	4	1	13	10
3	66	64	0	0	2	0	13	16
4	57	70	0	2	6	1	13	10

Sin embargo, la diferencia no parece ser tan determinante cuando se observa el perfil discriminativo que producen todos los ítems en cada modelo. Para ilustrar este efecto, obsérvese la siguiente figura que muestra la discriminación de los ítems del modelo 1, obtenida mediante ambos métodos:

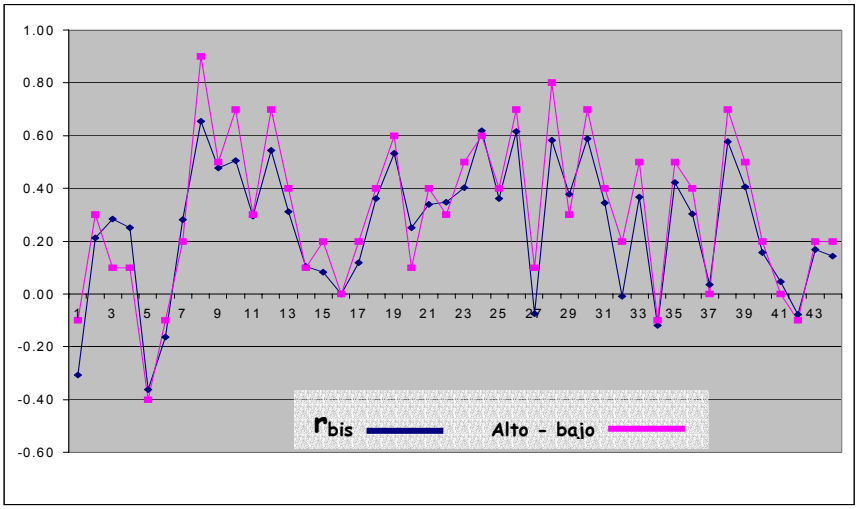

Figura Nº 5. Perfil de discriminación de los ítems del modelo 1, obtenido mediante los Métodos de correlación biserial (r_{bis}) y el de grupos extremos contrastados (alto y bajo)

Con las diferencias ya comentadas, en ambos casos el perfil general producido es similar; aunque se ve que el índice de discriminación tiene una ligera tendencia a beneficiar con valores más altos a los ítems del modelo.

- Para observar también como se distribuyó la discriminación de los ítems en los 4 modelos, en la página siguiente se presenta la correspondiente gráfica compuesta:

En este caso las diferencias son menos pronunciadas y, en general, en los 4 modelos la distribución de la discriminación de los ítems se aproximó más a la normal, con similares sesgos pero con diferente kurtosis. Considerando la posible equivalencia entre los modelos de examen, podemos decir que según este indicador psicométrico las distribuciones de los modelos 3 y 4 lucen más parecidas entre sí que con los otros modelos.

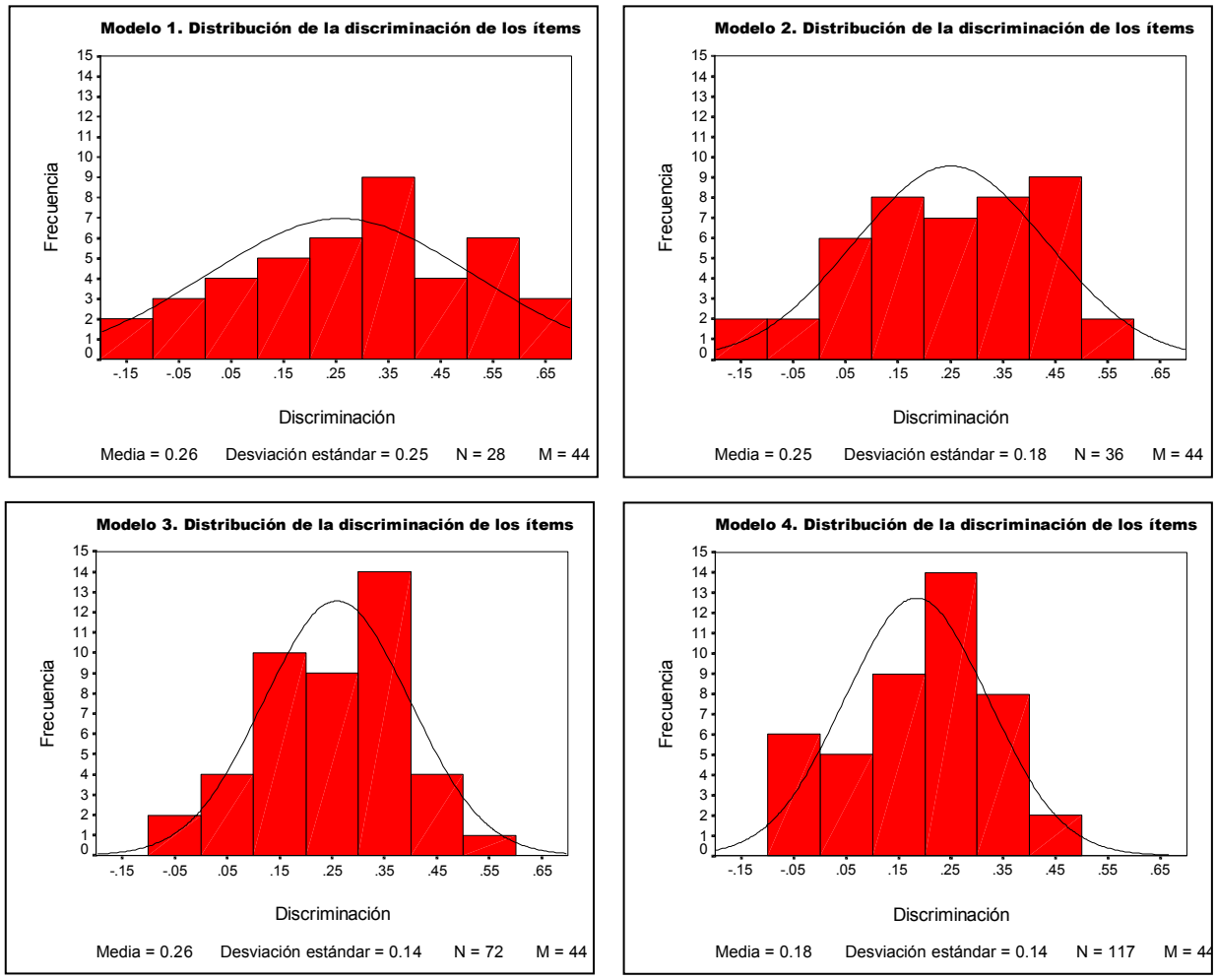

Figura Nº 6. Distribución del índice de discriminación de los ítems en los 4 modelos de examen

De hecho, puede decirse que un valor psicométrico clave de la discriminación radica en ponerle una bandera al ítem defectuoso. En particular, el coeficiente respectivo (o índice, en su caso) proporciona una buena orientación sobre el origen o el sentido de la falla; por ello, su principal ventaja se reveló durante el procedimiento de revisión de los ítems.

- No obstante, lo dicho hasta el momento en esta sección no oculta el hecho de que en todos los modelos de examen alrededor de un tercio de los ítems, un número considerable, presentaron fallas en cuanto a su poder discriminativo y requirieron de

una cuidadosa revisión para ser modificados o eliminados, como se verá más adelante.

- **Respecto a la confiabilidad de los modelos**

  En la parte inferior de la tabla compuesta (tabla Nº 13), se presenta también el valor de alfa que obtuvieron los modelos de examen; es decir, su coeficiente de confiabilidad. Como puede apreciarse, los 4 modelos quedaron cortos respecto al estándar de calidad definido para este indicador psicométrico, el cual establecía un mínimo de 0.85. Los modelos 1 y 3 son los que más se acercan al criterio con 0.80 de confiabilidad, cada uno de ellos; el modelo 2 obtuvo un coeficiente de 0.79 y el más alejado del estándar es el modelo 4, con 0.70 de confiabilidad.

