UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
IZTACALA

ANALISIS DE LA ESTRATEGIA "ACADEMIAS DE EVALUACION DEL LOGRO ESCOLAR" PARA LA EVALUACION MASIVA DE ESTUDIANTES UNIVERSITARIOS.

T E S I S

PROPUESTA PARA LA MAESTRIA EN
MODIFICACION DE CONDUCTA

P R E S E N T A :

ANDRES EDUARDO SANCHEZ MOGUEL

IZTACALA LOS REYES IZTACALA 2002

Campus Iztacala

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES IZTACALA

ANÁLISIS DE LA ESTRATEGIA "ACADEMIAS DE EVALUACIÓN DEL LOGRO ESCOLAR" PARA LA EVALUACIÓN MASIVA DE ESTUDIANTES UNIVERSITARIOS.

Tesis propuesta para la Maestría en

Modificación de Conducta

por

Andrés Eduardo Sánchez Moguel

Año 2002

Resumen

Las Academias de Evaluación Institucional son una estrategia metodológica para mejorar la manera en que se hacen los exámenes a los estudiantes de la FES Iztacala, así como para contribuir a formar una cultura de la evaluación entre sus profesores. Estas academias consisten en reunir a profesores de las diferentes corrientes de conocimiento de las carreras de la FES Iztacala (una academia por carrera) para que reflexionen en el sentido y práctica de la evaluación de estudiantes, proporcionándoles materiales para ello y fomentando el intercambio de ideas, tras lo cual el grupo se avoca a diseñar exámenes para determinar el nivel de conocimientos en la disciplina de los estudiantes que llevan cursada la mitad y el total de la carrera. Con la intención de analizar si en realidad genera ventajas este método de trabajo, se comparan exámenes de la época previa a la existencia de las Academias de Evaluación con exámenes diseñados por éstas. Las comparaciones son tanto estadísticas como cualitativas, y se integran en un cuerpo de conclusiones que muestra cómo las academias tienen algunas ventajas sobre las prácticas tradicionales.

TABLA DE CONTENIDO

Introducción

La evaluación del aprendizaje[1] es una práctica de suma importancia en la educación y, sin embargo, ha tenido poco espacio en la reflexión de los estudiosos en México. El estado de conocimiento sobre la evaluación del aprendizaje del II Congreso Nacional de Investigación Educativa (COMIE, 1993) señala que son muy pocos los trabajos realizados entre 1982 y 1992 al respecto (sólo se encontraron 81 estudios, tesis y artículos en revisión exhaustiva). La situación no cambió mucho durante la década de los 90. Ya en 1979 García Cortés señalaba el problema de la "realización paupérrima de estudios de evaluación educativa en nuestro medio" (p.37). La falta de interés teórico por el tema, y su poco análisis en las instituciones formadoras de maestros han llevado a la generalización de prácticas de evaluación del logro escolar con las siguientes características:

a) Falta de reflexión sobre las razones por las cuales se evalúa, dando prioridad al cumplimiento administrativo sobre la utilidad real de la información.

b) Procedimientos e instrumentos de evaluación poco planeados y mal estructurados.

c) Poco análisis de lo obtenido en las evaluaciones, priorizando los intereses crediticios ("Aprobé o no", "Quince alumnos reprobaron", etcétera) sobre los logros académicos ("He aprendido el cien por ciento de este contenido", "Ya hay un conocimiento generalizado de esta materia o no", etcétera).

---

[1] Una versión de esta introducción fue publicada en Básica, revista de la escuela y del maestro con el título "La validez de la evaluación del aprendizaje en el ámbito educativo", No. 15, enero – febrero de 1997.

d) Una serie de factores que distorsionan la medición del conocimiento real de los estudiantes, como relacionar la conducta con la calificación, las altas posibilidades de fraude, o las pruebas que privilegian lo memorístico sobre lo reflexivo.

En estas circunstancias, existen quienes se inclinan por realizar el procedimiento de evaluación de la manera más rápida y menos "estorbosa" para el profesor y los alumnos, sin importar que la información recabada se acerque a la realidad o no, y por otra parte quienes se quejan amargamente de que la evaluación es "tecnicista" y mecánica y no plantea un sentido ni bases metodológicas adecuadas a un proceso social (Cf., como primer ejemplo, Díaz Barriga, 1982). Los críticos de la evaluación educativa pueden ser a su vez diferenciados en una gama que va desde el que de manera rabiosa y amarga critica las evaluaciones mal hechas, y sobre generaliza que, si así son las que él conoce, así deben ser todas, hasta el que señala con claridad dos tipos de problemas con la evaluación: los problemas de construcción incorrecta que se resuelven atendiendo a los aspectos técnicos con mayor cuidado y los problemas inherentes a la evaluación, de los cuales muchos son irresolubles actualmente (lo que no impide que el autor coherente haga, tras su crítica, una propuesta para resolver o disminuir la dificultad). Como ejemplos del tipo rabioso, tenemos a Labarca (1973), que pone un ejemplo de por qué le parece que la evaluación escrito - objetiva es un error, transcribiendo el siguiente ítem de la Prueba de Práctica de 1971 para la Prueba Nacional:

> "Louis Amstrong, rey del jazz, pasó los primeros años de su vida en un barrio popular de Nueva Orleans con gente menesterosa.
> Respuestas:
> 1) Vivió con gente que se ganaba la vida en diversos trabajos.
> 2) En un barrio de Nueva Orleáns vivían gentes que no tenían bastantes recursos para vivir.
> 3) Todos eran hombres que pertenecían a la raza negra.
> 4) La que conoció era gente descuidada en el vivir."

Nos parece claro que Labarca eligió un reactivo francamente pésimo. Una persona con unas pocas nociones de cómo deben redactarse ítems de opción múltiple

encontraría fácilmente varios errores en esta pregunta. Sin duda puede hacerse de mucha mejor manera. En donde diferimos con Labarca es cuando él utiliza este ejemplo para argumentar en contra de todos los reactivos de opción múltiple. Un reactivo defectuoso no es argumento para suprimir ítems bien hechos.

Una crítica equilibrada de la evaluación, que sin embargo no pierde totalmente el "halo" de indignación hacia todo intento sistemático de medición es la que hace Santos Guerras (1998), en que se señalan 22 errores comunes al evaluar (sin implicar que sean insalvables), entre los que, para nosotros, destacan los siguientes: "se evalúan sólo los conocimientos" (p. 18), "se evalúa principalmente la vertiente negativa" (p.20), "se evalúa descontextualizadamente" (p.22), "se evalúa competitivamente" (p.24), "los profesores repiten una y otra vez sus esquemas de evaluación" (p.25), "no se evalúa éticamente" (p.26), "no se hace metaevaluación" (evaluación de los métodos y procedimientos de evaluación) (p. 31). Todos estos puntos son atendidos en las Academias de trabajo cuyo funcionamiento se explica en el capítulo 3. El autor comete también algunas exageraciones, desde nuestro punto de vista, al señalar por ejemplo que "sólo se evalúan los efectos observables" (p. 19. Nos parece difícil evaluar lo no observable), "se evalúa cuantitativamente" (p. 21. No vemos el problema. Todo lo que existe, existe en cierta cantidad. Negar la dimensión cuantitativa es perder información.). Sin embargo, esta crítica de la evaluación es en general propositiva, y señala la necesidad de atender más cuidadosamente a los diferentes momentos de la evaluación, sin apuntar a su erradicación.

Algunas veces, los artículos son utilizados de manera descontextualizada para apoyar la vertiente "apocalíptica" de la evaluación (la evaluación no sirve y es mala; debe eliminarse totalmente) a pesar de que su intención es más bien la de señalar los malos usos de la evaluación y no condenarla a muerte. Como ejemplo, diremos que en una conferencia en que participamos hace unos años una profesora esgrimió un artículo de Landsheere (1973) como la prueba de que el fundamento de nuestra disertación (que era una ponderación de los usos de la evaluación) era incorrecto. Al finalizar, solicitamos a la profesora una copia del artículo. Hacia su parte final éste mostraba no sólo mesura y razón, como se puede ver en el siguiente párrafo:

> "En resumidas cuentas, los exámenes tradicionales suelen presentar graves defectos de construcción. Su validez es dudosa. La evaluación de los trabajos, por su parte, padece grandes deficiencias. Además, los exámenes pueden perjudicar la salud física y mental de los alumnos. Finalmente, puede agregarse que los profesores, en algunos casos, dictan sus cursos en función de un examen y no al revés" (p. 50)

Estamos, en general, de acuerdo con este párrafo que, en todo caso, tiene como premisa que los exámenes tienen una serie de inconvenientes cuando no se organizan correctamente, lo cual no implica su erradicación. Pero lo más interesante es que el autor se disponía a continuación, en el artículo, a establecer argumentos a favor de los exámenes e, imagino, a explicar cómo evitar muchos de los inconvenientes enumerados: "Pero ha llegado el momento de otorgar la palabra a la defensa." (p. 50). ¡La profesora no me proporcionó las siguientes páginas del artículo! Al parecer, todo lo que ella quería saber estaba en las primeras páginas.

Entre el extremo apocalíptico y el desganado que hace la evaluación de cualquier forma para salir del paso, se encuentra la correcta aplicación de los principios de evaluación, con intenciones bien determinadas y una reflexión profunda sobre su sentido y metodología.

La evaluación educativa es una estrategia de recolección de información sobre los diferentes momentos, actores y auxiliares del proceso enseñanza-aprendizaje. Si bien es indispensable que cada profesor haga evaluaciones particulares y a profundidad de estos elementos al interior de su espacio de trabajo para poder actuar sobre éste de manera certera, es también necesario contar con perspectivas más generales de la labor académica que sirvan de monitor a los directivos y a la comunidad en general de las escuelas. Por tanto, la evaluación masiva se presenta como una práctica útil en el acopio de datos globales sobre la situación escolar. Sin embargo, múltiples circunstancias han hecho que en algunas ocasiones la evaluación se lleve a cabo con una perspectiva de conteo y control que tiene las siguientes características:

- Los datos se recaban con instrumentos que no han sido suficientemente depurados, o en ocasiones ni siquiera se ha planteado la necesidad de depurarlos.
- Se asignan calificativos por simple tradición numérica [2], que implican consecuencias administrativas importantes para los estudiantes pero pocos o ningún cambio en términos de apoyo a sus deficiencias.
- Se generan listados llenos de cifras a los que no se da un uso en el perfeccionamiento del esquema educativo, sino que sirven para "cubrir" un expediente y adornar un cajón.

La intencionalidad de quienes realizan el proceso de evaluación y quienes lo promueven es decisiva en el énfasis que se va a dar a unos u otros elementos del sistema de evaluación que se genere. Así, ante el único interés de cumplir de la manera más eficiente posible (pero no eficaz) con una exigencia administrativa o estatutaria, los evaluadores pueden preferir hacer instrumentos que estén listos para ser aplicados en muy poco tiempo, que sean lo más económicos posible en tiempo, dinero y esfuerzo, así como hacer análisis rápidos de los exámenes que no evidencien las carencias de estos instrumentos, aunque no tengan uso fuera de los expedientes.

En cambio, ante el interés de obtener información útil en el análisis de la situación educativa, las características de un buen sistema de evaluación deberán incluir:

- Objetivos claros previos al desarrollo del sistema (La pregunta ¿para qué evaluar? se responde antes de llevar a cabo otras acciones)
- Preferencia por desarrollar métodos e instrumentos con un nivel suficiente de validez, confiabilidad y pertinencia, aunque impliquen más costos en tiempo, recursos y esfuerzos.
- Interés en que los métodos e instrumentos sirvan para obtener datos que reflejen la realidad escolar de manera clara, aunque se evidencien las áreas

[2] Nos referimos especialmente a la conocida "escala de cero a diez", en que 6 o más significa "aprobado", es decir, adecuado, y 5 o menos significa "reprobado", es decir, inadecuado. Del mismo modo podemos hablar del sistema NA-MB.

débiles del sistema (o quizá precisamente buscando conocer estas áreas problemáticas).

- Interés en que los resultados sean conocidos por quienes toman las decisiones en la escuela y por la comunidad en general para que apoyen cambios sustanciales en las prácticas educativas, que lleven a disminuir las áreas débiles.

Este es el caso del proyecto que se presenta, insertado en un plan de evaluación que pretende obtener la mayor cantidad de información posible para monitorear el estado del logro académico en el *Campus* Iztacala. Sólo un segmento del sistema de evaluación es presentado y empleado para efectos de este trabajo (el subprograma IV.1, que se refiere a evaluación masiva de estudiantes); además, este trabajo se centra en el conjunto de métodos e instrumentos como objeto de estudio, no en los resultados obtenidos por la población escolar.[3]

El objetivo de este trabajo es exponer y analizar la estrategia llamada "academias de evaluación del logro escolar" en cuanto a su eficacia en desarrollar un sistema de evaluación masiva de estudiantes que permita diseñar instrumentos con índices de confiabilidad y validez adecuados, y que sea pertinente en lo que evalúa, cómo lo evalúa y el uso que se da a la información obtenida por el proceso. Se pretende responder a la pregunta: La estrategia de academias de evaluación del logro escolar ¿es más eficaz y eficiente que los métodos tradicionales para construir instrumentos de medición sumativa?

La resolución de esta pregunta requiere la integración de conocimientos de psicología educativa y de teoría de la medición, así como del punto de vista de la modificación de conducta. En efecto: la eficacia y eficiencia del método solamente son comprensibles y justificables en términos de cómo modificarán su conducta los actores educativos, tanto alumnos, como profesores y autoridades ante un método de evaluación que pretende conocer y socializar con la mayor exactitud posible el resultado de la práctica

---

[3] La importancia de esta aclaración ya se ha demostrado en una presentación del proyecto de este trabajo. Si bien para el trabajo institucional los métodos e instrumentos de evaluación no son más que los medios por los que se llega a la información del estado académico -la verdadera finalidad-, para el trabajo de maestría se centra la investigación en la búsqueda de procesos que mejoren la calidad de los medios de evaluación.

educativa, con la intención de que sea más fácil proponer acciones de cambio en su caso, y de mejoramiento. De hecho, la evaluación propuesta sigue claramente las características de la evaluación psicológica, y se considera un caso específico de la misma: pretende tener un enfoque científico, incluye la exploración y análisis de un grupo de sujetos, los niveles de complejidad de la observación son adaptables, y sirve a fines de detección, diagnóstico, investigación, clasificación, intervención, cambio y valoración, características que, en suma, llevan a un proceso de toma de decisiones, pues de otra manera, la evaluación sería estéril, ya que no tiene un fin en sí misma, sino en el mejoramiento.

El capítulo 1 examina las nociones de evaluación y de evaluación educativa. Deslinda ambos procesos de la simple medición y de la calificación, y explica por qué es importante que la evaluación tenga una razón de ser previa a su ejecución.

En el capítulo 2 se tratan a profundidad las cualidades de validez y confiabilidad que, si bien pueden asociarse a la teoría clásica de la medición, tienen vigencia al hablar de la evaluación debido a que la única manera de establecer juicios de valor útiles es basando éstos en datos válidos y confiables. Se explora también la cualidad de pertinencia, que no corresponde a las teorías de la medición, como las dos primeras, pero debe ser considerada al establecer un plan institucional de evaluación del logro escolar.

En el capítulo tres se presenta la estrategia de las Academias de Evaluación de Logro Escolar, y se le compara con la estrategia seguida por el CENEVAL para generar su exámenes EGEL-Psicología; también se enuncian las razones por las que se considera que la estrategia de las Academias conlleva procesos de validación de contenido.

El capítulo 4 explica el método que se usó para comparar estadísticamente algunos exámenes generados con la estrategia de las Academias y exámenes generados bajo estrategias tradicionales. El capítulos 5 describe los resultados de esta comparación.

Finalmente, en el capítulo 6 se establecen las conclusiones y se comentan los últimos avances que ha tenido la estrategia, así como los posibles desarrollos futuros.

Capítulo 1 LA EVALUACIÓN EDUCATIVA

En este capítulo se hará un breve recorrido teórico por las nociones de evaluación y evaluación educativa. Se considera importante porque popularmente se considera a la evaluación como idéntica a medición, examen, calificación. Tras la lectura, podrán entenderse claramente las diferencias. Además, señalamos las características que, a nuestro criterio, deberían tener los ejercicios de evaluación en el aula.

La ANUIES (1997) definió desde 1984 a la evaluación como "un proceso continuo, integral y participativo que permite identificar una problemática, analizarla y explicarla mediante información relevante. Como resultado, proporciona juicios de valor que sustentan la consecuente toma de decisiones. Con la evaluación se busca el mejoramiento de lo que se evalúa y se tiende a la acción" (p. 78). Miras y Solé (1990) señalan que la evaluación es una actividad encaminada a provocar modificaciones en un objeto, situación o persona.

Conviene aclarar que, los elementos específicos que, a nuestro juicio, permiten que un proceso pueda ser llamado de evaluación, son los siguientes:

1. Razones (puede bastar una) para realizar el proceso. Normalmente, implican el mejoramiento.
2. Un objeto de evaluación definido de antemano.
3. Una serie de pasos para obtener información relevante sobre el objeto de evaluación.
4. Análisis de la información relevante obtenida.
5. Juicios de valor que se emite tras considerar lo analizado en su contexto.
6. Acciones basadas en los juicios de valor, que dan sentido a las razones del punto 1, y que normalmente implican mejoramiento: "el proceso de evaluación debe aportar no sólo elementos de juicio del quehacer institucional que ofrezcan una crítica estática, sino también elementos de aprendizaje derivados de la propia experiencia, que favorezcan la crítica dinámica y estimulen la reorientación de una estrategia que permita corregir, en lo posible, el curso de desarrollo hacia la consecución de los objetivos planteados" (Casillas Muñoz, 1995, pp. 46, 47)

Desde nuestro punto de vista, estos incisos distinguen claramente a la evaluación de la simple medición (que contendría apenas los puntos 2, 3 y 4), así como de la acreditación, la certificación, la asignación de puntajes y la calificación, procesos todos que pueden estar vinculados a la evaluación pero que ni le son necesarios, ni la contienen.

En el contexto de la educación, la evaluación tiene una larga trayectoria, de siglos, en lo referente a los alumnos; si bien 'desde la antigua China pueden rastrearse los estudios primigenios, hace sólo aproximadamente cien años que se comenzaron a encarar sistemáticamente los problemas de medición psicológica'. (Nunally, 1970, p. 47). "El examen escolar estandarizado tal como se lo conoce actualmente llegó a generalizarse sólo en los últimos cincuenta años. Antes era habitual que el profesor calificara al estudiante por las impresiones que había recogido de él o basándose en exámenes orales" (Nunally, Op. Cit., p. 61). Más recientemente (apenas en los años 70) su uso se ha ido extendiendo a los profesores, a los planes y programas, a las instituciones como un conjunto, a sistemas o subsistemas educativos nacionales, e incluso a los egresados (aunque con estos últimos el proceso es conocido como "seguimiento", no como evaluación). En la educación superior, es apenas hace dos décadas que la evaluación cobra la importancia que actualmente se le da:

> "Los procesos de planeación y evaluación (...) se impulsaron con mayor interés por parte del gobierno federal a partir de la década de los ochenta, con el fin de reorganizar las tareas de las instituciones de educación superior e iniciar una revisión a fondo de la eficiencia y la eficacia con que desempeñan sus funciones. Esta necesidad surgió como un imperativo ante los –cada vez más frecuentes– cuestionamientos a la calidad de los servicios que ofrecen las instituciones, sobre todo las de carácter público." (Casillas Muñoz, 1995, p.5)

Como ejemplo de los principios que deben conducir el uso de la evaluación en la educación, tomemos las declaraciones de la Asociación Nacional de Universidades e Instituciones de Educación Superior: "La evaluación y la acreditación son procesos que a escala mundial han sido reconocidos como medios idóneos para el mejoramiento de los sistemas de educación superior" (p. 64), "la evaluación forma

parte de un proceso de planeación para mejorar la calidad del sistema de educación superior" (p. 59, ANUIES, 1997).

