Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

"Aplicación y validación de un método para determinar el punto de corte del Examen de Egreso de Inglés de la Universidad Autónoma de Baja California"

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS EDUCATIVAS

Presenta

Ma. Del Carmen Enriqueta Márquez Palazuelos

Ensenada B.C. Junio del 2005

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Maestría en Ciencias Educativas

"Aplicación y validación de un método para determinar el punto de corte del Examen de Egreso de Inglés de la Universidad Autónoma de Baja California.".

T E S I S

que para obtener el grado de

MAESTRO EN CIENCIAS EDUCATIVAS

Presenta

Ma. del Carmen Enriqueta Márquez Palazuelos

APROBADO POR:

Dra. Norma Larrazolo Reyna
Directora de Tesis

Dr. Eduardo Backhoff Escudero
Sinodal

M.C. Javier Organista Sandoval
Sinodal

M.C. Virginia Velasco Ariza
Sinodal

Ensenada B.C. Septiembre 2005

DEDICATORIA

A Victor,

Por compartir siempre mis sueños, por su apoyo incondicional e infinita paciencia

A mis queridos hijos,

Victor Gabriel, Luis Francisco y Carlos Alberto, por su comprensión a mis obligadas ausencias

AGRADECIMIENTOS

A mi Directora de Tesis, Dra. Norma Larrazolo Reyna, con especial estimación por compartir sus conocimientos y otorgarme su paciencia tiempo y dedicación para la culminación de este trabajo.

Al M.C. Javier Organista, M.C. Virginia Velasco y Dr. Eduardo Backhoff por sus valiosas aportaciones y reiteradas muestras de apoyo.

A todos los maestros del IIDE por su dedicación y pasión por transformar a los que llegamos a sus aulas, especialmente a la Dra. Lucia Aguirre y M.C. Luis Angel Contreras de quienes siempre recibí muestras de afecto y palabras de aliento.
Gracias.

Al Instituto de Investigación y Desarrollo Educativo por permitirme ser parte de sus alumnos, de lo cual me siento orgullosa.

A Rebeca Vidal y Saúl González de la Facultad de Idiomas por darme las facilidades para hacer realidad este sueño.

Con sincero afecto a mis compañeros y amigos de la Maestría por permitirme compartir sus vidas en la cotidianidad de nuestra aula, especialmente a Angie y Elisa, así como a Luz Elena y Marisela.

A mis entrañables amigas Paty, Any, Lily, y Claudia, por acompañarme siempre.

CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

CAPÍTULO I

INTRODUCCIÓN

La búsqueda de la calidad y la excelencia en materia de educación a nivel superior ha llevado a evaluar cada vez más los diferentes aspectos del proceso enseñanza-aprendizaje. Esta tendencia ha tenido grandes repercusiones a nivel nacional, que van desde la creación de organismos colegiados dentro de la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES), consejos nacionales, consejos regionales, consejo de universidades, hasta la creación en 1994 del Centro Nacional de Evaluación para la Educación Superior (CENEVAL) como organismo evaluador (Revista de la Educación Superior en México, Vol. XXVI, 1997).

La evaluación del aprendizaje se considera un proceso sistemático y permanente que comprende la búsqueda y obtención de información sobre los diferentes procesos del quehacer educativo (Fermín, 1971). La determinación de los factores que inciden en la valoración de las metas propuestas y la toma de decisiones son parte importante del proceso de evaluación. Para ello, la evaluación se vale de la medición para obtener información que permita la toma de decisiones en forma objetiva. La medición por su parte, utiliza las pruebas como instrumentos que indican en qué medida se han obtenido los resultados esperados.

Con este propósito, algunas universidades han iniciado diversas acciones que permitan consolidar su sistema de evaluación y acreditación. A mediados de los setentas una de las primeras universidades en desarrollar un banco de reactivos de opción múltiple para la titulación de la carrera de médico cirujano fue la Universidad Nacional Autónoma de México. Esta universidad cuenta también con exámenes de selección para el ingreso a bachillerato y licenciatura. Otras universidades, tanto públicas como privadas, han optado por crear sus propios exámenes de selección para el ingreso de sus estudiantes, como la Universidad Autónoma de Aguascalientes y la Universidad Iberoamericana, entre otras (Martínez, 2001).

La Universidad Autónoma de Baja California (UABC), al igual que otras universidades, no se ha quedado atrás en la búsqueda de la calidad y excelencia en la educación. Actualmente cuenta con el Examen de Habilidades y Conocimientos Básicos (EXHCOBA) (Backhoff y Tirado, 1992), desarrollado por la propia universidad para el ingreso de sus estudiantes, así como el Examen de Egreso del Idioma Inglés (EXEDII) (Larrazolo y Velasco, 2000), elaborado este último como una opción para cumplir con el requisito de certificación del dominio de una lengua extranjera a nivel intermedio.

En la evaluación a gran escala es posible identificar dos grandes categorías de pruebas: las referidas a una norma y las referidas a un criterio.

Popham (1990) indica que fue Glaser en 1963, quien hizo mención por primera vez de la diferencia entre medir con referencia a una norma (desempeño de un alumno en comparación con la ejecución de otros estudiantes) y medir con referencia a un criterio (desempeño con relación a la clara descripción de los contenidos que se evalúan). Estos dos tipos de pruebas difieren tanto en su construcción como en la interpretación de los resultados.

Para desarrollar y aplicar pruebas a gran escala se requiere de un grupo de especialistas, tanto en evaluación como en las áreas del conocimiento que se pretende evaluar. En la literatura se pueden encontrar algunos ejemplos sobre evaluación a gran escala en materia de educación a nivel internacional, sobre todo en Estados Unidos, donde se cuenta con una experiencia de aproximadamente 100 años (Cizek, 2001). En Hispanoamérica y particularmente en México, son muy recientes los esfuerzos que se han realizado en materia de evaluación a gran escala. A partir de 1990 algunas instituciones nacionales de nivel superior han incursionado en la elaboración de exámenes de este tipo (Martínez, 2001).

Varios autores (Popham, 1990; Hambleton, 1980; Klein, 1990; entre otros) están de acuerdo en que las pruebas referidas a un criterio son las más adecuadas cuando se utilizan con el propósito de certificación. Las pruebas con referencia a un criterio requieren que un individuo demuestre el grado de competencia adquirido en relación con los contenidos previamente establecidos

con respecto a un área específica de dominio. Por el contrario, en el caso de las pruebas con referencia a una norma, se requiere que el individuo demuestre su nivel de desempeño con relación al grado de competencia de otros individuos.

Los exámenes a gran escala se pueden aplicar con diferentes propósitos. Entre éstos se pueden mencionar los exámenes de diagnóstico utilizados para saber el nivel de conocimiento de los examinados para el ingreso a diversos niveles de instrucción y los exámenes de certificación utilizados con propósitos de promoción o egreso de algún programa de instrucción. Este tipo de exámenes se considera de alto impacto por la repercusión que tienen en la vida académica del estudiante: ingresan o no ingresan; o egresan o no egresan.

Cuando se utilizan pruebas con propósito de certificación, las exigencias en la calidad del examen aumentan considerablemente. Por ello, es sumamente importante contar con evidencias de la calidad del instrumento de medición, que permita apoyar la toma de decisiones de forma objetiva y confiable, ya que "*una prueba es un instrumento de medición que se crea para inferir una medida de las capacidades de los sujetos a través de las respuestas que éstos dan*" (Larrazolo y Velasco, 2000, p. 111).

El EXEDII es una prueba a gran escala con referencia a criterio. Las pruebas a gran escala usualmente tienen un gran impacto, tanto en el *currículo*, como en los profesores y alumnos, por ello, es necesario contar con

instrumentos confiables de evaluación que indiquen el grado de competencia que los estudiantes han adquirido. Por otro lado, en las pruebas con referencia a un criterio los resultados deberán compararse con un criterio determinado o punto de corte, que permita establecer, si el estudiante tiene el nivel de desempeño requerido o si no lo tiene.

El EXEDII se diseñó siguiendo el modelo de Hambleton (1988) para el desarrollo de pruebas referidas a un criterio y se complementó con los criterios de Popham (1978) y Nitko (1994) en lo que se refiere a las especificaciones de las preguntas y la elaboración de las pruebas con referencia a criterio. Este examen evalúa dos habilidades comunicativas y una en el manejo de la estructura del idioma: comprensión de lectura, comprensión auditiva y gramática. Estas áreas del conocimiento fueron seleccionadas por un comité de especialistas después de analizar y considerar el currículo del curso de inglés a nivel intermedio (nivel III) que ofrece la Escuela de Idiomas de la UABC así como las habilidades que un estudiante de licenciatura requiere para manejar el idioma inglés como lengua extranjera en este nivel.

Actualmente este examen se aplica por computadora para poder implementarse a gran escala y se considera de gran impacto, puesto que su acreditación se establece como uno de los requisitos de egreso de la licenciatura. Como se mencionó anteriormente, el EXEDII se diseñó como un examen con referencia a un criterio, ya que de acuerdo con Popham (1990) las

pruebas con referencia a un criterio son las más adecuadas cuando el propósito es de certificación. Por otro lado, la aplicación de pruebas por computadora mejora la eficiencia de la evaluación (Larrazolo, 2002) al permitir mantener un mayor nivel de seguridad en la aplicación de las mismas. Asimismo, con relativa facilidad se puede implementar un proceso de selección aleatorio de los reactivos, de tal forma que cada estudiante cuente con una versión diferente de acuerdo al nivel de conocimientos mostrado, lo que apoya la seguridad en la aplicación.

El EXEDII cuenta con un punto de corte establecido en forma normativa (*a priori*), con base en el desempeño promedio de un grupo piloto de estudiantes de nivel intermedio de la Escuela de Idiomas de la UABC en las tres unidades académicas (Ensenada, Tijuana y Mexicali). Como medida inicial se puede implementar un punto de corte con información real sobre el desempeño mostrado por los grupos instruidos. Sin embargo, la población a la que se destina el examen es diferente ya que son aquellos estudiantes que consideran tener el dominio del idioma y no se conoce la forma en que adquirieron sus conocimientos. Por ello, se debe asegurar que efectivamente tengan el dominio de los contenidos requeridos. Esto se logra estableciendo un punto de corte con referencia a un criterio.

1.1 Planteamiento del Problema

Cuando se utilizan pruebas referidas a un criterio para otorgar un certificado de estudios, la importancia que adquiere la construcción y su interpretación es considerable. El riesgo de cometer errores de interpretación de resultados afectará el futuro académico de los examinados. En las pruebas referidas a un criterio hay dos aspectos de vital importancia: 1) la definición de las áreas de dominio y 2) la definición de los estándares de desempeño (Meskaukas, 1976).

El punto de corte de una prueba se debe establecer con base en los objetivos que se persiguen. Las pruebas que utilizan puntos de corte establecidos de manera normativa usualmente persiguen el objetivo de mejorar la instrucción, establecer programas de tipo remedial, etc. Al establecer el punto de corte de manera normativa se utilizan porcentajes para interpretar los resultados de un estudiante en relación con las calificaciones de otros estudiantes, lo que proporciona resultados relativos. En cambio, cuando se busca establecer puntos de corte para pruebas con propósito de certificación se busca diferenciar a los examinados en dos categorías: los que tienen el dominio y los que no lo tienen. Por ello, es necesario establecer estándares de ejecución, los cuales se obtienen únicamente a través de la implementación de puntos de corte de tipo absoluto (Payne, 1992).

Si consideramos que el EXEDII es una prueba con referencia a criterio y que su punto de corte debe ser de tipo absoluto y no relativo, se plantea la necesidad de realizar un nuevo estudio para establecer el punto de corte con base en un criterio absoluto que permita diferenciar a los estudiantes que tienen el dominio del idioma inglés de los que no lo tienen. Este nuevo punto de corte permitirá valorar el nivel de competencia del estudiante en relación con las áreas de dominio establecidas en el programa de inglés para el nivel intermedio de la Escuela de Idiomas de la UABC.

La legitimidad del punto de corte es fundamental para dar validez a las pruebas elaboradas con propósitos de certificación. La validez del punto de corte depende, tanto de la construcción de la prueba, como del procedimiento que se haya seguido para establecer este parámetro. Desde la década de los 50's, han surgido varios métodos para establecer puntos de corte (Nedelsky, 1954; Angoff, 1971; Ebel, 1972; Jaeger, 1978; Livingston y Zieky, 1982; por mencionar algunos). Sin embargo, no se ha encontrado el método ideal para establecer un punto de corte confiable. La mayoría de los métodos se basan en las decisiones y/o juicios que emite un grupo de expertos sobre un examen. Varios investigadores (Glass, 1978; Shepard, 1979; entre otros) han querido eliminar el factor de juicio por considerarlo arbitrario. Sin embargo, en la actualidad existen más de 38 métodos que buscan minimizar el factor de arbitrariedad del juicio de valor (Hambleton, 1978; Livingston y Zieki, 1982; Berk, 1986; Jaeger, 1989; entre otros).

El problema que se plantea en este trabajo de investigación es el de determinar un nuevo punto de corte utilizando el método más adecuado para pruebas con referencia a un criterio. De esta manera se podrá comparar el punto de corte obtenido, con el punto de corte vigente. Así validar o modificar el punto de corte para el EXEDII, examen que utiliza la Universidad Autónoma de Baja California.

1.2 Objetivo General

El objetivo principal de este trabajo es determinar el punto de corte con base a un criterio para el Examen de Egreso del Idioma Inglés (EXEDII). Para lograr esto, habrá que atender los siguientes objetivos particulares:

1) Buscar y seleccionar el método más adecuado para establecer el punto de corte de entre los métodos existentes para pruebas con referencia a un criterio considerando la objetividad, claridad y facilidad de aplicación.

2) Implementar el proceso con el método seleccionado para establecer el punto de corte.

3) Reunir evidencias para validar el proceso de establecimiento del punto de corte.

1.3 Justificación

El hecho de que algunos estudiantes de la UABC opten por presentar el EXEDII para cumplir con el requisito de egreso de la universidad hace que el punto de corte del examen sea de suma importancia para la institución y para la comunidad universitaria de estudiantes. Establecer el punto de corte de una prueba que determina la certificación del dominio del idioma inglés de los estudiantes para efecto de egreso lleva implícitos aspectos éticos, políticos y sociales difíciles de soslayar (Cizek, 2001).

Como Popham (1990) sugiere, se requiere que el examen haya sido aplicado en varias ocasiones para establecer el punto de corte. Esto quiere decir que por lo menos se deberá contar con una muestra del desempeño de grupos de examinados con el fin de tener información que pueda guiar el proceso de determinación de un punto de corte con una visión real del desempeño de los estudiantes y del programa a nivel intermedio del idioma inglés de la Escuela de Idiomas de la UABC. A continuación se hace una breve descripción de la importancia de este trabajo en el contexto de la relevancia académica y social que implica determinar el nuevo punto de corte del EXEDII con base a un criterio.

a) Relevancia Académica.

En la literatura existente sobre los métodos disponibles para establecer puntos de corte se ha encontrado que éstos se basan en juicios de valor que se cuestionan por el elemento de subjetividad que tienen implícito. A pesar de que existe el elemento arbitrario de juicio, se han realizado varias investigaciones y se han aplicado una gran variedad de métodos que prueban que se puede determinar el punto de corte de las pruebas con referencia a un criterio de una manera más científica (Popham, 1978).

b) Relevancia Social.

La importancia que tiene el presente estudio para los objetivos del Examen de Egreso del Idioma Inglés de la UABC es clara. Desde el punto de vista social y académico, a la UABC le interesa que los estudiantes que han cumplido con todos los créditos requeridos por sus facultades de adscripción puedan egresar y no se vean detenidos por el requisito de presentar el examen de un idioma extranjero que en la mayoría de las veces es el inglés.

Por lo anterior, es conveniente contar con investigaciones para establecer un punto de corte adecuado al Examen de Egreso del Idioma Inglés (EXEDII) que permitan efectuar una comparación, con el fin de validar o modificar el punto de corte existente. Esto evitará que se determine un punto de corte demasiado

alto que penalice un gran número de examinados, o por el contrario, que sea tan bajo que permita egresar a aquellos que no tienen el dominio adecuado (Zieky, 2001, p.46). De esta manera, se podrá contar con la información que permita separar con objetividad y justicia, quienes tienen el dominio del idioma de los que no lo tienen y así poder tomar la mejor decisión sobre la certificación del nivel mínimo de competencia que debe establecer el examen de inglés para efecto de egreso.

1.3 Limitaciones del estudio

Generalmente, el tipo de estudios que se llevan a cabo para determinar puntos de corte involucran la valoración de juicios de expertos sobre el contenido de un examen o sobre el desempeño de los examinados.

Cuando se utiliza la valoración de expertos comúnmente llamada "jueceo", se recomienda contar con un número grande de especialistas o jueces que permita desarrollar diversas estrategias para recabar mayores evidencias de validez del proceso de valoración. Una de estas estrategias es formar subgrupos que reciban el mismo entrenamiento y procedan a valorar el examen en forma independiente, de tal manera que se puedan obtener puntos de corte por subgrupo para contrastar la homogeneidad de los juicios.

Una limitación para este estudio fue la imposibilidad de contar con un grupo grande de especialistas conformado por académicos de las tres sedes (Tijuana, Mexicali y Ensenada). La razón principal fue la disponibilidad de tiempo de los académicos para llevar a cabo reuniones consecutivas en las diferentes sedes. Otro obstáculo fue la complejidad que representaba implementar el proceso a distancia, por lo cual el estudio se llevó a cabo sólo con académicos de la Escuela de Idiomas de Ensenada.

Organización de la tesis

En el primer capítulo de esta tesis se presenta una breve reseña de la evaluación con relación a las pruebas educativas con el propósito de definir el contexto en el cual se desarrolla este estudio. Asimismo, se presenta tanto el objetivo general como los objetivos particulares a realizar.

En el capítulo 2 se hace mención del desarrollo histórico de las pruebas educativas y los conceptos metodológicos y psicométricos relacionados con la evolución de los estándares y el desarrollo de puntos de corte. Por otra parte, se revisa la construcción del EXEDII, ya que es parte primordial del presente estudio.

El capítulo 3 describe las diferentes etapas del proceso y el desarrollo de la guía de instrucción así como el entrenamiento de los especialistas. Esta parte

del estudio es sumamente importante ya que la descripción cuidadosa es parte de las evidencias de validez. En el capítulo 4 se analizan los resultados obtenidos y por último, el capítulo 5 presenta las conclusiones y recomendaciones propuestas.

CAPITULO 2

FUNDAMENTACIÓN TEÓRICA

En este capítulo se presentan los aspectos psicométricos más importantes de la evaluación educativa en relación con los estándares y el contexto en el que se aplican así como el "estado del arte" en que se encuentra la implementación del punto de corte. Para ello, se revisará lo que existe hasta la fecha sobre como establecer criterios para el punto de corte, las pruebas y los propósitos para los cuales se desarrollan y por último, se presentará una breve descripción del contenido del EXEDII, examen para el cual se desarrolló el punto de corte en este estudio.

La literatura dice que los primeros en utilizar la evaluación fueron los chinos por el año 2,200 A.C. Ellos emplearon pruebas con el propósito de calificar al personal que serviría como oficiales al reino chino y lo siguieron haciendo por más de 3,000 años. A principios del siglo XIX, Voltaire trajo este sistema a Francia y fue adoptado posteriormente por los británicos. Fue en 1860 cuando este sistema se llevó a los Estados Unidos. En 1897, Thorndike aplicó pruebas a gran escala por primera vez a 33,000 niños para determinar su desempeño en ortografía. Para 1940 y 1950 se hacía uso frecuente de la evaluación a través de la aplicación de pruebas con el objetivo de identificar las diferencias entre los estudiantes con respecto a sus habilidades, desempeño, etc. (Popham, 1990).

Es así como surgen las pruebas educativas que intentan medir a gran escala el grado mínimo de competencia. A partir de los años 50, el uso de pruebas estandarizadas se ha extendido con diversos propósitos, entre los que podemos mencionar el de mejorar la instrucción y el de otorgar diplomas o certificación de estudios a los estudiantes.

2.1 Estándares

La toma de decisiones es una tarea cotidiana. Diariamente un gran número de personas se ven en la necesidad de tomar decisiones y desean elegir la más acertada. Para tomar la mejor decisión se requiere contar con información objetiva. Para que la información sea objetiva debe poderse medir y comparar con un estándar o criterio específico. En la actualidad se tienen estándares en todos los órdenes de la vida diaria de la sociedad. Por ejemplo: en el ramo de la construcción, en seguridad industrial, en la salud, en las telecomunicaciones, en el campo del comercio y la economía, los estándares se establecen y se mejoran para promover una mejor calidad de vida de los ciudadanos.

Los estándares surgieron en el área de la industria en respuesta a los cambios en la tecnología y a la búsqueda de la calidad. En 1946 nace la Organización Internacional para la Normalización ISO, (*International Standard Office* por sus siglas en inglés), con el propósito de promover estándares o

normas internacionales para la manufactura, el comercio y la comunicación. Este organismo creó una serie de normas (ISO) que proveen los requerimientos necesarios para asegurar la calidad en la producción industrial internacional (Ravitch, 1995).

En el contexto de la educación sucede exactamente lo mismo. Se requiere contar con criterios claros y precisos de las habilidades y destrezas que deben ser aprendidas por los estudiantes, complementados a su vez con indicadores que definan con claridad lo que el estudiante sabe o puede hacer.

La evolución de los estándares en la educación ha tenido una larga historia. En los Estados Unidos su inicio se remonta a 1892, con la creación del Comité de Asociación Americana de Psicología (APA), creada con el fin de investigar la factibilidad de estandarizar la medición física y mental existente. Posteriormente, surgen otros grupos y organizaciones, tanto públicas como privadas, destinadas a la medición en educación como son: *The National Assessment of Educational Progress* (NAEP), *The National Council of Measurement in Education* (NCTM), por mencionar algunos. En la década de los ochenta, se dio un resurgimiento de los estándares en el campo educativo con la publicación del informe "Una Nación en Riesgo" (1983) de la Comisión Nacional sobre Excelencia en Educación. En este documento se pusieron en evidencia las deficiencias de la educación en Estados Unidos. Esto dio lugar a una gran

preocupación por la calidad de la educación y dio inicio a un movimiento hacia el rendimiento de cuentas en la educación (Zieky 2001).

A pesar del desarrollo de los estándares en los últimos tiempos, aún queda una pregunta por responder: ¿Qué es un estándar? Un estándar es tanto una meta (*lo que debe hacerse*) como una medida de progreso hacia esa meta (*cuán bien fue hecho*). Se puede decir que los estándares ofrecen una perspectiva realista de evaluación. Un estándar real debe estar siempre sujeto a observación, evaluación y medición.

El error más común que se comete con respecto a los estándares es el de confundir estándar con evaluación. Aún cuando van de la mano, no son lo mismo. El estándar se vale de la evaluación para proveer información objetiva sobre qué tan bien se ha alcanzado el estándar. Al confundir estándar con evaluación se hace hincapié en las pruebas, en particular si son confiables, si miden lo que deben de medir, si son justas, etc., olvidando el aspecto de que el estándar es una meta, hacia el fin específico de lo que la educación debe de ser.

Por otra parte, es importante sobre todo para este estudio, distinguir la diferencia entre un estándar o criterio y el proceso o método de implementación de este criterio. El estándar es un resultado. El proceso o método es el medio a través del cual se llega a este resultado. Un método para establecer estándares requiere que se identifique un punto en la escala de puntuaciones que defina la

competencia o no competencia de un examinado para la toma de decisiones (Hambleton, 1998). De esta manera, se puede concluir que la meta de un método para establecer estándares es identificar la puntuación en la escala de resultados que represente el nivel de conocimientos y habilidades requeridos para que un individuo pueda ser considerado como competente.

Los estándares pueden ser voluntarios, obligatorios o de hecho. Los estándares voluntarios pueden ser establecidos por organizaciones privadas o profesionales y lo que los distingue, es que son de uso accesible para todos. En el caso de los estándares obligatorios, estos son impuestos por ley. Por otra parte, los estándares de hecho son aquellos que son impuestos por la costumbre o convenciones sociales (Ravitch, 1995).

En el campo educativo, el efecto de la globalización se muestra en la tendencia que existe hacia la internacionalización de los estándares. Algunas materias, como las matemáticas y la ciencia, son susceptibles de permitir esta estandarización, ya que no están relacionadas con el género o el país de donde provienen los estudiantes. La validez de las ciencias geográficas, biológicas o físicas son independientes de la identidad de quienes requieren su estudio. Buscar la internacionalización de los estándares tiene muchas implicaciones. Una de ellas, es lograr un acuerdo de lo que deben saber los estudiantes de diferentes países. Con ello, se logrará que compitan en igualdad de condiciones, sin importar de qué país provengan (Ravitch, 1995).