  Se sabe que la confiabilidad se mejora incrementando el número de ítems de un examen. Sin embargo, esa no parece ser la solución en este caso. Sin ser muy grave la situación, más bien los problemas de confiabilidad están relacionados con las fallas del poder discriminativo de los ítems anotadas en el punto anterior, las cuales tendrían que ser resueltas primero.

Para tener una visión global de los resultados obtenidos, a continuación se presenta un resumen de la información derivada del análisis de ítems y modelos que se presentó en los puntos anteriores.

Tabla Nº 15. Resumen de la información derivada del ítem análisis

Modelo	Porcentaje de ítems con dificultad apropiada	Porcentaje de ítems con discriminación apropiada (r_{bis})	Porcentaje de ítems con discriminación apropiada (altos vs bajos)	Confiabilidad (alfa)
1	89	64	70	0.80
2	98	60	66	0.79
3	98	66	64	0.80
4	100	57	70	0.70

Como puede observarse, tras la primera aplicación del examen el análisis mostró que tanto los reactivos, como los modelos el examen, aún no alcanzan los niveles de calidad requeridos para su uso extensivo como instrumento para monitorear el aprendizaje que logran los niños en el área de español de la educación primaria. El punto fuerte del examen es que, con pocas excepciones, los ítems tuvieron el nivel de dificultad apropiado para los examinados. En cambio, sus puntos débiles consisten en que cerca de un tercio de las preguntas, en cada modelo, no discriminaron apropiadamente entre los examinados que dominan los contenidos respectivos y quienes no lo hacen; y en que los modelos mismos no obtienen los puntajes del test de manera suficientemente consistente. Desde luego, hasta este punto, los datos mostraron que no se está lejos de lograr los estándares especificados en esos rubros. Justamente, en esa dirección estuvieron encaminados los esfuerzos en el siguiente procedimiento.

7.5 Revisión de Ítems y estructuración de la prueba.

Con base en los resultados del análisis estadístico de los ítems, el **Comité Diseñador** y el **Comité Elaborador** revisaron cuidadosamente los ítems que no cumplieron con los estándares de calidad establecidos, a fin de determinar sus fallas y decidir cuáles de ellos podían ser corregidos y cuáles deberían ser reelaborados.

Para la revisión de cada ítem se tomó en consideración, además de los índices de dificultad y discriminación y del coeficiente de discriminación respectivos, el análisis de los distractores. Es decir, de manera complementaria se observó la ejecución de los niños ante cada una de las opciones de respuesta. De esta manera, pudieron ser identificadas varias clases de fallas en los ítems, respecto a las cuales se adoptaron las siguientes decisiones: su conservación, modificación o eliminación. Los tipos de fallas más comunes que fueron detectadas y las decisiones adoptadas ante ellas, se describen en la tabla siguiente:

Tabla Nº 16. Taxonomía de fallas más comunes y decisiones adoptadas para el mejoramiento de los ítems

Tipo de falla	Ejemplos	Tipo de decisión adoptada para el mejoramiento
Complejidad cognitiva	Muy difícil	Cuidadoso análisis del contenido del ítem y, en su caso, redacción más categórica, hacer más o menos atractivas las opciones o sustituirlas
	Muy fácil	
Discriminación errónea	Discriminación negativa	Cambiar la opción que atrae a los que saben, cambiar distractores poco elegidos
	Discriminación baja	Cambiar la opción que parece esta creando el problema
Edición	Escritura confusa	Corregir errores mecanográficos, hacer dibujos más claros
Redacción	Conceptos complejos	Simplificar las conceptualizaciones, cambiar la opción confusa
Mixta	Respuesta al azar	Hacer más categóricas la base y la respuesta correcta, o sustituir distractores

De hecho, para revisar cada ítem y con ello detectar las posibles fallas que tenía a fin de proceder posteriormente a su correspondiente modificación, se efectuó una especie de análisis estructural cuyos elementos fueron las evidencias disponibles a partir del análisis de reactivos y del análisis de los distractores. A continuación se presenta un diagrama que ilustra el proceso general que se siguió para analizar y modificar el ítem Nº 20 del modelo 4:

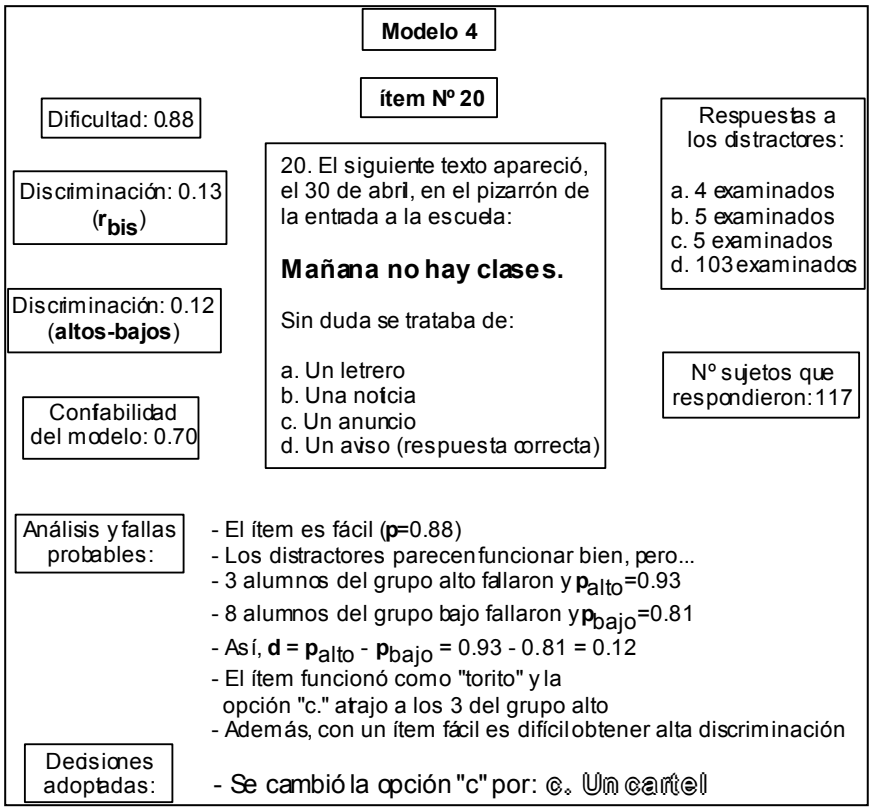

Figura Nº 7. Ilustración del proceso para analizar y modificar un ítem

La idea principal al presentar el diagrama es mostrar el poder que tiene el uso conjunto de los indicadores psicométricos para detectar y corregir fallas en los ítems y modelos de examen.

Una vez corregidos o reelaborados los ítems que presentaron fallas en cada modelo, el **Comité Coordinador** procedió a estructurar una muestra de ítems que fuera representativa del dominio curricular del área de español; es decir, conformó de nueva cuenta los modelos de examen de tal manera que, mediante una especificación y un muestreo adecuados, mostraran los resultados del curriculum con un peso y una proporción apropiados.

Con esta última acción, se dieron por concluidos el diseño, la elaboración y el pilotaje de la prueba y, con ello, se cumplen los propósitos del presente trabajo. Sin embargo, de conformidad con la metodología propuesta en el capítulo 3, aún faltan por desarrollarse fases y procedimientos que resultan necesarios para pasar el instrumento al nivel de gran escala y probarlo en esa dimensión a fin de garantizar su calidad integral para monitorear permanentemente el aprendizaje de los niños. Tales acciones se describen más adelante en el capítulo de conclusiones.