Actualmente, la evaluación del logro escolar (es decir, del aprovechamiento de los estudiantes) es una de las tareas que ocupan el tiempo y los esfuerzos no solamente de profesores y alumnos, sino de grupos de investigación e incluso de grupos multidisciplinarios, que hacen estudios transnacionales que permitan comparar el nivel de logro de estudiantes de diferentes partes del mundo. "Durante los últimos 40 años se han realizado [...] estudios, principalmente dirigidos por la International Association for the Evaluation of Educational Achievement (IEA, Asociación Internacional para la Evaluación del Rendimiento Académico) y la International Assessment of Educational Progress (IAEP, Asociación Internacional para la Evaluación del Progreso Educativo) del Education Testing Service (Servicio de Evaluación Educativa)" (Madrid: Ministerio de Educación, Cultura y Deporte, INCE, 2000). A continuación se reseñan dos experiencias recientes y muy conocidas de evaluación educativa de escala mundial:

El estudio Education at a Glance, llevado a cabo cada dos años desde 1992 por la Organización para la Cooperación y el Desarrollo Económicos (OECD, por sus siglas en inglés) incorpora los esfuerzos de políticos, servidores públicos, investigadores, educadores y estadísticos de 30 países (nuestro país empezó a participar desde 1995) con la intención de obtener indicadores de los sistemas educativos que lleven a una reflexión que permita mejorar el nivel de las escuelas, ya que 'la educación es una inversión que puede ayudar a promover el crecimiento económico, contribuir al desarrollo personal y social y reducir la inequidad social' (OECD, 2000, p. 9). Entre los diferentes tipos de dominios que Education at a Glance incorpora[4], se encuentra el logro escolar (student achievement) en disciplinas básicas, específicamente

[4] Otros dominios que se incluyen en Education at a Glance son, por ejemplo y para abrir el apetito intelectual de los interesados en educación: el contexto demográfico de los sistemas educativos; los costos de la educación y los recursos humanos y financieros dedicados a ella; el acceso a la educación; el ambiente escolar y los procesos educativos; egresados; alfabetización...

matemáticas y ciencia. Debido a su naturaleza comparativa entre dominios y naciones, los resultados obtenidos en este estudio son muy interesantes. Por ejemplo, se encontró que Japón y Corea puntean muy alto comparado con el promedio de los demás países en matemáticas (resultado que es más impresionante al relacionarlo con el hecho de que ambos países son más bien moderados con respecto a la media en cuanto a su gasto en educación); que en todos los países los niños obtienen mejores puntuaciones que las niñas en ciencias; y que los niveles de logro escolar no parecen estar directamente relacionados con los indicadores de alfabetización de los países (cfr. OECD, Op. Cit., pp. 198-217). En el contexto de la OCDE, estos hallazgos tienen importancia porque las economías modernas dependen en gran parte de los descubrimientos científicos y la innovación tecnológica.

Otro programa de la OCDE, el PISA (OECD, 2002)(Programme for Internacional Student Assessment), también de amplio espectro (cubre 32 países, a través de la evaluación de 265,000 estudiantes: México es uno de los participantes) se concentra más en las cuestiones del logro escolar y no tanto en las contextuales. Este estudio, que se llevó a cabo por primera vez en el año 2000, para muestras de estudiantes de 15 años, y pretende replicarse cada 3 años, tiene como centro la evaluación del logro escolar en los siguientes términos:

- No evalúa lo que los estudiantes aprenden en la escuela, sino qué tan bien se desempeñan más allá de los programas escolares.
- Sus dominios primarios de evaluación son la alfabetización, la noción matemática y la noción científica, desde el enfoque de las habilidades adquiridas, más que de los conocimientos, aplicándolas a problemas de la vida real.
- Además, PISA incluye la evaluación de las actitudes de los estudiantes y sus maneras de acercarse al conocimiento, así como datos contextuales de sus escuelas.
- En cada país las muestras deben ser de entre 4,500 y 10,000 alumnos.

PISA tiene, como resultados básicos, los siguientes (Madrid: Ministerio de Educación, Cultura y Deporte, INCE, 2000):

- Un perfil básico de los conocimientos y destrezas de los alumnos al término del periodo de escolarización obligatoria.
- Indicadores contextuales que relacionan los resultados con las características de los alumnos y los centros educativos.
- Indicadores de tendencias que muestran los cambios en los resultados a lo largo del tiempo.

Probablemente, el fin más claramente definido y aplicado de la evaluación del aprendizaje es el administrativo: conocer el número de estudiantes que aprueban y reprueban, sacar promedios de calificación y guardar expedientes de los resultados a examen de los estudiantes son tareas que todas las instituciones educativas, aun las más liberales[5], tienen que llevar a cabo. Sin embargo, existe una serie de funciones que puede cumplir la evaluación del aprendizaje y que en muchos casos no se conocen. Carreño Huerta (1991) plantea al menos 9 funciones distintas de la evaluación, entre las que se encuentran, por ejemplo "Conocer los resultados de la metodología empleada en la enseñanza y en su caso, hacer las correcciones de procedimientos pertinentes" y "Dirigir la atención del alumno hacia los aspectos de mayor importancia, conclusivos o centrales en el material de estudio" (ambas citas, en p. 3). Nuestra cita favorita con respecto a las funciones de la evaluación, debido a que desde nuestro punto de vista deja claro cuál es el objetivo primordial de la evaluación, es la siguiente: "Sin embargo, teóricamente, la evaluación habría de ser compañera inseparable de cualquier acción educativa. No por la razón que antes señalaba Buisson (la presión hacia la excelencia), sino por razones cibernéticas: conocer la marcha de los procesos para poderlos ajustar a las diversas situaciones y necesidades" (Zabalza Beraza, 1990). Saber cómo van las cosas y qué se puede hacer para mejorarlas es la

[5] A este respecto, hemos notado que en escuelas activas y Montessori de educación básica, aunque no se haga énfasis en las calificaciones frente a los alumnos, la SEP exige que exista un expediente indicando la "calificación" de cada alumno en cada materia.

razón de ser de la evaluación. Sin ella, no tiene sentido llevar a cabo un proceso tan largo y costoso.

Varios puntos de decisión se plantean en el proceso de evaluación. Estos han sido, tradicionalmente, resueltos con base en criterios poco claros, e incluso sin considerar siquiera los problemas teóricos y de aplicación, planteando las decisiones en términos meramente técnicos y/o administrativos. García Cortés (1979) explica la gran importancia que tiene determinar, para cada caso específico, las respuestas a ¿para qué evaluar? y ¿qué evaluar? Responder estas dos preguntas señala criterios que generalmente sirven de gran ayuda para tomar decisiones sobre la manera de operar un programa de evaluación, y las subsiguientes preguntas, tales como: ¿quién debe decidir las áreas a evaluar, los contenidos de las áreas a evaluar, los métodos de evaluación? ¿quién debe crear los instrumentos de evaluación? ¿qué características deben tener los ítems de los instrumentos de evaluación? ¿en qué momento se pueden considerar adecuados los instrumentos de evaluación? ¿qué criterios deben tomarse en cuenta para analizar la información obtenida? ¿qué segmentos y agrupaciones de la información obtenida son más útiles, y a quién?

La pertinencia y utilidad del proceso dependen de que se den respuestas razonables a estas preguntas, en los momentos oportunos, perfilando un sistema. Estas "respuestas razonables" varían según el contexto: "La evaluación tiene un carácter relativo a cada institución , al tener como eje sus propios objetivos y metas, así como las políticas y estrategias para conseguirlos" (ANUIES, 1997, p.78).

Como último punto para alimentar nuestra reflexión, citamos seis axiomas en la evaluación educativa, de Bertoni, Poggi y Teobaldo (1997, p.103) que establecen un marco de realidad con el que trabajar:

- "no se pueden evaluar realidades que nos son totalmente desconocidas.
- No se pueden evaluar, tampoco, realidades que nos son indiferentes.

- La evaluación es imposible si no se tienen los medios de discernir los diferentes aspectos de la realidad implicada en el acto educativo.
- Todo evaluador es "portador" de un modelo de evaluación, ya sea explícito o implícito.
- Los criterios de evaluación siempre proponen expectativas, positivas o negativas, sobre los resultados esperados en las producciones de los alumnos ("se espera que el alumno realice...)
- Como la evaluación se refiere a una norma o criterios, ya sean preconstruidos, ya sea construidos durante el proceso mismo de evaluación, por lo menos en principio esos criterios deberían ser comunes al evaluador y al evaluado."

Tomando en cuenta los materiales y la reflexión enunciados en este capítulo, nuestro criterio ante las preguntas fundamentales señaladas párrafos arriba (¿para qué evaluar? y ¿qué evaluar?, y todas sus ramificaciones), para el caso de la evaluación del aprendizaje en el contexto de la educación superior, es el siguiente:

- evaluar el aprendizaje no significa valorar a los alumnos, sino a las habilidades y contenidos que éstos han adquirido en una situación de aprendizaje específica. Es importante clarificar esto pues al obtener información sobre lo que el alumno sabe, aprendemos sobre la interrelación del medio con el estudiante, y no sólo de la capacidad intrínseca del mismo.
- el objetivo primordial de la evaluación deberá ser el de hacer evidentes las carencias e incluso, de ser posible, sus causas, para mejorar las condiciones y procesos en que se lleva a cabo el proceso de enseñanza aprendizaje.
- las áreas a evaluar y los contenidos de dichas áreas deben ser determinados por quienes se encuentran en estrecha relación con los evaluados durante su proceso de aprendizaje (es decir, sus profesores).

- los métodos e instrumentos de evaluación se deben discutir entre expertos en evaluación y expertos en lo que se va a evaluar (estos últimos son, otra vez, los profesores).
- los ítems de los instrumentos de evaluación deben tener características que, desde un punto de vista científico, permitan ponerlos a prueba y ser mejorados constantemente, y que simultáneamente ofrezcan información completa y útil sobre el objeto de estudio (por ejemplo, creemos que las preguntas directas de opción múltiple siguen siendo la mejor elección para evaluar masivamente el logro escolar de los estudiantes).
- Se deben utilizar los análisis estadísticos como herramienta de análisis de resultados, pero no como única herramienta; no debe subestimarse el valor de la reflexión colegiada como la más valiosa herramienta de análisis.
- Debe tratarse de involucrar a los interesados (los actores educativos y los tomadores de decisiones al interior de las escuelas) lo más posible en todo el proceso.

Con esta declaración de principios, damos por terminada nuestra exposición de lo que es la evaluación educativa y lo que quisiéramos que fuera. El siguiente capítulo analiza algunos elementos técnicos que deben ser tomados en cuenta para que un sistema de evaluación posea cualidades que le permitan cumplir sus objetivos cabalmente.

Capítulo 2 VALIDEZ, CONFIABILIDAD Y PERTINENCIA EN LAS PRUEBAS DE LOGRO ESCOLAR[6]

Es conocido el hecho de que un instrumento de medición (desde el más simple al más complejo) que va a ser utilizado en repetidas ocasiones y para sacar conclusiones haciendo comparaciones, debe cumplir ciertos criterios de confiabilidad y validez, así como ser pertinente. En este capítulo se abordarán algunas nociones básicas para comprender las propiedades que debe tener cualquier instrumento de medición, de la disciplina que sea. Estas propiedades garantizan que el uso del instrumento sea adecuado, y que los datos que arroje sean aproximaciones válidas a la realidad.

Confiabilidad

La confiabilidad, según definición clásica, se refiere a la estabilidad del instrumento a través del tiempo y de las muestras. Sabemos que las condiciones de los procesos educativos son dinámicas, así que esta primera definición no parece sernos muy útil en la escuela, dado que será muy difícil pedir estabilidad en los resultados de desarrollos que de por sí van a ser cambiantes. Una segunda aproximación refiere que la medida confiable es aquella que se encuentra libre de error. (cfr. estas dos definiciones con National Academy of Sciences, 1991, pp. 116-117, o con Aiken, 1996, p.87) Sin embargo, aunque esto parece ser suficientemente exacto (nunca totalmente) en las ciencias naturales (por ejemplo, en la medición del contenido de sodio en un compuesto), en las ciencias sociales es muy ingenuo pensar en alcanzar la exactitud (puede incluso plantearse la duda de la posibilidad o la necesidad de ella a nivel filosófico). Una definición que nos parece más viable para la tarea que nos ocupa es la de considerar semejante a lo que es semejante y diferente a lo que lo es, con rangos razonables de error. Lo razonable de los rangos de error dependerá de la naturaleza de lo evaluado y de los objetivos para los cuales es evaluado. Consideramos que esto

[6] Una versión de este capítulo fue presentada en el "Foro de Análisis sobre los procesos de evaluación" en la ENEP

acerca la noción cuantitativa de confiabilidad a la noción cualitativa de imparcialidad. Para efectos prácticos, la confiabilidad se ha medido a partir de uno de tres aspectos (Dawson, s.d, pp. 4-6): a) **Confiabilidad como estabilidad**: si los individuos responden de manera consistente entre una aplicación de la prueba y otra, la correlación entre ambas aplicaciones será alta. Este es el conocido método test-retest para evaluar la confiabilidad; b) **Confiabilidad como equivalencia**: Utilizando las mismas estrategias estadísticas, se aplican dos pruebas que se pretenden equivalentes a los mismos individuos, en tiempos cercanos. Si los individuos responden de manera equivalente a una prueba y otra, la correlación entre los puntajes de las dos pruebas para cada par de individuos será alta[7]; c) **Confiabilidad como consistencia**: Debido a que en la práctica es difícil aplicar la misma prueba a los mismos individuos más de una vez, así como disponer de dos versiones presumiblemente equivalentes de la misma prueba, puede medirse la consistencia interna de una prueba como índice de su confiabilidad. Un método muy utilizado es el de dividir al instrumento en dos mitades, y considerar cada mitad como una versión de la prueba, de tal manera que pueda ser tratada como en los método de test-retest y los de pruebas equivalentes. Un método alternativo, que considera todas las posibles divisiones en mitades de una prueba es el Alpha de Cronbach (de cuyo modelo existen versiones más limitadas, como el método Inter – item de Kuder – Richardson (Aiken, 1996)). Aunque los métodos basados en la consistencia tienen la limitante de que no son capaces de medir el error causado por cambios en las condiciones o tiempos de aplicación, el Alpha de Cronbach es el método que se preferirá para medir confiabilidad en este trabajo, debido a que es muy reconocido; en nuestro diseño no hay la posibilidad de hacer test – retests ni versiones alternativas de las pruebas, y evita el sesgo de hacer una partición por mitades afortunada o desafortunada.

Iztacala, 8 y 12 de Junio de 1998, con el título "Criterios de Evaluación Educativa"

[7] Una versión de este método es empleada por los psicólogos clínicos y experimentales, al utilizar varios observadores de una sola serie de hechos y luego ver el índice de coincidencias – divergencias entre observadores.

Validez

De entre las características que permiten un análisis certero de la información obtenida por procedimientos evaluativos, probablemente la más importante sea la validez: "el conocimiento del grado de validez del instrumento es necesario para que los datos obtenidos con él puedan usarse significativamente" (Magnusson, 1975, p.77). Los Estándares para la Evaluación Educativa y Psicológica por Medio de Pruebas (APA, 1985) señalan que 'la validez es la consideración más importante en la evaluación por medio de pruebas. El concepto se refiere a la pertinencia, significación y utilidad de las inferencias específicas que se hagan de los puntajes de una prueba.' (p. 9)

Es muy difundida la definición clásica de validez en instrumentos de evaluación que indica que éstos son válidos cuando miden lo que pretenden medir (p. ej., la cita Magnusson, Op. Cit., p. 153). Sin embargo, el concepto de validez aparentemente tan simple se encuentra en el centro de una polémica que aún actualmente se lleva a cabo. Gray (1997), haciendo una pequeña revisión, señala que:

> 'en 1949 Cronbach declaró que la definición de validez como "la extensión con que una prueba mide lo que pretende medir" era comúnmente aceptada, aunque él prefería una ligera modificación: una prueba es válida en el grado en que sabemos qué mide o predice. Cureton (1951) provee una definición similar: La cuestión esencial de la validez en las pruebas es qué tan bien realizan la tarea para la cual se les esté usando. La validez es definida entonces en términos de la correlación entre los puntajes de una prueba y los "verdaderos" puntajes del criterio. La perdurable definición de Anastasi (usada desde 1954), "la validez es qué mide una prueba y qué tan bien lo hace", es también citada ampliamente.' (p. 1)

Gray (Op. Cit.) señala también que, aunque Cronbach no volvió a intentar definiciones del término después de 1949, en 1971 hizo un comentario que reavivó la controversia: 'validación es el proceso de examinar la precisión de una predicción o inferencia específica hecha a partir de los puntajes de una prueba' (p. 1), o bien, como señalan él y otros autores "la validez se refiere no a las puntuaciones o datos en sí mismos, sino a las inferencias que se hagan a partir de ellos bajo determinadas

circunstancias" (Cronbach, Vernon, cit. en Silva y Martorell, 1991, 0. 113); "lo que se valida no es el instrumento, sino la interpretación de los datos obtenidos por medio de un procedimiento especificado" (Aragón, 1990, p.108); 'la validez depende de la "adecuación y pertinencia de inferencias y acciones" basadas en los resultados de la evaluación' (Messick, 1989, en Linn y Baker, 1996, p. 5).

Finalmente, es importante señalar que, aunque muchos autores (p. ej., Rudner, 1993; Niemi, 1996; Aragón, 1990; Tourón, 1989; Burns, 1996; GAO, 1991) reportan al menos tres tipos "clásicos" de validez, e incluso en algunas versiones de los estándares de la APA llegaron a considerarse varias decenas de tipos de validez, actualmente existe una tendencia a considerar un tipo único de validez (cf. Gray, Op. Cit.; Silva y Martorell, Op. Cit., quienes incluso sugieren que el concepto de confiabilidad también es mucho más cercano al de validez de lo que se ha pensado), de la cual, eso sí, se obtienen distintos tipos de evidencias. Con respecto a la relación entre la validez y la confiabilidad, Siegel (1983) señala que "en teoría, la validez potencial máxima de una prueba se obtiene de la raíz cuadrada de su confiabilidad" (p. 151). Suena lógico considerar que la estabilidad de los datos, o el que estén lo más libres de error que sea posible si se prefiere, influye en la capacidad que tengan los datos de ser interpretados correctamente. A continuación, se comentan brevemente los tipos de validez (o de evidencia de validez, mejor dicho) más usuales:

Validez de contenido

Este es el aspecto de la validez que probablemente requiere ser más cuidadosamente trabajado cuando se planea un examen global de conocimientos. La validez de contenido se refiere a 'la extensión con que los ítems de una prueba o situación evaluativa son representativos de las habilidades del dominio que se intenta medir' (Rudner, 1993, p. 2), es decir, que debe demostrarse que las preguntas o problemas que se pide afrontar a los evaluados son las suficientes como para hacer inferencias sobre el dominio o "universo de generalización" al que pertenecen dichas preguntas. Esta noción cuestiona el diseño "al vapor" de las pruebas de aprovechamiento, dado que su fabricación apresurada puede causar (causa, de hecho) el sesgo de un examen hacia algunas áreas del contenido revisado en el periodo a evaluar y la exclusión por

olvido de otras áreas del mismo. Tener claro de antemano para qué va a servir la información que se va a generar (cf. García Cortés, 1979), no sólo da sentido a la evaluación, sino también apoya a la validez de contenido, ya que así se puede definir si la manera en que se está obteniendo una muestra de la información posible es adecuada para considerar que representa a todo el universo de generalización al que se quiere hacer extensiva. Una práctica común que ayuda a mejorar la validez de contenido en una prueba es el uso de tablas de especificaciones durante la planeación de una prueba.