Algunos países, como Gran Bretaña, Francia y Japón, han establecido estándares nacionales con el propósito de asegurar una educación de calidad y un mejor rendimiento académico. En esos países, los alumnos presentan exámenes nacionales al término de su enseñanza secundaria para garantizar su ingreso a la educación superior. Una de las ventajas de establecer estándares nacionales es asegurar la misma oportunidad de educación para todos los estudiantes.

En cambio, en los Estados Unidos aún se siguen utilizando estándares a nivel estatal. Desde su creación en 1892, el Comité de la Asociación Americana de Psicología (APA), se ha dedicado a investigar la factibilidad de estandarizar la medición física y mental existente. Con la publicación en 1983 del informe de la Comisión Nacional sobre la Excelencia en Educación, cada estado de la unión americana empezó a implementar pruebas a gran escala para conocer el desempeño de sus estudiantes. La publicación de los resultados ha permitido establecer comparaciones entre ellos y a nivel nacional.

En 1993, se creó el Consejo Nacional sobre Estándares y Pruebas en Educación (NCEST por sus siglas en Inglés "*National Council on Education Standards and Testing*"), con el objetivo de establecer un consenso nacional sobre los estándares. Posteriormente, surgieron otros grupos y organizaciones, tanto públicas como privadas, enfocadas a la medición en educación como son:

The National Assessment of Educational Progress (NAEP), *The National Council of Measurement in Education* (NCTM), entre otras.

Como consecuencia de lo anterior, surge un énfasis en la medición educativa o Psicometría. Aunque se puede afirmar que los estándares en educación se han utilizado por mucho tiempo, el concepto ha estado sujeto con el paso del tiempo a diversas interpretaciones y a diversas controversias dada su subjetividad. El término estándar se ha utilizado para describir un gran número de situaciones educativas, tanto generales como específicas. Por ejemplo, cuando se habla de propósitos básicos de la educación como son los programas de competencia mínima, los objetivos curriculares, la medición del desempeño de los estudiantes, etc. (Noonan, 1996).

En México, es relativamente reciente el interés por la evaluación a gran escala. En 1970, se creó la Subdirección de Evaluación y Acreditación en la Secretaría de Educación Pública (SEP) con el propósito de estudiar las características y la calidad del sistema educativo de México. Esta organización ha efectuado algunos esfuerzos a través de los años para evaluar de manera estandarizada la educación en México, enfocándose principalmente en la educación primaria y en la evaluación del profesorado, para lo cual se implementaron diversos estudios y acciones. Por ejemplo, en 1970 se llevó a cabo un estudio de aptitud con niños de sexto grado de primaria. Con base en

los resultados, se construyó un examen de ingreso a la educación secundaria (Himmel, 2000).

Durante el período de 1983 a 1988 este organismo (SEP) desarrolló y aplicó exámenes a los egresados de las escuelas de capacitación de profesores. De 1995 a 1999 aplicó exámenes a aproximadamente 600 mil docentes y a 7 millones de alumnos entre el tercer grado de educación primaria y el tercer grado de educación secundaria en casi todas las escuelas secundarias y primarias del país. Sin embargo, los resultados de estos exámenes no han impactado la toma de decisiones en el contexto educativo, ya que se utilizaron primordialmente para promover una mejora al salario de los maestros con la implementación del programa de Carrera Magisterial. El objetivo primordial de los exámenes era el de mantener actualizados a los maestros, pero este fin se desvirtuó al enlazar los resultados a un apoyo económico adicional al salario de los maestros.

En 1994 se creó el Centro Nacional para la Evaluación de la Educación Superior, A.C., (CENEVAL) (Backhoff *et al.,* 2000). Este organismo fue fundado como una dependencia gubernamental descentralizada dedicada al diseño y validación de los exámenes nacionales (ingreso a bachillerato, a licenciatura y los exámenes generales de calidad profesional (EGCP) para egresados de licenciatura). El CENEVAL ha generado a la fecha diversos exámenes, como son: los exámenes nacionales de ingreso a la educación media superior (EXANI-

I); a la educación superior (EXANI-II); al posgrado (EXANI-III); 23 exámenes generales para el egreso de licenciaturas en diversas áreas, por ejemplo: psicología, ingeniería electrónica, medicina, etc.; exámenes de certificación de competencias laborales en medicina veterinaria y zootecnia (ECEP); exámenes de la Universidad Pedagógica Nacional; exámenes del acuerdo 286 para la acreditación del nivel licenciatura y técnico profesional así como para la acreditación de estudios de bachillerato.

Congruentes con el tema de evaluación y conscientes de la necesidad de contar con indicadores más objetivos que guíen este proceso, Martínez *et al* (2000) auspiciados por el CENEVAL, publicaron un documento sobre estándares de calidad para instrumentos de evaluación educativa.

En el campo de la evaluación, el más reciente esfuerzo a nivel nacional ha sido la creación del Instituto Nacional para la Evaluación de la Educación (INEE), por decreto presidencial del 8 de agosto del 2002. Esta es una dependencia gubernamental descentralizada cuyas funciones son las de implementar indicadores de calidad del sistema educativo nacional y de los subsistemas estatales, de pruebas de aprendizaje y de evaluación de escuelas.

En el contexto de la educación se manejan tres significados dentro del término de estándares:

1) *Estándares de contenido o estándar curricular.* Estos estándares tienen que ver con lo que los maestros deben enseñar y los alumnos aprender. Estos estándares están enlazados directamente al contenido y deben contar con descripciones claras y precisas de los conocimientos que los alumnos deben adquirir de tal forma que a través de la medición de estos contenidos los estudiantes puedan demostrar el nivel de dominio de la habilidad o conocimiento adquirido

2) *Estándares de oportunidades de aprendizaje.* Estos estándares están relacionados con la disponibilidad de los programas, del personal y los recursos con que cuentan las escuelas e instituciones dedicadas a la enseñanza, para asegurar la equidad y la igualdad de oportunidad para todos.

3) *Estándares de ejecución y niveles de logro.* Estos últimos representan el grado en que se logró el estándar de contenido o sea un desempeño aceptable, superior o inadecuado (Noonan, 1996). Es muy importante distinguir, sobre todo para este estudio, la diferencia que existe entre estándares y el establecimiento de un criterio para este tipo de estándar.

El hecho de establecer estándares dentro del campo de la evaluación educativa está íntimamente ligado con la medición. Uno de los roles de la evaluación educativa es la de proveer información sobre el nivel de competencia que los individuos han alcanzado en relación con los diferentes campos de

estudio. Para ello, la evaluación educativa hace uso de la medición. La medición a su vez, hace uso de las pruebas como "un método para la obtención de datos con el propósito de efectuar comparaciones entre individuos" (Payne, 1992).

Sin embargo, la medición sin la comparación con un estándar o criterio es inútil, ya que la medición por sí misma no indica nada. La medición es la evaluación expresada en términos cuantitativos. Para hacer que ese número sea significativo es necesario compararlo con algo que indique con qué extensión se han logrado los objetivos especificados (Woolfolk, 1996). Ese algo recibe el nombre de criterio o estándar, el cual nos indica "qué" y "cuánto" sabe el estudiante.

Una decisión con base en la medición que es muy común en el campo de la educación es la de clasificar a los examinados en dos categorías: los que obtuvieron una puntuación alta y aquellos cuya puntuación fue baja. La toma de decisión se utiliza para diversos aspectos, por ejemplo: para decidir quienes tomarán un curso remedial y quienes no; para recibir un diploma o certificado; para practicar una profesión; para recibir entrenamiento, etc. Son muchos los propósitos que requieren una toma de decisión. Las pruebas proporcionan una muestra del desempeño de los examinados a través de las puntuaciones. Lo que se obtiene es una distribución del conocimiento que requiere definir cuánto de este conocimiento es suficiente para demostrar un dominio aceptable de acuerdo con el propósito de la prueba.

Así, las puntuaciones se comparan con un criterio que defina qué tanto es suficiente. Estos criterios pueden ser absolutos o relativos. El criterio será relativo cuando la puntuación de un individuo se compara con las puntuaciones obtenidas por otros examinados. Así, si las puntuaciones de ese grupo son altas, el criterio será alto, o viceversa. Por otro lado, también se habla de un criterio absoluto cuando las puntuaciones de un individuo no dependen del desempeño de otros. Para saber si el criterio del que se habla es absoluto o relativo, será necesario conocer la interpretación que se da a las puntuaciones. Por ejemplo, si se dice que el criterio que se utilizó es 60 de 100, esto no indica mucho. Por otro lado, si este 60 indica un 60 por ciento de las preguntas correctas, indica un criterio absoluto. Ahora bien, si este 60 representa 60 por ciento mejor que los otros examinados, este valor muestra un criterio relativo (Popham, 1990, p.27).

Como se menciona en el capítulo uno, la medición con referencia a un criterio introdujo la idea de que se pueden medir varios aspectos, tales como: el conocimiento, la habilidad, la actitud, etc., a lo que Glaser (1963) le dio el nombre de estándar de desempeño para describir este criterio (citado por Noonan 1996). Se puede decir que las pruebas con referencia a un criterio se definen como aquellas que evalúan el nivel de desempeño del individuo contra objetivos específicos, o áreas de dominio de conductas bien definidas. Por lo tanto, una prueba con referencia a criterio deberá contar con un estándar de desempeño claramente especificado y un punto de corte bajo el cuál se evaluarán los resultados.

2.2 Pruebas con referencia a un criterio o a una norma.

No se puede hablar de estándares sin hablar de las pruebas con referencia a una norma y las pruebas con referencia a un criterio. Antes de los años 70's, las pruebas con referencia a una norma se utilizaban ampliamente en el campo de la evaluación educativa. Estas pruebas surgieron como opción para medir la inteligencia. Según Popham (1990), fue Glaser en 1963 quien con su artículo *"Instructional Technology and the Measurement of Learning Outcomes: Some Questions"* hace notar por primera vez la diferencia de medir con referencia a una norma y medir con referencia a un criterio.

Estos dos tipos de pruebas difieren tanto en su construcción como en la interpretación de los resultados. Así, una prueba referida a una norma se construye con un número relativamente grande de reactivos, midiendo en forma más general las habilidades, aptitudes o conocimientos del examinado. En cambio, un examen basado en un criterio se enfoca a medir áreas de dominio en forma específica, dando así resultados absolutos. Se puede decir entonces, que el uso de pruebas referidas a una norma tiene resultados óptimos cuando se quiere medir, en forma amplia, un área del conocimiento y que las pruebas referidas a un criterio se utilizan, con buenos resultados, cuando se desea evaluar la competencia en un área específica de dominio (Klein, 1990). Los tres aspectos que diferencian a las pruebas con referencia a una norma y las pruebas con referencia a un criterio, son: 1) la manera en que se construye. 2) la

interpretación de los resultados y 3) los propósitos a que se destina (Hambleton, 1988). En la tabla I, se presenta una descripción detallada de cada uno de los aspectos que diferencian a este tipo de pruebas.

Tabla I. Diferencias en pruebas con referencia a una norma y con referencia a un criterio

Características	Pruebas con referencia a una Norma	Pruebas con referencia a un Criterio
Construcción	Utiliza un amplio número de reactivos para medir un campo de conocimiento en forma global.	Utiliza un menor número de reactivos para medir con la mayor precisión posible áreas de conocimiento previamente delimitadas de acuerdo con los contenidos específicos.
Interpretación	a) Provee interpretación relativa del desempeño del individuo en relación con un área amplia de dominio	a) Provee interpretación absoluta del desempeño del individuo en relación a áreas de dominio menos amplias y más específicas.
	b) Los resultados se reportan en forma de percentiles o grado equivalente en relación con el desempeño de otros individuos (*que tanto sabe el estudiante en comparación con otros compañeros*)	b) Los resultados se interpretan en porcentajes, o sea en relación con el desempeño en áreas de dominio (*Que tanto sabe el estudiante del curriculum especificado.*)
Propósitos *a) Decisiones para Selección*	Cuotas fijas. Se requiere elegir a los individuos de acuerdo con sus habilidades relativas	Habilidad necesaria Requiere detectar las habilidades necesarias para el buen desempeño en un área específica de dominio.
b) Asesoría	Mide las habilidades en relación a las de los demás	Posee o no posee las habilidades requeridas
c) Evaluación de Programas (detección de deficiencias para mejoramiento)	No se recomienda	Relaciona los contenidos con el curriculum a implementar y de esta forma puede detectar las deficiencias
d) Diagnóstico y diseño de instrucción (Cuando se requiere saber qué contenidos no domina el individuo)	No se recomienda	Este tipo de examen puede detectarlo por la forma específica en que se construye
e) Asignación de recursos en gran escala (Requiere información sobre el desempeño general de los alumnos)	Obtiene información general para comparar con mayor facilidad los resultados de los programas	No se recomienda

Fuente: Basado en Woolfolk (1996) Psicología Educativa, 6a. Ed. Díaz, J. Traductor, Prentice Hall Hispanoamericana, S.A.

2.3 Punto de corte

A continuación, se presenta la definición del punto de corte según la opinión de tres de los autores que más han contribuido al estudio, aplicación y conocimiento del punto de corte.

Hambleton (1980, p. 369) especifica el punto de corte como: "...un punto en la escala de puntuación que se usa para separar a los examinados en dos categorías: 1) *competente*: a las personas que se sitúan en la categoría de puntuación alta y 2) *no competente*: a los que obtuvieron una baja puntuación".

Jaeger (1989, p.490) a su vez, define el punto de corte como: "una medida establecida para evaluar el desempeño del estudiante con relación a una competencia o área de dominio; donde el área de dominio o competencia es una manifestación del desempeño deseado que demuestra la habilidad del estudiante para aplicar conocimientos, capacidades y habilidades en situaciones reales de la vida".

Popham (1990, p.343) por otra parte, define al punto de corte como: "una medida de lo que el examinado puede realizar (desempeño) con respecto a un propósito definido".

Con base en las opiniones de los expertos, es importante diferenciar un estándar de desempeño del punto de corte para evitar confusión. El primero describe lo que un examinado sabe y puede hacer a cierto nivel de dominio. El punto de corte es el nivel operacional de este estándar, es decir; determina el punto que separa el desempeño en dos o más áreas de dominio (Hambleton, 1998). Esta última idea sobre la concepción del punto de corte es la que se va a utilizar en el desarrollo del presente estudio.

Los antecedentes en la determinación del punto de corte en pruebas educativas nos indican que su uso ha existido por varios años. Sin embargo, tomó gran relevancia en los Estados Unidos a mediados del siglo XX. En 1970 se inició en esta nación una etapa de falta de credibilidad en el sistema educativo debido a que algunos estudios revelaron que existía una "inflación de los diplomas". Los resultados de las pruebas educativas no concordaban con el conocimiento que los estudiantes tenían y las industrias y negocios estaban recibiendo graduados que tenían certificados pero no sabían leer o escribir.(Cizek, 2001, p.7). Este fue el detonador que permitió buscar mejores alternativas que permitieran establecer criterios o puntos de corte que determinaran la competencia real de los examinados y dio lugar a la conceptualización de la medición con referencia a un criterio en lugar de la medición con referencia a una norma. La medición con referencia a un criterio, enfoca la ejecución del sujeto con relación a áreas de dominio específicas y no a

la comparación de la ejecución del sujeto contra la ejecución de un grupo de examinados, como lo hace la medición normativa.

Con la medición referida a un criterio surgieron diferentes métodos para definir el punto de corte en las pruebas educativas. La mayoría de estos métodos utilizaban el factor de juicio de paneles de expertos para establecer los puntos de corte, dándole un carácter científico y profesional a todo el proceso.

En esta época aparecieron métodos como el de Nedelsky (1954), Angoff (1971), Ebel (1972), Jaeger (1978), Livingston y Zieky (1982) que utilizaban el juicio de expertos para la conceptualización del punto de corte para que el estudiante con el conocimiento mínimo indispensable pueda acreditar.

Con la publicación del artículo de Glass (1978) *"Standards and criteria"* en el *"Journal of Educational Measurement"* se inicia la controversia sobre la aplicación de estos métodos para definir puntos de corte. En este trabajo Glass, comenta que los métodos utilizados para definir el punto de corte son completamente inadecuados por el grado de subjetividad del juicio de expertos lo que da como resultado puntos de corte arbitrarios. En la misma edición de la revista, Popham como respuesta a los comentarios de Glass, admite que los procedimientos para definir puntos de corte utilizan el juicio de expertos pero considera que darle un *sentido arbitrario* a estos juicios es completamente erróneo y cita lo siguiente:

"Incapaces de confiar en el juicio humano como el ingrediente básico para el establecimiento de estándares, algunos investigadores han calificado todos los esfuerzos como arbitrarios y aún más, los han considerado inaceptables.

Sin embargo, el diccionario Webster's nos ofrece dos definiciones del término arbitrario: La primera de ellas es positiva, describiendo el término como un adjetivo que refleja selección o discreción, esto es, determinado por un juez o tribunal. La segunda definición le da una connotación negativa, describiendo arbitrario como un adjetivo que denota capricho, esto es, seleccionado al azar y sin razón. A mi parecer, las gentes que optan por designar al establecimiento de punto de corte como arbitrario, están claramente empleando la segunda definición con connotación negativa, del diccionario Webster's.

Pero la primera definición refleja con más precisión la seriedad de los esfuerzos para establecer estándares. Estos representan intentos genuinos de hacer un buen trabajo al decidir cual tipo de estándares se van a emplear. Que éstos estén basados en juicios es inevitable. Pero el designar a estos valores de juicio como caprichosos resulta absurdo".

A pesar de la controversia, los procedimientos para definir puntos de corte que utilizan el juicio de expertos se siguieron implementando y para 1983, la definición de puntos de corte con base en el juicio de expertos ya se había consolidado como un nuevo método en el campo en la medición y convertido en dogma para los especialistas en medición.

2.4 Métodos para establecer el punto de corte.

Desde 1960 se han implementado una serie de métodos para establecer puntos de corte y se han realizado estudios comparativos tratando de determinar el mejor (Hambleton, 1978; Livingston y Zieky, 1982; Berk, 1986; Jaeger, 1989; entre otros).

El debate iniciado por Glass (1978b) deja en claro que determinar estándares conlleva el factor de arbitrariedad, ya que al establecer un punto de corte se produce una falsa dicotomía en la distribución del conocimiento. Sin embargo, cuando hay necesidad de tomar decisiones, en cuanto al dominio o no dominio para fines de certificación, la toma de decisión es necesaria.

Actualmente existe un acuerdo entre los especialistas en medición en cuanto a que los procedimientos para establecer puntos de corte utilicen el jueceo. Jaeger (1989 p.492), apoya esta decisión al aseverar que "...ninguna cantidad de recolección, análisis y modelado de datos puede reemplazar el acto de decidir cuáles desempeños son aceptables o adecuados y cuáles son inaceptables o inadecuados; lo que varía es el juicio de aproximación al desempeño real".

Se han tratado de categorizar los métodos para establecer puntos de corte de distintas maneras. Desde el punto de vista del aprendizaje, los métodos se dividen en dos categorías: los modelos estáticos y los modelos continuos (Meskaukas, 1976; Shepard, 1984).

En el modelo estático se considera el aprendizaje como un "todo" o "nada"; se tiene o no se tiene. La mayor crítica que se ha hecho a este modelo es que requiere que las pruebas tengan un alto grado de homogeneidad en los reactivos, lo cual no es factible para materias que no sean del área de las

matemáticas, ya que son las únicas que pueden tener este grado de homogeneidad. Por este motivo, se considera que los modelos estáticos no son apropiados para situaciones de evaluación fuera de este contexto, porque en este modelo, el punto de corte se asume como el 100% de las respuestas correctas.

Por otro lado, en los modelos continuos se considera que el dominio de una habilidad se distribuye en forma continua, de acuerdo a las teorías de adquisición del conocimiento, por lo cual se puede ver como un intervalo que se circunscribe a los límites de dominio y competencia. Determinar el punto de corte para la toma de decisiones sobre un examinado, es una necesidad imposible de soslayar a pesar de lo arbitrario que pudiese considerarse separar artificialmente el continuo del aprendizaje (Shepard,1984). Por esto, cuando se requiere establecer puntos de corte se descartan los modelos estáticos y únicamente se toman en consideración los modelos continuos.

Berk (1986), propone tres categorías para agrupar los procedimientos para determinar puntos de corte:

1) *De Jueceo*. Basado en el juicio de una o varias personas, al cual se puede llegar en forma independiente o a través de un panel de discusión. Los jueces no tienen acceso a información sobre desempeño.

2) *Jueceo-Empírico*. Basado principalmente en el juicio de una o varias personas tomando en consideración la información que se les proporcionó sobre el desempeño.

3) *Empírico-Jueceo*. Se basa en el análisis estadístico del desempeño de uno o más grupos de examinados. El factor de juicio se fundamenta utilizando esta información para definir el criterio de dominio o no dominio de los examinados.

Los métodos de Nedelsky (1954), Angoff (1971), Ebel (1972), Jaeger (1978), Livingston y Zieky (1982) entre otros, son los métodos que han causado más impacto y que han sido utilizados más ampliamente para establecer puntos de corte (Cizek, 2001). Los primeros tres métodos se localizan en la categoría de métodos de Jueceo. El método de Jaeger está clasificado en la categoría de Jueceo-Empírico y por último, el método de Livingston y Zieky corresponde a la categoría de los métodos de Empírico-Jueceo. A continuación se presenta una breve descripción de estos métodos.

Método de Nedelsky (1954). Es el más antiguo de los métodos y se utiliza únicamente con pruebas de opción múltiple. Este método se basa en la idea de que un estudiante con conocimiento mínimo, primero elimina las respuestas que considera incorrectas y después utiliza la adivinación para escoger de entre las opciones restantes, la respuesta que considera correcta. El juez calculará cada

una de las respuestas de este examinado de la siguiente manera: Si el examinado descartó dos de cinco opciones, tendrá una oportunidad en tres de obtener la respuesta correcta, así para esa pregunta su puntuación será la que resulte de dividir una entre tres (1/3), o sea .33. Si descartó tres de cinco opciones, le quedará una oportunidad en dos 1/2 o sea .50 y así sucesivamente para todas las preguntas. La valoración de cada juez se suma y se divide entre el número de jueces para encontrar el punto de corte.

Shepard (1984), en su artículo *"Setting Performance Standards"* menciona que este método tiene varios defectos: primero, no existe una buena razón o evidencia que apoye la idea de que los examinados que no conocen la respuesta, adivinen al azar, ya que las pruebas se elaboran de manera tan específica, que los distractores son atractivos para el examinado que no domina el concepto. Segundo, los jueces raramente asignan probabilidades de 1 (equivalente al 100 %), por lo regular los expertos se ven forzados a asignar valores de .5 (equivalente al 50%) a muchos de los reactivos. Esto produce un punto de corte más bajo que los otros métodos que utilizan jueceo. Este método también requiere de un mayor entrenamiento de los expertos, lo que provoca la falsa idea de que este método es más científico que los otros métodos.

Método de Angoff (1971). El método de Angoff nació en 1971, como parte de un capítulo que detallaba el estado del arte de las escalas, normas y las equivalencias para las pruebas. Angoff en ese capítulo, dedicó únicamente 23

líneas de texto y una nota al pie de página a "... un proceso sistemático para decidir sobre la mínima puntuación para pasar un examen" (Zieky, 2001). Angoff no proporcionó detalles sobre cómo llevar a cabo la implementación del punto de corte y no hizo mención alguna sobre cómo seleccionar y entrenar a los participantes, o si se debe ó no se debe permitir la discusión y revisión de los juicios entre los participantes. Debido a la falta de especificación del método original, han surgido nuevos procedimientos que permiten la inclusión de otro tipo de información. Entre los métodos que surgieron se pueden mencionar los siguientes: interactivo de Angoff (Saunders y Mappus, 1984), interactivo de dos opciones de Angoff (Jaeger 1978, 1982; Cross, Impara, Frary y Jaeger, 1984) modificado de Angoff de opción múltiple (ETS,1976).

El procedimiento de Angoff utiliza un panel de especialistas, donde cada uno de ellos, revisa los reactivos de la prueba para estimar la probabilidad (una proporción de 0 a 1) de que un examinado pueda contestar correctamente la pregunta. Una vez efectuada esta operación, se suma la estimación proporcionada por cada especialista y posteriormente se determina el promedio de la suma de todas las estimaciones. El punto de corte se establece con este promedio. La ventaja que este método ha presentado es que el punto de corte es específico a las preguntas de la prueba y a los juicios de valor emitidos por los especialistas en la materia evaluada (Geinsinger, 1992).

Método de Ebel (1972). Con este método se hace un análisis de la dificultad y relevancia de cada pregunta usando una cuadrícula que tiene, en un eje la dificultad y en otro la relevancia. El método hace una estimación del porcentaje de reactivos de cada celda que puede contestar correctamente un examinado con un mínimo de competencia. El promedio se obtiene al multiplicar el número de reactivos de las celdas por el porcentaje respectivo. Se suman los resultados de todas las celdas y se divide entre el número total de reactivos. Este proceso lleva a un porcentaje promedio ponderado (Hambleton, 1978; p. 105). Berk (1986), manifiesta que el procedimiento de valoración consume mucho tiempo y que es difícil para los jueces determinar las dos dimensiones, por lo cual tienden a simplificar la tarea del examinado.