Tabla Nº 18. Porcentaje de aciertos en los ejes curriculares del área de español, por modelo y escuela

Áreas curriculares					Modelos aplicados a escuelas												Ítems
					Modelo 1		Modelo 2			Modelo 3			Modelo 4				
	Área	Sub-área	Línea de formación	Contenidos	5	13	3	4	6	1	7	12	2	8	10	11	
					N=13	N=15	N=16	N=3	N=17	N=23	N=24	N=25	N=36	N=30	N=21	N=30	
	Lengua hablada	Conocimientos, habilidades y actitudes	Exposición y entrevistas **(37.10)**	Uso de vocabulario adecuado para situaciones específicas: diferencia entre términos cotidianos y especializados **(51.96)**	30.77	53.33	25.00	33.33	47.06	86.96	62.50	52.00	58.33	60.00	47.62	66.67	21
				Práctica del debate **(26.60)**	15.38	40.00	31.25	00.00	23.53	52.17	29.17	40.00	30.56	23.33	23.81	10.00	43
				Planeación de exposiciones y presentaciones orales y elaboración de esquemas para apoyar la exposición **(32.75)**	23.08	60.00	68.75	33.33	23.53	47.83	37.50	40.00	19.44	16.67	09.52	13.33	44
Subtotal		**37.10**			**23.08**	**51.11**	**41.67**	**22.22**	**31.37**	**62.32**	**43.06**	**44.00**	**36.11**	**33.33**	**26.98**	**30.00**	
Total área	**37.10**																
	Lengua escrita	Conocimientos, habilidades y actitudes	Manejo de letras y sílabas: ortografía **(48.88)**	Consolidación de la aplicación de normas ortográficas relativas al uso de X, S, Z, V, B, H, así como de las sílabas Ce, Ci, Ge, Gi, Gui, Güi **(57.30)**	42.31	84.45	69.79	27.78	45.10	64.49	54.86	72.67	66.67	55.56	53.97	50.00	8, 9, 10, 11, 12, 13,
				Consolidación del reconocimiento de la sílaba tónica y la aplicación de las reglas de acentuación **(40.45)**	41.02	82.22	56.25	33.33	29.41	39.13	48.61	36.00	36.11	40.00	22.22	21.11	38, 39, 40
			Redacción y elaboración de resúmenes **(37.21)**	Elaboración de resúmenes sobre temas de otras asignaturas, detectando las ideas centrales de un texto **(31.39)**	15.39	26.67	50.00	50.00	47.06	23.92	18.75	30.00	37.50	33.33	19.05	25.00	14, 15
				Localización de las ideas principales con base en la estructura formal del texto: introducción, desarrollo y conclusión **(43.02)**	23.08	46.67	56.25	66.67	29.41	47.83	29.17	40.00	50.00	33.33	57.14	36.67	32
			Manejo de materiales de consulta	Manejo e identificación de las partes del diccionario **(43.25)**	61.54	93.33	75.00	00.00	41.18	08.70	25.00	20.00	66.67	46.67	47.62	33.33	18
			Comprensión de instrucciones, normas de biblioteca y elaboración de fichas	Conocimiento de normas de uso de bibliotecas: solicitud, inscripción, uso de catálogos, préstamo en sala y a domicilio **(32.10)**	34.62	33.33	37.50	33.33	26.47	23.91	25.00	20.00	38.89	38.33	40.48	33.33	41, 42
Subtotal		**45.23**			**37.44**	**67.56**	**59.58**	**33.33**	**38.43**	**43.77**	**41.11**	**46.93**	**51.85**	**45.11**	**40.95**	**36.67**	
		Situaciones comunicativas	Comprensión de ilustraciones y textos	Lectura: comprensión y seguimiento de instrucciones para armar un objeto, jugar, experimentar, etc., identificando los tipos de texto usados en la escuela y la calle (letreros, listas, avisos, anuncios). Comparación del periódico con otros materiales escritos **(67.70)**	65.39	86.67	87.44	50.00	64.71	80.44	62.05	60.00	69.45	73.34	59.53	53.34	19, 20
			Redacción de preguntas, cartas, anuncios y solicitudes	Redacción individual y colectiva de textos, considerando título, secuencia y relación entre ideas, atendiendo a la exposición de relaciones causales	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	45
Subtotal		**67.70**			**65.39**	**86.67**	**87.44**	**50.00**	**64.71**	**80.44**	**62.05**	**60.00**	**69.45**	**73.34**	**59.53**	**53.34**	
Total área	**56.46**																
	Recreación literaria	Conocimientos, habilidades y actitudes	Comprensión y creación de géneros populares **(65.28)**	Apreciación y exploración del significado de trabalenguas, adivinanzas, dichos, chistes, canciones, versos y leyendas populares y tradicionales **(82.80)**	82.69	83.33	85.94	83.34	92.65	79.35	78.12	86.00	87.50	90.00	57.14	87.50	2, 3, 4, 5
				Creación de cuentos y poemas **(47.76)**	69.23	93.33	56.25	33.33	29.41	39.13	08.33	24.00	63.89	66.67	42.86	46.67	36
Subtotal		**75.79**			**80.00**	**85.33**	**80.00**	**73.34**	**80.00**	**71.31**	**64.16**	**73.60**	**82.78**	**85.33**	**54.28**	**79.33**	
		Situaciones comunicativas	Creación oral y escrita de géneros populares **(48.64)**	Recreación de finales de cuentos **(43.73)**	07.69	63.34	46.88	66.67	23.53	34.78	37.50	30.00	55.56	53.34	57.15	48.34	1,7
				Creación: elección de un tema o cuento para elaborar historietas **(62.72)**	57.69	63.34	75.00	16.67	47.07	63.05	70.84	58.00	73.61	80.00	69.05	78.34	6, 35
				Creación de rimas a partir de palabras dadas y redacción individual y colectiva de cuentos ilustrándolos **(56.34)**	76.92	66.67	81.25	00.00	47.06	78.26	75.00	52.00	72.22	60.00	33.33	33.33	34
				Creación: redacción colectiva de guiones para presentar obras de teatro basadas en la lectura de narraciones; creación de personajes para obras de teatro y elaboración colectiva de cuentos **(31.76)**	15.38	20.00	50.00	66.67	35.29	60.87	25.00	20.00	22.22	30.00	19.05	16.67	37
			Representación Literaria	Escenificación: simulación de entrevistas con personajes de obras elegidas por los alumnos **(45.34)**	46.15	66.67	37.50	33.33	35.29	47.83	41.67	40.00	52.78	53.33	42.86	46.67	33
Subtotal		**49.48**			**38.46**	**58.10**	**58.93**	**38.10**	**36.98**	**54.66**	**51.19**	**41.14**	**57.94**	**58.57**	**49.66**	**50.00**	
Total área	**62.63**																
	Reflexión sobre la lengua	Conocimientos, habilidades y actitudes	Tiempos verbales: matices de significado entre el copretérito y el pospretérito	Los tiempos verbales: matices de significado entre el copretérito y el pospretérito **(33.92)**	00.00	00.00	62.50	33.33	11.76	39.13	12.50	40.00	58.33	60.00	42.86	46.67	16
			Desarrollo de vocabulario mediante campos semánticos	Ampliación del vocabulario a través de la formación de campos semánticos a partir de términos poco usuales y de tecnicismos **(41.24)**	46.15	46.67	68.75	33.33	47.06	13.04	12.50	40.00	41.67	40.00	52.38	53.33	17
			Uso de palabras en la oración: sujeto, verbo y predicado **(60.76)**	Observación del orden de las palabras en la oración **(76.16)**	84.61	93.33	93.75	66.67	64.71	86.96	79.17	72.00	69.44	56.67	66.67	80.00	22
				Identificación del sujeto y el predicado en las oraciones **(62.35)**	55.38	84.00	71.25	40.00	40.00	80.00	49.17	68.00	71.67	66.00	60.00	62.67	24, 25, 26, 27, 28
				Reconocimiento y uso adecuado de artículos, adjetivos y pronombres **(57.76)**	38.46	66.67	87.50	55.55	52.94	55.07	55.56	54.67	65.74	54.44	47.62	58.89	29, 30, 31
			Aporte de otras lenguas al español: galicismos y anglicismos	Localización de aportes de otras lenguas al español: galicismos y anglicismos **(51.53)**	53.85	80.00	62.50	33.33	41.18	60.87	58.33	60.00	44.44	50.00	23.81	50.00	23
Subtotal		**57.32**			**48.07**	**70.00**	**75.52**	**44.44**	**43.63**	**63.77**	**47.92**	**59.67**	**64.12**	**58.33**	**52.38**	**60.00**	
Total área	**57.32**																
Total escuelas	**55.44**				**48.74**	**69.79**	**67.19**	**43.57**	**49.19**	**62.71**	**51.58**	**54.22**	**60.37**	**59.00**	**47.30**	**51.56**	
Total modelos					**59.27**		**53.32**			**56.17**			**54.56**				
Total examen	**49.79**																