Validez de criterio

Este aspecto de la validez se refiere "a las relaciones de una determinada prueba con otras variables, medidas o criterios" (Tourón,.1989, p. 738). Y contiene una lógica muy clara: si digo que tener determinada cualidad debe reflejarse en el instrumento que yo he creado, pero que también se refleja en otros instrumentos o genera determinadas consecuencias, es lógico intentar probar dicho aserto observando las relaciones existentes entre los puntajes de mi instrumento y los otros instrumentos, o las consecuencias. Esta evidencia de validez se encontraba en el pasado dividida en validez predictiva (cuando, con los puntajes de un instrumento, se predice con cierto grado de certeza la ejecución futura de los evaluados en una tarea distinta) y validez concurrente (cuando los puntajes de un instrumento se asocian a los de otros instrumentos o sucesos actuales), pero a partir de la edición de 1966 de los Estándares para la Evaluación Psicológica y Educativa por medio de Pruebas ambos tipos se agruparon (cf. American Psychological Association, 1985).

Reflexionar sobre la validez concurrente y predictiva en el ámbito educativo es útil porque nos permite encontrar el sentido de las evaluaciones diagnósticas[8], bajo la pregunta: la información que estoy obteniendo sobre los alumnos ¿me permite hacer inferencias razonables sobre su aprovechamiento posterior? ¿me permite planear una

[8] Todo el planteamiento es aplicable también a la evaluación para selección de alumnos, pero dado que no creo que ésta sea adecuada, al menos en educación básica, decidí no poner ejemplos de este tipo de evaluación, para no favorecer que sea considerada.

estrategia adecuada para enfrentar las limitaciones de los estudiantes con predicciones acertadas sobre su éxito posterior? Si no se puede contestar afirmativamente a estas preguntas, probablemente la evaluación diagnóstica se está haciendo solamente para llenar un expediente.

Otro caso en que la validez de criterio debe ser tomada en cuenta es el de utilizar pruebas que fueron concebidas, por ejemplo, como evaluaciones sumarias de un curso para otros usos, por ejemplo, como prueba diagnóstica del siguiente curso. 'Un tema importante en ambientes educativos es el grado en que la evidencia de validez obtenida (o construida, N. del A.) en una situación puede ser generalizada o transportada a otra situación sin mayor estudio de la validez en la nueva situación' (Raflison, 1991, p. 1). En el caso de exámenes globales, puede constatarse la validez de la medida comparando, por ejemplo, los resultados obtenidos con las calificaciones promedio de los estudiantes. Estas dos mediciones deberían correlacionarse, puesto que ambas dan cuenta del nivel aprendizaje que ha tenido el estudiante, aunque han sido tomadas por diferentes medios y en diferentes momentos educativos.

Validez de constructo

> 'Se ha sugerido que la validez de constructo abarca tanto a la validez de criterio como a la de contenido. Sheperd anotó que la validez de constructo incluye los requisitos teóricos y empíricos de la validez de contenido y de criterio. Anastasi (1986) coincide en que la validez de constructo subsume los requisitos de la validez de contenido y de criterio.' (Stapleton, 1997, p. 3)

La validez de constructo se refiere a 'la medida en que un método o instrumento de medición representa con precisión a un constructo, y produce una observación distinta a la causada por la medición de otros constructos' (GAO, 1991, p. 92). Esto requiere, por supuesto, que exista un "constructo" en las explicaciones que queremos dar a un puntaje. Un constructo es un atributo, usualmente inobservable, que da explicación teórica a las razones por las cuales se dan ciertos fenómenos. Un ejemplo conocido de

constructo es el átomo que, antes de la invención de los microscopios electrónicos con los que puede percibirse, ya permitía explicar una serie de fenómenos físicos basándose en la suposición de su existencia. Para el caso de la educación, "aprendizaje de los contenidos curriculares", "aprovechamiento", "logro escolar", son algunos constructos que pretenden explicar las razones por las cuales algunos estudiantes obtienen mayores puntajes en las pruebas y algunos egresados logran desempeñarse mejor en estudios posteriores o en empleos. Debe notarse la circularidad del planteamiento en que determinada prueba mide el logro escolar; logro escolar que no es otra cosa que aquello que hace que los resultados de la prueba sean diferentes para distintos alumnos; es por esto que deben buscarse métodos alternativos de medir el mismo constructo, y relacionar sus resultados y debe hacerse un análisis muy cuidadoso de por qué se considera que un ítem da evidencias de la existencia de determinado constructo. Como puede observarse, esto da zonas de traslape entre la validez de constructo y los otros dos tipos "tradicionales" de validez, explicados anteriormente. Posiblemente por ello, se considera que las diferentes evidencias de validez se deben apoyar unas a otras y se tiende a unificar el otrora muy disgregado concepto de validez.

Tradicionalmente, se ha validado una explicación referida a un constructo 'determinando el grado en que ciertos conceptos explicatorios o constructos afectan la ejecución de una prueba' (Niemi, 1996, p. 21).

Reflexionar en la validez de constructo en su relación con las pruebas escolares es útil porque nos lleva a procurar definir con la mayor claridad los constructos educativos enunciados arriba, y a tratar de diseñar las situaciones evaluativas de tal modo que sea razonable suponer que las variaciones en los puntajes obtenidos por los alumnos sean

debidas a la ausencia o presencia de distintos grados de dichos constructos, y no a otras variables poco relacionadas con el aprendizaje[9]. Por ejemplo, Williams (1988) y Koretz (1988) (citados en Williams, 1989.) señalan que los cambios en la ejecución de los alumnos ante una prueba repetida tras un periodo de tiempo pueden no deberse a cambios en la adquisición de conocimientos, sino a materiales defectuosos de preparación para una prueba, instrucción dirigida a los ítems de la prueba, entrenamiento "especial", o fraudes; los autores acotan que los instrumentos y las circunstancias en que se aplican deben diseñarse e implementarse de modo que, si existen incrementos en los puntajes, éstos representen un verdadero cambio en el desempeño de los evaluados.

En resumen con respecto a la validez en general, debemos considerar como una cualidad primordial de las pruebas la posibilidad de extraer de manera correcta y verdadera el significado de sus puntajes. Dado que esto no depende sólo de la prueba sino también de las circunstancias de aplicación y los objetivos de la misma, diferentes aspectos de esta cualidad pueden ser considerados. Aunque esto puede parecer sencillo cuando los instrumentos de medición son muy cercanos a la realidad física, la tarea se complejiza conforme el objeto de evaluación se vuelve abstracto o difícil de observar directamente. Tal es el caso de la evaluación del aprendizaje.

Pertinencia

La pertinencia del procedimiento de evaluación no es una categoría clásica para valorar a los instrumentos y métodos; al menos no con ese nombre. Sin embargo, podemos observar que una serie de autores hacen señalamientos en el sentido de que no se pierdan de vista las razones por las cuales se llevan a cabo los ejercicios concretos de evaluación en la valoración de éstos, así como una apreciación de si se está dando respuesta adecuada a estas razones. Por ejemplo, los Estándares para la Evaluación

[9] ¿Quién no ha tenido, por ejemplo, un profesor que "baje puntos" en el examen de matemáticas por estar hablando, o comiendo en clases?

Educativa y Psicológica a través de pruebas (APA, 1985) señalan que 'el uso correcto de pruebas bien construidas y validadas otorga una base mejor para tomar decisiones sobre los individuos y los programas que otros métodos' (p. 1). Este mismo escrito dedica un apartado completo (la parte II: Estándares profesionales para el uso de pruebas) a definir cuándo las pruebas se usan correctamente y cuándo no.

La pertinencia es el aspecto más cualitativo de la valoración de un instrumento de evaluación, y apunta a la participación de la comunidad implicada.

Existen tres puntos de especial importancia en cuanto a la pertinencia de un procedimiento de evaluación: 1.- Que el tipo de información arrojada sea realmente un indicador útil sobre los conocimientos y/o habilidades de la población; 2.- Que existan criterios fundamentados para interpretar las cifras obtenidas en la evaluación masiva; 3.- Que la información obtenida llegue a los destinatarios que pueden darle utilidad, es decir, los profesores, planificadores académicos al interior de la escuela y los propios estudiantes. A continuación, se desarrollan estos puntos:

1.- Que el tipo de información arrojada sea realmente un indicador útil sobre los conocimientos y/o habilidades de la población:

Existe una discusión importante con respecto a los instrumentos de evaluación que se utilizan en educación. En realidad, el origen de la discusión está en el pseudoproblema de lo cuantitativo vs. lo cualitativo[10]. Algunos autores señalan que la evaluación no debe hacer uso de la tecnología de medición generada por la psicometría y perfeccionada constantemente pues "se minimiza tanto el proceso mismo de la evaluación del aprendizaje como la noción de aprendizaje y la de docencia." (Díaz Barriga, 1982 p.23); otros plantean problemas técnicos en el uso de ciertos tipos de

[10] Al respecto, Silva y Aragón (1997) señalan: "lamentablemente, esta confrontación entre cualitativistas y cuantitativistas ha obstaculizado el avance de las ciencias sociales, puesto que la disputa se ha centrado en 'cualitativo sí, cuantitativo no' o bien 'cuantitativo sí, cualitativo no', más que en desarrollar un cuerpo sistemático de conocimientos que den una coherencia tanto a los hallazgos cualitativos como cuantitativos" (p. 174). El artículo apunta muy claramente a ver cómo, al menos desde los puntos de vista filosófico y metodológico, las dificultades de plantear una unificación de los dos puntos de vista son salvables.

evaluación "objetiva", por ejemplo que sólo se mide lo que el alumnos memoriza, o la posibilidad de acertar por azar (cfr. Fermín, 1971, p. 9 y 10); finalmente, otros autores, reconociendo los problemas de "el hiato indudable entre la medida y lo que pretendemos medir" (Tourón, 1989, p. 735) confían sin embargo en el uso del método científico para la valoración escolar y generan estrategias cada vez más refinadas para salvar los problemas mencionados (Cf. Tourón, Op. Cit.; Tirado y Serrano, 1989; Rodríguez y García, 1982). Nosotros consideramos importante rescatar nociones de cada uno de estos planteamientos que equilibren una práctica evaluativa eficaz, eficiente y útil a partir de una autocrítica. Consideramos necesario a) estudiar a profundidad, desde puntos de vista no técnicos (sino desde las teorías del desarrollo y el aprendizaje) las características que debe tener un reactivo que elicite procesos de reflexión más que de memorización; b) generar un conjunto de sugerencias y ayudas para los diseñadores de reactivos que les permita hacer uso de las conclusiones del punto anterior; y c) hacer bancos de reactivos, para los instrumentos de evaluación del aprendizaje, que cumplan con estas características.

2.- Que existan criterios fundamentados para interpretar las cifras obtenidas en la evaluación masiva:

Una problemática común entre los que atacan el problema de la evaluación desde un punto de vista social y/o filosófico, que en cambio es poco tocado por quienes tienen el punto de vista únicamente técnico, es el criterio de pase/reprobación en el caso de evaluaciones con fines de acreditación, o el criterio de "aceptabilidad/inaceptabilidad" en caso de evaluaciones para toma de decisiones. Sabemos que existen en este sentido juicios "por criterio" y juicios "por norma". En el primer caso, se establece de antemano el mínimo aceptable, que depende de una discusión teórica de lo que se va a evaluar, y en el segundo se juzga cada caso individual con base en la cercanía o lejanía que tenga con la media (por ejemplo, número de desviaciones estándar), y el sentido de esta distancia (positivo o negativo). En papel, estos criterios parecen fáciles de aplicar, pero en la práctica, vale la pena reflexionar profundamente en los motivos y las consecuencias de permitir, por ejemplo, que sean acreditados estudiantes de medicina con calificaciones apenas pasables, además de relativas (dado que ante el examen de

una escuela podrían sacar altas calificaciones y ante el de otra podrían sacar calificaciones bajas). En efecto, no hay una estandarización en la dificultad que deben tener este tipo de pruebas, ni normas o consejos de uso generalizado para establecer los criterios. Por todo ello, el conjunto de la comunidad escolar debe dedicar tiempo a la reflexión de este problema, aterrizándolo en programas concretos en que se trabaje y tomando decisiones con respecto a los criterios a emplear en ellos. La relatividad llegó a la física –una de las ciencias más duras y clásicas– hace unos ochenta años; tal vez ya es tiempo de que llegue a la educación: no existen criterios ni fórmulas universales para llevar a cabo las tareas evaluativas, ni deben existir. Cada sociedad escolar debe definir los suyos propios. "Las interpretaciones válidas del significado y la verdad son hechas por gente que comparte decisiones y las consecuencias de las decisiones", escribe Steinar Kvale a propósito del conocimiento (trad. de Carrascosa, inédita). Estos términos, llevados a la evaluación educativa implican el compromiso y la reflexión de todos los participantes de la educación.

3.- Que la información obtenida llegue a quienes pueden darle utilidad:

El último problema que planteamos para reflexionar en cuanto a la pertinencia de la evaluación, es el de decidir la manera de presentar la información y el análisis realizados con base en la aplicación de los instrumentos, así como los modos e instancias de distribución de estos datos. Consideramos útil discutir de antemano estos elementos, y evaluar la certeza de nuestras decisiones tras cada experiencia de divulgación, mejorando cada vez las estrategias de difusión con base en las observaciones que se hagan. También consideramos útil consignar el proceso de búsqueda de las mejores estrategias en escritos que pueda ser de utilidad a otros en su práctica evaluativa.

De todas estas consideraciones, surge el diseño de la estrategia "Academias de Evaluación" de la FES Iztacala, que se explicará en el siguiente capítulo.

Capítulo 3 LAS ACADEMIAS DE EVALUACIÓN[11]

Una vez establecida la necesidad de llevar a cabo evaluaciones sistemáticas, es necesario determinar una estrategia de trabajo que nos permita cumplir con las condiciones de confiabilidad, validez y pertinencia señaladas en el capítulo anterior. Ante la tarea de mejorar estas cualidades en las evaluaciones sistemáticas a alumnos en Iztacala, diseñamos el modelo de Academias de Evaluación como una técnica experimental para este fin; este modelo se presentará en la primera parte de este capítulo. En la segunda parte, la estrategia se comparará con la empleada por el CENEVAL para realizar el Examen General de Egreso de Licenciatura en Psicología (EGEL – Psicología).

Definición de las Academias de Evaluación

Las Academias de Evaluación son espacios de trabajo colegiado entre profesores de una misma carrera, pero de áreas y corrientes teóricas diversas en que se discuten los problemas básicos de la evaluación escolar (¿para qué evaluar?, ¿qué evaluar?, ¿cómo evaluarlo?) y se instrumentan herramientas de evaluación (esquemas de contenidos a evaluar, bancos de reactivos, reportes de análisis).

Para decirlo en otros términos, una Academia de Evaluación es un grupo heterogéneo de profesores de una carrera que se reúne para reflexionar sobre principios y políticas evaluativos generales y después plantear y llevar a cabo acciones específicas en la evaluación del logro escolar de los alumnos. El grupo es heterogéneo porque se considera necesario para la tarea que haya profesores de todas las corrientes, áreas, y temáticas generales de interés, de manera que la Academia llegue a conclusiones con el menor sesgo ideológico posible, debido a que en una Academia plural los diferentes sesgos ideológicos se neutralizan entre sí al hacerse las reuniones de colegio. Para cada Academia, se consideran al menos cinco fases de desarrollo:

---

[11] Parte de este capítulo fue tomado de un escrito publicado en las Actas del V Congreso de Metodología de las CC. Humanas y Sociales (Sánchez Miguel, 2000).

a) **Preparación:** en la que el grupo revisa materiales y contenidos básicos de evaluación con intenciones formativas. Más que expertos en evaluación, se prefirió que la Academia se integre con buenos profesores expertos en la enseñanza de su disciplina, por lo que hay que darles en esta primera fase un pequeño taller para que profundicen en sus conocimientos teóricos de la evaluación. La coordinación de la Academia corre a cargo de una persona ajena a la carrera en que se inserta, pero que en cambio tiene mayores conocimientos de evaluación que la mayoría de los otros participantes, y conoce y promueve el modelo de Academias de Evaluación.

b) **Definición de contenidos:** Como primer producto de las academias de evaluación, se plantea el tener un examen de mitad y otro de final de carrera, de los que se hablará más adelante. En esta fase de desarrollo se pide a los participantes que de manera colegiada decidan las temáticas que deberán abarcar dichos exámenes, y las expliciten en una "lista de nodos básicos estructurales de su disciplina". Esto no es otra cosa que construir la tabla de especificaciones para los instrumentos de evaluación (cfr. Por ejemplo, Aiken, 1996, p. 27). Para lograr esta tarea, se pide que se reflexione y conteste como grupo a las preguntas ¿Cuáles son los contenidos curriculares básicos, los más importantes, los indispensables? ¿Cuáles son los temas que es imperativo que sean parte de la formación de los alumnos? ¿Qué cosas no pueden faltar en el bagaje de conocimientos de los estudiantes?. A continuación se muestra un fragmento de una lista de nodos, a manera de ejemplo:

A- Filosofía e historia de la Biología (segundo semestre).
- T- Teoría de la Ciencia: la Biología y sus métodos.
- N- Teoría de la ciencia.
  - SN-Falsacionismo en Popper.
  - SN-Estructura de la ciencia según Kuhn.
- N- La Biología y sus Métodos (conceptos).
  - SN-Método inductivo.
  - SN-Método deductivo.
  - SN-Método analítico.
  - SN-Método experimental.
    - SSN- Hipótesis.
    - SSN- Verificación.

Las listas de nodos se estructuran en el modelo conocido como "diagrama de árbol", puesto que cada área (A) puede tener uno o varios temas (T), cada tema puede tener uno o varios nodos (N) y cada nodo puede tener uno o varios subnodos. Las bondades de este esquema para representar las temáticas que tendrá un examen son varias: 1.- Los contenidos se estructuran por grupos, de tal manera que es fácil observar la manera en que un contenido se relaciona con otros; 2.- Los contenidos se estructuran jerárquicamente de tal manera que mientras mayor sea la jerarquía de un elemento (A > T > N > SN), más genérico será éste, y viceversa. Esto es útil para controlar la profundidad y detalle con que se hacen los reactivos dedicados a una temática; 3.- Los elementos que corresponden a las "ramas" de otro elemento (por ejemplo, los SN de un N) ayudan a clarificar el sentido y extensión de éste.

c) **Desarrollo de un banco de reactivos:** A partir de la tabla de especificaciones generada por el grupo en la fase anterior, los participantes construyen preguntas de opción múltiple, con cuatro opciones, relacionadas necesariamente a uno o varios de los contenidos definidos anteriormente, y haciendo énfasis en que se intente que las preguntas incluyan lo menos posible la simple memorización para ser contestadas correctamente (aunque esto es extremadamente difícil: incluso Bloom señala que, tras muchos años de haber establecido su famosa categorización de contenidos educativos, más del 95 % de los reactivos siguen siendo memorísticos (cfr. Bloom, cit. En Díaz Barriga, 1982, p 24). En general, se pide en las Academias de Evaluación que las preguntas se atengan a los 15 criterios establecidos en "Improving the clasroom test" de Sherman y Tinkelman (1964), material que se revisa en la primera fase de trabajo colegiado. A manera de ejemplo se reproduce a continuación un fragmento de dicho documento, en traducción libre:

> **"5. SITÚE EN EL CUERPO DE LA PREGUNTA CLARA Y COMPLETAMENTE EL PROBLEMA PRINCIPAL.**
>
> La pregunta o situación introductoria de un ítem debe ser significativa en sí misma y contener claramente el problema principal. No debe ser necesario que el estudiante lea y relea

todas las respuestas antes de entender las bases sobre las cuales deberá hacer su selección. El siguiente ítem, por ejemplo, se relaciona con un sólo problema pero no hay nada en la pregunta que indique cuál es, sino que deberá inferirse de la lectura de todas las respuestas.