Método de Livingston y Zieky (1982). A este método se le llama también "Método de contraste de grupos". Este es un método de jueceo que se basa principalmente en los examinados, a diferencia de los otros métodos que se basan en el examen. En este método los jueces clasifican a los examinados en dos grupos; los "aptos" y los "no aptos", de acuerdo con los conocimientos y habilidades reflejados en la prueba. Una vez clasificados los estudiantes en las categorías mencionadas, se combinan las dos distribuciones de las puntuaciones, los aptos, los no aptos y el punto en el cual se intersectan ambas distribuciones será considerado como el punto de corte (Cizek, 2001), como se muestra en la figura 1.

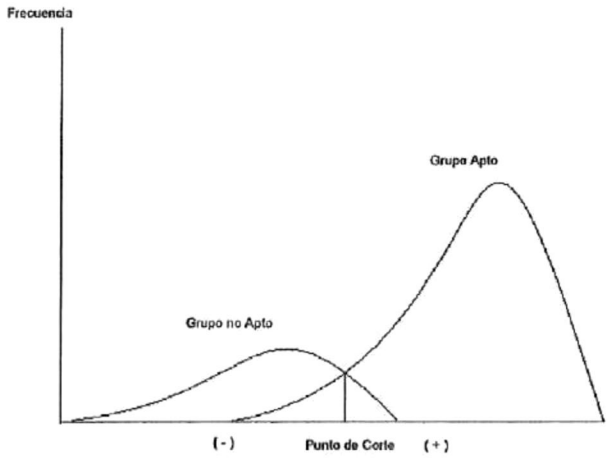

Fig. 1. Ilustración hipotética de la distribución de frecuencia de dos grupos que utiliza el método de contraste de grupos.

A primera vista, este método podría considerarse como una forma fácil y elegante para terminar con los problemas de determinar puntos de corte. Sin embargo, el punto donde los dos grupos se entrelazan puede verse afectado por muchísimos factores, como podrían ser: la proporción de los no calificados en la muestra; la severidad con que los jueces definen el dominio; la forma en que se clasifican los casos inciertos (dudosos), es decir; si éstos se eliminan, si se asignan como no calificados o se toman como calificados, etc. Según Shepard (1984), este procedimiento únicamente puede identificar la región, en la continuidad del desempeño, donde no se puede diferenciar los que tienen el dominio de los que no lo tienen.

Shepard (1984) recomienda que para fines de certificación se utilicen los métodos de jueceo o jueceo-empírico. En este trabajo Shepard llevó a cabo una revisión de los métodos más utilizados para establecer puntos de corte y en sus

conclusiones recomienda que se utilice el método de Angoff (1971) cuando se persiguen fines de certificación, ya que este método permite tomar decisiones en una línea continua de probabilidades. Esto produce puntos de corte más directos y verosímiles.

Método de Jaeger (1982). Este método se considera interactivo ya que utiliza un conjunto de jueces con diferentes marcos de referencia (maestros, padres, administradores, etc.) e incorpora información normativa sobre el desempeño de los examinados. Este método en lugar de buscar un examinado con competencia mínima, requiere que el grupo de expertos conteste la pregunta, ¿puede cada graduado contestar correctamente esta pregunta? y de no ser así, ¿debe negársele el diploma?. En este método se calcula el promedio de las respuestas de los jueces y este promedio se designa como el punto de corte. La diferencia entre este método y los otros mencionados, es la incorporación del proceso interactivo en la toma de decisión.

De acuerdo con el esquema de clasificación proporcionado por Berk (1986), existen más de 38 métodos para establecer el punto de corte. Sin embargo, se puede decir que muchos de ellos son modificaciones surgidas a partir de los métodos clásicos arriba mencionados. A continuación se presenta una reseña de algunos de estos métodos tomando en consideración las tres categorías (jueceo, jueceo-empirico, empírico-jueceo) (ver Apéndice A).

Entre los métodos de Jueceo Berk (1986) menciona el Método Interactivo de Angoff (1976). Este método fue utilizado en el estudio realizado por el *Educational Testing Services* y se basó fundamentalmente en el método de Angoff con algunas variaciones. En lugar de dejar abierta la probabilidad de la estimación por los expertos, se solicitó a éstos que realizaran las estimaciones de acuerdo con una escala de 7 puntos con porcentajes fijos para la valoración. La escala de porcentajes fue la siguiente: 5, 20, 40, 60, 75, 90, 95 y una opción de "no lo sé". Se utilizó esta modificación por considerar que es sumamente difícil para los expertos dar una estimación de los porcentajes en forma abierta, como lo pide el método de Angoff.

En el método de 2 opciones de Angoff (Nassif, 1979), se pide a los jueces que determinen si una persona con el dominio mínimo aceptable contestará correctamente la pregunta. Se contemplan como posibles respuestas "Si", "No" y "No sé". El acuerdo entre las contestaciones "SI" determinadas por los jueces se tomará como el número de reactivos mínimos aceptables. El porcentaje de estos reactivos será el estándar inicial. Este estándar se ajusta después para evitar errores de medición. Las ventajas de este método son: que es fácil de implementar, entender, calcular y requiere sólo una respuesta Si-No por cada juez. Sin embargo, el formato SI-NO limita las probabilidades de 0% a 100 % y el método para estimar el acuerdo entre jueces no se especifica.

Por otra parte, una variación del método de Ebel es la Taxonomía de Relevancia de Ebel implementado por Skakun y Kling, en 1980. Este método es similar al de Ebel con la diferencia de que utiliza tres dimensiones de relevancia (esencial, importante y aceptable) en lugar de las dos que utiliza Ebel. Se considera que la clasificación previa de ambas dimensiones simplifica la labor de los expertos. Por otro lado, no considera los niveles de dificultad del reactivo. Las dimensiones de la taxonomía pueden ser difíciles de definir para ciertos contenidos y la clasificación puede ser ambigua.

Otro de los métodos considerados por Berk en el área de jueceo es el de Compromiso I (absoluto-relativo) llevado a cabo por Beuk en 1983. Aquí los jueces especifican dos valores: 1.- el mínimo porcentaje de aciertos que un estudiante debe obtener para pasar (k) y 2.- el porcentaje de estudiantes que se espera que pasen (v). Se puede utilizar una ecuación que relacione (k) y (v) para ajustar los valores si son inconsistentes con la distribución actual de las puntuaciones. Se grafica la media y la desviación estándar de los valores (k) y (v) de la muestra de los jueces, junto con el porcentaje de estudiantes que pasarán (y) como un decremento de la función de la puntuación del examen (x). Se dibuja la curva para identificar el punto donde el estándar y el valor de pase del examen representan el mejor compromiso entre los estándares relativo y absoluto.

Una de las ventajas de este método es que considera sistemáticamente los juicios sobre el desempeño estimado y el desempeño real. La medida en la cual los jueces acuerden sobre (k) o (v), identifica si el grupo se orienta hacia el examen o hacia el examinado. Sin embargo se requiere que los jueces consideren dos estimaciones las cuales pueden cambiar a medida que los jueces tengan oportunidad de practicar el hacer estimaciones, por lo tanto se puede considerar que este método es más complejo de realizar por los datos estadísticos y los cálculos que se requieren y la interpretación del estándar puede ser difícil de comprender para los que no son especialistas.

En cuanto a los métodos de Jueceo-Empírico, Berk menciona el de compromiso II Absoluto-Relativo (Hofstee 1983). Los jueces especifican cuatro valores: (a) el porcentaje máximo correcto (K_{max}) —el estándar que sería satisfactorio aún cuando cada estudiante alcanzara ese puntaje (b) el porcentaje mínimo correcto (K_{min}) —el estándar al cual nadie podría acceder aún cuando ningún estudiante obtuviese ese puntaje (c) el porcentaje máximo aceptable de errores (f_{max}) y (d) el porcentaje mínimo aceptable de errores (f_{min}). Usando una curva de la distribución acumulada del puntaje del examen, se determina una relación empírica entre (k) y (f). El estándar es el punto de intersección entre el modelo (valor absoluto) y la curva (valor relativo). Este método da a los jueces la oportunidad de revalorar sus estimaciones basadas en los tres tipos de información. El juicio se obtiene de diversas muestras de jueces representativos de las partes interesadas. Se considera que el formato SI-NO limita la

probabilidad de los reactivos de 0 a 100 y requiere que los jueces proporcionen cuatro estimaciones lo que tiende a consumir mucho tiempo y es más complejo que otros métodos.

En el método de Juicio informado de Popham, el estándar se fija basado en información sobre la estadística de los reactivos y las puntuaciones del desempeño probado e información sobre las preferencias reactivo por reactivo. El juicio final se otorga con la información reunida de los grupos de asesores. Las ventajas que presenta este método son su facilidad de implementación y que es fácil de entender y de calcular. Por otra parte, incorpora la información sobre el desempeño y la preferencia de grupos diversos. Utiliza también el mismo formato para la valoración de las preguntas (Si-No) de Angoff, además considera los juicios de todos los especialistas. Entre las desventajas que presenta se puede mencionar que la recopilación de datos puede consumir mucho tiempo y ser costosa. El análisis de los jueces sobre la información es poco sistemática y requiere además la elaboración de guías para utilizar los diferentes tipos de información.

Otro de los métodos considerados en los de jueceo-empírico es el de Grupos contrastantes y compuestos de Angoff, implementado por Shepard en 1983 y 1984. En este método los jueces estiman la "probabilidad" por área de competencia y no por reactivo. Una vez efectuada la valoración, se provee a los jueces con información sobre la dificultad de los reactivos según el desempeño

real de un grupo de examinados. El estándar será la suma de los promedios una vez que los jueces revisen nuevamente sus estimaciones. Entre las ventajas de este método se encuentra que incorpora el nivel de dificultad real en el proceso de valoración, que es fácil de implementar, entender y calcular y enlaza el estándar con las áreas de competencia. Por otro lado, como desventajas se puede mencionar que la valoración por área de competencia y no por reactivo no es adecuada, ya que aún cuando debe existir un alto grado de correspondencia entre áreas y reactivos no puede asumirse que sean idénticos.

Se han realizado otros estudios utilizando diferentes métodos, como por ejemplo el estudio comparativo implementado por William Donnoe y Roy Amato (1997), *"Supportive Data and Guidelines for Using the Angoff, Ebel and Nedelsky Cut Off score Methods",* en un contexto ocupacional para la selección de personal. Los tres métodos mencionados utilizan grupos de expertos para valorar las preguntas de las pruebas y requieren determinar el perfil referencial del nivel mínimo de dominio de los sujetos, pero difieren en la manera de llevar a cabo el jueceo.

El resultado de este estudio favorece al método de Angoff en contraposición con los métodos de Ebel y Nedelsky. Donnoe (1997) recomienda utilizar el método de Angoff para establecer puntos de corte, ya que demostró tener una alta correlación entre el nivel de dificultad de las preguntas y la valoración de los jueces. Otro resultado interesante de esta investigación es que

el método de Angoff no presenta gran variabilidad en los resultados aún cuando se utilice un grupo de expertos menor a 5.

En la investigación realizada por Impara y Plake (2000), se comparan tres métodos para implementar puntos de corte. a) el método de Angoff modificado b) el método de grupo limítrofe y c) el método de impacto avanzado (Dillon, 1996). Este estudio se implementó en la primavera de 1999, con un grupo de 22 maestros que utilizaron los métodos en forma continua. La hipótesis de este trabajo fue que el método de Angoff proporciona puntos de corte más bajos, que los que se establecen cuando se utiliza el método de grupo limítrofe. Los resultados de este estudio confirmaron que efectivamente el método de Angoff proporciona un punto de corte más bajo.

Otros investigadores también han obtenido un resultado similar. Livingston y Zieky (1989), Jaeger (1989). Hurtz y Hertz (1999), realizaron otro estudio comparativo utilizando el método de Angoff con diferente propósito. En ese estudio, se utilizaron los resultados de ocho aplicaciones distintas del método en pruebas de certificación, con objeto de aplicar la teoría de generalización para estimar el número óptimo de expertos que se deben de utilizar al aplicar el método de Angoff. Hurtz y Hertz, en su estudio, mencionan la importancia y aceptación que tiene este método en el campo de medición relativo a puntos de corte.

2.5 Criterios de validez en los procedimientos de puntos de corte.

Cuando se habla de aspectos de validez para un estudio de punto de corte, se debe considerar más bien validar el argumento en el cual se basan las decisiones. De acuerdo con la definición proporcionada por la nueva versión de los Estándares para la medición psicológica y educativa, citada por Kane (2001) es la siguiente: "validez se refiere al grado en que las evidencias y la teoría apoyan la interpretación de las puntuaciones de acuerdo con los diferentes propósitos de las pruebas" (p. 56).

La prueba en sí no se valida, ni la puntuación de la prueba, sino más bien la validez es una propiedad de la interpretación de dicha puntuación. Bajo este esquema Kane (2001), propone cuatro aspectos como evidencia de validez en puntos de corte y su correspondiente estándar de desempeño:

1) La congruencia en la conceptualización del procedimiento de punto de corte.
2) La descripción del procedimiento y las políticas consideradas.
3) La consistencia interna.
4) El acuerdo con criterios externos.

En el primer aspecto se deberá cuidar que el procedimiento seleccionado sea congruente con el tipo de examen y el concepto de desempeño que subyace en la toma de decisión.

El segundo aspecto que se relaciona con la descripción del procedimiento tiene gran importancia en la determinación del punto de corte, ya que una forma de validez ampliamente aceptada entre los expertos en medición (Hambleton, 1980; Berk, 1986; Shepard, 1984) es precisamente la detallada descripción del proceso que demuestre que el grupo de expertos involucrados en el estudio conocen el propósito del estudio y entienden el procedimiento.

Para asegurar la validez del procedimiento, Hambleton (1980) sugiere que se lleve a cabo una descripción cuidadosa de cada una de las fases que involucran el proceso: 1) definición de los propósitos del estudio; 2) selección del panel de expertos; 3) entrenamiento de los participantes; 4) definición del concepto de desempeño y 5) procedimientos para la captura de los datos.

El tercer aspecto tiene que ver con acumular evidencias de consistencia interna, o sea la posibilidad de obtener el mismo punto de corte si el estudio se repitiera. Estas evidencias se acumulan con estudios de validez de los reactivos, así como de evidencia empírica sobre la homogeneidad de los juicios de los expertos. De acuerdo con Kane (2001), se espera un cierto grado de discrepancia entre las estimaciones de los jueces, pero no tan grande que pueda

menoscabar el proceso. Existen varias opciones para demostrar este tipo de evidencia, por ejemplo:

1) Con el uso de la teoría de la generalización para encontrar la varianza de los participantes y de las preguntas o tareas llevadas a cabo. Kane (2001) indica que este enfoque tiene la desventaja de que requiere utilizar más de un procedimiento para determinar puntos de corte, lo cual es costoso e involucra demasiado tiempo. Por otra parte, la ventaja que ofrece este enfoque es que puede proveer de indicadores claros de la diferencia entre uno y otro estudio y se puede utilizar con cualquier tipo de método.

2) Análisis del nivel de las preguntas. Aquí existen dos formas para llevar a cabo este análisis: a) La primera requiere que los expertos apliquen el concepto del examinado de competencia mínima a áreas específicas para generar puntos de corte por cada área, los cuales se agregarán después al punto de corte global. Una vez realizada esta tarea, se analizará el desempeño de los examinados con conocimiento mínimo en estas áreas para probar la consistencia interna de los juicios emitidos por los expertos. Si el desempeño de estos examinados se encuentra muy por arriba o debajo del punto de corte establecido para esa área, se podrá pensar que existe inconsistencia en los resultados. Cabe mencionar, que se espera que existan pequeñas diferencias. Sólo las diferencias muy grandes podrán mostrar que existe un problema de inconsistencia. b) El

segundo análisis de consistencia interna utiliza dos grupos de examinados. Uno con puntuaciones un poco arriba del punto de corte y el otro con puntuaciones un poco por debajo del mismo. En este caso las puntuaciones de ambos grupos deberán ser consistentes con el punto de corte establecido. Es decir, que sus puntuaciones deberán mantener esta diferencia al ser comparadas con el punto de corte. De esta forma se demuestra la consistencia de los resultados.

El último aspecto que menciona Kane (2001) es el acuerdo con criterios externos. La literatura sobre este tema muestra una amplia diversidad en los tipos de análisis que pueden hacerse comparando aspectos cruciales, como serían por ejemplo: la comparación de otros procedimientos basados en la toma de decisiones, comparación de resultados de otros estudios de punto de corte, comparación de decisiones utilizando para ello otra prueba distinta, comparación de la distribución de muestras independientes, comparación con otros métodos de evaluación.

Con respecto a la comparación con los resultados de otro método de evaluación, se podría utilizar por ejemplo, los resultados de un examen diagnóstico cuyo propósito sea similar al del examen que nos ocupa, identificar aquellos estudiantes que tengan el conocimiento suficiente, para cumplir con el requisito esperado, sin pasar por un curso de instrucción, el resultado de este

tipo de examen se pudiese utilizar como una evidencia de acuerdo con un criterio externo.

Cuando se habla de evaluar el desempeño es necesario entender que se cuenta con dos modelos: El holístico y el analítico.

El *modelo holístico* asume que la única forma de evaluar es observando las conductas en su totalidad, las cuales no se pueden separar en pequeñas tareas ya que ello destruye el significado de desempeño. Los métodos que más se adaptan a este modelo son los que se centran en el examinado.

El segundo es el *modelo analítico* que considera que el desempeño se puede evaluar utilizando grupos de tareas específicas como indicadores de la conducta total. Si se traslada esto al campo de la medición, se puede decir que se evalúan las tareas específicas en forma independiente y sus puntuaciones se combinan para dar una sola puntuación que represente el desempeño completo del individuo. Este modelo favorece al uso de los métodos centrados en las pruebas.

La coherencia entre el procedimiento elegido, el concepto de desempeño seleccionado y el propósito de la decisión, no asegura la validez de los resultados, pero sí ayuda a tener confianza en el proceso para la toma de decisión.

2.6 Examen de egreso del idioma Inglés (EXEDII)

Se considera pertinente para este estudio conocer la manera en que fue elaborado el examen en el que se basa el punto de corte, ya que esto es primordial para determinar el método a utilizar. A continuación se hace una descripción de los elementos más importantes del EXEDII.

El EXEDII nace como una respuesta a una necesidad de la UABC de contar con opciones para que sus estudiantes puedan cubrir el requisito de egreso de la Universidad que a la letra dice: "los estudiantes deberán acreditar el conocimiento de un idioma extranjero, por lo menos a nivel intermedio, como requisito necesario para egresar de la universidad" (Reglamento General de Admisión, Inscripción, Evaluación de los Alumnos y su seguimiento en los Planes de Estudio de la UABC, 1995, pp 55).

En un principio este examen fue elaborado en una versión de lápiz y papel, posteriormente se adecuó al formato computarizado utilizando la plataforma SICODEX (Backhoff, Ibarra y Rosas, 1995) por contar con mayores ventajas como son: facilidad de aplicación, reducción de errores de interpretación de resultados, flexibilidad en la presentación de los reactivos, estandarización de las condiciones de aplicación de la prueba y seguridad en el manejo de los datos.

El EXEDII se considera un examen de alto impacto ya que tiene consecuencias relevantes para el estudiante. Esto es, estar en posibilidad de egresar o no egresar de la Universidad. (Por otra parte, el EXEDII es un examen de certificación que evalúa las habilidades que los estudiantes han desarrollado en el idioma inglés independientemente de la instrucción recibida. Cuando el propósito de un examen es certificar niveles de competencia, se denomina como un examen con referencia a un criterio. Esto quiere decir que se evalúa al estudiante con base en un criterio de calidad establecido de acuerdo a un programa o área de conocimiento y los resultados se interpretan con base en el dominio del contenido que el estudiante demuestra conocer.

Una parte central de cualquier prueba con referencia a un criterio la constituye la especificación de los contenidos que definen la habilidad a evaluar. Por lo tanto, la calidad de una prueba criterial comienza con el proceso mismo de su construcción. Para la elaboración del EXEDII se utilizó el modelo de Hambleton (1988) para el desarrollo de pruebas referidas a un criterio que contempla 9 etapas y fue complementado con los trabajos de Popham (1978) y Nitko (1994). A continuación se detallan las etapas sugeridas por Hambleton (1998) para el diseño de una prueba con referencia a un criterio. Larrazolo y Velasco (2000).

Tabla II. Plan de trabajo diseñado para la construcción del EXEDII.

1. Consideraciones preliminares.
   a) Especificar el propósito de la prueba.
   b) Especificar los grupos a evaluar y cualquier requisito especial de la prueba.
   c) Determinar el tiempo disponible para producir la prueba.
   d) Identificar a los especialistas en contenido y evaluación)
   e) Estimar el tamaño de la prueba.
2. Revisión de contenidos.
   a) Analizar el contenido y los objetivos a evaluar.
   b) Especificación de preguntas:
      i. Descripción general breve y concisa del contenido y/o conductas.
      ii. Ejemplo de las indicaciones de la prueba y un modelo de pregunta.
      iii. Atributos de la pregunta.
      iv. Atributos de las respuestas incorrectas que se deberán elaborar.
3. Elaboración de las preguntas.
   a) Elaborar un número suficiente de preguntas para la prueba piloto.
   b) Editar las preguntas elaboradas.
4. Evaluación de la validez del contenido.
   a) Revisar las preguntas de la prueba para determinar si cumplen con las especificaciones de contenido.
   b) Revisar las preguntas para determinar su adecuación técnica.
   c) Con base en los dos puntos anteriores, corregir las preguntas o cambiarlas.
   d) Escribir preguntas adicionales (si es necesario y repetir el paso 4).
5. Administración de la prueba de campo.
   a) Organizar las preguntas de la prueba en un formato para la prueba piloto.
   b) Administrar el formato de prueba a grupos seleccionados adecuadamente.
   c) Realizar el análisis de las preguntas, estudios de validez y estudios de sesgo (tendencias inadecuadas de las preguntas).
   d) Corregir o eliminar las preguntas de la prueba cuando se juzgue pertinente, con base en los resultados del punto anterior.
6. Ensamble de la prueba.
   a) Determinar el tamaño de la prueba, el número de formatos que se necesitan y el número de preguntas por objetivo.
   b) Seleccionar las preguntas de la prueba del banco de preguntas válidas.
   c) Preparar las instrucciones, preguntas de práctica, formato, clave de respuestas, etc.
7. Selección de un estándar.
   a) Iniciar un proceso para determinar el punto de corte.
8. Preparación de manuales.
   a) Peparar un manual para la administración de la prueba.
   b) Preparar un manual técnico.
9. Colección de datos técnicos.
   a) Conducir investigaciones de confiabilidad y validez.

La metodología para la elaboración de pruebas con referencia a un criterio utiliza el juicio de un grupo de expertos o especialistas. En el caso del EXEDII, se formaron tres comités de trabajo: El comité coordinador (CC) formado por investigadores del Instituto de Investigación y Desarrollo Educativo (IIDE) y dos comités integrado por profesores de las tres sedes (Mexicali, Tijuana y Ensenada) de la Escuela de Idiomas (EI) de la Universidad Autónoma de Baja California; el Comité de especialistas (CE) y el Comité encargado de elaborar preguntas (CP), (Larrazolo, 2000).

De acuerdo con el modelo de Hambleton (1988), se siguieron 9 etapas para la construcción del EXEDII. El primer paso que el comité de especialistas tuvo que realizar fue definir la población a evaluar. Esta población está formada por aquellos estudiantes que consideran tener el conocimiento del inglés necesario para desempeñarse a un nivel intermedio. La siguiente etapa fue determinar el tiempo de elaboración de esta prueba, el cual se determinó que fuese de un año. Asimismo, se utilizó la estrategia de análisis curricular por reticulado, del contenido a evaluar para determinar la muestra representativa del currículo (Contreras, 1998 Tesis de Maestría) de la EI basado en el Método de Scott Foresman (Purpura, J. y Pinkley, D., 1991). Para la elaboración del EXEDII se tomó como base el nivel 1 de *On Target*, que corresponde al nivel intermedio-alto de los estándares de la American Council of Teachers of Foreign Languages (ACTFL, 1997), (Larrazolo, y Velasco, 2000).

Un aspecto importante en una prueba criterial es la cantidad de reactivos que debe medir cada una de las habilidades. En el EXEDII se determinó medir tres habilidades: comprensión auditiva, comprensión de lectura y gramática. Ambos comités decidieron elaborar 100 reactivos de opción múltiple distribuídos en las áreas mencionadas. En el caso de las pruebas criteriales es necesario elaborar dos tipos de especificaciones.

a) Especificación de la prueba

Esta especificación debe incluir: i) el tema y los subtemas del dominio que será evaluado, ii) el número de los reactivos necesarios para evaluar el dominio, iii) el número y tipo de reactivos que será necesario elaborar para cada especificación. Asimismo contiene información acerca de la forma de interpretación de los resultados; como la estimación de la ejecución, administración de la prueba y el tiempo estimado para su aplicación.

b) Especificación de los reactivos.