Capítulo 8. Análisis de los aprendizajes logrados en el área de español

En este capítulo se presenta un análisis de los resultados obtenidos en la prueba por los niños que fueron examinados. Cabe señalar que el análisis solo tiene como propósito ilustrar el uso de los procedimientos que son necesarios para dar cuenta de la ejecución de los examinados en los ejes, subejes, líneas de formación y contenidos específicos en que se clasificó el contenido del área de español de la educación primaria. Con rigor, puede decirse que se trata de un ejercicio analítico; que es auténtico, en la medida en que los resultados que se presentan son los que realmente obtuvieron en el examen los niños que lo respondieron y que el proceso analítico que se describe es el mismo que debe emplearse para evaluar dichos resultados. Sin embargo, solo tiene un carácter ilustrativo porque en la etapa de desarrollo en que se empleó el instrumento, la meta primaria era efectuar una primera calibración de los ítems y los modelos de examen y no monitorear el aprendizaje de los niños. Ese uso está reservado para el momento en que el test exhiba sus bondades, a la luz de los estándares de calidad especificados y con suficientes datos empíricos.

8.1 Características generales de la ejecución de los examinados

Después de haber sido analizados psicométricamente los ítems y modelos del examen, se procedió a analizar los resultados que obtuvieron en la prueba los niños que formaron parte de la muestra. Para ello, la base de datos que se generó cuando las hojas de respuesta fueron leídas por un lector óptico, se pasó a una hoja de cálculo para calificar el examen y posteriormente se procesó la información con el apoyo de dos paquetes estadísticos. Cuando se obtuvieron los puntajes logrados por los niños, se procedió a caracterizar su ejecución. En la siguiente tabla se muestran los resultados generales de ese análisis.

Tabla Nº 17. Características generales de la ejecución de los estudiantes de la muestra, en los modelos de examen.

Modelo número	Escuela número	Nº de examinados		Media de aciertos	Desviación estándar	Rango de aciertos	Porcentaje de aciertos	
		Escuela	Modelo				Escuela	Modelo
1	5	13	28	24.93	6.26	15 – 37	45.63	56.66
	13	15					66.21	
2	3	16	36	23.83	6.56	9 – 39	66.19	54.17
	4	3					41.67	
	6	17					45.05	
3	1	23	72	23.17	6.28	8 – 39	57.02	52.65
	7	24					48.20	
	12	25					52.91	
4	2	36	117	23.73	5.44	11 – 35	59.41	53.92
	8	30					55.91	
	10	21					46.86	
	11	30					50.30	
Totales	**12 escuelas**	**253**		**23.92**	**6.14**	**8 - 39**	**52.95**	**54.35**

Varios comentarios pueden hacerse sobre los resultados que se muestran en la tabla:

- Si el dominio mostrado en el examen refleja el dominio que tienen los niños sobre el contenido del curriculum del área de español (lo cual es de esperarse que suceda en alguna medida, dados los diversos esfuerzos realizados para garantizarlo), entonces podemos decir que su dominio es medio. De ningún modo es despreciable este resultado, y parece contradecir diversos pronunciamientos, más y menos formales y más y menos documentados, en el sentido de que dicho dominio es mucho menor en el país.

- En promedio (54. 35% de aciertos), los cuatro modelos parecen reflejar los mismos resultados y, como ya se comentó, solo destacan a la alta los alumnos de las escuelas primarias privadas (planteles Nº 3 y 13), que entre

ambas tienen un promedio de aciertos de 66.20% y a la baja las rurales (planteles Nº 4 y 6) y una urbana marginal (plantel Nº 5), que tienen las tres un promedio de aciertos de 44.11%.

- Aunque el rango de aciertos en los cuatro modelos presenta disparidades importantes en su límite inferior, en realidad la dispersión de los puntajes es similar, como lo muestran los valores de la media y la desviación estándar.

Si consideramos la frecuencia de respuestas correctas en cada modelo de examen, observamos el siguiente comportamiento general:

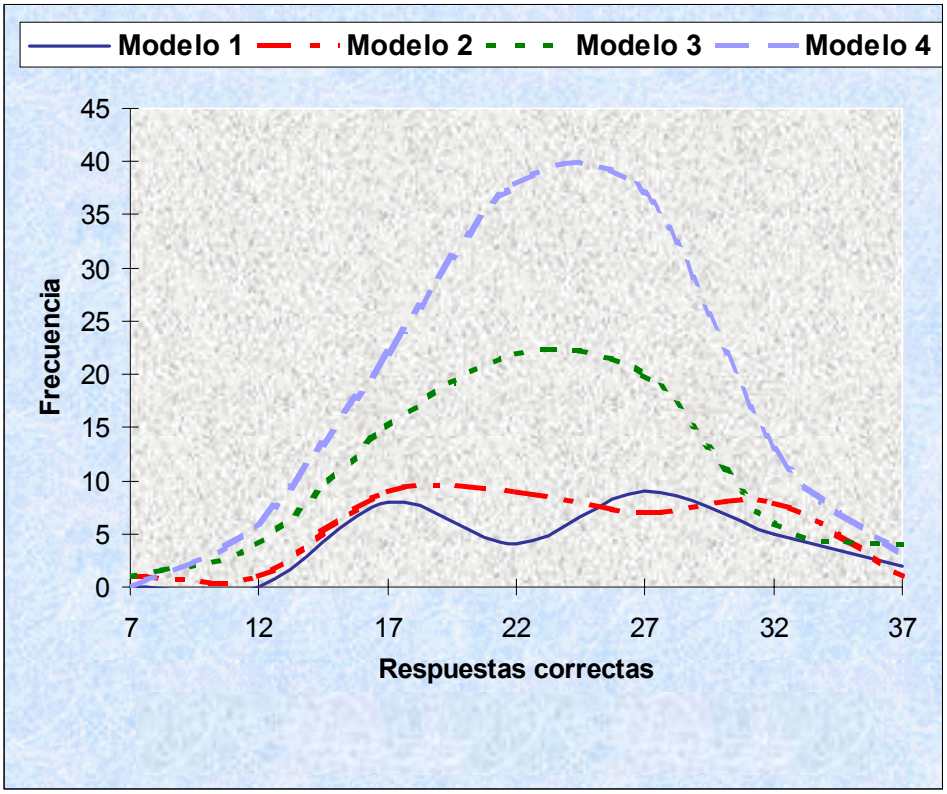

Figura Nº 8. Frecuencia de respuestas correctas en los cuatro modelos de examen

Una característica común en las distribuciones, aunque aquí no se aprecia bien en el caso del modelo 2, es que las frecuencias de respuestas correctas están moderadamente sesgadas hacia las puntuaciones bajas. La diferencia en la forma de las curvas de distribución de frecuencias y en las Kurtosis de las

distribuciones, obedecen al número desigual de estudiantes que respondieron el examen (117 examinados en el modelo 4; 72 examinados en el modelo 3; 36 examinados en el modelo 2; y 28 examinados en el modelo 1).