EL BECERRO:
1. Es una historia del pie de las colinas de Ozark
2. Describe la vida de la región boscosa de Washington
3. Tiene como asiento las montañas de Allegheny
4. Tiene la Florida como su sitio de origen

El ítem puede ser mejorado por la inclusión del problema en la pregunta de tal modo que el estudiante sepa lo que tiene que buscar antes de leer cualquiera de las respuestas. Un examen hecho con ítems así arreglados permite que el estudiante proceda con mayor rapidez y confianza.

EL SITIO DE ORIGEN DE "LA LEYENDA DEL BECERRO" ES:
1. El pie de las colinas de Ozark
2. La región de los bosques de Washington
3. Las montañas de Allegheny
4. La Florida"

En 15 recomendaciones como la anterior, este escrito previene a los redactores de reactivos para que eviten los errores más comunes y redondeen preguntas de opción múltiple de tal modo que queden listas para ser piloteadas.

d) **Revisión de las primeras experiencias:** Una vez conformados bancos de reactivos suficientemente grandes como para sustentar la elaboración de un examen, éste se aplica a una población piloto, y se lleva a cabo un análisis de los resultados. Haciendo énfasis que en la primera aplicación de un conjunto de preguntas es más importante analizar críticamente el instrumento y no tanto a los sustentantes, se examina la distribución de puntajes de la población piloto, la dificultad y discriminación de cada pregunta y la pertinencia de cada distractor. El trabajo de hacer cálculos y gráficos es llevado a cabo por la Unidad de Programación y Evaluación, pero se lleva a las Academias para que sean sus integrantes quienes analicen y extraigan conclusiones de los simples datos numéricos. Tras esta experiencia, las Academias reformulan, avalan o modifican tanto las listas de nodos como los reactivos utilizados y el conjunto

específico de preguntas empleado. Este trabajo se hace grupalmente, a través de la discusión ordenada. La meta propuesta para esta fase es conseguir dos exámenes, uno que se aplicaría a mitad de la carrera y otro al final; estos exámenes tendrían la función principal de servir como instrumentos para crear índices de avance de las poblaciones escolares, ponderando más el extraer significado de los puntajes (por ejemplo, damos cuenta de cuáles contenidos curriculares que deberían ser dominados por la población que ya terminó la mitad de la carrera no están siendo adquiridos de manera suficiente) que utilizarlos para fines administrativos (por ejemplo, asignar calificaciones). En este punto, la Academia ya tiene claro que su labor como evaluadora tiene poco que ver con la asignación de calificaciones, y en cambio mucho qué decir con respecto a las acciones regulatorias para mejorar el desempeño de la población escolar.

e) **Independización de las Academias:** En esta última fase, se plantean algunas rutas que la Academia podría seguir si son de su interés, por ejemplo, elaborar un examen diagnóstico que se aplicaría al inicio de la carrera, apoyar la elaboración de exámenes departamentales (que son responsabilidad de cada área de módulos o asignaturas), desarrollar métodos de evaluación de los contenidos no- escritos de la carrera... La intención es que la Academia lleve a cabo una discusión sobre lo que a sus miembro le interesa trabajar, para que ésta cobre un sentido único y específico en cada carrera. Al quedar claros los objetivos que una Academia en particular decide hacer suyos a partir de esta fase, la dirección de la Academia deja de estar en manos de quienes promovemos la estrategia, y pasa a alguno de los participantes pertenecientes a la carrera de marras. El papel de quienes dirigimos la Academia en un principio se transforma, y pasa a ser de coordinación a simple asesoría (a petición de la Academia) y de apoyo técnico, pues se ofrece a la Academia seguir realizando los análisis estadísticos que ésta ha empleado en fases anteriores, para que pueda tener el mismo tipo de discusión al que para este entonces ya estará acostumbrada.

Academias de evaluación y otras estrategias colegiadas

La preocupación de fundamentar cuidadosamente el método de aplicación e interpretación de los instrumentos de evaluación educativa a través de un grupo de personas que forman el equipo de desarrollo de la prueba, o "test development team" es común a todas las experiencias serias de este tipo. En una búsqueda en línea de bibliografía, utilizando el conocido buscador Google (www.google.com) se encontró, al cruzar reliability (confiabilidad), validity (validez) y "test development team", que hay 49 trabajos referidos que asocian los tres conceptos. Por ejemplo, en el proyecto PISA, reseñado en el Capítulo 1, se sigue esta secuencia para desarrollar sus marcos conceptuales:

> "– el desarrollo de una definición de trabajo para el área de conocimiento y la descripción de los supuestos sobre los que se basa dicha definición;
> – la evaluación de la organización de las tareas construidas de modo que sirvan para proporcionar información a los encargados de la política educativa y a los investigadores, sobre el rendimiento en cada área de evaluación de los alumnos de 15 años en los países participantes;
> – la identificación de una serie de características básicas a considerar a la hora de construir tareas de evaluación para uso internacional;
> – la operacionalización de un grupo de características básicas que se emplearán en la construcción de las pruebas, con definiciones basadas en la bibliografía existente, en la experiencia y en la realización de otras evaluaciones a gran escala;
> – la validación de las variables y la evaluación de la contribución de cada una de ellas para comprender la dificultad de las tareas en los distintos países participantes;
> – la preparación de un esquema de interpretación de los resultados."
>
> (Madrid: Ministerio de Educación, Cultura y Deporte, INCE, 2000)

Todo este trabajo no es realizado por un solo experto, o un grupo de expertos cerrado, sino con la participación de expertos en evaluación de todos los países participantes,

dirigidos por el Australian Council for Educational Research (ACER, Consejo Australiano de Investigación Educativa). (Madrid: Ministerio de Educación, Cultura y Deporte, INCE, Op. Cit.)

Shadish (1998) reporta 4 pasos lógicos en el desarrollo de un trabajo de evaluación, tomados de los escritos de dos décadas de trabajo de Scriven. Estos pasos, en resumen, son los siguientes:

1. Seleccionar los criterios de mérito, aquello que el evaluado debe hacer para ser considerado "adecuado".

2. Establecer estándares de ejecución de esos criterios; niveles referidos al criterio o a la norma que deban ser igualados o superados para ser considerado "adecuado".

3. Recolectar información relativa a la ejecución del evaluado en cuanto a los criterios referidos en los estándares.

4. Integrar los resultados en un juicio de valor final.

Shadish comenta que, dado que la evaluación se da en muchos ámbitos y con multitud de variables y características especiales, estos cuatro pasos son lo único en común que tendría cualquier desarrollo de un sistema de evaluación que tenga como meta establecer juicios de valor sobre objetos, sucesos o individuos. Desde nuestro punto de vista, falta al menos un paso, en el que se hacen racionalizaciones sobre la manera correcta de recolectar información pertinente y en su caso establecer los mecanismos, sistemas e instrumentos necesarios. Otra información importante que no establece este procedimiento es la referida a quiénes deben llevar a cabo los distintos pasos; aunque dado lo general del escrito, es comprensible que sea difícil establecer un estándar que se adecue a los distintos contextos de evaluación real.

Más completa es la ruta que propone Crawford (2002), con 11 pasos para desarrollar la evaluación de una experiencia educativa. Estos son:

1. Hacer un análisis de la tarea o tareas que se quieren mejorar.
2. Establecer la validez de contenido de los objetivos a cubrir.

3. Escribir ítems basados en los objetivos que hayan demostrado validez de contenido en la fase anterior.
4. Establecer la validez de contenido de los ítems.
5. Hacer un piloteo inicial de los ítems.
6. Hacer un análisis de los ítems basado en los resultados del piloteo.
7. Revisar los ítems que lo requieran y establecer formatos paralelos.
8. Establecer la puntuación mínima de aprobación de la prueba.
9. Llevar a cabo un segundo piloteo para establecer la validez y confiabilidad del instrumento.
10. Reportar los resultados de las aplicaciones.
11. Llevar a cabo trabajos de seguimiento y mantenimiento del instrumento de evaluación.

Este procedimiento, si bien nos parece útil e incorpora puntos que nosotros también desarrollamos, tiene el defecto de estar demasiado centrado en la construcción de un instrumento de evaluación, y no en la lógica general de un sistema de evaluación. En la versión extensa de los puntajes, se menciona sólo a un tipo de participante en el desarrollo de la evaluación: el grupo de expertos que juzgará si los objetivos están en consonancia con las tareas a mejorar (validación de los objetivos) y que juzgará si los ítems están en consonancia con los objetivos (validación de ítems). En defensa del procedimiento, se debe aclarar que su contexto está más orientado a la evaluación de experiencias educativas en grandes compañías (por ejemplo, capacitación) que en escuelas.

Un procedimiento más cercano a nosotros es el del Programa de Exámenes de Diagnóstico para los alumnos que ingresan a la licenciatura, de la UNAM. Este programa se inicia en febrero de 1994, con el propósito de "identificar el nivel de conocimientos y habilidades de los estudiantes, detectar aquellos que permitan predecir su desempeño escolar en los primero semestres, aportar información al bachillerato y a las licenciaturas para la revisión de sus planes de estudio y planear acciones propedéuticas" (Valle Gómez-Tagle y Cols., 1995, p. 113). Para desarrollar los instrumentos de evaluación de este proyecto, se formaron dos comisiones de trabajo: la primera, para definir la estructura de los exámenes y la segunda para

elaborar y revisar los reactivos. Ambas fueron formadas principalmente por profesores de bachillerato y licenciatura designados por las autoridades de sus dependencias. En este proyecto, la valoración de los ítems tiene gran importancia, y se ha hecho una comparación de tres diferentes métodos de calibración: el análisis convencional de reactivos, el análisis de Rasch y el análisis de modelo de tres parámetros; los dos últimos partes de la teoría de respuesta al ítem. Estos dos últimos modelos están fuera del marco de esta tesis, pues todos los análisis que se llevan a cabo en ella se refieren a la teoría clásica. Es importante señalar que entre las conclusiones de la comparación de los tres métodos, Valle Gómez-Tagle y Cols. (op. Cit.) dicen: "la información que se obtiene cuando se examinan tanto el reactivo como sus opciones de respuesta con el modelo convencional, resulta muy útil para explicar a los expertos que lo elaboraron cómo pueden mejorarlo, aunque no se debe perder de vista que sus estadísticas no son generalizables ya que dependen de la muestra de la cual se obtuvieron y que, además, cualquier modificación que se haga al examen también afecta dichas estadísticas [...] su elección dependerá en gran medida de sus objetivos y necesidades" (p. 120).
En todo caso, probablemente la apreciación de Herman, Morris, y Fitz-Gibbon (1987) sea la más sensata: 'No hay un método que sea la solución a todos los problemas de evaluación. El mensaje es este: algunos necesitarán una aproximación cuantitativa, algunos necesitarán una aproximación cualitativa, probablemente la mayoría se beneficiaría con una combinación de ambas'.
Regresando a nuestro estudio, diremos que aunque la fase d), en la que se validan los instrumentos no es en si una estrategia novedosa pues se ha hecho con grupos de personas desde hace años en diferentes instituciones, como se muestra en los párrafos anteriores, las Academias de Evaluación son diferentes a otras experiencias en los siguientes puntos:

- Se ha preferido para formar los grupos de trabajo a personas que destaquen en la enseñanza de las disciplinas por sobre quienes han destacado en las disciplinas propiamente. Se considera a estas personas de las Academias de Evaluación no sólo expertos en su disciplina, sino también en la *enseñanza* de su disciplina. No se requiere que sean

expertos en evaluación, ya que el plan de trabajo implica dedicar parte del trabajo colegiado a la formación en este sentido.

- Las opiniones de los expertos se trabajan de manera colegiada, haciendo que el grupo construya las propuestas desde el inicio. La discusión e intercambio de puntos de vista entre los miembros de las Academias se considera crucial en todas las fases de trabajo, y se da prioridad a alcanzar acuerdos en cuanto a cuestiones de base (¿qué evaluar, cómo hacerlo? ¿por qué evaluar eso y no otra cosa? ¿quién decide esto y con qué derecho?). En otras experiencias, los expertos trabajan de manera individual o con poco intercambio real
- Los expertos de las Academias son formados, como parte de los trabajos de las Academias, en nociones de evaluación y se trabaja con ellos constantemente en el por qué y para qué de ésta evaluación específica, de modo que su trabajo está muy contextualizado. En otras experiencias, se presupone que los expertos ya tienen todos los conocimientos mínimos necesarios para llevar a cabo su tarea.
- Finalmente, el trabajo de los expertos de las Academias está en constante revisión, de ellos mismos y de los efectos de sus productos, de tal manera que la retroalimentación y la noción científica de que el ensayo y error nos conducen a modelos más adecuados permean el trabajo de los grupos. En otras experiencias, los trabajos de los expertos se concretan a episodios aislados, sin hacer una revisión grupal de lo alcanzado, dejando esta tarea al supervisor o coordinador del examen.

Para dejar claras las implicaciones de estas diferencias con otros modelos de desarrollo de instrumentos de evaluación, se contrastará a continuación la estrategia de Academias de Evaluación con la empleada por el CENEVAL para la realización del Examen General para los Egresados de Licenciatura (EGEL) – Psicología. El CENEVAL se ha preocupado desde su fundación en 1994 por utilizar los métodos de evaluación educativa más desarrollados a nivel mundial (CENEVAL, 2002). Prueba de

ello es el hecho de que sus publicaciones y cursos en general están de acuerdo con el *mainstream* de instituciones tan consolidadas como el Educational Testing Services (ETS, 2002). Por todo ello, comparar el método empleado por nosotros con el del EGEL – Psicología tiene la finalidad de resaltar las similitudes y diferencias de las Academias de Evaluación con las corrientes principales del desarrollo de pruebas educativas. El autor de este trabajo fue convocado para la fase de validación de reactivos del EGEL - Psicología, lo cual nos permitió ver la manera en que se llevó a cabo el procedimiento. El autor entrevistó además a la coordinadora del proyecto, para completar la información de las fases previas, y ella nos facilitó amablemente algunos documentos sobre el mismo, de tal manera que consideramos tener suficientes elementos para comparar el procedimiento del EGEL – Psicología con el de las Academias de Evaluación.

El EGEL - Psicología es un instrumento de evaluación "cuya finalidad es identificar la medida en la que se tienen los conocimientos y las habilidades que debe mostrar un recién egresado de sus estudios de licenciatura para iniciar su vida profesional". (CENEVAL, 1998, p. 5). Su desarrollo puede dividirse en tres fases. En la primera, los cuerpos colegiados del EGEL conforman el Perfil Referencial de Validez del Recién Egresado de Psicología, así como los contenidos, bibliografía y taxonomía de evaluación. Estos cuerpos colegiados son un Consejo Técnico y un Comité Académico. Éstos están conformados por "representantes de instituciones de educación superior, profesionistas, especialistas y empleadores de psicólogos, tanto del sector público como del privado" (Op. Cit., p. 5). El perfil referencial que se menciona es un documento en que se describen "las competencias y subcompetencias genéricas que el Consejo Técnico [...] ha establecido como esenciales y comunes para iniciar una práctica profesional de calidad" (Op. Cit., p. 8). La Taxonomía de Evaluación es un documento que "describe cada uno de los niveles considerados y las operaciones cognoscitivas que los constituyen" (Op. Cit.,p.8), es decir, describe puntualmente los niveles de profundidad y los tipos de habilidades que se espera con los reactivos que se generen para evaluar las competencias y subcompetencias señaladas en el perfil referencial. Por ejemplo, los niveles de profundidad del EGEL – Psicología son tres:

1.- comprender y organizar lo aprendido; 2.- Aplicar lo aprendido; y 3.- Resolver problemas.

En la segunda fase, la coordinadora del EGEL – Psicología convoca a todos los profesores interesados de varias instituciones educativas que imparten la Licenciatura en Psicología a que realicen reactivos que se apeguen al perfil referencial. Se ofrece pagar los reactivos que sean aceptados, pero fuera de un instructivo para elaborar reactivos, no se da ninguna preparación especial a los redactores. Cada profesor puede enviar dos o más reactivos (cada reactivo debe ir con un "gemelo", para ser utilizado en las versiones del examen que son piloteadas), de aquellas competencias y subcompetencias que prefiera, a través de su institución.

En la tercera fase, se invita a personas con experiencia en evaluación educativa, y de preferencia pertenecientes a áreas relacionadas a la psicología a que juzguen los reactivos enviados por los profesores en la segunda fase, a partir de una lista de verificación (check list) muy extensa: 57 aspectos a evaluar, más un dictamen final ¡para cada reactivo!. Para cada aspecto, se solicitaba que usáramos la escala 0=No o nunca, 1= Parcialmente, 2= Sí o totalmente. Nos parece que este método de trabajo es muy disfuncional, en principio por la tarea de juzgar cientos de reactivos por ese método. En segundo lugar, había observaciones que uno quería hacer a determinado reactivo, que no estaban contempladas en los aspectos a verificar. También había casos en que uno sabía perfectamente qué cambios debían hacerse al reactivo para que resultase mucho mejor pero ¡eso no se preguntaba! Y por lo tanto, el reactivo tenía que ser juzgado en su forma original, y tal vez desechado, a pesar de que ligeras modificaciones lo hubieran salvado. En algunos casos, por muy experto que fuera uno, se volvía muy difícil evaluar con justicia algunos de los aspectos solicitados (por ejemplo el punto 3: "evita usar la información literal que se encuentra en la bibliografía recomendada"... ¿acaso podemos recordar de memoria los párrafos de los libros recomendados? O el punto 13: "evita fraseología estereotipada" a la fecha, no estamos seguros de qué se nos pide que juzguemos en ese aspecto). Finalmente, había aspectos demasiado particulares a ciertos contextos, que por lo tanto eran incontestables en muchos casos (Considérese por ejemplo el punto 21: "Si usa una batería de reactivos, cada uno de ellos mide, en lo individual una sola idea". Todo esto causó mucho tedio

en los que fuimos convocados a revisar los reactivos. Nuestro trabajo al revisar reactivos en otros casos nos había mostrado que ello es una tarea agradable y creativa, pero la mecanización establecida por el EGEL despojó a la tarea de todo interés. La "solución" fue que cada reactivo no sea juzgado más que por un individuo, de manera que los diferentes expertos se reparten el trabajo y nunca hay interacción entre ellos. Fue muy desmotivante leer un reactivo, darse cuenta de que era muy malo y debía desecharse casi en el acto, y en lugar de poder expresar las razones de tal rechazo, tener que llenar un formato vasto, que en el mejor de los casos tardaba cerca de siete minutos en ser llenado. En el anexo 1 se incluye un original del formato, para que el lector juzgue por sí mismo la magnitud de la tarea.

Suponemos que después de esto, los formatos de la escala para la revisión de reactivos fueron pasados por algún tratamiento estadístico (nos parece irreal pensar que fueron analizados cualitativamente uno por uno) que, bajo ciertos criterios, estableció cuáles reactivos debían ser aceptados, cuáles reformulados y cuáles rechazados.