Incluye información sobre el diseño de los reactivos. Esta especificación se elaborará con sumo detalle, de tal forma que pueda guiar el proceso de la elaboración de nuevos reactivos. La especificación contempla el área, la subárea, el objetivo, las instrucciones para elaborar la pregunta, los atributos de ésta, los contenidos que se incluyen y los que no se incluyen y un modelo de la pregunta. Asimismo, se especifican las

respuestas correctas y los distractores, sus atributos, y estructura (Larrazolo 2000). A continuación se proporciona un ejemplo de la especificación de un reactivo.

Ejemplo extractado de Hambleton (1988, p. 279).

Descripción.

El estudiante identificará los tonos o emociones expresados en los párrafos.

Instrucciones muestra y reactivo muestra

Instrucciones: *Lee el párrafo de abajo y contesta la pregunta y encierra en un círculo la letra que aparece junto a tu respuesta.*

Jaimito había estado jugando y nadando en la playa todo el día, pero ahora ya era tiempo de regresar a su casa. Jaimito se sentó en el asiento trasero del carro de su papá; con dificultad podía mantener sus ojos abiertos.

¿Cómo se sentía Jaimito?

a) Asustado b) Amigable c) Cansado d) Amable

Límites del contenido.

1. Los párrafos describirán situaciones que le son familiares a estudiantes de tercer grado.
2. Los párrafos deberán contener entre dos y tres enunciados. El nivel de legilibilidad deberá ser el adecuado para tercer grado (Fórmula de Dale-Chall).
3. Los tonos y emociones expresados en los párrafos deberán ser elegidos de la siguiente lista:

triste	enojado	furioso	amable
cansado	asustado	amigable	exaltado
feliz	afortunado	listo	orgulloso

Límites de la respuesta

1 Las opciones de respuesta deberán tener una sola palabra.
2 Se usarán cuatro opciones de respuesta para cada reactivo.
3 Las respuestas incorrectas deberán ser tonos o emociones que resulten familiares a los estudiantes de tercer grado y que son frecuentemente confundidas con la respuesta correcta.

Fuente: Larrazolo y Velasco (2000).

El EXEDII cuenta con las especificaciones mencionadas, con base en las cuales se desarrollaron los reactivos. Posteriormente se elaboró un manual conteniendo esta información, lo cual permite que puedan crearse los reactivos y

pruebas paralelas con mayor facilidad al seguir las instrucciones para la creación de nuevos reactivos que contemplan la construcción de los mismos paso a paso, asegurando que se elaboren con un grado de dificultad similar al de la primera versión.

El proceso de validación de ambas especificaciones se llevó a cabo por jueceo, de acuerdo con las indicaciones del Modelo de Hambleton. En este caso el comité de expertos y el comité coordinador valoraron las especificaciones. En primera instancia se evaluó si los contenidos del EXEDII son una muestra representativa del universo de contenido a medir para el nivel intermedio de inglés y por otro lado se llevó a cabo un trabajo exhaustivo de valoración de los reactivos elaborados.

Una vez que se tuvo el formato de preguntas del examen se procedió a elaborar una prueba piloto la cual se aplicó (segundo, tercero y cuarto nivel) siguiendo las indicaciones del modelo de Hambleton (1988). En el caso particular del EXEDII, esta prueba piloto se aplicó a estudiantes de las tres sedes de la Escuela de Idiomas (Mexicali, Tijuana y Ensenada). Con la información vertida por esta aplicación de la prueba se procedió a realizar el siguiente paso, el análisis empírico de las preguntas.

El EXEDII cuenta con estudios específicos realizados para determinar los indicadores psicométricos que proporcionen evidencias de calidad de la prueba,

como son el índice de dificultad y discriminación, etc. Con este tipo de estudios se puede conocer si las preguntas tienen una dificultad adecuada y a su vez permiten detectar aquellas preguntas que no reúnen los criterios establecidos y que requieren modificarse o eliminarse.

Para calcular la dificultad de un reactivo se divide el número de personas que contestó correctamente entre el número total de personas que contestó dicho reactivo ya sea correcta o incorrectamente, lo cual dará el índice de dificultad, el cual es una relación inversa. Mientras más se acerque el valor a 1 el reactivo tendrá un menor nivel de dificultad y si el valor se acerca a 0 su dificultad será mayor. Por otro lado, el aspecto de discriminación se refiere a si las preguntas permiten diferenciar adecuadamente los estudiantes que obtienen bajas calificaciones de aquellos que obtienen altas calificaciones. Esto quiere decir que mientras más alto sea el índice de discriminación de las preguntas, diferenciará mejor a las personas con altas y bajas calificaciones (Backhoff, E., *et al*: 2000). Dentro del campo de la estadística existen varios métodos para calcular el índice de discriminación. Para el EXEDII se utilizó el método de grupos extremos que proporciona el índice de discriminación.

Este índice se calcula considerando el número de aciertos del 27% de las personas con puntuaciones altas menos el número de aciertos del 27% de las personas con puntuaciones bajas. El resultado se divide entre el número de personas del grupo más numeroso, lo cual proporcional el índice de

discriminación. Mientras más alto sea el índice el reactivo diferenciará mejor a las personas con altas y bajas calificaciones (Backhoff, 2000). Se obtuvo también el coeficiente de correlación biserial (Henrysson, 1971), que calcula el grado de coincidencia entre lo que mide la prueba y lo que mide la pregunta.

Estos estudios se llevaron a cabo con base en los lineamientos establecidos por diversas asociaciones como la Asociación Americana de Psicología (APA), la Asociación Americana de Investigación Educativa (AERA) y el Consejo Nacional para la Medición en Educación (NCME), encontrándose que los índices de dificultad y el coeficiente de discriminación se encuentran dentro de los criterios aceptables que marcan dichas instituciones De acuerdo con el estudio realizado por Larrazolo y Velasco (2000), el 98% de las preguntas del EXEDII tienen un grado de dificultad adecuado y un coeficiente de discriminación del (0.39%), los cuales son criterios aceptables para una prueba.

El EXEDII cuenta actualmente con diversas evidencias de la calidad y confiabilidad del instrumento. Sin embargo, no cuenta aún con un punto de corte de tipo absoluto, establecido con referencia a un criterio. Por ello se decidió implementar este estudio, con el fin de contar con un criterio adecuado para todos los estudiantes que presenten este examen. De esta manera, se asegurará que la calificación obtenida por el estudiante sea representativa del conocimiento que el estudiante domina y no de la comparación de su calificación con la calificación obtenida por el grupo sustentante.

Indudablemente la determinación de estándares en la evaluación del aprendizaje es un campo en constante evolución que ha tenido grandes avances. Este camino no ha estado exento de aspectos controversiales, sobre todo porque la mayoría de procedimientos para establecer estándares o puntos de corte se basan en la utilización del juicio de expertos para determinar esta puntuación. A la fecha el debate ya no se centra en permitir o no un proceso de 'jueceo" o emisión de juicios por expertos, sino más bien en las formas de implementar los procedimientos. Aún no se ha definido un método específico que se considere el mejor para determinar este tipo de puntuación. Es por ello que han surgido un gran número de procedimientos. Sin embargo, existe un acuerdo generalizado entre los especialistas de medición (Livingston y Zieky, 1982; Berk, 1986; Jaeger, 1989,) que cuando se permiten oportunidades de interacción entre los expertos y se provee información adicional, se apoya el proceso de "jueceo" o valoración, para lograr puntos de corte más objetivos.

CAPÍTULO 3

MÉTODO

3.1 Descripción del Método

En este apartado se describe el proceso que se llevó a cabo en esta investigación, así como los sujetos y los instrumentos empleados para realizarla. Para la implementación del proceso se consideraron tres etapas: a) selección del panel de expertos, b) entrenamiento para efectuar el proceso y c) proceso de valoración del examen por el panel de expertos.

Esta investigación se circunscribe en un enfoque cuantitativo, pero también se enriquece con la utilización del enfoque cualitativo, debido a la utilización del juicio de expertos para llegar a una toma de decisión y a la evaluación del proceso mismo por el panel de expertos.

Para llevar a cabo este estudio se utilizó el método de Angoff (1971) complementado con un proceso interactivo de valoración por expertos con el fin de asegurar la objetividad de los juicios. Hambleton en 1980 recomienda tomar en consideración una serie de fases o etapas en la implementación de cualquier método encaminado a la determinación de estándares o puntos de corte en educación, con el fin de asegurar los aspectos de validez y confiabilidad que

todo proceso debe tener. En el 2000, Hambleton elaboró un modelo de 11 pasos para tal efecto (Tabla III).

En este estudio se aplicó este modelo con el fin de incrementar la oportunidad de que el proceso produjera un punto de corte más defendible y válido, siguiendo los lineamientos que sugieren (Livingston y Zieky, 1982; Berk, 1986; Jaeger, 1989).

Estos once pasos se aplicaron en tres etapas. La primera etapa, *selección de los jueces,* comprende únicamente el paso 1. En la segunda etapa, *entrenamiento para efectuar el proceso*, se utilizaron los pasos del 2 al 4 y en la tercera, *proceso de valoración por los expertos*, se aplicaron los pasos del 5 al 10. Por último, el paso 11 consiste en la documentación del proceso.

Tabla III. Modelo de Hambleton para el establecimiento de estándares.

Pasos para la implementación de estándares en educación.

1. Seleccionar un panel de expertos
2. Seleccionar el método, preparar material de entrenamiento y la agenda de reuniones
3. Preparar la definición del concepto de un examinado con competencia mínima dentro del nivel intermedio de inglés.
4. Entrenar a los expertos en el uso del método (incluyendo la práctica en efectuar valoraciones de reactivos).
5. Realizar la valoración de las preguntas.
6. Realizar la primera reunión de retroalimentación.
7. Efectuar la segunda reunión de retroalimentación
8. Determinar el punto de corte
9. Presentar información sobre consecuencias esperadas.
10. Revisión y evaluación del proceso
11. Documentar técnicamente el proceso.

Fuente: Hambleton (2001). Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process en Cizek, G (Comp.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (Pp. 89-115) Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey

Antes de describir estas etapas se definirá a los sujetos de la investigación (panel de expertos y estudiantes) así como los instrumentos (examen de egreso EXEDII, guía de instrucción, formato de registro) y por último se explicará la definición del concepto de competencia mínima.

3.2 Sujetos de la investigación.

Para llevar a cabo la selección del panel de expertos se tomaron en consideración las sugerencias de Jaeger (1981), quien define a los expertos como aquellos que sobresalen en el dominio de la materia, que poseen una alta capacidad de decisión con relación a la misma, así como capacidad de análisis para ver los problemas y que poseen además, la habilidad de automonitoreo. Raymond y Reird (2001) por su parte, expresan que los jueces deben tener amplio conocimiento de la materia y deben entender a la población examinada. Y por último, se consideró la investigación realizada por Hurtz y Hertz (1999) en cuanto al número de expertos requeridos. Estos investigadores sugieren la participación de 6 a 11 expertos cuando se utiliza el método de Angoff para la determinación de puntos de corte, ya que este número permite realizar estudios de generalización de los resultados que sean confiables.

Los criterios que se tomaron como base para la elección de los 7 especialistas del área de inglés de la Escuela de Idiomas Unidad Ensenada, fueron los siguientes:

- Contar con grado de licenciatura como mínimo grado de estudios.
- Experiencia impartiendo el nivel intermedio por más de 4 semestres.
- Capacidad para el análisis de información.
- Conocimiento de la población estudiantil.

El estudio utilizó el resultado de 710 estudiantes que presentaron el examen de inglés como requisito de egreso (EXEDII) en Febrero del 2000 en las tres unidades académicas de la Universidad en Mexicali, Tijuana y Ensenada. Todos ellos eran estudiantes de licenciatura de las 62 carreras que ofrece la Universidad, cuyo conocimiento del idioma inglés fue adquirido de diversas formas, tanto formal como informal. Como único requisito para la presentación de este examen se consideró que los estudiantes estuviesen inscritos en el séptimo semestre de su carrera, que se considera el inicio de la etapa Terminal (Larrazolo, 2002). Estos resultados fueron utilizados para contar con información real sobre el desempeño de los examinados.

Asimismo se utilizaron los resultados de dos grupos de segundo y cuarto nivel, compuesto cada uno por 30 estudiantes de la escuela de idiomas de la unidad en Ensenada, para llevar a cabo un análisis de contraste de grupos como evidencia de validez para el presente estudio.

3.3. Instrumentos

Uno de los instrumentos que se utilizó en este estudio es el EXEDII, el cual fue desarrollado como una prueba de opción múltiple con referencia a un criterio que se aplica por computadora y consta de 100 preguntas divididas en tres áreas; comprensión de lectura, comprensión auditiva y gramática.

La estructura del examen consiste en áreas, subáreas, objetivos y preguntas que miden diferentes habilidades (ver Tabla IV). Tanto el área de gramática como la comprensión de lectura constan de 34 preguntas cada una y el área de comprensión auditiva de 32. Para cada una de las áreas se eligieron los contenidos que los expertos consideraron como los más importantes, de acuerdo con el análisis curricular que efectuaron en cada área (Larrazolo y Velasco, 2000).

Tabla IV. Estructura temática del EXEDII.

Áreas	Subáreas	Número de Preguntas
Comprensión de Lectura	❖ Comprensión de palabras ❖ Comprensión de enunciados ❖ Comprensión de párrafos ❖ Razonar con palabras ❖ Razonar con enunciados ❖ Razonar con párrafos	6 5 6 7 5 5
Gramática	❖ Presente ❖ Pasado ❖ Adverbios ❖ Pronombres	13 7 10 4
Comprensión Auditiva	❖ Comprensión de palabras ❖ Comprensión de frases ❖ Comprensión de diálogos ❖ Razonar con frases ❖ Razonar con diálogos	2 7 10 5 8

Fuente: Larrazolo, N. (2002)

Otro instrumento que se elaboró fue la guía de instrucción para expertos con el fin de apoyar el proceso de entrenamiento de los expertos que participaron en la determinación del punto de corte. Esta guía describe la estructura temática del EXEDII (Tabla IV) ejemplos similares a los reactivos del examen, instrucciones para la ponderación de las preguntas, así como ejercicios prácticos de la manera como se deberían llevar a cabo sus juicios (ver Apéndice B). El objetivo de la guía fue la de proporcionar a los expertos toda la información pertinente para facilitar su tarea.

Una parte primordial del proceso de entrenamiento es elaborar la definición del concepto de competencia mínima por los expertos. Como etapa inicial del proceso se guió a los expertos en una reflexión sobre las destrezas y habilidades que un individuo con competencia mínima debería tener para contestar el EXEDII, con el fin de llegar a esta definición. Posteriormente se les proporcionó una definición previa del concepto, como punto de partida para la creación por los jueces del concepto de competencia mínima que utilizarían para guiar el proceso de valoración de los reactivos.

Otro instrumento que se utilizó fue un formato para facilitar a los jueces o expertos la valoración y registro de cada una de las preguntas del EXEDII, siguiendo los lineamientos establecidos por el método de Angoff para la evaluación de cada pregunta (ver Apéndice C). En este formato los jueces o

expertos registraron qué proporción de los examinados con competencia mínima serían capaces de contestar cada pregunta del examen correctamente.

También se utilizó un formato para evaluar el proceso de valoración efectuado por los jueces desde su entrenamiento, hasta el proceso mismo del establecimiento del punto de corte. Este formato es una adaptación de la forma de evaluación que aparece en el libro *Handbook for setting standards on performance assesment* de Hambleton, Jaeger, Plake y Mills (1998). El propósito de este formato es tener la opinión de los jueces sobre el desarrollo del proceso, de tal manera que sirva como referencia para valorar la calidad del proceso mismo (ver Apéndice D).

3.4. Procedimiento

A continuación se describen las etapas del modelo de Hambleton (2001). Estas fases cuentan con tareas principales y actividades específicas, las cuales se muestran en la tabla V.

Tabla V. Etapas del procedimiento para establecer puntos de corte.

Tareas principales del procedimiento	Actividades a realizar
PRIMERA ETAPA Selección del método Selección del grupo de Expertos o jueces	❑ Revisión de Literatura para analizar los métodos existentes y los resultados obtenidos a fin de seleccionar el método para este estudio. Factores para la selección: ❑ Conocimiento de la materia a evaluar. ❑ Experiencia o contacto con la población de interés ❑ Capacidad de análisis (comprensión, razonamiento, capacidad de decisión, sensibilidad hacia el problema.
SEGUNDA ETAPA. **Entrenamiento de los jueces o expertos (Se efectuará en varias reuniones).** Reunión Informativa.- Desarrollo de Definición de competencia minima. Segunda reunión.- Práctica de Valoración	❑ Aplicación del EXEDII a cada uno de los jueces. a) Se les dará información sobre el tipo de examen y su construcción, información sobre el método, sobre el propósito del examen y la manera en que se llevará a cabo el proceso de valoración de los reactivos . b) Se procederá al desarrollo de la definición del concepto de un examinado con conocimiento mínimo dentro del nivel intermedio de Inglés, con el apoyo de una Guía de Instrucción elaborada para este efecto, y por último una práctica de cómo efectuar este proceso.
TERCERA ETAPA: Valoración de los reactivos Primera Reunión de Retroalimentación Segunda Reunión: Retroalimentación Tercera Reunión: Establecimiento del punto de corte y evaluación del proceso.	c) Los expertos procederán a efectuar la valoración de los reactivos en el formato elaborado ex profeso. d) Se llevará a cabo el cálculo del promedio de cada uno de los jueces y se proporcionará esta información. Se hará una comparación entre los puntos de corte propuestos por cada juez y se informará del punto de corte obtenido por el grupo de expertos. Se proporcionará información sobre la dificultad de los reactivos y se procederá a reconsiderar la valoración . e) Se informará nuevamente a los jueces los promedios obtenidos por cada uno y se proporcionará información sobre el desempeño real de los examinados. f) Se informará a los jueces sobre el punto de corte determinado por cada uno de ellos y el punto de corte obtenido con la media de su valoración. Se llevará a cabo la evaluación del proceso con la aplicación del formato elaborado para ello.

a) Primera Etapa. Selección de los jueces y del método.

Con base en lo recomendado por Raymond y Reid (2001) se seleccionaron los jueces o expertos de la planta de 60 maestros de la Escuela de Idiomas de la UABC que cumplieron con las características mencionadas. Del grupo de 60 maestros se seleccionaron 7 de los 15 maestros que cumplían con el perfil mencionado. Estos 7 maestros cuentan con un grado académico de licenciatura o superior que los hace idóneos para la tarea a realizar. Por razones prácticas, al efectuar la selección de la muestra de expertos se consideraron únicamente maestros de la Unidad Ensenada, sin considerar las demás unidades localizadas en Tijuana, Mexicali y Tecate.

Esta muestra se llevó a cabo en forma no probabilística, ya que por el tipo de investigación que se realiza, se requiere contar específicamente con maestros que estén familiarizados con el nivel intermedio de inglés utilizado como base para la elaboración del EXEDII.

b) Segunda Etapa. Entrenamiento de los expertos.

Antes de iniciar el entrenamiento formal para la valoración de las preguntas se programó a cada uno de los integrantes del panel de expertos para que respondieran el EXEDII, tal como lo hacen los examinados. Esta es una recomendación que se sugiere, ya que permite al experto conocer el examen tal

cual es, lo que le brinda una mejor percepción de las áreas de contenido a evaluar (Hambleton, 2001).

Para facilitar la instrucción, se creó una guía (Apéndice B) que se entregó a los jueces con anticipación a la primera reunión. Esta guía contiene información sobre el proceso. Sin embargo, es sólo un complemento del entrenamiento, ya que el entrenamiento completo involucra actividades a realizar en conjunto con el panel de expertos, guiados por el responsable del proyecto.

Para llevar a cabo la etapa de entrenamiento se efectuó una primera reunión con el panel de expertos y se inició el entrenamiento propiamente dicho. En esta primera etapa se hizo una presentación a los expertos con material audiovisual para instruirlos sobre la importancia del estudio y darles a conocer información importante para capacitarlos en la labor a desarrollar. Se les proporcionó información sobre las dos categorías de pruebas existentes; con referencia a una norma y con referencia a un criterio. Asimismo, se les instruyó sobre las tres grandes diferencias que existen entre ellas, como son: a) el propósito de la prueba, b) la especificación del contenido y c) la generalización de los resultados.

Con la ayuda de la información contenida en la guía de instrucción y el apoyo de los responsables del proyecto, se llevó a cabo el proceso de elaborar una definición consensuada de un estudiante con competencia mínima, como

eje primordial del proceso de valoración de las preguntas. Para ello se guió a los expertos en una reflexión sobre las destrezas y habilidades que un individuo con competencia mínima debe tener para contestar el EXEDII con el fin de llegar a esta definición.

El proceso de definir las características de un estudiante con competencia mínima, es decir, el conocimiento mínimo necesario para pasar, es un factor muy importante en la implementación del método. La forma en que los especialistas conceptualicen este tipo de estudiante tiene implicaciones para la valoración del nivel de competencia, ya que el método de Angoff requiere que al momento de valorar cada pregunta los especialistas tengan en mente este tipo de examinado.

Para llevar a cabo este entrenamiento se presentó a los especialistas una definición de un estudiante con competencia mínima, para efecto de este trabajo se utilizará a partir de este momento las siglas ECCM para un estudiante con competencia mínima para aprobar el examen. La definición utilizada fue la siguiente:

> "El ECCM es alguien que tiene el conocimiento mínimo necesario del idioma Inglés para contestar correctamente los reactivos del examen a nivel intermedio".
>
> Cizek, 2001 p. 142

Para facilitar la discusión y con el fin de dirigir la atención de los especialistas hacia las habilidades y destrezas del ECCM se hicieron las siguientes preguntas:

Comprensión Auditiva. ¿Cuáles consideran ustedes que son los factores que facilitan o dificultan la comprensión auditiva de un ECCM?

- En fonología: ¿Reconoce los sonidos? ¿Sabe donde terminan y comienzan las palabras?

- En entonación: ¿Puede distinguir si es una pregunta, una orden, una petición? ¿Reconoce el énfasis y el propósito?

- En contenido: ¿Reconoce las palabras? ¿Puede determinar el propósito de acuerdo con el contexto? ¿Reconoce cuando existe ambigüedad? ¿Puede realizar inferencias del significado correcto? ¿Puede efectuar discriminaciones a partir de la información poniendo a un lado el concepto literal?

Comprensión de Lectura. ¿Cuáles son los factores que hacen fácil o difícil la lectura?

- ¿Puede reconocer entre la idea principal y las ideas secundarias?

- ¿Puede hacer inferencias?

- ¿Puede encontrar el significado de las palabras utilizando el contexto?

- ¿Entiende la organización del escrito, es decir, puede entender las conexiones lógicas (ej. but, otherwise, therefore?)

- ¿Capta la conexión entre las ideas?

En general ¿cuáles características tiene un ECCM? ¿Qué hace este tipo de estudiante al enfrentarse ante un examen?

Estas preguntas se realizaron por áreas de competencia y generaron una lista de factores que hacen fácil o difícil la tarea para los examinados. Estos factores se fueron registrando en el pizarrón a medida que se iban describiendo. Posteriormente, al final de la discusión se borró la lista y se solicitó a los especialistas que cada uno de ellos escribiera su propia definición de un ECCM. Con el fin de ayudarlos en esta etapa del proceso, se les sugirió que pensaran en uno de sus estudiantes que se ubicara en esta categoría.

Con la ayuda de la información contenida en la guía de instrucción y el apoyo de los responsables del proyecto se llevó a cabo el proceso de elaborar una definición consensuada de un estudiante con competencia mínima, como eje primordial del proceso de valoración de las preguntas.

Una vez que cada uno de los jueces escribió su propia definición, se procedió a la comparación de las mismas, con el fin de establecer una sola definición para el grupo de especialistas, la cual se utilizó en la valoración de los reactivos del EXEDII.

Con la ayuda de los ejemplos contenidos en la guía de instrucción y otros ejemplos adicionales, se procedió a la valoración de las preguntas siguiendo los pasos que se indican en la guía de instrucción. 1) Examinar una a una las preguntas del examen. 2) tener en consideración la definición de un examinado con competencia mínima para valorar las preguntas. 3) Estimar la probabilidad de que un estudiante "hipotético" con un conocimiento aceptable, pueda conocer la respuesta correcta sin hacer uso del factor de adivinación y 4) Registrar este porcentaje en el formato elaborado para ese efecto.

Angoff (1971), sugiere que para hacer más fácil la determinación de la probabilidad, el experto imagine un grupo de (10 ó 100) individuos que tengan una competencia mínima y después determine cuántos de ese grupo pueden contestar correctamente la pregunta.

a) Tercera etapa. Valoración de las preguntas.