8.2 Características de la ejecución de los examinados en los ejes curriculares

De manera congruente, para analizar la ejecución de los niños en los modelos de examen se empleó la misma clasificación del contenido que se desarrolló durante el análisis curricular. Es decir, se observó su dominio en el grupo de ítems que apuntaban a cada contenido y que fueron elaborados a partir de las especificaciones asociadas a cada línea de formación, subeje y eje curriculares del área de español. La ejecución en esos cúmulos de ítems se describe en la tabla Nº 18, que se presenta en la página siguiente.

Aunque es posible hacer múltiples análisis sobre la abundante información que contiene la tabla, algunos de los principales comentarios se presentan a continuación:

- El porcentaje de aciertos es la unidad de medida típica para reportar la ejecución de los examinados en los exámenes criteriales, tradición que se mantiene en este caso.

- Como corresponde, los porcentajes que aparecen en la tabla están ponderados según el número de ítems que evalúan el dominio de cada contenido. Sin embargo, los porcentajes que están anotados como totales en los ejes curriculares son promedios directos de los subtotales. Por ello, no cuadran con los subtotales y totales por escuela y modelo que si son promedios ponderados. Por la misma razón, tampoco hay coincidencia con los datos que se mencionan en la tabla Nº 17, mismos que no están ponderados.

Insertar aquí tabla No 18

- Con excepción del eje de Lengua Hablada donde el porcentaje de aciertos fue de 37.10, en general, el dominio que mostraron los examinados de los contenidos correspondientes a los ejes curriculares del área de español es similar. En recreación literaria se observa la mejor ejecución con 56.63% de aciertos. Seguido de cerca por lengua escrita con 54.48% y por reflexión sobre la lengua, donde se obtuvo un 53.83% de aciertos. Quizás la baja ejecución mostrada en el eje de Lengua Hablada no es muy relevante, dado el escaso número de reactivos que incluye, por las razones que ya se comentaron.
- Por su parte, las líneas de formación que obtuvieron mejores resultados fueron la comprensión de ilustraciones y textos con un porcentaje de 67.70 y la comprensión y creación de géneros populares con 65.28.
- En cambio, las líneas que obtuvieron porcentajes más bajos fueron los tiempos verbales (33.92), la exposición y entrevistas (37.10) y la redacción y elaboración de resúmenes (37.21).
- Sin duda, uno de los resultados más importantes es que los contenidos que fueron juzgados por el Comité Diseñador como los más relevantes del área de español y que, por ello, en las especificaciones se estableció como necesario elaborar un mayor número de ítems que exploraran su dominio, son justamente los que están en el grupo de contenidos que mejor dominaron los niños. Este resultado requiere de algunos comentarios:

  - Se juzgó que la apreciación y exploración del significado de adivinanzas, dichos, versos y otras producciones populares son fundamentales para la comprensión verbal y se especificaron cuatro ítems para evaluarlas. Aquí, el porcentaje de aciertos fue 82.80.
  - Lo mismo sucedió con el contenido que explora la identificación del sujeto y predicado en las oraciones, mismo que se juzgó como particularmente importante para la comprensión escrita, por lo que se

especificaron cinco ítems. En este contenido el porcentaje de aciertos fue de 62.35.

- Se consideró que la consolidación de la aplicación de diversas normas ortográficas es fundamental para el desarrollo de la habilidad para redactar y se especificó elaborar seis ítems para explorarla. En este contenido, el porcentaje de aciertos fue 57.30.
- Lo anterior significa que solo para evaluar estos tres contenidos se utilizó un tercio de los ítems del examen. Por ello, resulta relevante que los niños mostraran buen dominio en las partes del curriculum del área de español que se juzgaron más importantes.

- Parcialmente, lo anterior también sucedió con otros contenidos juzgados como relevantes y para los cuales se especificaron tres ítems, como fue el caso del uso adecuado de artículos, adjetivos y pronombres y de la consolidación de las reglas de acentuación, en los cuales los porcentajes de aciertos fueron 57.76 y 40.45, respectivamente.
- En el caso de los cuatro contenidos, donde se especificaron dos reactivos para evaluar cada uno de ellos, los resultados son extremos. Así, en el seguimiento de instrucciones e identificación de tipos de textos y en la creación de historietas a partir de un tema, los porcentajes de aciertos son altos con 67.70 y 62.70, respectivamente. En cambio, el conocimiento de normas de uso de bibliotecas y la elaboración de resúmenes sobre temas de otras asignaturas, se encuentran entre los que obtuvieron porcentajes más bajos con 32.10 y 31.39, respectivamente.
- El contenido con el porcentaje de aciertos menor fue la práctica del debate, con 26.60.

Puestos en la perspectiva adecuada, estos resultados y otros similares que es posible obtener a partir de un análisis riguroso de la información contenida en la tabla, pueden resultar de gran valor para las diversas instancias que intervienen en la formación de los niños con funciones de planeación, operación, evaluación

y control educativos, mismas que de hecho son los usuarios naturales de la información que proporciona el examen. En la tabla que se presenta a continuación, se incluyen algunos posibles destinatarios de la información que deriva del examen, así como los tipos de información que resultan más relevantes para apoyar la toma de decisiones en el nivel que les corresponde a cada uno de ellos.

Tabla Nº 19. Destinatarios de los reportes de resultados del examen

Destinatario	Tipo de información a ser reportada
Secretario de Educación del Estado	1. Distribución de promedios de las escuelas del estado, en cada eje curricular 2. Distribución de promedios de las escuelas, por municipio, en cada eje curricular 3. Tendencia del municipio y promedios estatales en cada eje curricular, durante los cinco años anteriores 4. Gráfica que compara simultáneamente los resultados de logro contra acciones educativas determinadas (perfil de profesores, recursos de la escuela, etc.)
Director de Planeación Educativa	1. Promedio de ejecución de estudiantes en grupos de preguntas que evalúan blancos de aprendizaje claves, por eje curricular 2. Promedio de ejecución de estudiantes en cada eje curricular 3. Resultados en 1 y 2, por municipio
Inspector	1. Promedio de resultados por escuela dentro de la región de servicio, para cada eje curricular 2. Tendencia de la ejecución promedio, durante los cinco años anteriores
Director de Escuela	1. Promedio general y por género, de resultados de los estudiantes de la escuela, por eje curricular 2. Porcentaje de aciertos por eje curricular, para cada egresado 3. Comparación del promedio general y por género, de los resultados de los estudiantes de la escuela, por eje curricular, contra las correspondientes medias de las escuelas del estado

En la página siguiente, y de nueva cuenta con propósitos ilustrativos, se presenta la tabla Nº 20, que contiene la parte medular del informe que se podría proporcionar a uno de los destinatarios de la información que se obtuvo al aplicar el examen, que en este caso se trata de un inspector escolar.

Cabe señalar que la información que se presenta en la tabla, está contenida en la tabla Nº 18 y que solo se arregló para efectos de presentación al usuario.