Como puede observarse, una primera e importante diferencia entre el método del EGEL – Psicología y el de las Academias de Evaluación es la distribución de los participantes para las diferentes tareas. En el EGEL, los participantes de las diferentes fases no son los mismos, lo que hace que no necesariamente estén contextualizados con la tarea en conjunto, mientras que en las Academias los participantes en la fase de decisión de contenidos también realizan las tareas de redactar y valorar los reactivos, pilotearlos y hacer los análisis y las modificaciones pertinentes; además, en al menos una fase del EGEL – Psicología (la tercera, de la que somos testigos directos) se fomenta el trabajo individual, mientras que en las Academias se prefiere siempre la discusión colegiada.

Otras diferencias importantes son las siguientes:

- En el EGEL, los participantes no reciben formación sobre evaluación, sino sólo algunos escritos instructivos, que se presupone serán leídos.
- Los participantes de algunas fases (al menos la segunda y la tercera) no tienen otras razones para realizar la tarea que las de la conveniencia personal (el pago

en el caso de los redactores de reactivos, un documento con valor curricular, en el caso de los revisores).

- La revisión de reactivos nos dejó claro que un examen de psicología a nivel licenciatura que incluya a diferentes instituciones educativas requiere una fase de homogenización que no parece haberse dado: muchos reactivos resultaban ridículos en el contexto de Iztacala, y suponemos que lo mismo habrán pensado los participantes de otras instituciones. Los esfuerzos de las Academias de Evaluación se reconocen locales, y su pretensión es sólo la de hacer buenos instrumentos de evaluación en el contexto específico de la FES.
- La revisión mecánica e individual de reactivos, a través de una lista de verificación, del EGEL es muy diferente a la discusión grupal que se da en las Academias de Evaluación, en la que todos los participantes revisan cada reactivo a la vez y cada quién indica lo que le parecen sus puntos fuertes y débiles, así como modificaciones sugeridas, tratando de convencer a los demás a través de argumentos.

En conclusión, consideramos que la estrategia de las Academias de Evaluación del Logro Escolar se aparta un tanto de la corriente principal de metodología para elaborar instrumentos de evaluación, aunque tienen el punto en común de que las diferentes fases de trabajo son semejantes. Consideramos también que la estrategia de las Academias es más adecuada en un contexto local, puesto que permite que los participantes se desarrollen como evaluadores (habilidad que nunca está de más en los profesores), establece un clima de intercambio respetuoso y académico, permite el conocimiento más profundo de los contenidos de áreas ajenas a la propia y ayuda a generar una cultura de evaluación entre los profesores. Ignoramos si la estrategia utilizada por el EGEL – Psicología es la mejor en su contexto de examen interinstitucional, pero dado que esa discusión no es parte de los objetivos de este trabajo, dejamos hasta aquí los comentarios al respecto.

En el siguiente capítulo se explicará el método por el cual se compara la estrategia presentada hasta aquí con un modo de trabajo tradicional.

Capítulo 4 MÉTODO

El objetivo general de este trabajo es revisar la estrategia "Academias de Evaluación" en cuanto a su capacidad para generar y desarrollar instrumentos de evaluación válidos y confiables. Por principio, la estrategia se considera exitosa en cuanto da origen a instrumentos de evaluación que muestren mejores indicadores estadísticos que los exámenes elaborados con estrategias tradicionales, y en cuanto reporta una mejor comprensión y reflexión de los resultados por parte de los interesados. Como objetivos particulares tenemos los siguientes: a) elegir un grupo de métodos estadísticos que puedan ayudar en el análisis de diferencias entre los exámenes antes y después de las Academias de evaluación; b) contrastar los exámenes y analizar si las academias en realidad aportan ventajas psicométricas a los mismos; c) analizar cualitativamente las aportaciones de las Academias a la vida académica de cada Carrera.

El método es sencillo: se tomó la base de datos de un examen profesional que se aplicó antes de la existencia de las Academias de Evaluación de las carreras de Enfermería y Cirujano Dentista, así como un examen profesional de Optometría que se realizó al inicio del trabajo de su Academia de Evaluación[12] y se compararon algunos estadísticos que dan cuenta de las cualidades del cuestionario en general y de las preguntas en particular contra pruebas equivalentes (exámenes de final de carrera) realizados por las Academias de Evaluación (en el caso de Optometría, un examen realizado por la Academia en una fase más avanzada de trabajo). Si los resultados de las comparaciones se muestran consistentemente a favor de los exámenes

[12] La razón por la que en el caso de Optometría no se utilizó un examen profesional de antes de que existiera la Academia es sencilla: no existen exámenes profesionales en esa carrera, la más joven de Iztacala, previos a la Academia de Evaluación. Para facilitar las explicaciones al agrupar a las tres carreras, a partir de este momento se dirá genéricamente "los exámenes previos a la Academia", o "los exámenes de Antes de la academia", pero en el caso de Optometría esto significará el examen que hizo la Academia en sus inicios.

desarrollados con las Academias, podremos afirmar que el empleo de éstas mejoró los exámenes con respecto a las prácticas anteriores.

Los datos estadísticos que se analizaron fueron: la distribución de las calificaciones y su ajuste a una curva normal (según análisis visual y, posteriormente, usando el método de la x^2 de Pearson, explicado por Downie y Heat, 1973, caps. 6 y 14), la dificultad y el índice de discriminación promedio de los reactivos (según el método clásico, explicado por ejemplo por Matlock-Hetzel, 1997), el alpha de Cronbach (explicado, por ejemplo, por Dawson, s.d.) y la correlación entre resultados del examen y el promedio durante la carrera de los alumnos (bajo el índice producto-momento de Pearson, explicado, por ejemplo, en Aiken, 1996, p. 434).

A continuación se explica someramente el sentido y la lógica de cada análisis. Una descripción más detallada se encuentra en el apartado correspondiente del capítulo siguiente, por considerarse así más didáctico para un hipotético estudiante lector.

Distribución de calificaciones y bondad de ajuste a la curva normal.

Un primer análisis, visual, de las cualidades psicométricas de los exámenes, es la distribución de calificaciones. Se considera que, mientras el examen esté mejor elaborado, la distribución tendría que ser más simétrica y tener una forma aproximadamente normal.

La distribución de calificaciones se graficó por carrera, colocando en la abcisa los posibles puntajes (por intervalos) y en la ordenada el porcentaje de alumnos que obtuvo cada uno de ellos; de este modo, pueden compararse en una sola gráfica las distribuciones del examen "tradicional" y el realizado por las Academias de Evaluación.

Para confirmar el análisis anterior, de una manera menos apreciativa, se hizo un análisis numérico que da cuenta de la bondad de ajuste que tiene cada examen a la distribución normal. Para ello se obtiene, a partir de la conversión a valores z de la distribución de los resultados del examen, la distribución que se esperaría obtener (normal). Posteriormente, se comparan con una x^2 los valores obtenidos y los esperados para constatar la probabilidad de que las diferencias sean debidas únicamente al azar. Se considera que la bondad de ajuste será mayor en los exámenes

mejor elaborados. Se prefirió la x^2 como prueba de bondad de ajuste por encima de otras como la prueba de Kolmogorov-Smirnov o la de Anderson-Darling debido a limitaciones de éstas tales como el tipo de distribuciones a las que se aplican o la falta de sensibilidad en la zona de las colas (para una revisión completa, puede consultarse Chakravart, Laha y Roy, 1967, pp. 392-394).

Distribución de la dificultad

La distribución de la dificultad de los reactivos de un examen da cuenta de qué tanto éste es adecuado para determinada población. Un examen demasiado difícil para una población dada dará como resultado una distribución de la dificultad de los reactivos sesgada a la izquierda; por el contrario, un examen demasiado fácil mostrará una distribución de la dificultad de los reactivos sesgada a la derecha. Se considera que un examen mejor elaborado dará una distribución de la dificultad de los reactivos más centrada y simétrica que uno mal diseñado.

Distribución de la discriminación

Por otra parte, la discriminación de un reactivo da cuenta de su consonancia con el examen en su conjunto; si bien el Alpha de Cronbach, que se verá en el apartado siguiente, indica el promedio de esa consonancia entre reactivos y el total de la prueba, puede ser útil ver a detalle la distribución de la discriminación de cada reactivo. Se considera que un examen bien elaborado tendrá un número mayor de valores de discriminación altos que un examen mal diseñado.

Alpha de Cronbach

Este coeficiente da cuenta de la consistencia interna de una prueba, es decir, de la covariación que tienen sus partes con el todo. Esta consistencia interna, como se vio en el capítulo anterior, es un indicador de confiabilidad del instrumento. Se considera que un examen bien diseñado debe tener una mayor consistencia interna que uno mal planeado, por lo que el valor alpha del primero deberá ser más cercano a 1 que el del segundo.

Correlación entre los resultados del examen y los promedios durante la carrera

Como una manera de comprobar la validez de una medición, se plantea la covarianza de los resultados de esa medición y los de otra medición independiente, referida a la misma característica (validez predictiva o concurrente, explicada en el capítulo anterior). A un examen profesional le subyace la idea de que un alumno con buen aprovechamiento escolar deberá obtener buenas notas en el examen. Una medición independiente del aprovechamiento escolar es el promedio obtenido por el alumno durante su carrera. Tomando en cuenta todo esto, se considera que un examen bien diseñado debe tener una correlación mayor con las calificaciones promedio de la carrera de un alumno que un examen mal elaborado.

Utilizando todos estos indicadores, se llevaron a cabo los análisis correspondientes. Es muy importante aclarar que para todas las comparaciones el examen diseñado por las Academias de Evaluación que se utilizó es reportado en su primera aplicación masiva. Esto nos pareció justo debido a que los indicadores que se comparan son precisamente buscados activamente como trabajo de las Academias para pulir los exámenes; de este modo, sería injusto comparar la confección de los exámenes no realizados por las Academias si se les contrastara con exámenes ya calibrados al máximo por las mismas. Por ello, elegimos resultados de exámenes de las Academias que, si bien ya habían pasado varios piloteos, se encontraban en su primera aplicación formal. En el siguiente capítulo se da cuenta de los resultados de estos análisis, tras lo cual, en el capítulo 6 se establece una discusión de los resultados cuantitativos y se establecen algunos elementos cualitativos de valoración de las diferencias y similitudes entre los exámenes antes y después de las Academias de Evaluación del Logro Escolar.

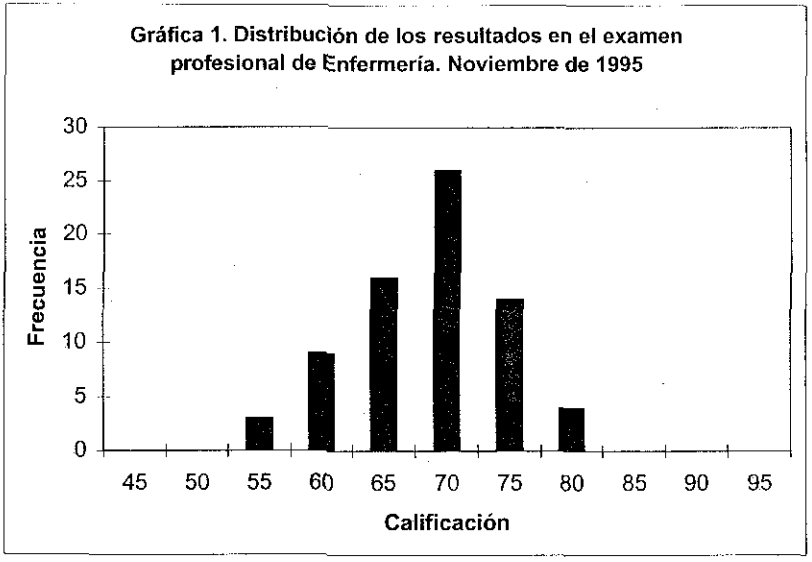
Los resultados se reportarán en el mismo orden y bajo los mismos apartados que el método, de manera que sean más fáciles de relacionar con las consideraciones del capítulo anterior.

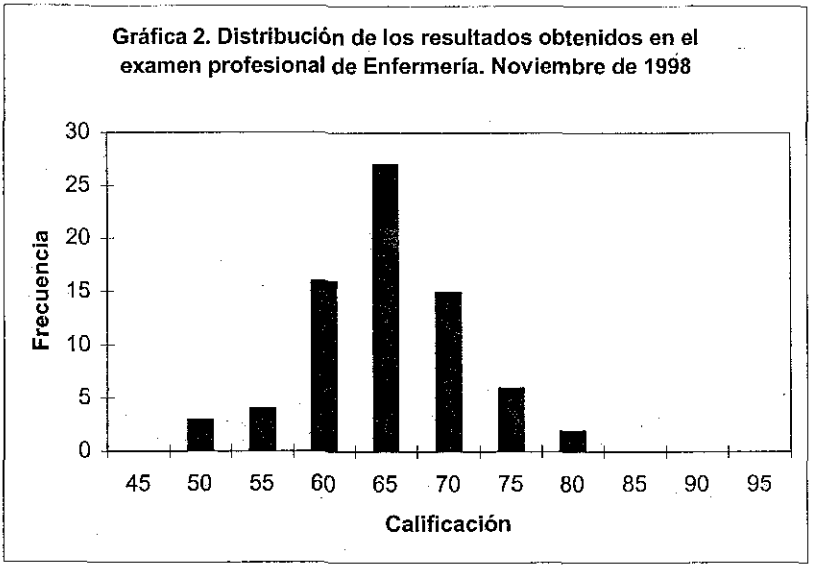
Distribución de calificaciones y bondad de ajuste a la curva normal.

Estos dos análisis se llevaron a cabo para comparar exámenes de la carrera de Enfermería entre sí y de la carrera de Cirujano Dentista entre sí; la exploración visual (distribución de calificaciones) se llevó a cabo también con la carrera de Optometría, no pudiendo realizarse la prueba de bondad de ajuste debido a que el tamaño de la población es insuficiente. La frecuencia esperada para los análisis de bondad de ajuste se obtuvo, en todos los casos, convirtiendo el valor superior de cada rango de clase (calificación) en el valor z correspondiente y verificando en tablas cuántos casos deberían presentarse teóricamente para ese intervalo en una distribución normal que tuviera la misma n que la distribución obtenida.

Carrera de Enfermería:

Se compararon exámenes profesionales aplicados en noviembre de 1995 y 1998. En el primero, aún no se formaba la Academia de Evaluación de Enfermería, por lo que el examen se diseñó conforme a estrategias tradicionales. En las gráficas 1 y 2 pueden observarse las distribuciones de los resultados de estos exámenes.





Se puede observar que en ambos casos la distribución de los resultados es aproximadamente normal ante el análisis puramente visual. Sin embargo, con el afán

de redondear esta observación, se decidió aplicar una x^2 comparando, para cada uno de los casos, la distribución obtenida con la distribución esperada. Los resultados se muestran en la tabla 1 y la tabla 2:

Tabla 1. Comparación entre resultados obtenidos y esperados. Enfermería 1995.

Clase (calificación)	Frecuencia obtenida(fo)	Frecuencia esperada(fe)	$(fo-fe)^2/fe$
44	0	0.0	0.01
48	0	0.0	0.04
52	1	0.4	0.85
56	3	2.1	0.36
60	8	7.0	0.13
64	11	14.7	0.92
68	17	19.4	0.29
72	19	16.2	0.48
76	12	8.6	1.31
80	1	2.8	1.17
84	0	0.6	0.59
88	0	0.1	0.08

Sumatoria de $(fo-fe)^2/fe$: **6.23**

Tabla 2. Comparación entre resultados obtenidos y esperados. Enfermería 1998.

Clase (calificación)	Frecuencia obtenida(fo)	Frecuencia esperada(fe)	$(fo-fe)^2/fe$
40	0	0.01	0.0073
44	0	0.09	0.0876
48	1	0.54	0.39136623
52	2	2.29	0.03724843
56	4	6.50	0.96421114
60	16	13.09	0.64745725
64	18	17.47	0.01614682
68	15	16.47	0.13099761
72	9	10.42	0.19463138
76	6	4.45	0.53743746
80	2	1.35	0.31236598
84	0	0.27	0.2701
88	0	0.04	0.0365

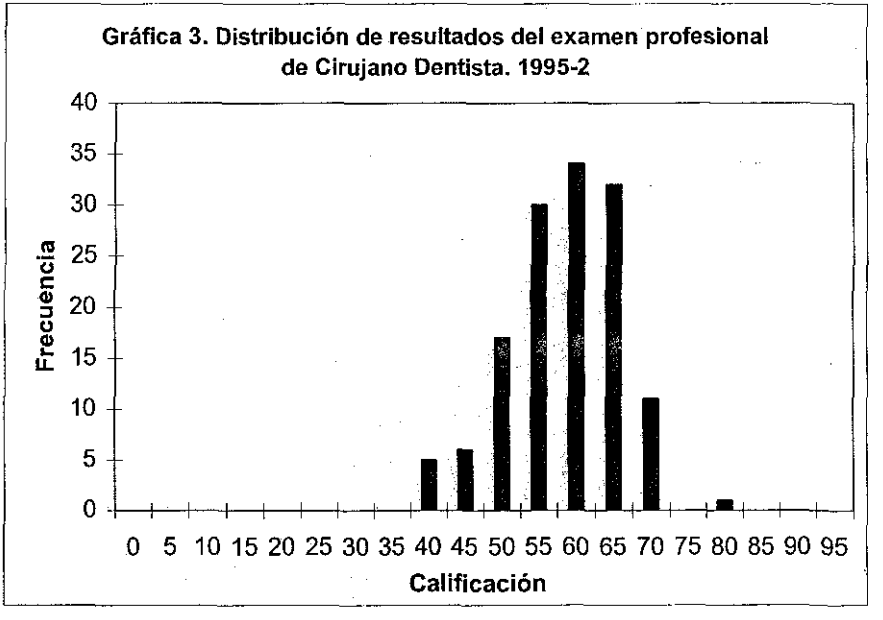
Sumatoria de $(fo-fe)^2/fe$: **3.63**

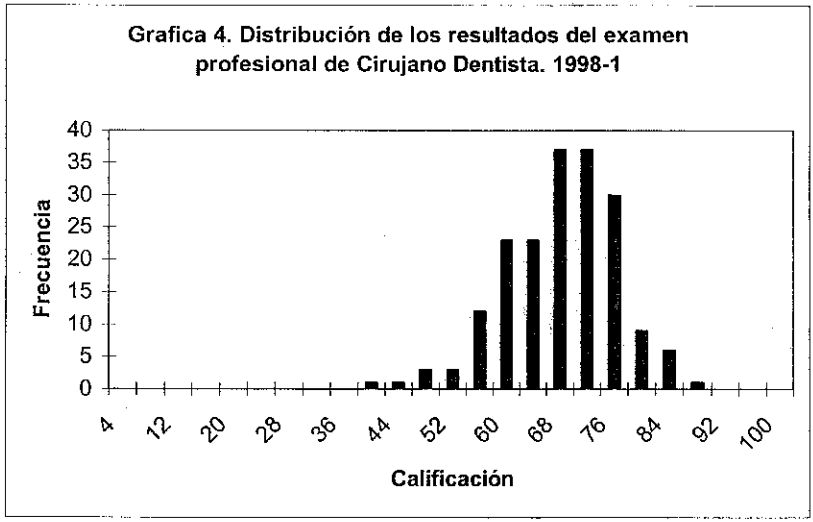
La sumatoria de $(fo-fe)^2/fe$ es precisamente la x^2. Al comparar estos valores con los de tablas (con 11 y 12 grados de libertad, respectivamente), encontramos que en ambos casos el valor obtenido es menor que el de tablas (valores de tablas: 19.67 y 21.02,

respectivamente), por lo que se acepta la hipótesis nula: no hay diferencias significativas entre las frecuencias obtenidas y las frecuencias esperadas, es decir, los resultados de ambos exámenes profesionales se distribuyen de manera aproximadamente normal.