Los expertos procedieron a la valoración de las preguntas en forma individual (estimación del porcentaje de probabilidad de que un examinado con competencia mínima pueda contestar correctamente las preguntas). La valoración de los expertos a cada pregunta se registró en el formato especialmente diseñado para ese efecto (Apéndice D) con una escala de intervalos de posibilidades que van desde 0 hasta 100. De esta manera los jueces pudieron establecer un juicio de una forma más fácil y confiable. A la

entrega de las respectivas valoraciones de los jueces se llevó a cabo el cálculo del promedio de valoraciones por cada experto, así como el promedio ponderado de la valoración del grupo de expertos.

Primera reunión de retroalimentación.

En esta primera reunión de retroalimentación y comparación de la clasificación entre jueces, se les proporcionó información sobre la media obtenida de las puntuaciones, sobre la valoración de las preguntas. Se analizaron aquellas estimaciones que se encontraban en los extremos, o sea, cuando un juez le daba a una misma pregunta una valoración de "muy fácil" y otro de "muy difícil". Se escogieron algunas preguntas que estuvieron en esa situación y se solicitó a cada uno de los jueces involucrados que explicara el argumento que utilizó para su interpretación con el fin de que tuvieran oportunidad de contrastar sus puntuaciones.

Posterior a esta discusión se mostró información empírica al grupo de expertos sobre el desempeño real del grupo de 710 estudiantes para cada una de las 100 preguntas del EXEDII, de tal manera que el grupo pudiera contar con información adicional para la siguiente valoración. Todo ello con el fin de apoyarlos en el proceso de evaluación como varios estudios recomiendan (Popham 1990; Hambleton 1998; Jaeger 1981). Al finalizar esta sesión se procedió a una nueva valoración individual.

Segunda reunión de retroalimentación

Se les proporcionó información sobre el punto de corte determinado por la media de sus valoraciones. Posteriormente se proporcionó información sobre el número de estudiantes que pasarían de aplicarse el punto de corte recomendado. Asimismo, se les hizo notar que de aplicarse el punto de corte que ellos determinaron el número de estudiantes que aprobarían el examen tendería a disminuir, ya que el punto de corte establecido por ellos era más alto (72) que el actual (55). Una vez proporcionada esta información se solicitó a los jueces que procedieran a realizar una última valoración.

Tercera reunión de valoración

Esta reunión fue la que concluyó el proceso de valoración. En esta ocasión se presentó a los jueces nuevamente el punto de corte obtenido con la media de sus valoraciones en la sesión anterior con el fin de que determinaran si se requería una nueva sesión de valoración. Después de analizar el punto de corte obtenido los jueces acordaron por consenso aceptar este resultado como el punto de corte.

3.5 Evaluación del proceso.

Como última etapa de este procedimiento de investigación se proporcionó a los expertos una forma (Apéndice C) para evaluar todo el proceso. En este formato los jueces externaron su opinión sobre todas las etapas de la investigación.

3.6 Evidencias de validez

Para obtener evidencias de validez para este estudio se utilizaron diversos análisis:

- Coeficiente de variación. El coeficiente de variación es un análisis de la variabilidad relativa que se utiliza como técnica estadística en los estudios que involucran jueceo. Este tipo de análisis sintetiza mejor la información sobre si los jueces han manifestado o no la misma tendencia en sus valoraciones, sin ser estrictos respecto al acuerdo total en el número que hayan utilizado para evaluar cada reactivo (Jornet, J y Suárez, J., 1989b).

Este coeficiente es una medida, en términos de porcentaje, acerca de que tan homogéneos son los juicios de los especialistas y se obtiene con la siguiente fórmula:

$$Cv = \frac{\sigma}{\mu} * 100 \quad (1)$$

Donde:

Cv= coeficiente de variación

σ = desviación estándar de las valoraciones de los jueces

μ = media de las valoraciones de los jueces

Los parámetros aceptables para este coeficiente son los siguientes: De 0 a 31 % se considera que se tienen juicios homogéneos, de 32 a 35% normal, se considera que se tienen juicios heterogéneos si el coeficiente es mayor a 35%. En este tipo de investigación se busca obtener juicios homogéneos para apoyar la validez del proceso.

Se realizó otro análisis de consistencia utilizando "grupos de contraste". En este caso se utilizaron dos grupos de examinados; uno considerado con un nivel de competencia alto y el otro con nivel de competencia bajo. Se llevó a cabo un análisis de la distribución de las puntuaciones de ambos grupos y posteriormente se graficaron estas puntuaciones para encontrar el punto de intersección, el cual se comparó con el valor obtenido en esta investigación y el punto de corte actual del EXEDII. En la medida en que este valor se acerque al valor establecido como punto de corte de este estudio, apoyará la validez de este estudio.

CAPITULO IV

RESULTADOS

En este capítulo se presentan los resultados distribuidos en cuatro apartados. El primero corresponde a la selección y entrenamiento de los expertos, para lo cual se elaboró una guía de instrucción (Ver Apéndice B). El segundo presenta las estadísticas de la valoración de los reactivos de las tres sesiones, tanto por área como por sesión. El tercero muestra los indicadores estadísticos de la tercera sesión en particular y por área de habilidad, utilizando para ello las medias y desviaciones estándar. Asimismo se calculó el nivel de acuerdo entre expertos utilizando el índice de variabilidad (Jornet, 1998). También se utilizó un análisis de contraste de grupos (Livingston 1972). Por último, se analizó la evaluación del proceso por los expertos utilizando una escala tipo Likert.

4.1 Selección y entrenamiento de los expertos.

Se seleccionaron siete maestros de la planta docente de la Escuela de Idiomas, quienes fungieron como expertos en el presente estudio. Para preparar a los expertos en el proceso de valoración de los reactivos se llevaron a cabo varias reuniones de entrenamiento, para lo cual se utilizó la guía de instrucción elaborada para ese efecto (Apéndice B). Una parte importante del proceso de entrenamiento fue que los expertos determinaran el perfil de referencia de un

estudiante con competencia mínima, la cual se utilizó para llevar a cabo la valoración de los reactivos. Esta definición es la siguiente:

> "Un ECCM es capaz de comprender textos y conversaciones sencillas, cortas y claras en situaciones cotidianas, con manejo de vocabulario y estructuras gramáticas básicas de presente, pasado y futuro simples, con dificultad para distinguir tiempos compuestos. Apoyándose en el contexto puede inferir algunos significados e información implícita. Generalmente es capaz de distinguir la idea principal. Sin embargo, tiende a confundirse en un texto o conversación compleja. Puede distinguir entre hecho y opinión y encontrar información específica, así como seguir instrucciones".

Posteriormente, los expertos procedieron a la valoración de los reactivos del EXEDII. En el Apéndice E se presenta el análisis de los reactivos de la tercera sesión.

Para la estimación de las preguntas del examen se efectuaron tres reuniones cuyos resultados se presentan en la Tabla VI. Cabe mencionar que, varios especialistas en medición (Berk, 1986; Jaeger, 1989; Livingston y Zieky, 1982 y Hambleton 2001) apoyan la recomendación de incluir información adicional en el proceso de valoración para la determinación del punto de corte. Por lo anterior, se llevaron a cabo varias sesiones de evaluación y retroalimentación con el fin de hacer más objetivas las valoraciones de los expertos y proporcionar información adicional que apoyara dicho proceso. Esto se logró al permitir un espacio de discusión entre cada sesión de valoración.

Tabla VI. Estadísticas básicas para cada una de las áreas en las tres sesiones.

AREA	SESION 1		SESION 2		SESION 3	
	$\bar{X}$	*s*	$\bar{X}$	*s*	$\bar{X}$	*s*
Auditiva	70	16	73	12	70	13
Gramática	73	19	70	16	70	14
Lectura	72	16	74	13	72	13
TOTALES	72	17	72	13	71	9

En la tabla VI se puede observar la variación en la estimación de acuerdo con las diferentes sesiones. Se esperaba cierta variación debido a la introducción de información adicional para ayudar a los jueces a aproximarse a una valoración real, según el método utilizado.

La valoración de la primera sesión proporcionó un punto de corte de 72. Después de la primera sesión se proporcionó información sobre el desempeño real de un grupo de examinados (muestra de 710 estudiantes) con la intención de que los jueces tomaran en cuenta la puntuación promedio de este grupo (58). Esta puntuación se considera como el punto de corte para dicha muestra. La información proporcionada no afectó grandemente la valoración de los jueces, ya que al final de la segunda sesión el promedio se situó nuevamente en el valor de 72. La única variación que se observa es una menor dispersión de los datos (*s*) en todas las áreas de la primera a la segunda sesión, como lo muestra la tabla VI arriba mencionada.

Para complementar el análisis de las tres sesiones, se analizó la valoración de los jueces por áreas (comprensión auditiva, comprensión de lectura y gramática) con el fin de mostrar la tendencia de dichas estimaciones, como se puede observar en la Tabla VII.

Tabla VII. Promedio de la valoración de las áreas en las tres sesiones.

AREA	JUECES						
	1	2	3	4	5	6	7
Auditiva	73	79	81	52	68	79	69
Gramatica	74	76	82	71	69	66	68
Lectura	72	73	81	77	72	55	64
TOTALES	73	76	81	67	70	67	67

4.2 Valoración de la tercera sesión.

Según los expertos en medición (Berk, 1986; Jaeger, 1989; Livingston y Zieky, 1982 y Hambleton 2000) en el tipo de modelo donde se utiliza información adicional sobre el desempeño real de grupos de examinados, la última sesión refleja la decisión final de los jueces (Aguilar, 2004, comunicación personal). Por tal motivo en este estudio se consideró la tercera sesión para determinar el punto de corte.

En la tercera sesión se proporcionó a los jueces una proyección del impacto del punto de corte partiendo de la información real sobre el desempeño de los estudiantes en el examen y comparando esta información con el punto de corte establecido por ellos en la sesión anterior. Con esta información fue posible dar a conocer a los jueces las consecuencias que podría tener la aplicación de este nuevo punto de corte. Posteriormente, los jueces procedieron a efectuar la siguiente valoración. Se puede observar el efecto que esta información tuvo en el jueceo, ya que el punto de corte se modificó de 72 a 71 como se muestra en la Fig. 2.

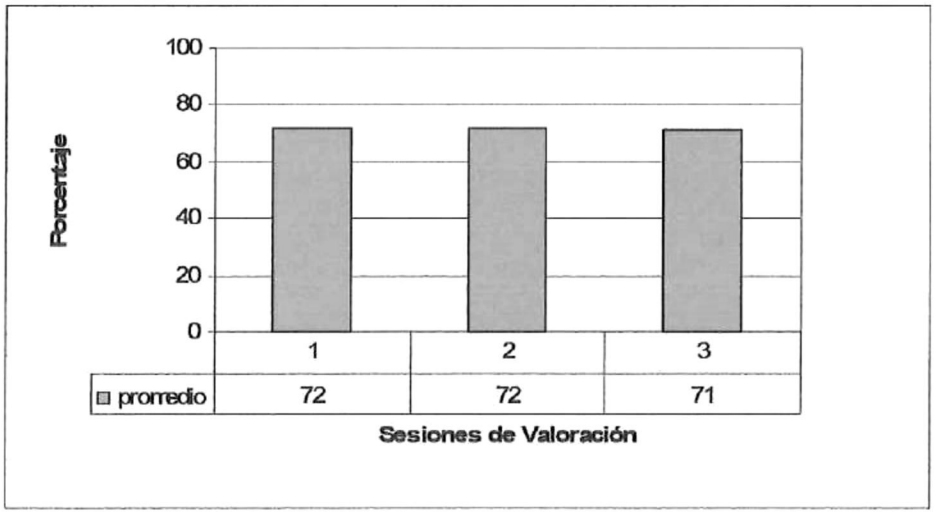

Fig. 2. Valoración de la tercera sesión.

En esta sesión se hizo un análisis de la estimación de los reactivos del examen por cada uno de los jueces (Apéndice E), para lo cual se construyó una gráfica de distribución de los datos que refleja la manera en que fueron evaluados cada uno de los reactivos (Fig. 3).

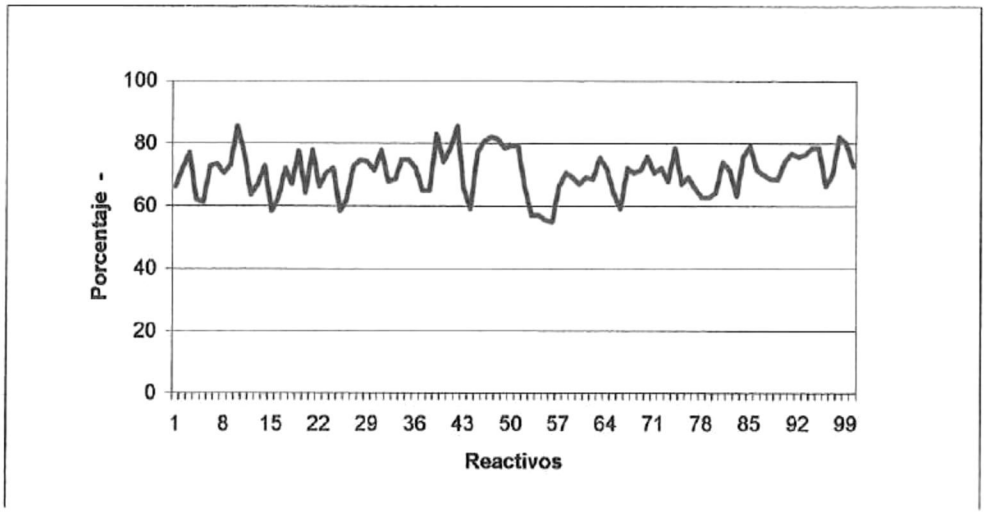

Fig. 3. Valoración de los reactivos de la tercera sesión

En relación a la estimación de los reactivos, cuando el valor se acerca a 100 se considera que es menos difícil, ya que la mayoría de los estudiantes puede contestar correctamente ese reactivo. En cambio, cuando el valor se acerca a 0 significa que el reactivo es difícil de contestar por los estudiantes (Crocker y Algina, 1986).

Al efectuar la valoración de los reactivos, los expertos le dan un valor que puede interpretarse como la dificultad que este reactivo representa para el examinado con competencia mínima. En la Fig. 3 se puede observar que la mayoría de los datos caen en el rango de 70-80.

En la Tabla VIII se presenta el promedio de la valoración de cada uno de los jueces por área para la tercera sesión.

Tabla VIII. Valoración de los jueces por área en la tercera sesión.

ÁREA	JUECES						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Auditiva	69	70	78	79	75	53	66
Gramática	74	74	81	75	64	54	70
Lectura	73	76	82	78	77	58	58

Asimismo se obtuvieron los indicadores estadísticos como la media ($\bar{X}$), la desviación estándar (s) y la varianza (s^2) por área temática, los cuales brindan información sobre la manera en que los jueces estimaron el nivel de dificultad de dichas áreas (Ver Tabla IX). A simple vista podría decirse que hubo diferencias en la valoración individual de los jueces por áreas. Sin embargo, en su mayoría los valores fluctúan en el rango de 70 a 80 con excepción del juez 6 que mantuvo la valoración de las tres áreas en el rango de 50 y el juez 7 que sitúa la valoración del área auditiva y lectura en el rango de 60 y 50 respectivamente. Por otra parte, si consideramos el promedio de valoración de las áreas por los jueces, ésta se sitúa en el rango de 70 (ver Tabla IX).

Tabla IX. Promedio Valoración por área en la tercera sesión

AREA	SESION 3		
	$\bar{X}$	s	s^2
Auditiva	70	13	174
Gramática	70	14	207
Lectura	72	13	173
TOTALES	71	9	74

Como se observa en la tabla IX, la tendencia de las valoraciones promedio por área es muy uniforme: comprensión auditiva (70), gramática (70) y comprensión de lectura (72). La interpretación de las áreas temáticas del examen es un aspecto importante a considerar para obtener mayores evidencias de validez.

En la Fig. 4 se puede observar el promedio de la valoración de cada uno de los jueces para la tercera sesión (72, 73, 81, 77, 72, 55, 67) respectivamente.

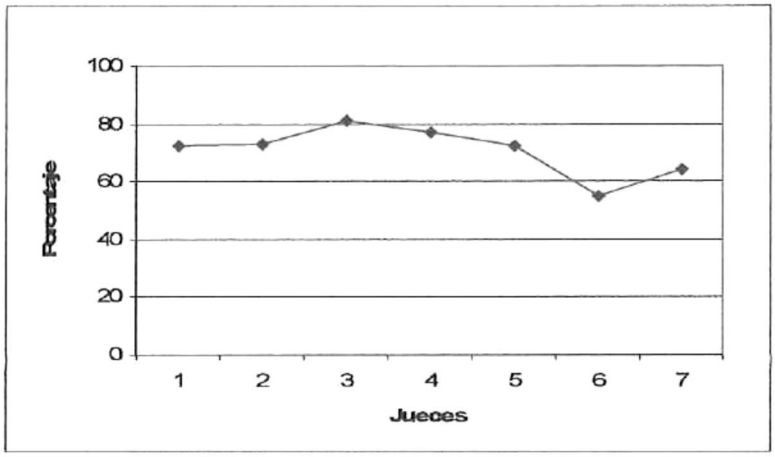

Fig. 4. Valoración de los jueces en la tercera sesión

Observando los análisis anteriores y de acuerdo con el método utilizado y con la opinión de los expertos, en la tercera sesión se determinó el punto de corte para el EXEDII en 71.

4.3 Evidencias de validez del punto de corte

Una evidencia de validez para estudios de punto de corte es la homogeneidad en la valoración de los expertos. El coeficiente de variación es un análisis de la variabilidad relativa que se utiliza como técnica estadística en los estudios que involucran jueceo. Este coeficiente es una medida, en términos de porcentaje, de qué tan homogéneos son los juicios de los especialistas. Para obtener dicho coeficiente se utiliza la fórmula 1 (Pág.79).

Con objeto de reunir información para realizar un análisis de la variabilidad de las valoraciones de los expertos, se obtuvo la media y la

desviación estándar de las valoraciones de la última sesión por los jueces, dichos resultados se muestran en la tabla X.

Tabla X. Índice de variación por juez.

Juez	Media	Desviación Estándar	Coeficiente de variabilidad
1	72	11	10.6
2	73	6	7.9
3	81	11	13.8
4	77	6	7.15
5	72	17	24.2
6	55	9	16.3
7	64	13	20.6

Este coeficiente de variabilidad muestra que tan homogéneos son los juicios de los especialistas. Sin embargo, para estar en posibilidad de darle una interpretación adecuada a dicho coeficiente, es necesario que la desviación estándar sea menor a un tercio de la media. Esto quiere decir que la dispersión de los datos será menor a una desviación estándar, lo que minimiza el error de medición (Jornet, 2003, comunicación personal). En la tabla XI se muestra la comparación del tercio de la media con la desviación estándar obtenida.

Tabla XI. Comparación Desviación Estándar y Tercio de la media.

Juez	Media	Desviación Estándar	Tercio de la media
1	72	11	24
2	73	6	24
3	81	11	27
4	77	6	25
5	72	17	24
6	55	9	18
7	67	13	22

Con base en la información obtenida, se observa que los valores de la desviación estándar son menores al tercio de la media para todos los jueces. En la tabla XII se muestran los valores asociados al grado de acuerdo de los juicios entre jueces (Jornet y Suárez, 1989).

Tabla XII. Porcentajes para interpretación del coeficiente de variabilidad (Cv)

Valores de Cv	Interpretación
De 0 a 31 %	Juicios Homogéneos
De 32 a 35 %	Normal
Mayor a 35%	Juicios Heterogéneos

De acuerdo con los datos obtenidos en la tabla XI y XII se muestra que los juicios observados se encuentran dentro de los valores homogéneos. Esto indica, que a pesar de las diferencias en los valores utilizados por los jueces, su valoración es aceptable

Otra evidencia de validez se relaciona con las recomendaciones de Kane (2001) de utilizar un análisis de contraste de grupos. Este análisis se efectuó con información real sobre dos grupos de examinados (*instruídos* y *no instruídos*), utilizando una muestra de 30 estudiantes de segundo y 30 de cuarto nivel de la escuela de idiomas a los cuales se les aplicó el EXEDII. De acuerdo con el procedimiento se llevó a cabo la distribución de las puntuaciones de ambos grupos, se graficaron los resultados y el punto en el cual ambas distribuciones se intersectaron se consideró como el punto de corte de esa muestra. En la siguiente gráfica se presentan los resultados de los grupos de contraste.

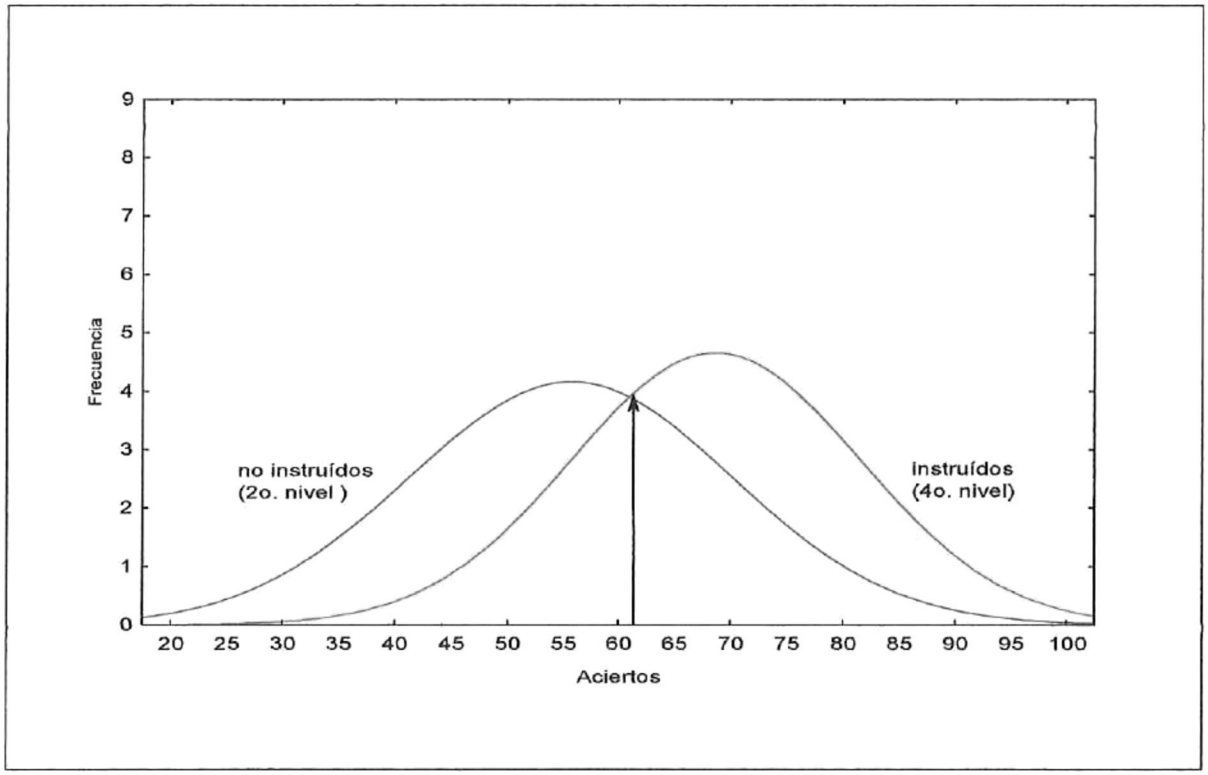

Fig. 5. Distribución de grupos de contraste.

Como se puede ver en la gráfica el punto de corte para estos grupos se situó en 61. Sin embargo, este resultado deberá tomarse con reserva ya que la muestra es muy pequeña. Cabe mencionar que el procedimiento de grupos de contraste es en sí un método para determinar puntos de corte. Sin embargo, en este estudio se está utilizando únicamente como evidencia de validez para estimar la tendencia de las valoraciones. No es sorprendente encontrar diferencias en el punto de corte establecido con el grupo de contraste y el determinado por el método de Angoff, ya que varios estudios han demostrado que existen discrepancias cuando se utilizan métodos diferentes (Andrew and Hecht, 1976; Donnoe W. 1976; Cross, et al.1984, Jaeger, R.M. 1989).

4.4 Evidencias de validez del proceso.

Otro de los aspectos de validez (Hambleton, 1998), se basa en la evaluación del proceso desde el punto de vista de los jueces. Para ello se utilizó un formato de evaluación (Apéndice D), en el cual los expertos dan su opinión sobre el procedimiento y su propio desempeño en la aplicación del mismo.

Este formato está compuesto por una serie de preguntas de opinión que consideran diversos aspectos del proceso de valoración. Para su análisis se dividió en cinco secciones:

- La primera sección corresponde a la presentación del proceso de clasificación.
- La segunda se relaciona con el proceso de entrenamiento.
- La tercera se refiere a la percepción de los jueces de su propio desempeño.
- La cuarta sección al tiempo de implementación del proceso.
- En la quinta sección a lo relativo al procedimiento.