Además de otros posibles comentarios evaluativos, como la tendencia de la ejecución promedio de los examinados durante los años anteriores, que podrían hacerse al inspector escolar en un informe sobre la ejecución de los examinados en las escuelas que tiene bajo su control administrativo, en la tabla se proporciona el promedio de resultados por escuela dentro de la región de servicio, por eje y subeje curriculares, así como por línea de formación.

Además, con propósitos de contextualización y posibles comparaciones para efectos de planeación del desarrollo de las escuelas, gestión escolar, retroalimentación educativa y otras funciones propias de su cargo, en este caso se le proporciona información sobre el porcentaje de aciertos que lograron los examinados en el municipio de Ensenada. Desde luego, estos elementos de referencia podrían aumentar a medida que el examen alcanzara nuevos niveles de escala. Así, se incorporarían los correspondientes datos comparativos completos por municipio y en el estado.

Tabla Nº 20. Ejemplo de informe a un inspector escolar, sobre el porcentaje de aciertos por área curricular que obtuvieron en el examen de español los egresados de las escuelas primarias de la zona escolar

Áreas curriculares			Porcentajes de aciertos por escuela, en la zona escolar										Porcentaje aciertos del municipio
Eje	**Subeje**	**Línea de formación**	**5**	**4**	**6**	**1**	**7**	**12**	**2**	**8**	**10**	**11**	
Lengua hablada	Conocimientos habilidades y actitudes	Exposición y entrevistas	23.08	22.22	31.37	62.32	43.06	44.00	36.11	33.33	26.98	30.00	**37.10**
Lengua escrita	Conocimientos habilidades y actitudes	Manejo de letras y sílabas: ortografía	41.67	30.56	37.26	51.81	51.74	54.34	51.39	47.78	38.10	35.56	**48.88**
		Redacción y elaboración de resúmenes	19.24	58.34	38.24	35.88	23.97	35.00	43.75	33.33	38.10	30.84	**37.21**
		Manejo de materiales de consulta	61.54	00.00	41.18	08.70	25.00	20.00	66.67	46.67	47.62	33.33	**43.25**
		Comprensión de instrucciones, normas de biblioteca y elaboración de fichas	34.62	33.33	26.47	23.91	25.00	20.00	38.89	38.33	40.48	33.33	**32.10**
	Situaciones comunicativas	Comprensión de Ilustraciones y textos	65.39	50.00	64.71	80.44	62.05	60.00	69.45	73.34	59.53	53.34	**67.70**
		Redacción de preguntas, cartas, anuncios y solicitudes	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	Ensayo	**Ensayo**
Recreación literaria	Conocimientos habilidades y actitudes	Comprensión y creación de géneros populares	75.96	58.34	61.03	59.24	43.23	55.00	75.70	78.34	50.00	67.09	**65.28**
	Situaciones comunicativas	Creación oral y escrita de géneros populares	39.42	37.50	38.24	59.24	52.09	40.00	55.90	55.84	44.65	44.17	**48.64**
		Representación Literaria	46.15	33.33	35.29	47.83	41.67	40.00	52.78	53.33	42.86	46.67	**45.34**
Reflexión sobre la lengua	Conocimientos habilidades y actitudes	Tiempos verbales: matices de significado entre el copretérito y el pospretérito	00.00	33.33	11.76	39.13	12.50	40.00	58.33	60.00	42.86	46.67	**33.92**
		Desarrollo de vocabulario mediante campos semánticos	46.15	33.33	47.06	13.04	12.50	40.00	41.67	40.00	52.38	53.33	**41.24**
		Uso de palabras en la oración: sujeto, verbo y predicado	59.48	54.07	52.55	74.01	61.30	64.89	68.95	59.04	58.10	67.19	**65.42**
		Aporte de otras lenguas al español: galicismos y anglicismos	53.85	33.33	41.18	60.87	58.33	60.00	44.44	50.00	23.81	50.00	**51.53**

Capítulo 9. Conclusiones y recomendaciones

En esta sección, se reconsideran los procesos metódicos que se siguieron al construir el examen, así como los productos y resultados obtenidos, a la luz de los objetivos planteados inicialmente y de la base conceptual propuesta. Al respecto, los aspectos más significativos del estudio evaluativo se comentan a continuación:

En la introducción, se estableció como primer objetivo de este trabajo el desarrollo de una prueba criterial de gran escala, para evaluar el aprendizaje que logran en el área de español los egresados de la primaria en Baja California. El proceso que se siguió para cumplir con este cometido fue descrito en los capítulos 2 al 7.

En el primero de ellos, se caracterizó este tipo de instrumentos y se comentaron los estándares de calidad a que deben sujetarse, debido al poderoso impacto social que tienen. Asimismo, se presentaron algunas experiencias evaluativas sobre el aprendizaje del lenguaje en el mundo y las surgidas en México en la década de los noventa, periodo en que surgieron más pruebas de lenguaje, que en toda la historia de la evaluación del aprendizaje en la educación básica nacional.

En el capítulo 3, se estableció la metodología general del proyecto. Es decir, se describió el modelo psicométrico de seis etapas que fue utilizado para desarrollar la prueba (tabla Nº 8), mismo que fue adaptado del propuesto por Anthony Nitko para crear exámenes nacionales alineados con el curriculum.

Por su parte, en el capítulo 4 se describió el primer grupo de acciones formales para desarrollar la prueba: el análisis del contenido de diversas fuentes que definen el curriculum del área de español, a fin de hacer explícito el dominio de resultados de logro pretendidos por el curriculum. El producto más importante que se obtuvo del análisis, fue el universo de contenido identificado que fue la base para construir la prueba, mismo que aparece en la retícula del contenido del área (figura Nº 1).

En el capítulo 5, se comentó la manera en que se determinó el universo de medida a partir del universo de contenido identificado. El resultado principal de estos procedimientos fue el contenido importante a evaluar, sobre el cual se elaboró el test, y que se estructuró en la retícula. Ahí, se identifican los contenidos que son fuente,

síntesis o enlace de servicios respecto a otros contenidos, y que fueron el criterio básico para determinar la importancia relativa de los contenidos.

En otra fase del proceso, comentada en el capítulo 6, se describió cómo se redujo el universo de medida al nivel de un examen de gran escala; y también la manera en que se procedió a diseñar las especificaciones de ítems del examen. En el primer caso, el producto fue la tabla de especificaciones de la prueba (tabla Nº 10), y en el segundo, las 30 especificaciones para producir los reactivos (anexo 6).

Finalmente, en el capítulo 7, se indicó la manera en que se procedió a desarrollar los reactivos y como fueron posteriormente analizados y corregidos mediante su comparación contra las especificaciones correspondientes. El resultado de dichas acciones fue el conjunto de 180 ítems que fueron estructurados en cuatro modelos de examen (uno de ellos se incluye en el anexo Nº 8).

Un segundo objetivo del estudio fue iniciar el proceso de validación empírica de los ítems y modelos. Para lograrlo, se operaron dos procedimientos descritos en el capítulo 7, que arrojaron los siguientes resultados:

Para aplicar el examen, se obtuvo una muestra intencional de 253 niños, de 12 escuelas de Ensenada, que presentaron diversas condiciones. Como resultado, se logro un perfil de los examinados que aparece en la tabla Nº 11. Cuando se correlacionaron las condiciones con los aciertos en el examen, se obtuvieron correlaciones estadísticas que son significativas (se muestran en la tabla Nº 12).

Se aplicó el examen a la muestra y se analizaron las respuestas a los ítems y la ejecución en los ejes curriculares del área. El análisis de los reactivos mostró que, con pocas excepciones, los ítems presentaron índices apropiados de dificultad en todos los modelos. En cuanto a su discriminación, alrededor de un tercio de ítems en cada modelo, no alcanzaron el criterio de calidad mínimo establecido. Otro tanto sucedió, pero en menor grado, con el índice de confiabilidad: los 4 modelos quedaron cortos en relación al estándar de calidad respectivo. Estos resultados se muestran con mayor detalle en las tablas Nº 13 y 15.