Carrera de Cirujano Dentista:

Se compararon exámenes profesionales aplicados en el segundo semestre escolar de 1995 y en el primero de 1998. En 95 aún no se formaba la Academia de Evaluación de Odontología, por lo que el examen se diseñó conforme a estrategias tradicionales, mientras que para 98 la academia ya llevaba trabajando un par de años y el examen se hizo conforme a sus criterios. En las gráficas 3 y 4 pueden observarse las distribuciones de los resultados de estos exámenes.





Se puede observar, en un análisis visual simple, que en ambos casos la distribución de los resultados es aproximadamente normal. Sin embargo, con la intención de completar esta observación, se decidió aplicar una x^2 comparando, para cada uno de los casos, la distribución obtenida con la distribución esperada. Los resultados se muestran en la tabla 3 y la tabla 4:

Tabla 3. Comparación de frecuencias obtenidas y esperadas. Cirujano Dentista, 1995.

Clase (calificación)	Frecuencia obtenida(fo)	Frecuencia esperada(fe)	$(fo-fe)^2/fe$
28	0	0.01	0.01
32	0	0.07	0.07
36	0	0.38	0.38
40	5	1.58	7.42
44	5	4.98	0.00
48	8	11.74	1.19
52	18	20.85	0.39
56	28	27.85	0.00
60	28	28.00	0.00
64	28	21.19	2.19
68	12	12.05	0.00
72	3	5.15	0.90
76	1	1.66	0.26
80	0	0.41	0.41
84	0	0.07	0.07

Sumatoria de $(fo-fe)^2/fe$: **12.83**

Tabla 4. Comparación de frecuencias obtenidas y esperadas. Cirujano Dentista, 1998.

Clase (calificación)	Frecuencia obtenida(fo)	Frecuencia esperada(fe)	$(fo-fe)^2/fe$
36	0	0.02	0.02
40	1	0.11	7.07
44	1	0.48	0.55
48	3	1.79	0.83
52	3	5.06	0.84
56	12	11.51	0.02
60	23	21.52	0.10
64	23	30.58	1.88
68	37	35.21	0.09
72	37	32.96	0.50
76	30	23.49	1.80
80	9	13.67	1.60
84	6	6.44	0.03
88	1	2.29	0.72
92	0	0.69	0.69
96	0	0.15	0.15

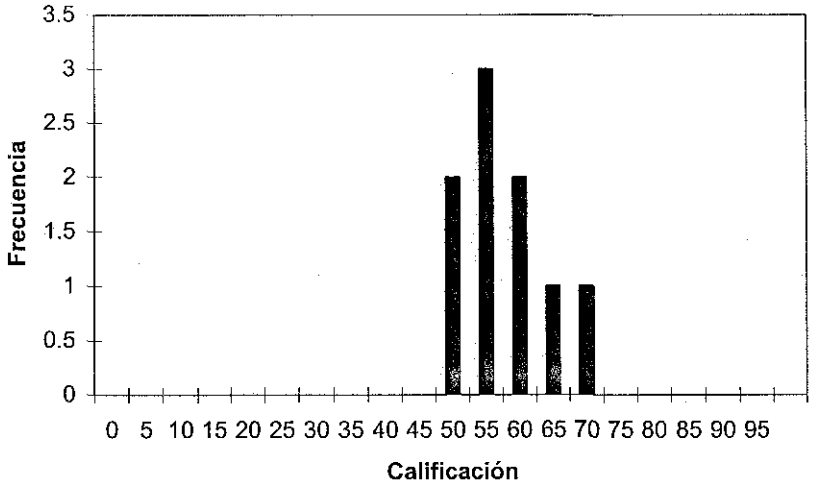
Sumatoria de $(fo-fe)^2/fe$: **16.86**

La sumatoria de $(fo-fe)^2/fe$ es la x^2. Al comparar estos valores con los de tablas (con 14 y 15 grados de libertad, respectivamente), encontramos que en ambos casos el valor obtenido es menor que el de tablas (valores de tablas: 23.68 y 24.99, respectivamente), por lo que se acepta la hipótesis nula: no hay diferencias significativas entre las frecuencias obtenidas y las frecuencias esperadas, es decir, los resultados de ambos exámenes profesionales se distribuyen de manera aproximadamente normal.

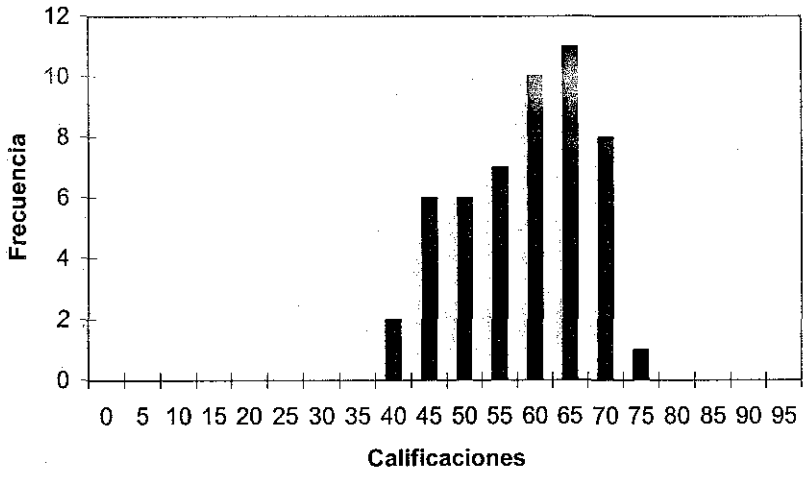
Carrera de Optometría.

Se compararon exámenes profesionales de esta carrera de noviembre de 1997 (en que aún no se trabajaba con la Academia de Evaluación de la Carrera) y de noviembre de 1999, con un examen ya conformado por la Academia. En las gráficas 5 y 6 pueden verse sus respectivas distribuciones.

Gráfica 5. Distribución de calificaciones. Optometría, 1997



Gráfica 6. Distribución de calificaciónes. Optometría, 1999.



Puede observarse en estas gráficas que la distribución de resultados, aunque sigue teniendo cierta aproximación con la normal, es poco clara. Incluso se observa una semejanza mayor a la distribución normal en la gráfica de 1997. Desafortunadamente, esto se debe, al menos en parte, a que el número de casos en estas dos aplicaciones es muy reducido, razón por la cual no se puede realizar una prueba de bondad de ajuste para confirmar la observación directa, ya que la x^2 involucrada tiene como regla, para considerarse válida, que no más del 20 % de las categorías esperadas deben tener valores inferiores a 5, lo que es imposible de cumplir en este caso. De este modo, debe considerarse que el hallazgo principal de este apartado de Optometría es que los datos son insuficientes para juzgar correctamente su distribución.

Distribución de la dificultad

Este análisis, al igual que el de discriminación que se comenta más adelante, se llevó a cabo en dos tiempos. En el primero, se estableció la dificultad y la discriminación de cada reactivo, y en el segundo, se generó una distribución de estos datos para cada examen.

La dificultad de cada reactivo se estableció a partir de la fórmula $p = \frac{C}{N}$ en que p es la dificultad, C es el número de personas que tuvieron la respuesta correcta y N es el número total de personas que contestaron la pregunta. El valor de este estadístico va de 0 a 1, en donde 0 indica que nadie contestó correctamente la pregunta, es decir, la dificultad es la máxima posible para el grupo que resolvió el examen; y 1 indica que todos los examinados contestaron correctamente la pregunta, es decir, la dificultad es la mínima posible para el grupo que resolvió el examen. Lo deseable, es que haya muy pocas, o no haya, preguntas que lleguen a estos dos extremos, pues de ellas no podemos extraer información que distinga y caracterice a los individuos de la población. Por tanto, una primera aproximación nos diría que los valores intermedios

(por ejemplo, P = 0.5), serían los mejores, pues maximizan las posibilidades de obtener información sobre los individuos de la población hacia ambos lados de la variable. Nótese que este estadístico da cuenta tanto de la dificultad del reactivo como de la habilidad o conocimiento de la población a que se aplica, es decir, la dificultad del reactivo no es constante, sino relativa a la población.

¿Cuáles son los valores de dificultad deseables en un instrumento de evaluación del logro escolar? Thompson y Levitov (en Matlock-Hetzel, 1997, p. 3) señalan que

> 'los ítems tienden a mejorar la confiabilidad de una prueba cuando el porcentaje de estudiantes que contestan correctamente el ítem es el valor medio entre el porcentaje esperado de respuestas correctas si se contesta la prueba al azar y el porcentaje de respuestas correctas que se tendría si todos supieran la respuesta (100 %).'

En nuestro caso, en que los reactivos previos a las academias y los generados por las mismas tienen en general 4 opciones, los valores deseables de dificultad estarían alrededor de .625 (dado que el porcentaje de estudiantes que contestarían correctamente por azar es de 25, $pd = 0.25 + \frac{1-0.25}{2}$)

Otra cuestión relacionada a la dificultad, es la distribución de la misma. No es deseable que todas las preguntas tengan exactamente la misma dificultad, sino que se distribuyan alrededor de este valor deseable, ya que el hecho de contar con algunas preguntas "extremas" que nos ayuden a distinguir a los que menos saben y a los que más saben, en distintos rangos, aumenta la capacidad de obtener información de un instrumento de evaluación. Una última consideración al respecto es que, cuando hablamos de exámenes de aprovechamiento que tendrán consecuencias en el futuro académico de los estudiantes, es razonable considerar la posibilidad de elevar el valor promedio deseable de aciertos, debido a que éste es idéntico a la calificación promedio que se obtendrá.

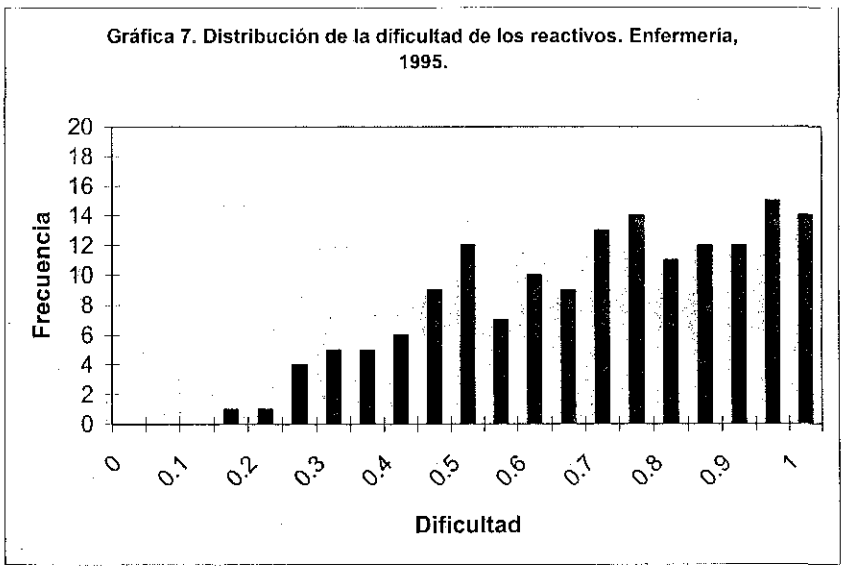
En resumen, una distribución adecuada de la dificultad de las preguntas de un examen implica que su valor promedio esté alrededor de 62.5 (quizá un poco más elevado para exámenes con consecuencias académicas importantes para los examinados), que haya preguntas en cada intervalo de dificultad, y que no haya preguntas con efecto de "piso" o de "techo" (es decir, que sean tan fáciles que todos puedan contestarlas, o tan difíciles que nadie pueda hacerlo).

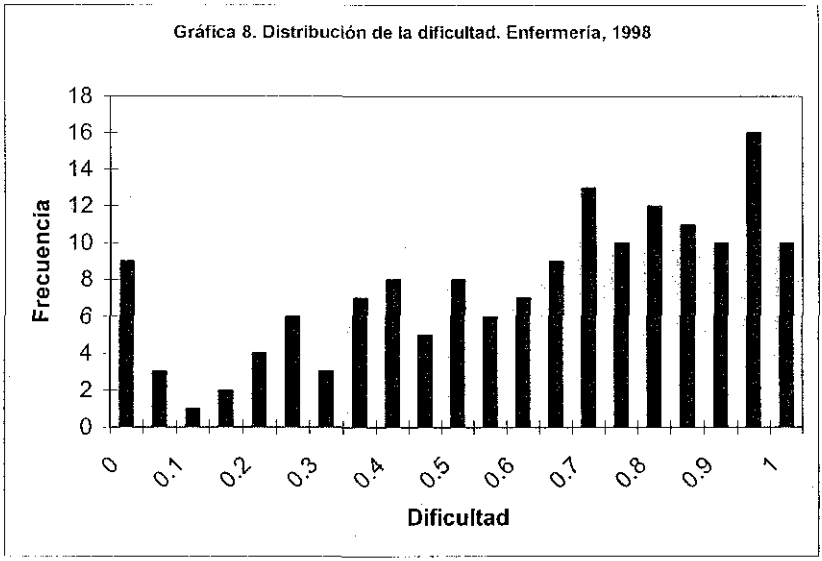
Así pues, veamos cuál es la distribución de la dificultad para los exámenes analizados, comparando, como en la sección anterior, el examen de cada carrera que se hizo sin la intervención de la Academia de Evaluación, con su análogo diseñado por la misma.

Carrera de Enfermería

A continuación, se presenta la distribución de la dificultad de los reactivos de esta carrera, en el examen de noviembre de 1995, realizado sin participación de la Academia de Evaluación, y la distribución de noviembre de 1998, en un examen realizado por la Academia.

Gráfica 7. Distribución de la dificultad de los reactivos. Enfermería, 1995.



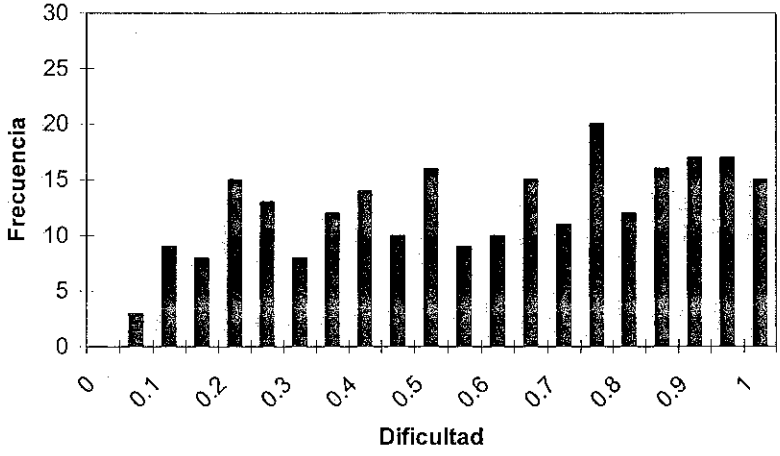


Puede observarse una distribución ligeramente más equitativa en el examen de 1998; es decir, a pesar de que en ambos se observa una tendencia a las preguntas fáciles (cercanas en dificultad a 1), parece haber más consideración en este segundo examen a tener preguntas entre los valores desde el intervalo .3 hasta el intervalo 1, con algunos intervalos especialmente elevados. La dificultad promedio, en el examen de 1995 es de 0.66 con 0.22 como desviación estándar. La dificultad promedio en el examen de 1998 es de 0.62, con una desviación estándar de 0.27. Como puede observarse, ambos promedios están alrededor del teórico deseable indicado más arriba; pero el de 1998 es más cercano, y la dificultad se encuentra más distribuida entre los diferentes valores. Un punto débil de la distribución del 98 es que tiene 10 preguntas que todos contestaron correctamente y 9 preguntas que nadie contestó correctamente, ambos casos que, como se indicó más arriba, son indeseables. La distribución del 95 tiene 14 preguntas que fueron contestadas correctamente por todos, y no tiene preguntas que nadie contestara correctamente.

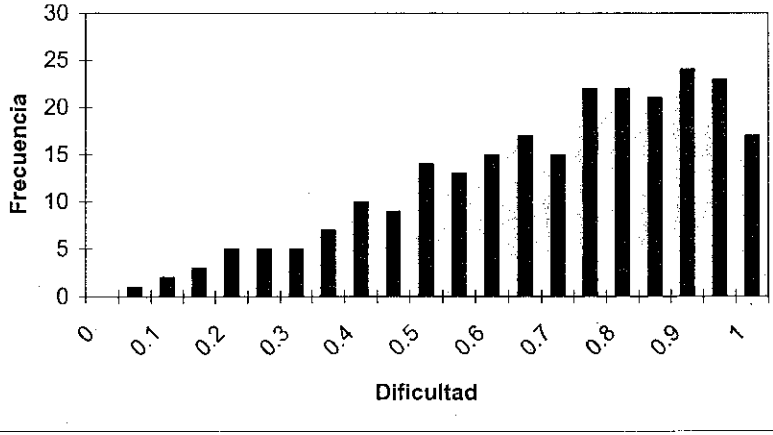
Carrera de Cirujano Dentista

Se presenta la distribución de la dificultad de los reactivos de esta carrera, en el examen de 1995, realizado sin participación de la Academia de Evaluación, y la distribución de 1998, en un examen realizado por la Academia.

Gráfica 9. Distribución de la dificultad. Cirujano Dentista, 1995.



Gráfica 10. Distribución de la dificultad. Examen de Cirujano Dentista, 1998

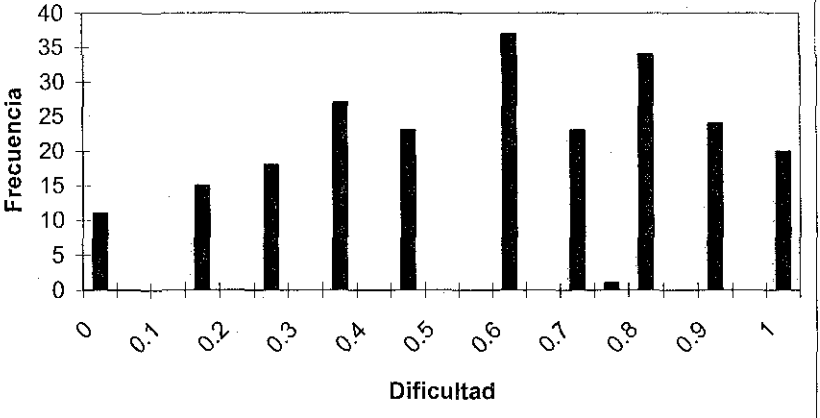


Puede observarse que el examen de 1995 tiene una distribución más pareja de las preguntas en cuanto a los diferentes niveles de dificultad. Sin embargo, el examen de 1998 también incluye al menos una pregunta en los niveles más difíciles, y la forma de su distribución es más cercana a la normal (al menos a tres cuartos de normal), ante el análisis visual. El promedio de la dificultad en el examen de 1995 fue de 0.56 con una desviación estándar de 0.28. En el examen de 1998, el promedio de dificultad fue de 0.66, con una desviación estándar de 0.15. Puede observarse aquí la aplicación del criterio de conveniencia señalado más arriba que indica que no es buena idea tener un promedio de ejecución de los alumnos reprobatorio en un examen profesional. Finalmente, el examen de 95 tiene 17 preguntas que fueron contestadas por todos los examinados, mientras que el de 98 tiene 15 preguntas en el mismo caso; ninguno de los dos exámenes tiene preguntas que no fueran contestadas por al menos un examinado.