Por último, se contemplan tres preguntas de respuesta abierta con el fin de obtener una reflexión de la participación de los jueces en el proceso de valoración. A continuación se presenta la tabla XIII que resume los resultados obtenidos en las primeras cinco secciones del proceso de evaluación.

Tabla. XIII Evaluación del proceso por los jueces

Presentación del proceso	**Importancia Relativa**
1 Introducción y presentación	85%
2 Entrenamiento en la clasificación	85%
3 Primera Sesión de discusión grupal	76%
4 Segunda Sesión de discusión grupal	85%
Proceso de entrenamiento	
1 Entrenamiento para preparación en la clasificación	85%
2 Tiempo utilizado en el entrenamiento	100%
Percepción de los jueces de su propio desempeño	
1 Dificultad del método	90.4%
2 Su propia experiencia	90.4%
3 Su primera clasificación individual	76.1%
4 Panel de discusión grupal	100%
Tiempo de Implementación del proceso	
1 Primera discusión grupal	100%
2 Segunda discusión grupal	92.8%
Relativo al procedimiento	
1 Confianza en que el procedimiento produzca un estándar de desempeño confiable	71.4%
2 Facilidades para llevar a cabo la valoración	78.5%

En esta tabla se muestran los resultados generales del proceso de evaluación. Para su análisis se utilizó una ponderación de importancia relativa (Organista, 1998). Para ese efecto se agruparon las opiniones expresadas por los jueces en la escala correspondiente. Por ejemplo: en la escala de 1 a 4, donde 1= menor y 4= mayor se hicieron dos columnas. En una columna se ubica la escala y en la otra la opinión de los jueces para cada una de las escalas. Posteriormente, se multiplica cada celda de la columna correspondiente para obtener el valor de la escala. Para calcular el valor máximo que la escala pudiese tener, se calculó el valor máximo de la escala (4) por el número de

jueces (7). Por último se obtuvo la importancia relativa mediante una regla de 3. Ejemplo:

(28) puntos es igual al (100%)

(18) puntos es igual a (x)

Al aplicar este procedimiento se obtiene la importancia relativa para cada uno de los aspectos considerados en las diferentes secciones. Cabe hacer notar que debido a que la encuesta considera diferentes escalas, esta fórmula se aplicó considerando estas diferencias.

Como ejemplo, la Figura 6 muestra el comportamiento de la evaluación por los jueces en la presentación del proceso. En la introducción del proceso, en el entrenamiento en la clasificación y en la segunda sesión de discusión grupal, tres de los jueces opinaron que fue exitoso y cuatro de ellos situaron su opinión en muy exitoso, lo cual se refleja en el 85% de la importancia relativa. En cuanto a la primera sesión de discusión grupal, dos de los jueces situaron su opinión en muy exitoso y cinco en exitoso, lo cual se refleja en el 76% de la valoración.

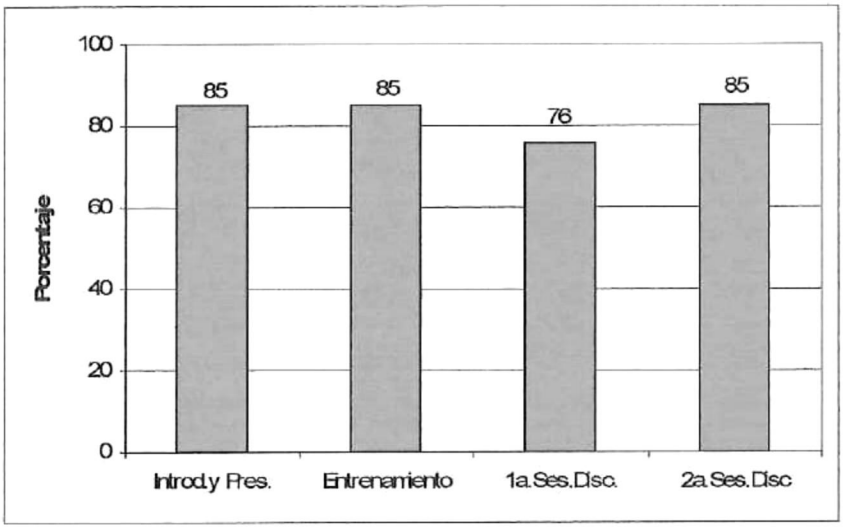

Fig. 6. Evaluación de la presentación del proceso por los jueces.

En la figura 7 se presenta la valoración que dan los jueces a su propio desempeño en la evaluación del proceso. En relación a la dificultad del método y su propia experiencia, cinco jueces seleccionaron la opción *muy importante* y dos jueces la opción *importante*, lo que se refleja en el porcentaje del 90.4%. Con respecto a su primera clasificación individual, cinco jueces consideraron que fue *importante* y dos *muy importante*, lo que nos da un 76% y en el panel de discusión grupal los siete jueces opinaron que fue *muy importante* lo cual da el 100% de importancia relativa.

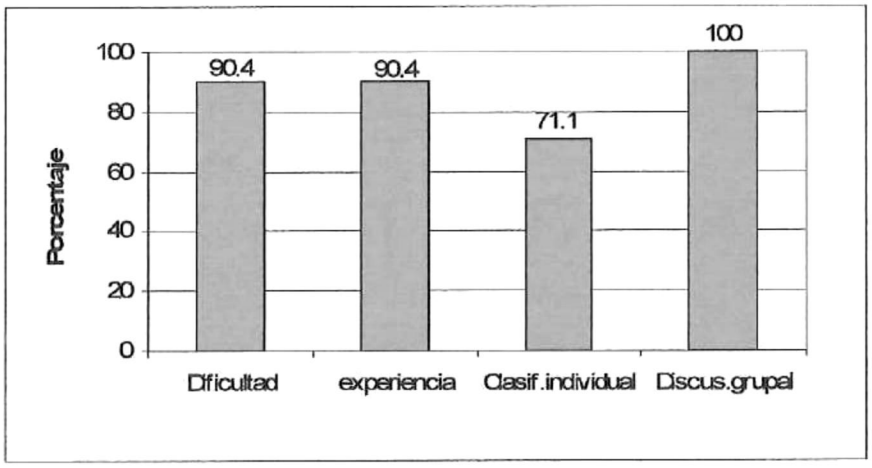

Fig. 7. Percepción de los Jueces de su desempeño.

Cabe mencionar que las cinco secciones que se evaluaron cuentan con diferentes categorías de evaluación. (Ver Apéndice D). Por ejemplo, en la primera sección (Presentación del proceso), las categorías son 4: *muy exitoso*, *exitoso, parcialmente exitoso* y *sin éxito*. (Ver Fig. 8). En esta figura se pueden observar dos categorías, esto significa que las valoraciones de los jueces se concentraron únicamente en estas opciones.

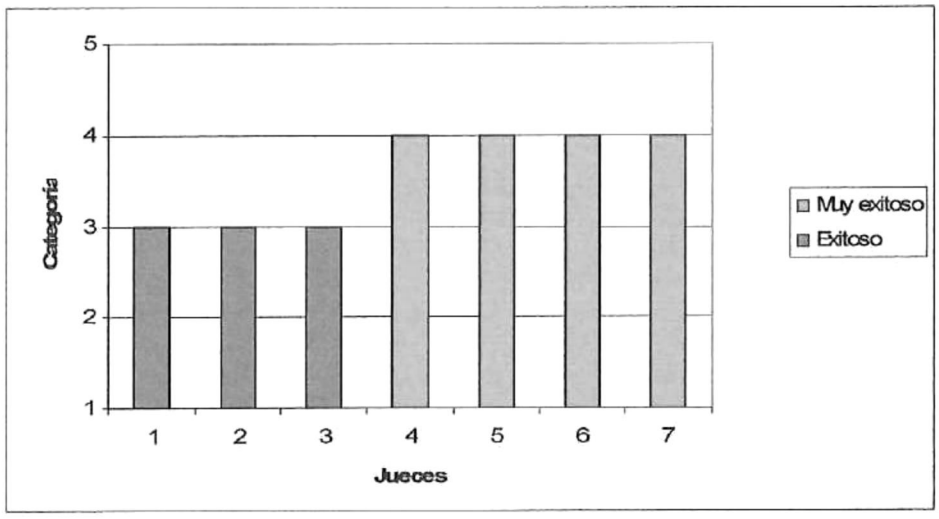

Fig. 8. Ejemplo de categoría de análisis del proceso.

Otro ejemplo de la diferente clasificación de opciones del cuestionario es lo relativo a la percepción de los jueces de su propio desempeño. Las categorías utilizadas fueron: *muy importante*, *importante, importante de alguna manera* y *sin importancia* (Ver Fig. 9).

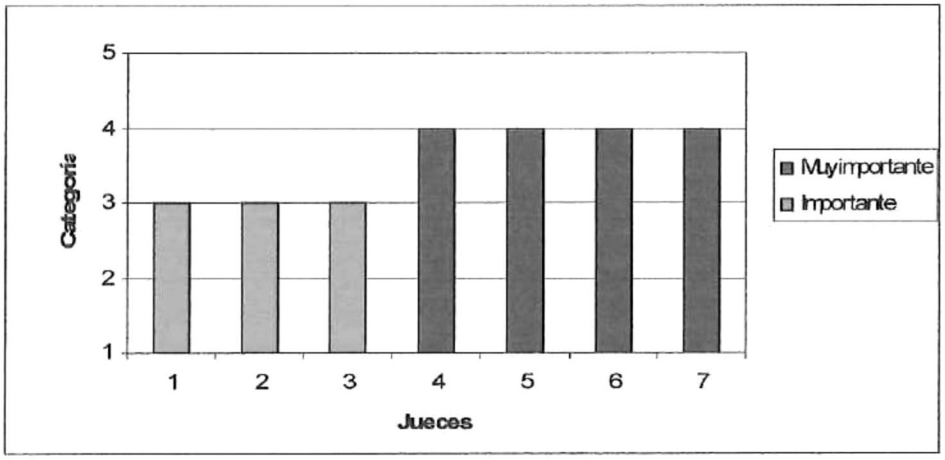

Fig.9. Ejemplo de categoría de análisis de los jueces de su propio desempeño.

Como se observa en la Fig. 9, la clasificación varía de importante a muy importante. Igualmente las opiniones de los jueces se concentraron en éstas dos categorías.

La segunda parte de la valoración del proceso tiene tres preguntas abiertas (Tabla XIV), cuyas categorías no están predeterminadas. Para su evaluación se identificaron las palabras clave y se valoraron las coincidencias de los jueces.

Tabla XIV. Preguntas abiertas del proceso de evaluación

Preguntas	Coincidencias
¿Qué estrategias utilizó para clasificar las preguntas?	▪ Perfil y definición ECCM ▪ Experiencia ▪ Dificultad del tipo de reactivo ▪ Conocimiento del programa de estudios ▪ Consideración de barreras afectivas del estudiante ▪ Opinión de otros expertos en la sesión de discusión grupal
¿Hubo problemas que influenciaran su decisión?	▪ Dificultad para considerar el estudiante con competencia mínima ▪ Comentarios de los compañeros expertos en las sesiones de discusión ▪ Influencia de la información real sobre desempeño ▪ La influencia del conocimiento sobre estudiantes formales
Sugerencias para mejorar el procedimiento	▪ La guía de instrucción y la presentación del instructor apoyan el proceso. ▪ Ninguna. Funcionó bien. Tiene confianza en el proceso. ▪ Efectuar la primera clasificación en la primera reunión de discusión grupal. ▪ Ampliar más las sesiones de discusión. ▪ Sobre el examen, poner la comprensión auditiva más adelante para darle al alumno oportunidad de acostumbrarse al idioma.

En la primera pregunta las palabras clave que contaron con mayores coincidencias fueron: el perfil y definición del ECCM (estudiante con conocimiento mínimo) con el 100% de acuerdo de los jueces. Con relación a la experiencia, 6 de 7 jueces consideraron ésta como un factor determinante. Le sigue en orden de importancia la dificultad del reactivo y el conocimiento del programa de estudios (4 de 7 jueces). Ahora bien, tres de siete jueces consideraron el aspecto afectivo del estudiante y por último, dos de siete jueces tomaron en consideración la opinión de otros jueces en la valoración. Con

relación a este último aspecto, los jueces consideraron que fue útil la discusión entre cada sesión de valoración.

En la segunda pregunta se investigó si hubo problemas que influenciaran su decisión. En este aspecto la mayoría de los jueces (5 de 7) coincidieron en afirmar que no visualizaron como problema la participación de los otros expertos en las reuniones de discusión, sino más bien lo vieron como algo positivo. Uno de los jueces reportó no haber tenido ningún problema que influenciara su decisión. Otro juez informó que su experiencia le ayudó a situarse en el lugar del estudiante con conocimiento mínimo. Sólo un juez comentó que le fue difícil visualizar al estudiante con conocimiento mínimo.

Con relación a las recomendaciones sugeridas para mejorar el proceso, en primera instancia se consideró que la guía y la presentación apoyaron el proceso. Uno de los jueces recomendó que el área de comprensión auditiva se cambie de lugar, ya que es la primera habilidad que se mide en el examen. Por otro lado, otro juez sugiere que se le otorgue más tiempo a las reuniones de discusión y por último varios sugieren que la primera sesión de valoración se efectúe en la primera reunión.

De acuerdo a las opiniones expresadas por los jueces o expertos, el proceso de valoración puede considerarse exitoso, lo cual proporciona una evidencia de validez del proceso.

CAPITULO V

CONCLUSIONES

En este capítulo se presentan las conclusiones derivadas de los resultados obtenidos en este estudio y se contrastan con los objetivos propuestos. Asimismo, se presentan las aportaciones que esta investigación puede generar en el contexto del establecimiento de estándares, o puntos de corte en el campo de estudio. Por otro lado, se exponen las limitaciones que se tuvieron en la realización de este trabajo, así como las recomendaciones para investigaciones futuras.

Como parte importante del proceso de determinación de puntos de corte, se analizan las implicaciones sociales, políticas y académicas que se tendrían al aplicar los resultados de este estudio en el contexto de la Universidad Autónoma de Baja California.

En relación al primer objetivo planteado en este estudio, sobre la selección del método más apropiado para determinar el estándar; se llevó a cabo una amplia revisión de la literatura. Se puede afirmar que a partir de 1970 se han realizado una gran cantidad de estudios para determinar puntos de corte sobre todo en Estados Unidos, país que le ha dado gran relevancia al rendimiento de cuentas en la educación, por lo cual la mayoría de los estados se someten a pruebas estandarizadas a gran escala que requieren establecer

puntos de corte. Por esta razón se han llevado a cabo diversos estudios con la intención de encontrar el método más apropiado para establecer puntos de corte, sin que hasta la fecha se haya determinado un método en particular. Sin embargo, de entre todos los métodos empleados por diversos investigadores, Harasym (1981), Cross e Impara, (1984), Livingston y Zieky, (1983), Berk (1986), Donnoe y Amato (1997), se ha destacado el método de Angoff (1971) por la facilidad de implementación y de cálculo, además de ser el más utilizado por el National Assessment of Educational Progress (NAEP) de Estados Unidos, organismo encargado de aplicar pruebas a gran escala a nivel nacional. Por lo anterior se eligió el método de Angoff para ser utilizado en la presente investigación, complementado con la inclusión de información adicional, ya que de acuerdo con estudios efectuados por Shepard, (1984), Hambleton (1998), Jaeger (1989), Kane, (2001), esto tiende a reducir la variabilidad en la distribución del estándar recomendado.

Con relación al segundo objetivo de este estudio, sobre la implementación del proceso para la aplicación del procedimiento seleccionado, se tuvo especial cuidado en la selección de los expertos y en su entrenamiento. El entrenamiento se llevó a cabo en cuatro sesiones. En una de ellas se realizó una presentación visual sobre el procedimiento y los objetivos del estudio a realizar. Asimismo, se creo una guía para el entrenamiento de los jueces con el fin de que tuvieran información de referencia que apoyara el proceso de valoración en el momento mismo de la estimación. Se brindaron también amplias facilidades de

retroalimentación para asegurar que los jueces comprendieran el método. Por recomendación del procedimiento mismo, se llevaron a cabo tres reuniones de retroalimentación en las cuales los jueces tuvieron oportunidad de comentar sus valoraciones. En estas reuniones se solicitó a aquellos jueces que presentaron valoraciones muy disímiles, que explicaran su razonamiento para ello. Con este tipo de interacciones se trató de uniformar los criterios de los especialistas. Sin embargo, de acuerdo con los análisis estadísticos de estas interacciones, el nivel de acuerdo de los jueces fue bastante moderado.

Respecto al último objetivo planteado, con relación a la acumulación de evidencias para validar el proceso, se realizaron varios análisis para verificar si las tendencias en la valoración de los jueces fueron similares. Para tal efecto se utilizaron análisis estadísticos cuyos resultados indican que los juicios de los expertos fueron homogéneos. Por otra parte, también se efectuó un análisis de la validez empírica del punto de corte utilizando para ello un análisis de grupos de contraste (instruídos y no instruídos). El resultado de este análisis fue un punto de corte de 61. Sin embargo, los grupos utilizados para este análisis fueron muy pequeños (30 estudiantes), por lo cual se recomienda tomar con precaución este resultado y continuar realizando análisis similares para corroborar la información.

Es interesante notar que aún cuando existieron diferencias en las valoraciones, la estimación del punto de corte tanto global como por áreas se

mantuvo dentro de un rango de puntuaciones de 70 a 72, quedando el punto de corte en 71. Si se toma en cuenta que el punto de corte para el programa de Inglés que ofrece la Escuela de idiomas es de 70 y que los expertos que participaron en este estudio provienen todos de la Escuela de Idiomas, podría caber la posibilidad de que el marco de referencia de los jueces sobre el nivel de conocimiento necesario para aprobar, pudiera haber influenciado la forma de evaluar cada uno de los reactivos haciendo sus juicios más estrictos sobre lo que un estudiante con competencia mínima fuese capaz de contestar correctamente.

Este marco de referencia pudo influenciar el proceso de valoración, de tal forma que fuese difícil para los expertos tomar en cuenta la información adicional que se les proporcionó sobre el desempeño real de los examinados (710 estudiantes), cuyo punto de corte se situó en 58. Por otro lado fue difícil para los expertos considerar la información sobre la proyección del impacto que tendría en los estudiantes el punto de corte establecido por ellos. Estos factores pudieron evitar que el proceso interactivo de las sesiones con información adicional tuvieran los resultados esperados. Por tal motivo se obtuvieron valoraciones diferentes entre los jueces. No obstante, la valoración global fue muy homogénea tanto en el examen completo como por áreas. Esto confirma lo observado por Shepard (1995) en relación a que los jueces tienden a presentar diferencias en sus valoraciones, sin embargo la valoración global es consistente.

La consistencia en los resultados es una forma de confiabilidad ya que si las áreas presentan resultados similares a la prueba completa, esto demuestra que se tiene consistencia interna. En el caso de este estudio el índice de variación demostró que los juicios fueron homogéneos.

Con objeto de tener una mejor proyección de la fuerza de asociación de las valoraciones de los jueces se decidió utilizar el análisis de variabilidad expresado a través del coeficiente de variación (Jornet, 1998), el cual provee información adicional sobre la semejanza de las valoraciones sin ser muy estricto en cuanto al acuerdo total en el número que hayan utilizado los jueces en su estimación (Jornet y Suárez, 1989b). De acuerdo con el coeficiente de variación encontrado, los juicios se sitúan en un nivel aceptable de uniformidad lo cual apoya el proceso de jueceo realizado en este estudio.

En cuanto a la valoración del proceso por los jueces, en su mayoría expresaron tener suficiente confianza en el proceso realizado. Asimismo consideraron que el entrenamiento así como el tiempo utilizado en la implementación del mismo fue adecuado.

Una de las aportaciones del presente trabajo es la elaboración de una guía de instrucción para dirigir el proceso de entrenamiento como parte de la implementación del procedimiento. Esta guía podrá ser útil como referencia para futuros estudios de punto de corte que involucren el jueceo, ya que en la

búsqueda de información para el proceso de entrenamiento no se pudo encontrar un instrumento similar que guiara dicho proceso.

Otra de las aportaciones importantes de este trabajo es contar con un punto de corte con referencia a un criterio, establecido de acuerdo a un método probado y adecuado para pruebas con referencia a criterio.

Una de las limitaciones del presente trabajo es no haber contado con un número mayor de jueces que permitiera desarrollar otro tipo de estrategias en el proceso de valoración, tales como formar subgrupos de 3 o 4 jueces cuya valoración pudiera ser contrastada con un grupo mayor de jueces, para ofrecer así una mejor comparación de la consistencia de las estimaciones. Otra de las limitaciones fue la dificultad de incluir otro tipo de jueces además de los profesores por ejemplo, administradores, padres de familia, miembros del sector productivo, etc. Esto podría generar mayor confianza en el punto de corte establecido (Jaeger, 1989). Sin embargo, puede suceder que el proceso de valoración se vuelva demasiado complejo, por la falta de conocimiento de estos grupos, ya que algunos de éstos son ajenos al medio educativo.

Las implicaciones del presente trabajo para la Universidad se pueden situar en el contexto social y político. En el aspecto social, aprobar el EXEDII es uno de los requisitos de egreso a nivel licenciatura que comprueba que los estudiantes tienen el dominio del idioma inglés a nivel intermedio. El cumplir o no

cumplir con el requisito para aprobar un examen de egreso de un idioma extranjero, en este caso el EXEDII, tiene un impacto directo en la vida del estudiante ya que de no aprobarlo se verá limitado en la posibilidad de obtener su título, el cual tiene un valor tanto social como económico para el estudiante.

En el aspecto político es indudable que para la Universidad es importante contar con un número significativo de egresados que le permitan tener un mayor nivel de aprobación ante los organismos evaluadores de las Instituciones de Educación Superior (IES), lo cual redundará en beneficios económicos para la propia Universidad, ya que el índice de titulados es un indicador de calidad y excelencia e incide de alguna manera en la obtención de recursos adicionales para la Universidad.

Otro aspecto que se debe analizar, es el factor de equidad y justicia del punto de corte de las diversas opciones que la Universidad ofrece al estudiante para cubrir el requisito del inglés como lengua extranjera. Si se compara el punto de corte que tiene el programa de Inglés de la Escuela de idiomas (situado en 70) y el punto de corte actual del Examen de Egreso de Inglés (situado en 55), hay una aparente falta de equidad entre el nivel de conocimientos que se les exige a los examinados que optan por cursar el programa de inglés y aquellos que presentan el EXEDII. De acuerdo a los resultados de este estudio, el punto de corte obtenido refleja más el criterio de lo que se considera conocimiento mínimo indispensable en el dominio del idioma inglés. En este sentido el punto

de corte actual del EXEDII pareciera exigir menos conocimientos para cumplir con el requisito del dominio del idioma inglés a nivel intermedio.

En el caso particular del EXEDII de la UABC, la toma de decisión en relación al punto de corte tiene implicaciones importantes. Si se aplica un punto de corte demasiado alto el costo social se incrementa, ya que menos estudiantes podrán cubrir el requisito para obtener su diploma y por otro lado, si se aplica un punto de corte muy bajo el nivel de conocimientos de una lengua extranjera será insuficiente para el desempeño apropiado del egresado en un ambiente profesional en relación con el inglés. Es por ello que es necesario tener en cuenta estos aspectos al efectuar una toma de decisión. Por otra parte, la literatura indica que una toma de decisión no se debe basar únicamente en el análisis psicométrico de los datos (Popham, 1990).

Recomendaciones

Una recomendación que cabe mencionar es la siguiente: si se toma en consideración el resultado del análisis de desempeño real, (58), el resultado del análisis de grupos de contraste (61) y el resultado del presente estudio donde se utilizó el jueceo (71), sería posible establecer un punto de corte que considerara los aspectos político económicos y sociales mencionados anteriormente, utilizando para ello la media de estos tres grupos. De esta manera sería posible establecer un punto de corte de 63. Con esta medida se evitaría establecer un

punto de corte demasiado alto que pudiera perjudicar a un gran número de estudiantes, así como evitaría que se determinara un punto de corte demasiado bajo que pudiera perjudicar la calidad del nivel de dominio de un segundo idioma de los egresados. Popham (1990) manifiesta que el punto de corte no debe considerarse definitivo sino que debe revisarse periódicamente. Para futuras investigaciones se hacen las siguientes recomendaciones

1. Llevar a cabo otro estudio de determinación de punto de corte para contrastar su resultado con el obtenido por el presente estudio.

2. En caso de realizar una nueva investigación con relación al punto de corte, tratar de incorporar a varios grupos involucrados: sociedad, directivos y expertos no obstante lo complejo del proceso de valoración.

3. Que la institución interesada realice estudios para determinar con claridad el nivel que requiere un estudiante universitario para desenvolverse adecuadamente en el ámbito profesional.