Las fallas de los ítems, detectadas mediante el análisis de reactivos, fueron corregidas y

los modelos de examen se reestructuraron de nueva cuenta. Actualmente se encuentran a la espera de su aplicación a gran escala.

Respecto a la descripción de la ejecución de los examinados en el curriculum, se efectuaron dos ejercicios analíticos a partir de la información derivada del examen. En uno de ellos se calculó el porcentaje de aciertos que presentaron los examinados, por escuela, por modelo, por eje y subeje curriculares y por línea de formación (tablas Nº 17 y 18). En otro análisis, se elaboró un informe a un inspector escolar sobre la ejecución de los examinados de las escuelas de la zona escolar (tabla Nº 20). En ambos casos, los análisis realizados y los resultados obtenidos solo tuvieron un carácter ilustrativo, pues la meta de la aplicación era efectuar una primera calibración de los ítems y los modelos y no monitorear el aprendizaje de los niños. Ese uso está reservado para cuando el test exhiba sus bondades, a la luz de los estándares especificados y con suficientes datos empíricos.

Cabe señalar que dos procedimientos produjeron resultados importantes que no se analizan en este trabajo. Uno de ellos, la aplicación del examen, produjo las respuestas a la pregunta 45 de cada modelo, la cual solicitó a los examinados la redacción de una descripción breve. Otro procedimiento, el cuestionario de opinión sobre la Lengua Hablada, fue aplicado y produjo los resultados correspondientes. En ambos casos, no se incluyen resultados porque se está elaborando un modelo para analizar las respuestas al ítem de ejecución escrita; y las respuestas al cuestionario de opinión no se han terminado de analizar aún.

En cuanto al objetivo de ensayar las condiciones necesarias para establecer la aplicación del instrumento a gran escala, como un mecanismo permanente para monitorear el aprendizaje en las primarias de Baja California, cabe señalar que el énfasis puesto a lo largo del trabajo en la idea de monitorear la calidad del aprendizaje, tuvo por finalidad dejar claro que el desarrollo y la aplicación del examen no deben ser interpretados como eventos coyunturales, sino que se trata de desarrollar una estructura y un proceso permanentes que formen parte de una estrategia descriptiva más amplia, destinada a saber lo que sucede en las escuelas en cuanto a operación curricular, la instrucción y el aprendizaje, con objeto de retroalimentarlos y mejorarlos.

Se trata de describir sistemáticamente, mediante cortes sincrónicos y aplicación en el campo, el aprendizaje que logran los niños, la capacidad de profesores y escuelas para promoverlo y otros aspectos igualmente importantes. ¿Qué tanto aprendieron los alumnos respecto a lo que señala el curriculum? ¿en cuáles áreas su ejecución es mejor? ¿en qué escuelas de la zona escolar se han observado problemas de aprendizaje con el tiempo? ¿cómo se dio el aprendizaje en los alumnos de un profesor determinado durante los últimos cinco años? ¿en qué municipios parece estarse promoviendo mejor el aprendizaje de los contenidos del área de español? Este es el tipo de preguntas que se quiere responder con el uso del examen. Sin duda, estas tareas le corresponden a las autoridades educativas del estado, quienes son los destinatarios finales de la metodología y de los instrumentos desarrollados, y quienes realmente tienen el poder para producir cambios.

En particular interesa proporcionar información válida, confiable y comprensiva, que resulte significativa y útil para las personas involucradas, protegiendo los derechos de quienes resultan afectados por la evaluación.

Recomendaciones para el desarrollo posterior del examen.

Una vez desarrollado y piloteado el test, y después de que los ítems y modelos de examen fueron revisados y corregidos con base en los resultados de su aplicación, el modelo que sirvió de guía durante la realización de este trabajo contempla probar de nueva cuenta el instrumento, pero ahora en el nivel de gran escala. Por su dimensión e impacto, las acciones que ello conlleva tendrían que ser el objeto de un estudio quizás más amplio y trascendente y de igual o mayor complejidad. Un estudio de tal naturaleza, necesariamente incluye ya efectuar un acuerdo formal con las autoridades educativas estatales, el involucramiento de nuevos grupos de personas con funciones de jueceo, estandarización y otras, así como la obtención de nuevos recursos financieros para apoyar las actividades. Entre otras, las fases y procedimientos que aún faltan por realizarse para contar con un instrumento de alta calidad, incluyen las siguientes:

Para concluir la fase de producción y validación de ítems, se requiere efectuar el ensayo empírico a gran escala de los ítems. Este procedimiento supone que, previo a la aplicación de la versión final del instrumento, sea seleccionada una muestra de

examinados representativa de los niños que egresan de la educación primaria en Baja California, que se lleve a cabo la capacitación del personal que habrá de aplicar la prueba de manera estandarizada.

Tras la aplicación, se requiere efectuar la consiguiente revisión y mejoramiento de los ítems y modelos, ya con suficientes datos empíricos. Sin embargo, esta fase requiere del uso de procedimientos para calificar y analizar grandes volúmenes de información, que deberán ser automatizados.

A diferencia de la fase en la que se quería principalmente detectar y corregir los problemas más gruesos de los reactivos para estar en condiciones de poder estructurar una primera versión de la prueba, el análisis de los ítems en este momento tendrá propósitos muy diferentes. Por un lado, se trata de efectuar un análisis de reactivos con suficientes datos empíricos para asegurar un control de calidad de los ítems a partir de los estándares definidos. Por otra parte, se trata de definir formalmente las cualidades que el examen deberá exhibir antes de ser utilizado oficialmente para monitorear de manera continua la calidad del aprendizaje en el área de español. Con ello, la idea es crear y mantener un programa de control de calidad y relevancia del examen, a cargo de la autoridad educativa.

Posteriormente, será necesario efectuar los correspondientes análisis de los resultados obtenidos por los sujetos en la prueba, por municipio y escuela; por eje temático del plan de estudios y por destinatario. A partir de estos análisis se deberán elaborar reportes de resultados del examen apropiados a las necesidades informativas de los usuarios, con el propósito de ir configurando un sistema de información permanente sobre los resultados del examen que apoye eficazmente la toma de decisiones en los diversos niveles donde corresponda.

A mediano y largo plazos, el horizonte de desarrollo del examen deberá incluir otras acciones formales orientadas a su continuo mejoramiento. En ese sentido serán necesarios, por ejemplo, estudios para ampliar las evidencias de validez de contenido logradas hasta el momento, y otros de validación criterial y de constructo, así como explorar la posibilidad de expandir el poder del examen para medir habilidades de ejecución que establece el curriculum y que fueron casi ignoradas en esta primera versión del examen.

Bibliografía

APA (1985). *Standards for educational and psychological testing*. Prepared by a joint committee of the American Psychological Association, American Educational Research Association, National Council of Measurement in Education. Washington, D.C.: American Psychological Association.

Backhoff, E., Tirado, F., Larrazolo y Antillón L. E. (1996). Desigualdad en la Calidad de la Educación Básica en México: Estudio Comparativo entre dos Entidades Federativas. *Revista Latinoamericana de Estudios Educativos (México), Vol. XXVI, Nº 3.*

Berk, R. (1984). *A Guide to Criterion- Referenced Test Construction*. Baltimore.The Johns Hopkins University Press.

Bond, L. (1994). *Reaching for New Goals and Standards: The Rol of Testing in Educational Reform Policy*. Policy Talks. Illinois, North Central Regional Educational Laboratory.

CAT/5. (s/f.). *An Inside Look at CAT/5.* Monterey. CTB. Macmillan McGraw-Hill.