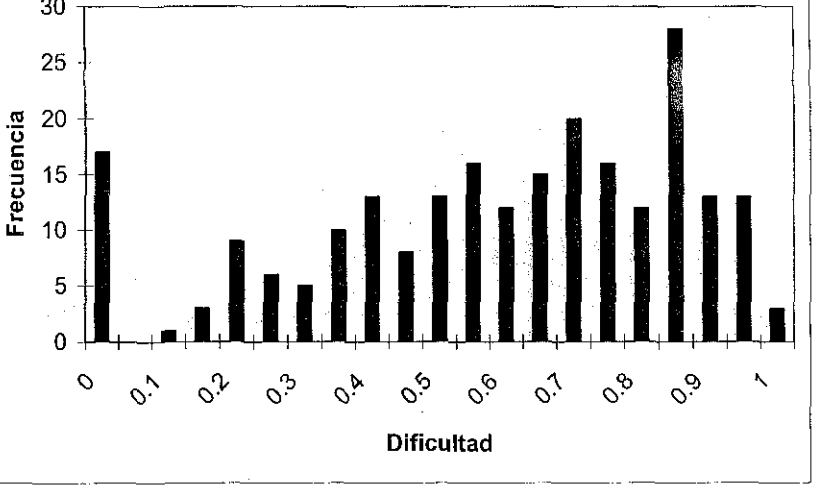
Carrera de Optometría

Se analizó la distribución de la dificultad del examen profesional de 1997, en que aún no existía la Academia de Evaluación, y del examen profesional de 1999, diseñado por la Academia de Evaluación.

Gráfica 11. Distribución de la dificultad. Examen profesional de Optometría, 1997



Gráfica 12. Distribución de la dificultad. Examen profesional de optometría, 1999.



Puede observarse en la gráfica 11 que la distribución del examen de 97 tiene intervalos vacíos. Esto se debe al tamaño de la población a que se aplicó el examen, que fue de sólo 9 personas. Por supuesto, este hecho limita mucho nuestro análisis, pero nos permite ver al menos que no todos los intervalos que matemáticamente era posible cubrir con esta población se cubrieron, que la dificultad promedio de las preguntas fue de 0.55 con una desviación estándar de 0.28 y que hubo 11 y 20 preguntas que tuvieron una dificultad de 0 y de 1 respectivamente.

En el examen de 1999 (Gráfica 12) puede observarse que sólo hay un intervalo de dificultad que no tiene al menos una pregunta, que la distribución tiende a las preguntas fáciles, y que hubo 17 y 3 preguntas que tuvieron una dificultad de 0 y 1 respectivamente. La dificultad promedio de las preguntas de este examen fue de 0.56 con una desviación estándar de 0.27. Aunque existe la tentación de señalar que en el caso de optometría es muy notoria la mejoría en este estadístico hacia el examen de 1999, debe tomarse en cuenta de que esta segunda aplicación tuvo a 51 examinados, y con estas poblaciones pequeñas es muy probable que muchas de las ganancias sean efecto del bajo número de la primera aplicación.

Distribución de la discriminación

La discriminación es un indicador de las preguntas sobre su capacidad de dar información acerca de los examinados. Su conceptualización es un poco más compleja que la de la dificultad, a pesar de que la manera de calcularla tiene una facilidad similar. La discriminación es la concordancia que tenga una pregunta con la prueba en su conjunto. En palabras de Matlock-Hetzel (1997, p.4) 'uno esperaría que las personas que se desempeñan bien con la totalidad de una prueba contesten correctamente un ítem en particular, y quienes se desempeñan mal en el total de la prueba contesten incorrectamente el ítem. Un buen ítem discrimina entre aquellos que se desempeñan bien en la prueba y los que no.' La discriminación es una medida de esa capacidad de discriminación del ítem.

La discriminación es un indicador cuyos valores pueden ir de -1 a 1, donde 1 significaría que el ítem discrimina a los que se desempeñan bien en la prueba de manera perfecta, 0 significa que el ítem no discrimina en absoluto entre buenos y malos examinados, y -1 significa que el ítem discrimina de manera completamente inversa (es decir, el ítem es contestado correctamente ¡sólo por quienes se desempeñan peor en el examen!). La fórmula por la cual calculamos la discriminación es $D = \frac{M - P}{n}$ en la que M es la suma de los examinados pertenecientes al 27 % de mejores calificaciones en el examen que tuvieron correcto el ítem; P es la suma de los examinados pertenecientes al 27% de peores calificaciones en el examen que tuvieron correcto el ítem; y n es el número de personas que hay en cualquiera de los dos grupos. En términos operativos, este valor se obtiene ordenando por el total obtenido en la prueba los resultados de los alumnos para cada pregunta (codificados como 1=respuesta correcta, 0= respuesta incorrecta), calculando cuántos son el 27 % de los alumnos y sumando, para obtener M del mejor hacia abajo hasta completar este porcentaje, y para obtener P, del peor hacia arriba. ¿Por qué el 27 %? Algunos autores señalan al 25 % como otra posibilidad, y en la práctica hemos visto utilizar un procedimiento muy semejante utilizando a la mitad de la población para el grupo M y a la otra mitad para el grupo P; sin embargo, Wiersma y Jurs (en Matlock-Hetzel, op. Cit.) señalan que la experiencia ha mostrado que este valor maximiza las diferencias entre los dos grupos en distribuciones normales, conservando al mismo tiempo suficientes casos para el análisis.

Para este análisis se procedió, igual que para el de dificultad, en dos tiempos: en el primero, se calculó el valor de discriminación para cada reactivo de cada prueba analizada. En el segundo, se obtuvo la distribución de la discriminación así como el valor promedio y la desviación estándar, por prueba. ¿Qué distribución es deseable con respecto a la discriminación? A diferencia de la dificultad, con la discriminación queremos que los valores de cada ítem se acerquen lo más posible a 1. Un criterio utilizado comúnmente dicta que los ítems que tengan valores de discriminación arriba de 0.40 son muy buenos, aquellos entre 0.30 y 0.39 son razonablemente buenos pero

aún pueden mejorarse; de 0.20 a 0.29 son ítems que deben revisarse, pero que tras ajustes pueden utilizarse de nuevo; y todo lo que esté abajo del 0.20 es un ítem malo. Por supuesto, los ítems con valores negativos son los peores, y lo deseable sería que no hubiera ninguno en nuestros exámenes. Tras hacer las distribuciones por rangos con saltos de 0.05, desde -1 hasta 1, nos dimos cuenta de que sería más descriptivo de las similitudes y las diferencias entre cada par de exámenes una tabla que indicara la frecuencia de los ítems para cada uno de los intervalos (irregulares, por cierto) que conforman el criterio arriba indicado. En todos los casos, cada examen tiene la misma cantidad de ítems que el par con el que se le compara.

Carrera de Enfermería

Se emplearon los mismos exámenes que en el análisis de dificultad del apartado anterior. Las distribuciones obtenidas son las siguientes:

Tabla 5. Distribución de la discriminación por rangos (irregulares, elegidos con base en los criterios de decisión) de los ítems de los exámenes de Enfermería de 1995 y 1998.

Rangos	95	98
mayores a 0.39	22	20
de 0.30a 0.39	35	37
de 0.20 a 0.29	20	24
de 0 a 0.19	60	65
de -1 a -0.01	23	14

Como puede verse, no hay diferencias notables entre las dos distribuciones; en algunos rangos "buenos" el examen de 1995 tiene más ítems, y en otros, el examen de 1998, y en todos los casos, la diferencia es pequeña. Sin embargo, el promedio de la discriminación sí muestra una pequeña tendencia: en el examen de 1995, el promedio de la discriminación es de 0.16, mientras que en el de 1998 es de 0.18.

Carrera de Cirujano Dentista

Los resultados obtenidos, utilizando los mismos exámenes que en los análisis anteriores de esta carrera son los siguientes:

Tabla 6. Distribución de la discriminación por rangos (irregulares, elegidos con base en los criterios de decisión) de los ítems de los exámenes de Cirujano Dentista de 1995 y 1998.

Rangos	95	98
mayores a 0.39	46	46
de 0.30a 0.39	44	53
de 0.20 a 0.29	47	66
de 0 a 0.19	94	78
de -1 a -0.01	19	7

Puede observarse que hay diferencias consistentes a favor del exámen de 1998. Los mejores ítems (los que tienen una discriminación igual o mayor a 0.40), casualmente se dieron en la misma cantidad en ambos exámenes; pero las siguientes dos categorías son más abundantes en ítems para la versión de 1998, mientras que las categorías indeseables (de 0.19 hacia abajo) son más abundantes para la versión de 1995. El promedio de discriminación confirma esta tendencia: el promedio de 1995 es de 0.18 con una desviación estándar de 0.16; el promedio de 1998 es de 0.21 con una desviación estándar de 0.15. Debe señalarse, de cualquier modo, que esta diferencia no es, en todo caso, espectacular.

Carrera de Optometría

Se utilizaron, como en las otras carreras, los mismos exámenes empleados en análisis anteriores. Los resultados fueron los siguientes:

Tabla 7. Distribución de la discriminación por rangos (irregulares, elegidos con base en los criterios de decisión) de los ítems de los exámenes de Optometría de 1997 y 1999.

Rango	97	99
mayores a 0.39	27	72
de 0.30a 0.39	69	22
de 0.20 a 0.29	0	64
de 0 a 0.19	101	51
de -1 a -0.01	36	24

Es en esta carrera donde se puede ver una diferencia mayor en el rubro de la discriminación. El examen de 1999 tiene una distribución notablemente mejor que la del examen de 1997: Los mejores ítems pasaron de ser 27 a ser 72; los ítems aceptables

también son más (si sumamos los del intervalo 0.30 a 0.39 con los de 0.20 a 0.29. Consideramos además que los ítems del intervalo de 0.30 a 0.39 son menos en la versión de 1999 porque precisamente de ese intervalo se alimentó el aumento en el intervalo superior). Puede observarse también que los ítems correspondientes a intervalos indeseables (discriminación igual o menor a 0) son bastantes más en el examen de 1997 que en el de 1999. Las medias confirman este análisis: en 1997 la discriminación tuvo un promedio de 0.13 con una desviación estándar de 0.32 (la mayor variabilidad observada en todo este ejercicio); en 199 la discriminación tuvo un promedio de 0.23, con una desviación estándar de 0.24. Puede observarse que esta diferencia es la más importante entre las diferentes carreras, en cuanto a discriminación.

Alpha de Cronbach

Debido a que la naturaleza de los exámenes profesionales aquí analizados no permite estrategias como el test-retest o las formas paralelas, no pueden utilizarse los coeficientes clásicos existentes para estas dos formas de estudiar la confiabilidad. La alternativa es el uso de coeficientes de consistencia interna. De entre éstos, se prefiere el Alpha de Cronbach, pues es una fórmula general que obtiene los mismos resultados que el método de Kuder-Richardson pero sin limitarse a entradas de datos binarios; por otra parte, comparando el Alpha con el método de división por mitades, resulta que el coeficiente del primero es la media de todas las posibles divisiones en mitades del segundo, por lo que sigue siendo la mejor elección.(cfr. Aiken, 1996).

El alpha de Cronbach 'es una medida de la correlación al cuadrado entre los puntajes obtenidos y los puntajes reales' (Yu, 2002) (basándose en la conocida idea de que el puntaje obtenido es igual al puntaje real mas el error de medida). Operativamente,

"puede definirse como $\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{S_i^2}\right)$ donde k es la cantidad de reactivos, s_i la varianza de calificaciones en el i y s^2 la varianza de las calificaciones totales de la prueba". (Aiken, Op. Cit., p. 90).

La confiabilidad inferida de esta forma puede tener valores entre 0 y 1, donde los más deseables son los cercanos a 1: esto indica mayor confiabilidad, que en el caso del Alpha de Cronbach implica mayor coherencia entre las preguntas de la prueba. Valores de 0.70 o superiores son ya aceptables en general, aunque no existe un criterio unificado para ello. Veamos cuáles fueron los valores alpha para cada par de exámenes. Debe hacerse la observación de que los valores alpha no fueron calculados a mano, utilizando la fórmula señalada arriba, sino utilizando el programa estadístico SPSS (versión 8.0).

Carrera de Enfermería

Se obtuvieron los siguientes valores alpha para los exámenes que ya hemos analizado en otros sentidos: 1995= 0.67; 1998= 0.75. Puede observarse que, aunque el alpha de 95 era bastante bueno, el de 1998 ya es aceptable según los parámetros generalmente utilizados.

Carrera de Cirujano Dentista

Se obtuvieron los siguientes valores alpha para los exámenes que ya hemos analizado en otros sentidos: 1995= 0.87; 1998= 0.90. Puede observarse que aunque el valor del alpha ya era muy bueno en el examen de 1995, todavía mejoró en el de 1998, y hay que recordar que, como en la mayoría de los indicadores numéricos, es más difícil mejorar una centésima en el extremo superior que en cualquier otra parte de la gama de valores posibles.

Carrera de Optometría

En esta carreta, el valor alpha no era aceptable en el examen de 1997 (0.54) y siguió sin acercarse lo suficiente al criterio aceptable para 1999 (0.60); sin embargo, puede observarse que, al igual que en las otras carreras, el alpha mejoró varias centésimas en el examen realizado por la Academia de Evaluación.

Correlación entre los resultados del examen y los promedios durante la carrera

Uno de los medios más comunes para establecer la validez de nuestras mediciones es el basado en el concepto de validez de criterio: en este caso, un criterio externo con el que pueden ser comparados los resultados de nuestros exámenes es la calificación promedio obtenida por los estudiantes a lo largo de su carrera. Como este es un dato difícil de construir, y no tuvimos acceso a la historia académica de cada alumno, se hizo el primer intento con la carrera en que nos era más fácil obtener la información: Cirujano Dentista. Sin embargo, los resultados no justificaron el esfuerzo, por lo que este ejercicio no avanzó hacia las otras carreras. Sin embargo, el uso de este método estadístico, recomendado por muchos autores para establecer la validez de criterio, no debe dejarse de lado en análisis que se hagan para comparar exámenes; simplemente, es mejor que se tomen previsiones (que nosotros no tomamos) para tener los datos del criterio de comparación antes de que éstos se vuelvan "históricos" y por lo tanto, difíciles de obtener.

Carrera de Cirujano Dentista

Para hacer una correlación entre los resultados del examen con el promedio durante la carrera de los alumnos, se solicitó a la jefatura de carrera la información correspondiente a los estudiantes; de 136 alumnos que realizaron el examen profesional en 1995, sólo se pudo recuperar el promedio durante la carrera de 72; en el examen de 1998 (que es más reciente) se pudo recuperar el promedio de los 221 casos que hicieron el examen. El procedimiento que se siguió es sencillo: se utilizó el coeficiente de correlación de Pearson para pares asociados para conocer la correlación entre la serie de promedios y sus correspondientes puntajes en el examen, tanto para el

examen de 1995 como para el de 1998. La correlación es "una medida de la relación entre dos variables" (Downie y Heat, 1973, p. 100). En el caso de el coeficiente de correlación producto-momento de Pearson, la definición matemática es la siguiente:

$r = \frac{\sum z_x z_y}{N}$ donde z_x y z_y son puntuaciones estandarizadas de X y Y, y N es el número de casos (Thorndike y Hagen, 1977, p. 628). Los mismos autores señalan que esta fórmula es impráctica para obtener el valor de correlación, por lo que proponen alternativas equivalentes. Nosotros, como en otros análisis, utilizamos el programa SPSS para obtener los resultados, que para nuestros datos fueron los siguientes:

Tabla 8. Comparación de las correlaciones de los puntajes en el examen con los promedios en la carrera entre el examen de 1995 y el de 1998. Carrera de Cirujano Dentista.

	1995	1998
r=	0.466	0.528
p<	0.000	0.000
No. De casos	72	221

Los resultados muestran que ambas correlaciones son significativas, con una bajísima probabilidad de equivocarse (p>0.000), lo cual habla muy bien de la validez de criterio de ambas pruebas. También parecen mostrar que el examen de 1998 es un poco mejor que el de 1995 en este rubro, dado que su índice de correlación es algo mayor. Sin embargo, a sabiendas de que no se pudieron utilizar todos los datos de 1995, se consideró que el uso de una muestra de los mismos podría estar sesgando de manera determinante esta aparente ventaja. Para analizar esto, se hicieron 10 muestreos aleatorios en que se extrajeron, en cada ocasión, 72 casos de la población de 1998, y se obtuvo su correlación. Los resultados fueron los siguientes:

Tabla 9. Resultados de 10 correlaciones realizadas en 10 muestras al azar. Examen profesional de Cirujano Dentista, 1998.

Muestra	r=	p<
Muestra 1	0.656	0.000
Muestra 2	0.486	0.000

Muestra 3	0.511	0.000
Muestra 4	0.313	0.007
Muestra 5	0.515	0.000
Muestra 6	0.421	0.000
Muestra 7	0.402	0.000
Muestra 8	0.549	0.000
Muestra 9	0.490	0.000
Muestra 10	0.529	0.000

Como puede observarse, los valores de las muestras en ocasiones son superiores a lo obtenido en el examen de 1995, pero en otros casos son inferiores (el caso extremo es la muestra 4). El promedio de las muestras es de 0.487. Por ello, creemos que es válido considerar ambos exámenes como bastante adecuados en términos de validez de criterio, pero sin jerarquizar a ninguno sobre el otro, dado lo sensible que parece ser r a la muestra elegida.

Dado que nos damos cuenta de esta sensibilidad, y ante la dificultad de obtener muestras completas de las otras carreras, consideramos que no era necesario hacer este análisis para ellas, ya que pensamos que es muy probable que el resultado sea tan indeterminante como en el caso de Cirujano Dentista.

Para la discusión de los resultados, consideremos antes que nada un resumen de lo obtenido, expresado en términos cualitativos. En la tabla siguiente, se muestra este resumen, separando las pruebas realizadas por carrera, e indicando en la columna de la extrema derecha una valoración sobre si el examen hecho por la Academia correspondiente mejoró sustancialmente, mejoró marginalmente, mantuvo igual, empeoró marginalmente o empeoró sustancialmente los valores del análisis correspondiente:

Tabla 10: Resumen y valoración de resultados de los análisis cuantitativos.

Distribución de calificaciones	Resultado	Valoración
Enfermería - sin Academia	Aproximadamente normal	Igual
Enfermería - con Academia	Aproximadamente normal	
Cirujano Dentista – sin academia	Aproximadamente normal	Igual
Cirujano Dentista – con academia	Aproximadamente normal	
Optometría – sin academia	Aproximadamente normal	Igual
Optometría – con academia	Aproximadamente normal	
Bondad de ajuste a la curva normal (x^2)		
Enfermería - sin Academia	Se acepta hipótesis nula	Igual
Enfermería - con Academia	Se acepta hipótesis nula	
Cirujano Dentista – sin academia	Se acepta hipótesis nula	Igual
Cirujano Dentista – con academia	Se acepta hipótesis nula	
Distribución de la dificultad		
Enfermería - sin Academia	Tendencia a las preguntas fáciles. Carece de preguntas en algunos intervalos de dificultad. Promedio=0.66	Mejoró marginalmente
Enfermería - con Academia	Tendencia a las preguntas fáciles. Promedio=0.62	
Cirujano Dentista – sin academia	Promedio=0.56	Mejoró marginalmente
Cirujano Dentista – con academia	Promedio=0.66. Ligera aproximación a la normal.	
Distribución de la dificultad	Resultado	Valoración
Optometría – sin academia	Intervalos vacíos. Promedio=0.55	Mejoró marginalmente
Optometría – con academia	Un solo intervalo vacío. Promedio=0.56	
Distribución de la discriminación		
Enfermería - sin Academia	Promedio=0.16	Mejoró marginalmente
Enfermería - con Academia	Promedio=0.18	

Cirujano Dentista – sin academia	Promedio=0.18	Mejoró marginalmente
Cirujano Dentista – con academia	Promedio=0.21	
Optometría – sin academia	Promedio=0.13	Mejoró sustancialmente
Optometría – con academia	Promedio=0.23	
Alpha de Cronbach		
Enfermería - sin Academia	0.67	Mejoró sustancialmente
Enfermería - con Academia	0.75	
Cirujano Dentista – sin academia	0.87	Mejoró sustancialmente
Cirujano Dentista – con academia	0.90	
Optometría – sin academia	0.54	Mejoró marginalmente
Optometría – con academia	0.60	
Correlación entre los resultados del examen y los promedios durante la carrera		
Cirujano Dentista – sin academia	Significativa a p>0.000	Igual
Cirujano Dentista – con academia	Significativa a p>0.000	

Así, en cuanto a la distribución de calificaciones podemos observar que las pruebas realizadas sin Academia se distribuyen de forma aproximadamente normal, lo cual es un buen indicador pues, asumiendo que los valores reales logro escolar de la población se distribuyen de manera normal, apunta a que el examen mide con bastante exactitud los datos de la realidad. Las pruebas realizadas por las Academias también se distribuyen de forma aproximadamente normal, lo cual nos indica que no hubo variación en este apartado, que ya de entrada era adecuado. La bondad de ajuste a la curva normal nos confirma este resultado.