4. Asegurar recursos financieros que permitan contar con un grupo mayor de jueces y que permitan también una dedicación completa de los expertos al proceso de valoración en un tiempo determinado, lo cual redundará en beneficio de la investigación.

5. Investigar si el punto de corte debe ser global o bien relativo en cuanto a las diversas áreas que contempla el examen. Esto tendrá que ver directamente con el tipo de competencias que la Universidad requiere para sus egresados.

6. Llevar a cabo estudios específicos que permitan seguir acumulando evidencia de validez tanto interna como externa para el punto de corte del EXEDII.

7. En cuanto a la organización del EXEDII se recomienda iniciar el examen ya sea con gramática o lectura, de tal manera que el estudiante vaya adecuando con mayor facilidad su contexto de percepción del idioma, antes de contestar la sección auditiva.

REFERENCIAS BIBLIOGRAFICAS

Angoff, W. H. (1971). Scales, norms and equivalent scores. En R. L. Thorndike (Ed.), *Educational Measurement* (2nd. ed., pp. 508-600). Washington, D.C.: American Council of Education.

Asociación Nacional de Universidades e Instituciones de Educación Superior ANUIES, (1997). La evaluación y acreditación de la Educación Superior en México. *Revista de la Educación Superior, XXVI*(1), 57-91.

Backhoff, E. y Tirado, F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos. *Revista de la Educación Superior,* 83, 95-117.

Backhoff, E., Ibarra, M.A. y Rosas, M. (1995). Sistema Computarizado de Exámenes (SICODEX). *Revista Mexicana de Psicología, 12*(1), 55-62.

Backhoff, E., Larrazolo N., y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa,* 2(1), 2-10.

Berk, R. A. (1986). A consumer's Guide to Setting Performance Standars on Criterion-Referenced Tests. *Review of Educational Research, 56*(1), 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Educational Measurement, 21*, 147-152.

Camilli, G., Cizek y Lugg, C. (2001). Psychometric Theory and the Validation of Performance Standards: History and future perspectives. En G. Cizek

(Ed.), *Setting Performance Standards: Concepts, Methods and perspectives.* (pp. 445-475).Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cizek, G. J. (2001). Conjectures of the Rise and Call of Standard Setting: An Introduction to Context and Practice. En G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 3-51). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Contreras, L. A. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California.* Tesis de Maestría en Ciencias Educativas. Universidad Autónoma de Baja California, México.

Crocker, L. y Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York, N.Y. Holt, Rinehart y Winston.

Cross, L.H., Impara, J.C., Frary, R.B., & Jaeger, R.M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement,* 21, 113-130.

Dillon, G. F. (1996). The expectations of Standard setting judges. *CLEAR Exam Review*, Summer 22-26.

Donnoe, W. E., Amato, R.P. (1997). *Supportive Data & Guidelines for Using the Angoff, Ebel and Nedelsy Cutoff Score Methods.* Presentado en the International Personnel Management Association Assessment Council (IPMACC) Conference, Newport Beach, Ca.

Ebel, R.L. (1972). *Essentials of educational measurement (2nd ed.)* Englewood Cliffs, NJ: Prentice Hall.

Fermín, M. (1971). *La evaluación, los exámenes y las calificaciones.* Buenos Aires, Argentina: Kapelusz, S.A.

Gallagher, J. J. (1980). Setting educational standards for minimum competency: A case study. En R. M. J. y. C. K. Tittle (Ed.), *Minimum Competency achievement testing: Motives, models measures, and consequences* (pp. 239-257). Berkeley, Ca: McCutchan.

Glass, G.V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.

Geisinger, K. (1992) Testing Limited English Proficient Students for Minimum Competency and High School Graduation. *2nd. National Symposium,* Washington, D.C. Consultado el 25 de Febrero del 2002, en http://www.ncbe.gwu.edu/ncbepubs/symposia/second/testing.htm

Hambleton, R.K. (1978) On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 15, 277-290.

Hambleton, R.K. (1980a). Test score validity and standard-setting methods. En R.A. Berk (Ed), *Criterion-referenced measurement: The state of the art.* (pp. 80-123) Baltimore, MD: Johns Hopkins University Press.

Hambleton, R.K., y Eignor, D.R. (1980b) Competency test development, validation, and standard setting. En R.M. Jaeger & C.K. Tittle (eds.) *Minimum achievement testing: Motives, models, measures and consequences (pp. 367-396).* Berkeley, CA: McCutchan.

Hambleton, R.K. (1988) Criterion-referenced Measurement. En: Keeves, J.P. (Ed.) *Educational Research Methodology and Measurement. An International Handbook.* Pergamon Press, E.U.A. (pp 277-282).

Hambleton, R.K., Jaeger, R.M., Plake, B.S., y Mills, C.N. (1998). Setting performance standards on achievement tests: Meeting the requirements of title I. In. L.Hansche (Ed.), Handbook for the development of performance standards (pp.87-114). Washington, D.C.: U.S Department of Educations and the Council of Chief State School Officers.

Hambleton, R.K.,(2001). Setting Performance Standards on Educational Assessment and Criteria for Evaluating the Process. En: Cizek (Ed). *Setting Performance Standards: Concepts, Methods, and Perspectives.* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Publishers.

Henrysson, S. (1971). Gathering, Analysing, and Using Data on Test Items. En R.L. Thorndike (Ed.), *Educational Measurement.* Washington, DC: American Council on Education.

Himmel, E. "Sistemas Nacionales de Evaluación Educativa (Recopilación)" Documento Web Sep. 2000. Consultado en Dic. 2004, de http://www.ifie.edu.mx/sistemas_nacionales_de_evaluacion_educativa.htm

Hopkins, K. y Glass, G. (1997). *Estadística Básica para las Ciencias Sociales y del Comportamiento.* (3ª. Ed.),México. Prentice Hall Hispanoamericana. Simon and Schuster Company.

Hofstee, W.K.B. (1983), The case for compromise in educational selection and grading. En: S.B. Anderson & J.S. Helmick (Eds), *On educational Testing* (pp. 109-127). San Francisco: Jossey-Bass.

Hurtz, G.M.,y Hertz, N.R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method: A generalizability theory study. *Educational and Psychological Measurement,* 59, 885-897.

Impara, J. y Plake, B. (2000). A comparison of Cut Scores using Multiple Standard Setting Methods. Documento presentado en *The Large Scale Assessment Conference.* Snowbird, UT.

Jaeger R. y Mill, C. (2001). An integrated Judgment Procedure for Setting Standard on Complex, Large-Scale Assessments. En Cizek *Setting Performance Standards. Concepts, Methods and Perspectives.* (pp. 313-338). Mahwah, NJ: Lawrence Erlbaum Associates.

Jaeger, R.M. (1978) *A proposal for setting a standard on the North Carolina high school proficiency test.* Documento presentado en the spring meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M. (1981). *An Interactive Structured Judgment Process for Establishing Standards on Competency Tests: Theory and Application.* Documento presentado en the Annual Meeting of the American Educational Research Association and the National Council on Measurement, Los Angeles, CA.

Jaeger R. M. (1989) Certification of Student Competence, en R.Linn (Ed.) *Educational Measurement* (3a.ed), (pp. 485-514) Washington, DC: American Council on Education.

Jaeger R. M. (1991) Selection of Judges for standard setting. *Educational Measurement: Issues and Practice* 10, 3-6.

Jornet, J. y. S., J. (1989). Revisión de Modelos y Métodos en la determinación de estándares y en el establecimiento de un punto de corte en la Evaluación referida al Criterio. *Bordón, 41*(2), 277-301.

Kane, M (2001) So Much Remains the Same: Conception and Status of Validation in Setting Standards. En: Cizek (Ed). *Setting Performance Standards: Concepts, Methods, and Perspectives.* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum Publishers.

Klein, T.W. (1990). Characteristics which Differentiate Criterion-Referenced from Norm-Referenced Tests. *Nevada Department of Education* at Carson City, Nevada. (ERIC No. ED 324 327)

Larrazolo N. y Velasco V.A., (2000). Examen de egreso del idioma Inglés EXEDII: Indices de dificultad y discriminación. *Memorias del Cuarto foro de Evaluación Educativa* (pp. 111-115). México.

Larrazolo, N. (2002) *Desarrollo y validación de un examen de inglés para egresar del Nivel Superior.* Tesis Doctoral. Sin publicar. Universidad Iberoamericana de Tijuana, México.

Livingston, S. A. (1972): Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9*, 13-26.

Livingston , S.A., y Zieky, M.J. (1982) *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Martínez, R., Backhoff, E., Castañeda S., De la Orden, A., Schmelkes, Silvia, Solano-Flores, G., Tristán, A., Vidal, R. (2000). *Estándares de Calidad*

para Instrumentos de Evaluación Educativa. México, D.F.: Centro Nacional de Evaluación para la Educación Superior, A.C., (CENEVAL).

Martínez, R. F. (2001). Evaluación Educativa y Pruebas Estandarizadas. Elementos para Enriquecer el Debate. *Revista de la Educación Superior, XXX*(4), 71-85.

Meskaukas, J. A. (1976). Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard-Setting. *Review of Educational Research, 46*(1), 133-158.

Nassif, P. (1979). *Standard Setting for criterion-referenced teacher licensing test.* Documento presenteado en the annual meeting of the National Council on Measurement in Education, Toronto.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement,* 14, 3-19.

Nitko, A.J. (1994). *A model for Development Curriculum-Driven Criterion-Referenced and Norm-Referenced Examination for Certification and Selection of Students.* Ponencia presentada en la Conferencia Internacional de Educación, Evaluación y Medición de la Asociación para el Estudio de la Evaluación Educativa en Sudáfrica (ASEESA.

Noonan, B. (1996). Setting Standards in Education: Some principles and Practices. Ponencia presentada en Saskatchewan Standards Symposium. Compilado and editado por Loraine Thompson. Information Testing Limited. SSTA Research Center Report # 96-2. consultado el 17 de Mayo del 2002 en www.ssta.sk.ca/research/evaluation_and_reporting/96-02.htm

Organista, J. (1998) *Desarrollo y Validación de un Sistema Computarizado para Administrar Tareas, Exámenes y Asesorías via Internet.* Tesis. Instituto de Investigación y Desarrollo Educativo, de la Universidad Autónoma de Baja California, Ensenada, B.C.

Payne, D.A. (1992). *Measuring and Evaluating Educational Outcomes.* MacMillan.

Popham, W.J. (1978). As always, provocative. *Journal of Educational Measurement,* 15, 297-300.

Popham, W. J. (1981). *Modern Educational Measurement.* Englewood Cliffs. NJ: Prentice Hall.

Popham, W. J. y. Yelow., E.S. (1984). *Standard-Setting options for teacher competency tests.* Paper presented at the Annual meeting of the American Educational Research Association, New Orleans.

Popham, W. J. (1990). *Modern Educational Measurement: A Practitioner's Perspective* (2a ed.). Needham Heights, MA: Allyn and Bacon.

Raymond y Reid (2001). Who made thee a judge? Selecting and Training Participants for Standard Setting. En: Cizek (Ed). *Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, NJ: Lawrence Erlbaum Publishers.

Ravitch, D. (1995) National Standards in American Education. A Citizen's Guide Disponible en .(http://www.campus-oei.org/calidad/ravitch.pdf.

Saunders, J. C. y. M., L.L. (1984). *Accuracy and Consistency of Expert Judges in Setting Passing Scores on Criterion-Referenced Tests: The South*

Carolina Experience. Ponencia presentada en the Annual Meeting of the American Educational Research Association, New Orleans.

Shepard, L.A. (1979) Setting Standards. En M. A. Bunda y J.R. Sanders (Eds.), *Practices and problems in competency-based measurement* (p. 72-88). Washington, DC.

Shepard, L. (1995). Implications for Standard Setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels. En the *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessment,* Vol. II (pp 143-160).

Skakun, E. N., y Kling, S. (1980). Comparability of methods for setting standards. *Journal of Educational Measurement, 17*, 229-235.

Universidad Autónoma de Baja California. (1995) *Reglamentos Universitarios* Dirección General de Servicios Escolares. Mexicali: Autor.

Woolfolk, A. (1996). *Psicología Educativa* 6a.Ed. Díaz, J. Traductor. Prentice Hall Hispanoamericana, S.A.

Zieky, M., (2001). So much Remains the Same: Conception and Status of Validation in Setting Standards En G. Cizek (Ed). *Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, NJ: Lawrence Erlbaum Publishers.

APENDICE A

Tabla de clasificación de Métodos

TABLA DESCRIPTIVA Y COMPARATIVA DE METODOS PARA ESTABLECER ESTANDARES

Evaluación de los métodos de JUECEO para establecer estándares (Listados en orden cronológico) Adaptado de Berk, R., (1986). A Consumer's guide to Setting Performance Standards on Criterion-Referenced Tests

Método y fuente	*Decisión para establecer estándares*	*Comentarios*
Nedelsky *(Nedelsky, 1954)*	*4*	Ventajas: *Existe un registro con los exámenes certificados en el área de salud.* Desventajas: *La suposición de que los examinados que no conocen la respuesta tenderán a adivinar no es muy firme. Permite únicamente un set de valores de probabilidad basados en el número de opciones de respuesta. Se puede forzar a los jueces a asignar un valor .5 a muchos reactivos. Prácticamente el método requiere considerable entrenamiento de los jueces y puede usarse solamente con profesores con experiencia..Poggio (1984) encontró que los jueces tienden a confundirse, reporta falta de confianza en el juicio y no son muy cuidadosos en el análisis de cada reactivo. El método tiende a producir un estándar muy por debajo de la mayoría de los otros métodos.*
Angoff *(Angoff, 1971)*	*2*	Ventajas: *Fácil de implementar , entender y calcular.* Desventajas: *Tiende a obtenerse una variación considerable entre las probabilidades de los reactivos proporcionados por cada juez. Muchos jueces tienen dificultad para determinar quienes tienen el conocimiento mínimo aceptable.. Algunos reactivos pueden ser más difíciles que otros de valorar adecuadamente.*
Método modificado de Angoff de opción múltiple (ETS, 1976)	*2*	Ventajas: *Fácil de implementar, entender y calcular.* Desventajas: *Proporcionar a los jueces una probabilidad puede desviar las respuestas (Livingston & Zieky, 1982, p. 25) Esta selección puede producir una distribución asimétrica de valores de probabilidad. Sin embargo, se puede usar una distribución simétrica que cubra el rango de 0 a 1. (Jaeger, en prensa).*
Angoff 2 opciones *(Nassif, 1978)*	*1*	Ventajas: *Fácil de implementar, entender y calcular. Requiere sólo una respuesta Si-No por cada juez.* Desventajas: *El formato SI-NO limita las probabilidades de 0% a 100%. Un formato de probabilidades continuas es más apropiado para la mayoría de los reactivos.* *El método para estimar el acuerdo entre jueces no se especifica. Aún cuando el punto de corte se ajustará en las mediciones de error, no se dá el método para estimar estos errores.*
Dificultad-Relevancia de Ebel (Ebel, 1979)	*3*	Ventajas: *Fácil de calcular, fácil de entender por la mayoría de los jueces.* Desventajas: *Los jueces parecen no tener bien definidas las dos dimensiones y el porcentaje por celda lleva al debate sobre la posición*

Método y fuente	*Decisión para establecer estándares*	*Comentarios*
		"cuestionable". La posición también provoca preocupación en los jueces sobre el método. El procedimiento de valoración de los reactivos consume mucho tiempo y el proceso se puede ver afectado por la fatiga y el aburrimiento. No hay control sobre cuáles reactivos fáciles realmente contesta el examinado, las importantes o no importantes. Los jueces tienden sistemáticamente a simplificar la tarea del examinado e ignorar algunas de las discriminaciones que el examinado tiene que hacer. Se dan estándares diferentes dependiendo de si es calculado por el juez o si se basa en los valores del grupo de celdas.
El método de opción múltiple de Angoff, modificado y ajustado (panel de estimación del conocimiento) *Bernkopf, Curry & Bashaw, 1979)*	*4*	Ventajas: *Fácil de implementar y entender. El estándar se ajusta por la clasificación y los errores de predicción.* Desventajas: *Las concernientes al método de opción múltiple de Angoff, excepto que la variación de los juicios de los diferentes jueces se considera al ajustar el estándar final.*
Dificultad-Taxonomia de Ebel (Skakun & Kling, 1980	*3*	Ventajas: *Similar al método de Ebel. La clasificación del nivel de dificultad se determina con la información sobre el desempeño. La clasificación previa en las dos dimensiones facilita la labor de los jueces.* Desventajas: *No considera la dimensión de relevancia e importancia. Las dimensiones pueden ser difíciles de definir para ciertas áreas de contenido. La clasificación puede ser ambigua*
Taxonomia de Relevancia de Ebel *(Skakun & Kling, 1980)*	*3*	Ventajas: *Similar al método de Ebel. La clasificación previa den ambas dimensiones simplifica la labor de los expertos. Se omite la dimensión "cuestionable" en relación con la relevancia.* Desventajas: *No considera los niveles de dificultad del reactivo. Las dimensiones de la Taxonomía pueden ser difíciles de definir para ciertos contenidos. La clasificación puede ser ambigua.*
Compromiso I (Absoluto-relativo) *(Berk, 1982, 1984)*	*4*	Ventajas: *Considera sistemáticamente los juicios sobre el desempeño estimado y el desempeño real. La medida en la cual los jueces acuerden sobre* k o v, *identifica si el grupo se orienta hacia el examen o hacia el examinado.* Desventajas: *Se requiere que los jueces consideren dos estimaciones las cuales pueden cambiar a medida que los jueces tengan oportunidad de practicar el hacer estimaciones. El método es más complejo de realizar por los datos estadísticos y los cálculos que se requieren. La interpretación del estándar puede ser difícil de comprender para los que no son especialistas.*
Especificación de reactivos *(Mills y Barr, 1983)*	*2*	Ventajas: *Fácil de entender y calcular.* *Enlaza el estándar a las áreas de dominio y al universo de reactivos que miden estas áreas. Requiere menos tiempo que el valorar cada reactivo. El Punto de corte se refiere al dominio no al formato del examen.* Desventajas: *El consenso de los jueces se usa para determinar el estándar . Aún cuando la relación entre la especificación y el reactivo es directa, la valoración de los reactivos puede no ser la misma que las especificaciones.*

Método y fuente	*Decisión para establecer estándares*	*Comentarios*
Estimación del nivel de dificultad (Cangelosi, 1984)	***3***	**Ventajas:** ***Fácil de entender y calcular. Enlaza el estándar con las especificaciones de las áreas de dominio .*** **Desventajas:** ***El consenso de los jueces por objetivo se usa para establecer el estándar. Aún cuando la relación entre los reactivos y sus objetivos es directa, debe asumirse que la valoración de los reactivos no necesariamente puede ser la misma que la de los objetivos.***
Combinación Angoff-Nedeslky ***(Reid, 1984)***	***5***	**Ventajas:** ***Las mismas que el método de Angoff*** **Desventajas:** ***Igual que las de Angoff y Nedelsky, más una posible combinación de diferentes fuentes de error cuando los dos estándares se promedian.***

Decisión del estándar como sigue: 1= por consenso de los grupos de jueces; 2= porcentaje promedio entre jueces; 3=porcentaje mínimo de la media de las muestras de los jueces; 4= punto de intersección entre los jueces y la evidencia empírica; 5= valor basado en juicios y evidencia empírica.

Evaluación de los métodos de** JUECEO-EMPIRICO **para establecer estándares (Listados en orden cronológico) **Adaptado de Berk, R., (1986). A Consumer's guide to Setting Performance Standards on Criterion-Referenced Tests**

Método y fuente	*Determinación de estándar*	*Comentarios*
Método interactivo de dos opciones de Angoff ***(Jaeger, 1978, 1982; Cross, Impara, Frary, & Jaeger, 1984)***	***3***	**Ventajas:** ***Le da a los jueces la oportunidad de reevaluar sus estimaciones basados en los tres tipos de información. El juicio se obtiene de diversas muestras de jueces que son representante de la población interesada. El proceso interactivo tiende a reducir la variabilidad en la distribución de los estándares recomendados y consecuentemente incrementa la confiabilidad del estándar (Jaeger & Busch, 1984).*** **Desventajas:** ***El formato SI-NO limita la probabilidad de los reactivos de 0% a 100%. Una continuidad de probabilidades es mas apropiada para la mayoría de los reactivos. Requiere que los jueces se reúnan en 3 ocasiones diferentes. Tiende a consumir mucho tiempo y es más complejo que otros métodos de jueceo. La interpretación del estándar puede ser difícil de entender para las personas no especializadas.***
Compromiso II (Absoluto-Relativo) (Hofstee, 1977, 1983)	***4***	**Ventajas:** ***Las mismas que el método anterior. Pero considera también*** K_{max} ***y reemplaza*** v ***por*** f_{max} ***y*** $f_{min.}$ ... **Desventajas:** ***Las mismas que el método anterior, excepto que en éste se requiere que los jueces proporcionen 4 estimaciones.***
Juicio informado ***Popham & Yalow, 1984; Yalow & Popham, 1983)***	***1***	**Ventajas:** ***Fácil de implementar, entender y calcular. Incorpora la información sobre el desempeño y la preferencia de diversos grupos. La probabilidad de cada reactivo en el formato de SI-No (véase el método interactivo de dos opciones de Angoff). Además considera los totales del juicio de todos los especialistas.*** **Desventajas:** ***La recopilación de datos fue consumir mucho tiempo y ser costosa. El análisis de los jueces sobre la información es poco sistemática. Es esencial la elaboración de guías para la utilización de los diferentes tipos de información. La discusión en grupo de opiniones individuales puede no ser recomendable por su sentido normativo. (Fitzpatrick, 1984).***
Grupos contrastantes y compuestos de Angoff ***(Sheppard, 1983, 1984)***	***5***	**Ventajas:** ***Se usan cuatro tipos de juicio y de información empírica para evitar imprecisiones en la determinación del estándar. Conserva las ventajas del método de Angoff y del método de grupos contrastantes.*** **Desventajas:** ***Los mismos de los dos métodos mencionados mas las dificultades para reconciliar las diferencias entre estándares y para ajustar el estándar basado en puntajes aceptable de pase y poca confiabilidad. Aún cuando la estrategia de triangulación tenga sus ventajas, se recomiendan guías específicas a fin de que se pueda llegar sistemáticamente a un estándar.***

Método y fuente	*Determinación de estándar*	*Comentarios*
Angoff Interactivo *(Saunders & Mappus, 1984)*	*1*	Ventajas: *Mismas que el método anterior, excepto que en este se emplea sólo una muestra de los jueces.* Desventajas: *Mismas que el método anterior, excepto por la limitación del uso del formato dicotómico (Si-No).*

Decisión del estándar como sigue: 1= por consenso de los grupos de jueces; 2= porcentaje promedio entre jueces; 3=porcentaje mínimo de la media de las muestras de los jueces; 4= punto de intersección entre los jueces y la evidencia empírica; 5= valor basado en juicios y evidencia empírica.