Congress of the US, (1992). *Testing in American Schools: Asking de Right Questions. Full Report*. Office of Technology Assessment. Washington. U.S. Government Printing Office.

CRESST. (1994). *Improving America's School: A Newsletter on Issues in School Reform.* National Center for Research on Evaluation, Standards and Student Testing.

CRESST. (1994b). *Assessment Profile-State Summary. Evaluation Comment*. National Center for Research on Evaluation, Standards and Student Testing.

CTP (1993). *Comprehensive Testing Program III*. Washington. Educational Testing Service.

Davies, A. (1990). Principles of Language Testing. En David Crystal and Keith Johnson (Ed). *Applied Language Studies*. Basil Blackwell.

ETS (1996). Descripciones resumidas de tests de la colección de instrumentos del Educational Testing Service, obtenidas vía INTERNET durante los meses de agosto y septiembre de 1996. Disponible en *http://www.cua.edu/www/eric_ae/alltext.html*

Gómez Palacio M., *et. al.* (1990). Evaluación de la Comprensión Lectora, SEP-OEA-Universidad de las Americas. Estudio comentado en *La Investigación Educativa en los Ochenta, Perspectiva para los Noventa. Estados del Conocimiento, cuaderno 9. pp 23-24.*

Greaney, V y Kellaghan. (1995). *Equity Issues in Public Examinations in Developing Countries*. Washington. The World Bank.

Haladyna, T. (1990). Advances in Item Design. Rasch Measurement Transasctions, 4 (2). Disponible en *´Gopher_root_eric_ae:[_rasch.back]rm42_52.txt:1´.*

Hambleton, R. (1993). Advances in the Detection of Differentially Functioning Test Items. *Laboratory of Psychometric and Evaluative Research Report No. 237.* Amherst, MA: University of Massachusetts, School of Education.

Hogan, T. P. (1992). *Prospects and Problems for a National Test: Some Reflections of a Test Author.* Ponencia presentada en el simposium □National Goals and National Testing□, en el encuentro anual del National Council on Measurement in Education. San Francisco.

INCE (1995). *Evaluación de la Educación Primaria 1995. Primer Ciclo. Lengua Castellana*. Madrid. Ministerio de Educación y Ciencia. Instituto Nacional de Calidad y Evaluación.

Jackson-Maldonado, D. (1993). Desarrollo del Español como Primera Lengua. En II Congreso Nacional de Investigación Educativa. *La Investigación Educativa en los Ochenta, Perspectiva para los Noventa. Estados del Conocimiento, cuaderno 9.*

Joint Committee on Testing Practices. (1994). *Code of Fair Testing Practices in Education*. American Psychological Association.

Jones, R.W. y Hambleton, R.K. *(1992).* Recent Advances in Psychometric Methods. *Laboratory of Psychometric and Evaluative Research Report No.233.* University of Massachusetts. pp. 2-16.

Linn, R. (1991). *Test Misuse*: *Why is it So Prevalent?*. University of Colorado. Center for Research on Evaluation, Standards and Student Testing (CRESST).

Linn, R. (1993). Discurso Inaugural en la Conferencia Anual del Center for Research on Evaluation, Standards and Student Testing (CRESST*).* En *Rothman, R. (1994). Evaluation Comment. Assessment Questions: Equity Answers. Proceedings of the 1993 CRESST Conference.* ERIC. ED 367 684.

Madsen, H. (1983). *Tecniques in Testing*. Oxford University Press.

Makey, W. (1992). Lenguas Maternas, Otras Lenguas y Lenguas Vehiculares: Su Significado en un Mundo que se Transforma. *Perspectivas. XXII (1) pp. 41-51*.

Martínez, F. (1993). Evaluación del Aprendizaje. Presentación del Estado del Conocimiento. En *II Congreso Nacional de Investigación Educativa. Síntesis del Congreso Nacional Tematico*. D.F.

Martínez, F. (1995). *Diseño e Implementación de un Sistema Integral de Monitoreo de la Calidad de la Educación Básica*. Universidad Autónoma de Aguascalientes. Documento.

NCTE/IRA, (1996). *Standards for the English Languaje Arts. Professional Summary*. International Reading Association (IRA) and the National Council of Teachers of English (NCTE).

Nitko, A.J. (1984). Defining Criterion- referenced Test. En *Berk, R.A. (Ed.). A guide to criterion-referenced test construction. Baltimore*.The Johns Hopkins University Press.

Nitko, A.J. (1994). *A Model for Developing Curriculum-Driven Criterion-Referenced and Norm-Referenced National Examinations for Certification and Selection of Students*. Ponecia presentada en la Conferencia Internacional sobre Evaluación y Medición Educativas, de la Asociación para el Estudio de la Evaluación Educativa en Sudafrica (ASSESA).

Nitko, A.J. (1995). Curriculum-based Continuos Assessment: a framework for concepts, procedures and policy. *Assessment in Education, Vol. 2, No. 3.*

Kellaghan, T. y Greaney, V. (1992). *Using Examinations to Improve Education. A Study in Fourteen African Countries*. Washington. The world Bank.

OCDE. (1997). *Exámenes de las Políticas Nacionales de Educación*. *México, Educación Superior*. París. Organización para la Cooperación y Desarrollo Económicos.

Ornelas, C. (1994). Reseña de libro. Syvia Schmelkes (Coord.). La calidad de la educación primaria: estudio en cinco regiones del estado de Puebla, Paris, Instituto Internacional de Planeación Educativa de la UNESCO, en prensa. *En Revista Latinoamericana de Estudios Educativos (México). XXIV (3 y 4) pp. 215-218.*

Poder Legislativo Federal. (1994). *Ley General de Educación*. México. Diario Oficial de la Federación.

Poder Legislativo Estatal. (1995). *Ley Estatal de Educación*. Mexicali. Diario Oficial del estado.

Popham, J. (1990). *Modern Educational Measurement. A Practitioner's Perspective*. MA. Allyn and Bacon.

Robredo, J.M., Ledezma, R. y Alvarado, J.F. (1983). *Reticulación: una estrategia para la elaboración de programas de estudio*. UNAM. Facultad de Psicología. Tesis para obtener el grado de licenciatura.

Rudner L. (1993). Test Evaluation. Disponible en Gopher ERIC/AE. 12/ 93.

Ruiz, G. y Martínez, F. (1996). *Evaluación de Niveles de Aprendizaje en Dos Grados de la Educación Primaria y en Cuatro Contextos Socioeconómicos del Estado de Aguascalientes*. Ponencia presentada en II Foro Nacional de Evaluación Educativa. Memorias. Centro Nacional de Evaluación para la Educación Superior.

Schmelkes, S, (1994). La desigualdad en la calidad de la educación primaria. *Revista Latinoamericana de Estudios Educativos (México). XXIV (1 y 2) pp. 13-38*.

SEP. (1996). *Políticas de Educación Superior en México. Informe Básico Preparado por las Autoridades Mexicanas.* México. Secretaría de Educación Pública.

SEP. (1993). *Educación Básica. Primaria. Plan y Programas de Estudio.* México. Secretaría de Educación Pública.

Weiping, W. (1993). *Universality vs Particularity in Chinese Teaching and Testing.* Conferencia presentada en el Simposium sobre Lenguajes y Lingüística. Washington. University of Pre- Georgetown.

Zorrilla, M. (1996). Comprender para Transformar. La Experiencia de Evaluación de Aprendizajes en la Escuela Primaria y Secundaria en Aguascalientes. Instituto de Educación de Aguascalientes. Ponencia presentada en *II Foro Nacional de Evaluación Educativa. Memorias. Centro Nacional de Evaluación para la Educación Superior.*