Con respecto a la distribución de la dificultad, lo deseable, como se comentó en el capítulo anterior, es que el valor promedio sea cercano al 0.62 y que la frecuencia se distribuya normalmente alrededor de éste. A este respecto, puede observarse que se encontraron en todos los casos mejoras ligeras en el promedio y/o la distribución de la dificultad de los exámenes sin Academia a los diseñados por las Academias.

La discriminación, a diferencia de la dificultad, tiene una distribución óptima cuando todos sus valores son iguales a 1. Esta distribución es utópica, y lo que podemos esperar en versiones mejoradas de los exámenes es que tengan menos preguntas con

discriminación negativa o con valores cercanos a cero, y más preguntas con valores superiores a 0.20; en nuestro cuadro, podemos observar que el promedio, utilizado como indicador de la tendencia a acercarse a la discriminación de 1, mejora un poco en todos los casos desde los exámenes realizados sin Academia a los diseñados por las Academias de Evaluación del Logro Escolar.

El alpha de Cronbach, un indicador claro de la confiabilidad de las pruebas, que también se relaciona con la validez (ya que mide la consistencia interna), mejoró en todos los casos desde los exámenes realizados sin Academias a los diseñados por las Academias, y en los casos de Enfermería y Cirujano Dentista, esta mejora no fue sólo marginal, sino bastante sustancial.

Finalmente, el estudio de validez utilizando la correlación entre los resultados del examen y los promedios durante la carrera, que se pudo realizar sólo con la Carrera de Cirujano Dentista no nos muestra diferencias entre el examen realizado sin Academia y el realizado por la Academia; en ambos casos la correlación es significativa a $p>0.000$, lo cual indica muy buenos niveles de validez concurrente en las pruebas.

Puede observarse que, en 6 casos, los análisis no mostraron diferencias entre un examen y otro; en 7 casos se encontraron diferencias marginales a favor del examen realizado por las Academias de Evaluación; y en 2 casos hubo mejoras sustanciales. Tomando esto en cuenta, nuestra primera conclusión es que la estrategia de las Academias de Evaluación puede considerarse una mejora a los métodos tradicionales de diseño de exámenes; si bien esperábamos que las diferencias en los indicadores cuantitativos fueran más espectaculares, y mostraran una superioridad aplastante de la estrategia, los indicadores no muestran esta gran diferencia. Habría que hacer notar, sin embargo, que en ningún caso hubo diferencias a favor de la estrategia tradicional.

La segunda conclusión que podemos extraer del análisis de la tabla 10 es que las diferentes carreras, al tener historias de desarrollo de la evaluación distintas, ponderaron más cuidadosamente algunos aspectos de la evaluación que otros, resultando que los indicadores no se movieron con la misma magnitud proporcional

en las distintas carreras: las Academias de diferentes carreras mejoraron los exámenes, pero no de manera regular; al parecer, cada Academia centró su interés en alguno de los aspectos de la construcción de los exámenes, y este aspecto se vio beneficiado en mayor medida que los otros.

Como tercera conclusión, relacionada con las dos anteriores, quisiera hacer notar que las Academias fungieron como un organismo regulador que ayudó a mejorar aquellos aspectos más deficientes de las evaluaciones previas: en efecto, los indicadores que ya se mostraban bastante buenos en las versiones de los exámenes realizadas bajo el método tradicional apenas se movieron (la excepción notable es el alpha de Cronbach en la carrera de Enfermería), en cambio, los indicadores que eran más deficientes en las versiones tradicionales son los que tendieron a mejorar a partir de las Academias.

Con respecto a la validez, que finalmente es el aspecto técnico más importante de la evaluación, se puede señalar que el indicador obtenido con la carrera de Cirujano Dentista muestra que tanto el examen tradicional como el realizado por la academia tienen validez de criterio, es decir, sus resultados son concordantes con los de otras mediciones independientes del logro escolar; queda la pregunta acerca de los otros dos tipos principales de evidencias de validez, cuya reflexión no fue dejada de lado:

La validez de constructo (el constructo es el logro escolar) escapa definitivamente de los alcances de este trabajo. Es un constructo utilizado diariamente por estudiantes, docentes y autoridades educativas en todas las instituciones de educación de todo el mundo. Recibe diferentes nombres, como aprovechamiento, conocimientos adquiridos, metas alcanzadas, logros, etc. Pero finalmente siempre se refiere a la capacidad que tienen los alumnos de demostrar lo que adquirieron a través de las clases por medio de un examen. Hacemos sólo la aclaración de que somos conscientes de que no todo lo adquirido por un estudiante es demostrable a partir de un instrumento escrito de opción múltiple, y en muchos casos sólo la parte menos importante de lo adquirido será demostrable. En nuestra descarga, diremos que en todos los casos hicimos conscientes de esta limitación a los participantes de las

Academias, de tal manera que quedaran sensibilizados para intentar hacer instrumentos de evaluación que pudieran medir otros aspectos del logro escolar, una vez que se terminara con la tarea inicial de dejar a punto los exámenes de mitad y final de carrera. Muchas discusiones en las Academias han girado en torno a este tema, aunque aún no existen productos concretos al respecto.

La validez de contenido está intrínsecamente unida al procedimiento de trabajo de las Academias de Evaluación del Logro Escolar. Recuérdese que la segunda fase de trabajo consiste en todos los casos en definir con claridad cuáles son los contenidos básicos estructurales que deben preguntarse para tener claro el nivel de logro de los estudiantes. Puede observarse que esto se atiene a los métodos sugeridos en general para manejar la validez de contenido. Por ejemplo, Martínez Arias (1995) señala que "en la práctica, la validación de contenido supone el examen sistemático del contenido del test, para determinar si es una muestra relevante y representativa del dominio comportamental que se pretende medir" (p.336). La misma autora indica que para ello debe seguirse el procedimiento citado a continuación:

> "1) Definición del universo de observaciones admisibles.
> 2) identificación de expertos en dicho universo.
> 3) Juicio de los expertos acerca del grado en que el contenido del instrumento es relevante y representativo de dicho universo, por medio de un procedimiento estructurado que permita emparejar los ítems con el dominio.
> 4) un procedimiento para resumir los datos resultantes de la fase anterior." (p. 337)

A estas alturas, el lector ya debe haber identificado esta propuesta con los pasos seguidos por las Academias de Evaluación para diseñar los exámenes. Así como la validez de contenido se puede evidenciar a partir del hecho de que los exámenes de las Academias incluyen un mecanismo que asegura el equilibrio entre diferentes asignaturas y profundidad de los contenidos, en los exámenes. También el alpha de

Cronbach nos da indicios de esta validez, debido a que un grupo de preguntas mejor emparentadas deberán dar mejores indicadores en cuanto a su consistencia interna. Recuérdese que en todos los casos, el alpha de Cronbach mejoró en los exámenes hechos por las Academias.

Finalmente, quisiéramos comentar la parte humana, cualitativa, de las Academias de Evaluación. Sostenemos que el logro más importante de las mismas no es el de mejorar algunos indicadores estadísticos en los exámenes, sino el de ser agente de promoción de una cultura de la evaluación en la que la comunidad participe y se comprometa con los principios, métodos, instrumentos y resultados de las evaluaciones.

Entrevistamos a algunos de los participantes más antiguos de varias Academias de Evaluación para que nos dijeran desde su punto de vista los beneficios e inconvenientes que éstas han traído a su trabajo docente. Ellos nos hicieron saber las opiniones en general que han escuchado de otros miembros de las Academias, y los comentarios parecen ser en general positivos: "los profes proponen sentarse en Academia de Evaluación para resolver problemas comunes cotidianamente. Comentan también que gracias a la Academia se logró unir a los jefes de distintos módulos"; "Los profesores de la Carrera creen que ésta es la manera natural de hacer la evaluación: una vez se comentaron experiencias de la Academia con la gente del Poli, que se interesó mucho y dijo no tener nada parecido, lo cual sorprendió a los profesores de la Academia"; "en sus inicios, lo veían como obligación, actualmente los asistentes [...] piensan que sus opiniones sí son importantes y sus experiencias y lecturas en evaluación les dan otra dimensión al quehacer del docente [...] tener una oportunidad de echar una mirada en un panorama integral a toda la carrera".

Puede observarse que algunos de estos beneficios percibidos subjetivamente van más allá de las intenciones de un grupo de evaluación. Algunos profesores lo ven como el único foro compartido en el que pueden conocer a sus compañeros de otras áreas y obtener una visión global de su carrera.

Por otra parte, con respecto a la utilidad que ven los participantes en las Academias, el balance también es positivo. Uno de ellos señala que "a partir de la Academia se creó una cultura de la evaluación en el personal docente de la carrera. El beneficio más tangible es que nos ayudó a crear el examen profesional desde su primera aplicación y a hacer un análisis de los resultados. Nos está ayudando a la evaluación curricular; nos ayudó a identificar los rubros en los que éramos más débiles, para hacer modificaciones y disminuir índices de reprobación". Otro profesor, de una Academia diferente, dice: "se ha logrado cimentar en la mayoría de los participantes que la evaluación es una actividad seria que repercute no sólo en alumnos sino en la institución e incluso en él mismo, porque puede hacer cortes y mediciones que orienten finalmente la toma de decisiones locales si se quiere a nivel de aula, a nivel de grupo, pero también a nivel de la curricula y más aún en las instituciones educativas en general". Consideramos que las Academias de Evaluación del Logro Escolar han logrado ser espacios en que "las lecturas y discusiones, y la oportunidad de ser escuchados y discutir sobre su módulo en particular y la carrera en general les nace y les da oportunidad a un sentido de pertenencia de grupo de trabajo y por otro lado en el sistema modular las posibles líneas de coincidencia o divergencia a pesar de ser un mismo tema común les da sentido o conocimiento mejor y mayor de su módulo en particular y de los demás en lo general".

Existen dos modificaciones a las Academias que los participantes de las mismas plantean como ideas a tomar en cuenta. La primera, es que las Academias se renueven cada cierto tiempo, para darle oportunidad a más profesores de participar en las mismas. La segunda apunta a formalizar las Academias en cuerpos colegiados oficializados, con un reglamento propio y con puntos de encuentro entre las diferentes carreras para conocer las similitudes y diferencias de la experiencia. Actualmente se está haciendo un trabajo de reflexión al respecto de estas sugerencias para tratar de incorporar lo que tengan de pertinentes en esta experiencia de las Academias de Evaluación del Logro Escolar que, desde nuestro punto de vista ha resultado enriquecedora y productiva.

Aiken, L.R. (1996) Tests Psicológicos y Evaluación. Prentice Hall.

American Psychological Association (1985) Standards for Educational and Psychological Testing. Washington, D.C., APA.

ANUIES (1997) "La Evaluación y Acreditación de la Educación Superior en México" En: *Revista de la Educación Superior*, ANUIES, Vol. XXVI (1) No. 101, enero-marzo.

Aragón, B. L. (1990) Elaboración de un instrumento de evaluación conductual, con validez de contenido y de tratamiento, para niños disléxicos. Tesis de Grado. ENEP Iztacala, UNAM.

Burns, W.C. (1996) Content Validity, Face Validity and Quantitative Face Validity. En R.S. Barrett (ed.), Fair employment strategies in human resource management. Quorum Books.

Carrascosa, C. (trad.)(s.d.) Traducción inédita de "Psicología Posmoderna: ¿Una Contradicción de Términos?" de Steinar Kvale.

Carreño Huerta, F. (1991) Enfoques y Principios Teóricos de la Evaluación. Ed. Trillas.

Casillas Muñoz, M. L. (1995) Los Procesos de Planeación y Evaluación. ANUIES.

CENEVAL (2002) Página Web Oficial. URL: http://www.ceneval.edu.mx/

CENEVAL (1998) Instructivo para la elaboración de reactivos. Documento interno.

Chakravart, Laha y Roy (1967). Handbook of Methods of Applied Statistics. Ed. John Wiley. Vol. I. Pp. 392-394

COMIE (1993) Martínez, F.F.; Fuentes Trejo, G.; Cepeda Hinojosa, B.; Burgos Fajardo, R. Estado de Conocimiento 8: Evaluación del Aprendizaje. Comité Organizador del Segundo Congreso Nacional de Investigación Educativa. COMIE.

Crawford, S. (2002) Test Development: Tips for Ensuring Reliability and Validity. URL: http://www.crawfordinternational.com/022002.pdf

Dawson, T.E. (s.d.) "Basic Concepts in Classical Test Theory: Relating Variance Partitioning in Substantive Analyses to the Same Process in Measurement Analyses. URL: http://cricae.net/ft/tamu/dawson.pdf

Díaz Barriga, A. (1982) "Tesis para una teoría de la evaluación y sus derivaciones en la docencia", en *Perfiles Educativos,* num. 15, CISE-UNAM, enero-mayo.

Downie, N. M. y Heath, R. W. (1973) Métodos Estadísticos Aplicados. Ed. Harper and Row.

ETS (2002). Página oficial de Educational Testing Services. URL: http://www.ets.org/international/index.html

Fermín, M. (1971) La Evaluación, los Exámenes, las Calificaciones. Ed. Kapelusz.

GAO (1991) Designing Evaluations. United States General Accounting Office.

García Cortés, F. (1979) "La Evaluación en la Educación". CISE-UNAM, *Revista Perfiles Educativos*, No. 3, enero-marzo.

Gray, B.T. "Controversies Regarding the Nature of Score Validity: Still Crazy After All These Years" Presentado en la reunión anual de la Southwest Educational Research Association, Austin, Enero, 1997.

Herman, J. L., Morris, L. L., & Fitz-Gibbon, C. T. (1987). Evaluators Handbook. Newbury Park, CA: Sage.

Labarca, G. (1973) "Un examen al examen: escuela secundaria en Chile". En: Comunicación y Cultura, No. 1. Buenos Aires, Galerna.

Landsheere, G. (1973) "Crítica de los exámenes". En: Evaluación Continua y Exámenes. Ed. Ateneo.

Linn, R.L. y Baker, E.L. (1996) Assessing the validity of the National Assessment of Educational Progress: NAEP technical review panel white paper. U.S. Department of Education.

Magnusson, D. (1975) Teoría de los Tests. Ed. Trillas, México.

Martínez Arias, R. (1995) Psicometría: teoría de los tests psicológicos y educativos. Madrid, Síntesis.

Matlock-Hetzel, S. (1997) Basic Concepts in Item and Test Analysis. Ponencia presentada en la reunión anual de la Southwest Educational Research Association, Austin, Texas, enero de 1997.

Madrid: Ministerio de Educación, Cultura y Deporte, INCE, (2000) Proyecto PISA. La medida de los conocimientos y destrezas de los alumnos: un nuevo marco de evaluación / OCDE — Madrid.

Miras, M. y Solé, I. (1990) La evaluación del aprendizaje y la evaluación en el proceso enseñanza aprendizaje. En: Coll, C., Palacios, J. Y Marchesi, A. Desarrollo Psicológico y Educación. Ed. Alianza, Madrid, Vol. II.

National Academy of Sciences (1991) Performance Assessment for the Workplace, Volume I. http://books.nap.edu/books/030904538X/html/116.html

Niemi, D. (1996) Instructional influences on content area explanations and representational knowledge: evidence for the construct validity of measures of principled understanding. National Center for Research on Evaluation, Standards, and Student Testing.

Nunally, J. C. (1970) Introducción a la Medición Psicológica. Ed. Paidós. Capítulo 2.

OECD (2000) Education at a Glance. OECD indicators. OECD.

OECD (2002) URL: http://www.pisa.oecd.org/pisa/summary.htm

Raflison, F. (1991) The Case of Validity Generalization. ERIC/TM Digest, ERIC Clearinghouse on Tests, Measurement, an Evaluation, Washington, D.C., July.

Rodríguez Cruz, H. M. y García González, E. (1982) Evaluación en el Aula. Ed. Trillas.

Rudner, L.M. (1993) Test Evaluation. ERIC /AE. (http://136.242.172.58/intass.htm).

Sánchez Moguel, A. (1997) "La validez de la evaluación del aprendizaje en el ámbito educativo", en *Básica revista de la escuela y del maestro*, num. 15, Fundación SNTE, enero - febrero.

Sánchez Moguel, A. (2000) "Las Academias de Evaluación: una Estrategia de Validación de Contenido para Pruebas de Rendimiento Escolar". En: Actas del V Congreso de Metodología de las CC. Humanas y Sociales, Universidad de Sevilla – Kronos, Vol. 2.

Santos Guerras, M. A. (1998) "Patología General de la Evaluación Educativa". En: La Evaluación: un proceso de Diálogo, Comprensión y Mejora. Ed. Aljibe.

Shadish, W. (1998) Presidential Address: Evaluation Theory is Who We Are. American Journal of Evaluation, 19, 1, 1-19.

Silva, A. y Aragón, L. (1997) "La Razón Última de la Naturaleza Cualitativa y Cuantitativa de la Investigación Social ¿Una Conjunción o una Disyunción?" *Acta*

Sociológica, Revista de la Coordinación de Sociología, Facultad de Ciencias Políticas y Sociales de la UNAM. No 19, enero-abril.

Silva, F. y Martorell, C. (1991) "Evaluación Conductual y Evaluación Tradicional: la Cuestión Psicométrica". En: V. E. Caballo (Ed.), Manual de técnicas de terapia y modificación de conducta. Siglo XXI, Madrid, España.

Stapleton, C.D. (1997) Basic Concepts in Exploratory Factor Analysis (EFA) as a Tool to Evaluate Score Validity: A Right-Brained Approach. Presentado en la reunión anual de la Southwest Educational Research Association, Austin, Enero, 1997.

Tirado Segura, F. y Serrano Carrillo, V. (1989) "En Torno a la Calidad de la Educación Pública y Privada en México" CONACyT, *Revista Ciencia y Desarrollo*, Vol. XV, num. 85, marzo-abril.

Thorndike, R. L. y Hagen, E. P. (1977) Measurement and Evaluation in Psychology and Education. E. John Wiley & Sons. Vol II.

Tourón, J. (1989) La Validación de Constructo: su Aplicación al CEED (Cuestionario para la Evaluación de la Eficacia Docente). En: *Revista Bordón*, V. 41 (3-4).

Valle Gómez–Tagle, R.; Meraz Ríos, P.; y Valenzuela Medina, M. (1995). Aplicación de tres modelos estadísticos en la calibración de un examen de español. En: Memorias del Foro Nacional de Evaluación Educativa, Colima, Col. CENEVAL.

Williams, Paul L. (1989). Using customized standardized tests. *Practical Assessment, Research & Evaluation*, 1(9). Available online: http://ericae.net/pare/getvn.asp?v=1&n=9.

Yu, Alex (2002). Using SAS for Item Analysis and Test Construction. URL: http://seamonkey.ed.asu.edu/~alex/teaching/assessment/alpha.html