EVALUACION DE LOS METODOS *EMPIRICO-JUECEO* ***para establecer estándares (Listados en orden conológico)*** **Adaptado de Berk, R., (1986). A Consumer's guide to Setting Performance Standards on Criterion-Referenced Tests**

Método y fuente	*Determinación de estándar*	*Comentarios*
Consecuencias educativas (Block, 1972)	*1*	Ventajas: *Fácil de implementar, entender y calcular.* Desventajas: *El criterio "aprendizaje futuro" debe especificarse claramente. La suposición del modelo no parece estar sustentada por la información. (Ver Glass, 1978; Shepard, 1980°); la relación entre las variables sugiere que no hay un punto de optimización del criterio para identificar el punto de corte (Sheppard 1984, pp. 183-184).*
Criterio con referencia a la norma (García Quintana & Mappas 1980)	*1*	Ventajas: *Fácil de implementar, de entender y calcular. La información criterial está disponible. Se considera una pérdida de función adicional a la relación directa entre las dos pruebas.* Desventajas: *el uso de un referente normativo de desempeño con un punto de corte seleccionado de forma arbitraria es técnica y conceptualmente inapropiado. ¿ Por qué no simplemente utilizar una decisión con respecto al dominio de la prueba en lugar de adecuar el dominio a la información real de desempeño?. La selección de un criterio de esta manera es ilógico y el punto de corte puede ser invalidado y poco confiable. Fijar puntos de corte bajo este contexto es inapropiado.*
Grupos límite (Livingston & Zieky, 1982)	*2*	Ventajas: *Fácil de implementar, entender y calcular.* *Requiere seleccionar una sola muestra de estudiantes* Desventajas: *Muchos maestros tienen dificultad en definir a los estudiantes "que tengan competencia mínima" (Poggio, 1984). No es claro que los jueces puedan limitar sus juicios a aquellos elementos del comportamiento que se pretende medir; pueden ser influenciados por otros factores cognoscitivos y no cognoscitivos (Jaeger, en prensa). Los jueces pueden colocar estudiantes en la categoría de competencia mínima cuando carecen de suficiente información para realizar una apropiada clasificación (Jaeger, en prensa). Una muestra adecuada puede ser difícil de obtener. El punto de corte no se obtiene al término de la prueba. Los legos tienden a confundirse acerca del método y a dudar de su legitimidad (Poggio, 1984).*
Grupos contrastantes *(Livingston & Zieky, 1982)*	*1*	Ventajas: *Las mismas del método de grupos criteriales.* Desventajas: *Las mismas del método de grupos criteriales. La poca confiabilidad de éxito en la determinación del punto de corte es virtualmente imposible de soslayar. (Ver Berk, 1984, pp 203-204).*

Grupos criteriales ***(Berk, 1976, 1984. cap. 6)***	***1***	**Ventajas:** ***Fácil de implementar, entender y calcular. Es una aplicación del método tradicional de "grupos conocidos" de (Hattie & Cooksey, 1984). Que tiene características estadísticas y validez de instrucción. Provee evidencia de validez de decisiones y de clasificación de probabilidades.*** **Desventajas:** ***No identifica el "Punto de corte real" (Shepard, 1984 p. 179; Van der Linden, 1984); A lo mucho localiza la región del desempeño donde los que tienen el dominio y los que no lo tienen se entrecruzan. Definir el dominio o no dominio en términos de grupos intactos es problemático (Hambleton, Swaminathan, Aalgina & Coulson, 1978; Shepard, 1980[3]) . Es difícil estimar la base real y obtener un tamaño adecuado de la muestra. El punto de corte no está disponible después que se termina la prueba. Los legos tienden a confundirse acerca del método y a dudar de su legitimidad (Poggio, 1984). Afirma que "Los maestros pueden decir quien es competente"***

1 = Puntuación que mejor discrimina; 2 = Media del grupo

APENDICE B

GUIA DE INSTRUCCIÓN

Universidad Autónoma de Baja California
Instituto de Investigación y Desarrollo Educativo

I
I
D
E

GUIA DE INSTRUCCIÓN

PARA EL PROCESO DE VALORACIÓN

DE LOS REACTIVOS DEL EXAMEN DE EGRESO

DEL IDIOMA INGLÉS (EXEDII)

DE LA

UNIVERSIDAD AUTONOMA DE BAJA CALIFORNIA

Escuela de Idiomas

Universidad Autónoma de Baja California

Preparada por:
Ma. del Carmen Márquez Palazuelos

Con la asesoría de
Dra. Norma Larrazolo Reyna

Como parte del Proyecto de Investigación "Aplicación y validación de un método para determinar el punto de corte para el Examen de Egreso del Idioma Inglés (EXEDII)

Apoyado por:

Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de B.C.

y

Facultad de Idiomas de la Universidad Autónoma de B.C.

Esta guía de instrucción fue elaborada para facilitar el entrenamiento del grupo de expertos (jueces) seleccionados para participar en los estudios que se requieren para determinar el punto de corte del Examen de Egreso del Idioma Inglés (EXEDII).

La guía constituye un recurso para apoyar a los jueces en dos etapas de este estudio: a) la definición de lo que es un examinado con competencia mínima del inglés y b) la valoración de cada una de las preguntas del EXEDII.

El diseño y la elaboración de los instrumentos, así como los datos recabados forman parte del proyecto de investigación "Aplicación y validación de un método para determinar el punto de corte para el Examen de Egreso del Idioma Inglés (EXEDII). El presente estudio se llevó a cabo con el patrocinio del Instituto de Investigación y Desarrollo Educativo (IIDE) de la Universidad Autónoma de Baja California y con la colaboración de la Escuela de Idiomas de la propia universidad.

INFORMACIÓN PARA GUIA DE INSTRUCCIÓN

El Exámen de Egreso del Idioma Inglés (EXEDII) es un examen criterial computarizado. Este examen fue diseñado por el Instituto de Investigación y Desarrollo Educativo con la colaboración de la Facultad de Idiomas de la UABC, como una opción para cumplir con el requisito establecido por la Universidad a sus estudiantes, de certificar el dominio de una lengua extranjera a nivel intermedio, para su titulación.

Un examen criterial es aquel que tiene como propósito medir el desempeño de un examinado en relación con las áreas de dominio previamente especificadas, donde dichas áreas están sujetas a un grupo de competencias perfectamente definidas.

Información sobre la Estructura del Examen

El examen evalúa tres áreas, cada una dividida en subáreas con distintos niveles de dificultad. Los objetivos tienen una secuencia de dificultad que va de lo más sencillo a lo más complejo.

El EXEDII consta de 100 preguntas y está dividido en tres áreas:

Lectura	34 preguntas
Gramática	34 preguntas
Comprensión auditiva	32 preguntas

(Ejemplo)

Áreas	Subáreas	Objetivos	Número de Preguntas
	Identificación y comprensión de	Palabras Enunciados Párrafos	6 5 6
Comprensión de Lectura	Razonar con	Palabras Enunciados Párrafos	7 5 5
	Identificación y comprensión de	Palabras Frases Diálogos	2 7 10
Comprensión Auditiva	Razonar con	Palabras Diálogos	5 8
Gramática	Manejo de verbos Adverbios, Pronombres y tiempos gramaticales	Presente Pasado Adverbios Pronombres Pasado/pasado progresivo	11 7 10 4 2

PROPÓSITO DEL PRESENTE ESTUDIO.

El objetivo del presente estudio es aplicar un método para establecer el punto de corte para el EXEDII.

Berk (1986) da la siguiente definición del punto de corte: " Es el punto en la escala que separa la distribución de reactivos en dos categorías mutuamente excluyentes: una categoría que contiene los aciertos que determinan el dominio del estudiante y otra categoría que determina el no dominio de la competencia" . Así podemos decir que el punto de corte es el punto en la escala que decide el dominio o no dominio del examinado, en áreas de contenido específicas.

METODO SELECCIONADO.

Se utilizará para este estudio el método desarrollado por William H. Angoff (1971). Este método consiste en determinar el punto de corte a través de la valoración de las preguntas por un grupo de expertos, a esta actividad se le llama jueceo.

El jueceo consiste básicamente en que un grupo de especialistas revisen los reactivos del examen para estimar la probabilidad (una proporción de 0 a 1) de que un examinado pueda contestar correctamente la pregunta. A medida que el valor se acerque más a 1 la pregunta será más fácil para el examinado y si el valor se acerca más a 0 la pregunta será más difícil de contestar para el sustentante.

Para facilitar la estimación de la probabilidad, Angoff recomienda a los especialistas que consideren cuántos examinados de un grupo de 100 podrán contestar cada pregunta. Asimismo, para realizar esta labor los especialistas tendrán que considerar a un examinado con el nivel mínimo de conocimientos necesarios para pasar el examen, o sea, aquel estudiante que se ubique en el límite entre el nivel bajo y el nivel intermedio medio.

A continuación se proporciona información de los estándares sobre el nivel intermedio.

INFORMACIÓN PERTINENTE PARA EL ESTUDIO.

Nivel intermedio de acuerdo con los estándares de la ACTFL (American Council of Teachers of Foreign Languages, 1988)

COMPRENSIÓN DE LECTURA

	Descripción de habilidades
Intermedio bajo	El lector será capaz de entender ideas principales y algunos hechos en textos simples e interconectados sobre necesidades personales y sociales básicas. Lingüísticamente sencillos y con una estructura interna clara (p.ej. secuencia cronológica). Proporcionan información sobre la que el lector tiene que hacer mínimas suposiciones y que son de su interés o conocimiento. Ej. Anuncios para audiencias amplias, instrucciones para el público en general.
Intermedio medio	**El lector será capaz de leer y comprender consistentemente textos simples e interconectados acerca de necesidades básicas y sociales. Los textos son lingüísticamente sencillos con una estructura interna clara. Proporcionan información sobre la que el lector tiene que hacer mínimas suposiciones y que son de su interés o conocimiento. Como por ejemplo descripciones de personas, lugares y textos escritos para una amplia audiencia.**
Intermedio alto	El lector será capaz de leer y comprender consistentemente textos medianamente complejos e interconectados; los temas se refieren a necesidades básicas, hechos sociales de su interés; el lector capta algunas ideas e información específica en textos de complejidad lingüística mediana que contienen descripciones o narraciones ; comprende parcialmente textos con estructuras complejas, tiene dificultad para entender factores cohesivos del discurso como igualar pronombres con sus referentes; necesita leer varias veces para entender.

COMPRENSIÓN AUDITIVA

	Descripción de habilidades
Nivel intermedio bajo	El lector será capaz de entender discursos que contengan frases largas que consistan de recombinaciones de elementos aprendidos o familiares en un número limitado de áreas de contenido, particularmente si son fuertemente apoyadas por el contexto situacional. El contenido se refiere a necesidades personales básicas, convenciones sociales y tareas rutinarias como obtener alimentos y recibir instrucciones o direcciones simples. Incluye conversaciones espontáneas cara a cara. La comprensión es frecuentemente dispareja e incompleta; puede necesitar repetición frecuente. Comprensión incompleta.
Nivel intermedio medio	**El lector será capaz de entender discursos que contengan frases largas, que consisten en combinaciones de discursos aprendidos, en una variedad de tópicos. El contenido continúa refiriéndose al contexto personal básico y sus necesidades, convenciones sociales y algunas tareas un poco más complejas relativas a habitación, transporte, compras y satisfacción de necesidades cotidianas. Algunas áreas adicionales de contenido pueden incluir interés y actividades personales, así como una mayor diversidad de instrucciones y direcciones. Incluye conversaciones cara a cara y telefónicas cortas y rutinarias; cierto tipo de discurso deliberado, como anuncios cortos e informes sencillos sobre el medio.**

	Comprensión incompleta.
Nivel intermedio alto	El lector será capaz de sostener la comprensión en fragmentos más largos de discursos, en una variedad de temas correspondientes a diferentes tiempos y lugares; sin embargo, la comprensión es todavía inconsistente debido a la dificultad para captar ideas principales y/o detalles. Por tanto, aunque los tópicos no difieren significativamente de los del oyente avanzado, la calidad de la comprensión es más pobre.

DEFINICIÓN DE UN EXAMINADO CON EL CONOCIMIENTO MÍNIMO NECESARIO PARA PASAR EL NIVEL INTERMEDIO DE INGLÉS.

De acuerdo con lo requerido por el método de Angoff, **para realizar la labor de valoración de las preguntas los especialistas tendrán que considerar a** un examinado con el nivel mínimo de conocimientos necesarios para pasar el examen, **o sea, aquel estudiante que se ubique en el límite entre el nivel intermedio bajo y el nivel intermedio medio.**

Para efecto de valorar las preguntas los expertos deberán considerar al examinado como

"Alguien que tiene el **conocimiento mínimo** necesario del idioma Inglés para contestar correctamente los reactivos del examen a nivel intermedio.

PROCESO DE DEFINICIÓN DE LA COMPETENCIA MINIMA DE UN EXAMINADO:

para realizar esta labor los especialistas tendrán que considerar a un examinado con el nivel mínimo de conocimientos necesarios para contestar correctamente las preguntas del examen, **o sea, aquel estudiante que se ubique en el límite entre el nivel intermedio bajo y el nivel intermedio medio.**

Tomando en consideración a un examinado con el conocimiento mínimo para contestar correctamente el examen, y los niveles de intermedio bajo e intermedio medio ¿Cuáles considera usted que son las habilidades que definen a ese estudiante en particular ?

Preguntas para reflexionar :

En comprensión de lectura ¿Cuáles factores podrían influir en la facilidad o dificultad en la comprensión de lectura para los examinados con conocimiento mínimo para contestar correctamente las preguntas del examen?.

Ayudaría por ejemplo que:

El texto contara con fotografías?

Los Pasajes cortos?
Las preguntas fueran sencillas y concretas?
La lectura utilizara vocabulario de alto nivel ?
Se diera una explicación indirecta de las palabras?
El texto tuviese términos difíciles de entender?
Los pasajes largos?
El texto tuviera preguntas de comprensión ?
Las preguntas fueran complejas?
Se tuviese que hacer inferencia de preguntas abstractas?
Se buscaran contenidos de interés para los examinados?
La forma en que está organizado el texto?
La idea principal fuera obvia?
El examinado tuviera conocimiento previo sobre el tema?

Etc.

En comprensión Auditiva ¿Cuáles factores podrían influir en la facilidad o dificultad en la comprensión para los examinados con conocimiento mínimo para contestar correctamente el examen?.

Dificultaría:
Que se utilizara vocabulario cotidiano?
El nivel del vocabulario utilizado?
Que se utilizaran términos no comunes?
Que hubiese necesidad de inferir información?
Que la información fuese directa o indirecta?

ENTRENAMIENTO EN LA VALORACIÓN DE PREGUNTAS

Para efecto de la valoración de los reactivos del examen. cada especialista registrarán en el formato qué proporción de los examinados con competencia mínima es capaz de contestar esa pregunta correctamente.

A continuación se presenta un ejemplo:

Reactivo	Especialista 1	Especialista 2	Especialista 3
1	.70	.65	.75
2	.50	.45	.55
3	.80	.75	.85
Total	**2.00**	**1.85**	**2.15**

La suma del promedio de las estimaciones de cada especialista dividido entre el número de especialistas representará el punto de corte. Ej.

$$2.00 + 1.85 + 2.15 / 3 = \boxed{2}$$

En este caso el promedio de los 3 jueces fue 2, es decir el 67%. Esto quiere decir que los estudiantes tendrán que tener una puntuación del 67% o más para pasar el examen.

PROCEDIMIENTO PARA VALORACIÓN DE LAS PREGUNTAS.

1. **Examine una a una las preguntas del examen.**
2. **Para evaluar las preguntas tenga en mente la definición de un estudiante con competencia mínima.**
3. **Estime la probabilidad de que ese estudiante pueda contestar correctamente la pregunta.**
4. **Registre este porcentaje en el formato elaborado para ese efecto (Ver anexo)**

EJEMPLOS PARA PRACTICAR LA VALORACIÓN DE PREGUNTAS

Ejemplos de reactivos en comprensión auditiva.

Reactivo 6.

Man: I haven't seen uncle Larry lately. I'm worried about him. He has been acting funny ever since he lost his business. It affected him a lot. He tells everybody that the business is going great even though it failed. He went to the bank to ask for a million dollars to continue his business, but they laughed in his face. The way he is acting now makes me fear he'll do something crazy.

Woman: Your are right. The other day I heard him say a lot of weird things that didn't make much sense.

What does funny mean in this conversation?

a) strangely
b) suspiciously
c) comically
d) amusingly

Reactivo 15.

Man: I saw Peter today and he looked a little strange, kind of nostalgic. Do you know why?

Woman: Well, he was studying biology at the library when he saw a picture of a lake like the one his father used to take him to when he was a boy. So, he went to the nearest pay phone to call his father, but he couldn't find him.

Man: Oh, so that's why he looked sad.

Woman: Yeah, I hope that doesn't distract him form his biology test.

Man: I hope so, too. But you know how sentimental he is. I'm sure he is still thinking about all that time he spent with is dad.

What sentence is the best summary of what happened?

a) peter remembered his childhood
b) Peter's father had died
c) Peter studied for a biology test
d) Peter didn't have money for a phone call.

Reactivo 29.

Woman: You said you were going to go to New York today, didn't you? What happened?

Man: It's a long story: Basically, I stayed up working on the New York project last night and getting up early in the morning was almost impossible. I arrived just in time to the airport only to find out that the flight was cancelled. So I had to go back in the office. There they told me that the New York Project was cancelled. Lucky, ha! I wasn't surprised when I heard in the news that the weather in New York was terrible, so I guess I am a pretty lucky guy.

Why does the man think he is lucky?

a) because the flight to New York was cancelled
b) because he overslept and lost his flight
c) because he has a secure job
d) because the weather was terrible in New York

Ejemplos de reactivos de comprensión en la lectura.

WISCONSIN'S DOOR COUNTY

The eastern coast of Door County, Wisconsin, known for its wildlife sanctuaries and state parks, is more sedate, its surf more dramatic, than what you find in the rest of the Door. Founded by fishermen in 1851, Bailey's Harbor is the largest town there, yet it feels like a small fishing village. It also boasts some of the peninsula's best eateries, such as the Sandpiper and the Common House restaurant. A few miles away is Ridge Sanctuary. It is a 1200-acre parcel set aside in 1937 as a haven for wildlife and native plants, and no one can hurt or disturb them. A nature

center there offers guided tours, and the chief naturalist leads visitors on early-morning bird walks.

Reactivo 1.

Which of the following options best expresses the main idea?

a) The eastern coast of Door County is known for its wildlife and parks.
b) Bailey's Harbor is the largest town on the eastern coast of Door County
c) The eastern coast of Door County boasts some of the peninsula's best eateries.
d) A nature center in Door County offers guided tours.

Reactivo 2.

Which of the following is true according to the text?

a) Ridge Sanctuary measures 1200 acres
b) Ridge Sanctuary was expanded to 1937 acres
c) Bailey's Harbor is inhabited by 1851 fishermen
d) Bailey's Harbor has 1200 parcels.

Reactivo 3.

In this text, "The eastern coast is more sedate and its surf more dramatic than what you find in the rest of the Door." Is:

a) an opinion of the author
b) a generally accepted fact
c) an opinion of someone other than the author
d) a verified fact

Ejemplos de reactivos de Gramática:

1. The earth _____ around the sun.

a) orbits
b) orbit
c) orbiting
d) orbited

2. Diego and Cristina compete in swimming. The _____every day.

a) practice
b) practices
c) practicing
d) practiced

3. _____Hermosillo the Capital of Sonora?

a) Is
b) Are
c) Do
d) Does

Registro para clasificación de preguntas

del EXEDII

	Pregunta	.0 ~ .10	.11 ~ .20	.21 ~ .30	.31 ~ .40	.41 ~ .50	.51 -.60	.61-.80	.81-90	.91 - 1.0
G	1									
R										
A	2									
M										
A	3									
T										
L	Texto 1									
	React. 1									
E										
	React. 2									
C										
T	React. 3									
U										
R										
A										
Subtotal										
A										
U	6									
D										
I										
T										
I	15									
V										
A										
	29									
Subtotal										
Gran Total										

APENDICE C

FORMATO PARA REGISTRO Y VALORACION

APENDICE C

Registro para clasificación de preguntas

Examen de Egreso EXEDII

	Pregunta	1=.0 ~	2=.11 ~	3=.21 ~	4=.31 ~	5= .41 ~	6= .51 -	7=.61-.80	8=.81-	9= .91 - 1.0
COMPRENSION AUDITIVA	1									
	2									
	3									
	4									
	5									
	6									
	7									
	8									
	9									
	10									
	11									
	12									
	13									
	14									
	15									
	16									
	17									
	18									
	19									
	20									
	21									
	22									
	23									
	24									
	25									
	26									
	27									
	28									
	29									
	30									
	31									
	32									
Subtotal										
GRAMATICA	33									
	34									
	35									
	36									
	37									
	38									
	39									
	40									
	41									
	42									
	43									
	44									
	45									
	46									
	47									
	48									
	49									

	Pregunta	1=.0 ~	2=.11 ~	3=.21 ~	4=.31 ~	5= .41 ~	6= .51 -	7=.61-.80	8=.81-	9= .91 - 1.0
	50									
	51									
	52									
	53									
	54									
	55									
	56									
	57									
	58									
	59									
	60									
	61									
	62									
	63									
	64									
	65									
	66									
Subtotal										
C O M P R E N S I O N D E L E C T U R A	67									
	68									
	69									
	70									
	71									
	72									
	73									
	74									
	75									
	76									
	77									
	78									
	79									
	80									
	81									
	82									
	83									
	84									
	85									
	86									
	87									
	88									
	89									
	90									
	91									
	92									
	93									
	94									
	95									
	96									
	97									
	98									
	99									
	100									
Subtotal										
Gran Total										

APENDICE D

FORMATO PARA EVALUACIÓN DEL PROCESO

APENDICE D

Versión editada de una forma de evaluación del libro Handbook for Setting Standards on Performance Assesment by Hambleton, Jaeger, Plake, and Mills (2000a)

Formato de Evaluación

El objetivo de este formato es recabar su opinión acerca del procedimiento para la determinación del punto de corte del examen de egreso EXEDII . Su opinión proporcionará una base para evaluar el entrenamiento y los métodos para la determinación del estándar mencionado.

Por favor no escriba su nombre en el formato. Su opinión será anónima. Gracias por tomarse el tiempo de llenar este formato.

1. Deseamos conocer su opinión en varios aspectos del estudio realizado. Por favor coloque una X en la columna que refleje su opinión acerca del nivel de éxito alcanzado.

Aspecto	Sin Exito	Parcialmente exitoso	Exitoso	Muy exitoso
a. Sesión de introducción y presentación del estudio				
b. Entrenamiento en la clasificación				
c. Primera sesión de discusión grupal				
d. 2a. sesión de discusión grupal con inf. sobre desempeño de examinados				

2. ¿Que tan adecuado fue el entrenamiento para prepararlo en la clasificación de las preguntas del exámen?

   a. Totalmente adecuado.
   b. Adecuado
   c. Parcialmente adecuado
   d. Inadecuado

3. ¿Como evaluaría el tiempo utilizado en el entrenamiento para prepararlo en la clasificación de las preguntas?

   a. Suficiente.
   b. Poco tiempo
   c. Muy poco tiempo

4. Indique la importancia de los siguientes factores en la clasificación de las preguntas.

Factor	Sin importancia	Importante De alguna manera	Importante	Muy importante
a. Su percepción de la dificultad del método de clasificación	_____	_____	_____	_____
b. Su propia experiencia en la materia.	_____	_____	_____	_____
c. Su primera clasificación individual	_____	_____	_____	_____
d. El panel de discusión grupal	_____	_____	_____	_____
e. La clasificación de otros expertos	_____	_____	_____	_____

5. ¿Como juzgaría el tiempo asignado para llevar a cabo la primera clasificación de las preguntas ?

a. Tiempo suficiente
b. Muy poco tiempo
c. Demasiado tiempo

6. ¿Como juzgaría el tiempo asignado para llevar a cabo la segunda discusión de la clasificación de las preguntas ?

a. Suficiente
b. Muy poco tiempo
c. Demasiado tiempo

7. ¿Que confianza tiene en que el procedimiento producirá un estándar de desempeño confiable ?

a. Mucha confianza
b. Suficiente
c. Con reserva
d. desconfianza

8. ¿Como evaluaría las facilidades para llevar a cabo la clasificación ?

a. Amplias
b. Suficientes
c. Insuficientes

Por favor conteste las siguientes preguntas acerca de la clasificación de los reactivos.

1. ¿que estrategias utilizó para clasificar las preguntas ?

2. ¿ Hubo algunos problemas en particular que influenciaran su decisión? Si es así, Cuáles ?

3. Por favor proporciónenos alguna sugerencia o maneras de mejorar el procedimiento de clasificación o el curso de entrenamiento de expertos.

Muchas gracias por completar este formato de evaluación.

APENDICE E

ESTIMACION DE LOS REACTIVOS DE TERCERA SESION

APENDICE E

ESTIMACION TERCERA SESION

Reactivos	Esp1	Esp2	Esp 3	Esp 4	Esp 5	Esp 6	Esp 7	MJUECEAL	treactivo
Pregunta 1	70	70	90	90	50	50	65	69	auditiva
Pregunta 2	70	70	75	80	90	50	70	72	auditiva
Prregunta 3	75	70	90	75	90	65	75	77	auditiva
Pregunta 4	70	70	65	75	50	60	45	62	auditiva
Pregunta 5	70	65	70	75	50	45	50	61	auditiva
Pregunta 6	70	70	80	80	80	60	70	73	auditiva
Pregunta 7	70	70	85	80	90	55	65	74	auditiva
Pregunta 8	75	65	80	70	85	50	65	70	auditiva
Pregunta 9	75	80	90	80	60	36	90	73	auditiva
Pregunta 10	85	85	90	85	90	70	90	85	auditiva
Pregunta 11	80	75	85	80	85	60	70	76	auditiva
Pregunta 12	75	70	70	75	50	45	60	64	auditiva
Pregunta 13	70	70	60	80	80	60	50	67	auditiva
Pregunta 14	70	75	75	75	90	50	70	72	auditiva
Pregunta 15	70	70	50	70	50	45	55	69	auditiva
Pregunta 16	75	70	50	80	50	45	70	63	auditiva
Pregunta 17	80	75	90	70	75	45	65	71	auditiva
Pregunta 18	75	70	80	80	55	45	65	67	auditiva
Pregunta 19	80	70	93	80	90	60	70	78	auditiva
Pregunta 20	70	60	70	80	65	50	55	64	auditiva
Pregunta 21	60	70	100	80	90	70	75	78	auditiva
Pregunta 22	65	65	65	80	80	50	55	66	auditiva
Pregunta 23	70	70	75	85	80	50	65	71	auditiva
Pregunta 24	60	70	90	80	85	50	70	72	auditiva
Pregunta 25	60	60	65	75	55	45	50	59	auditiva
Pregunta 26	70	60	70	70	60	55	50	62	auditiva
Pregunta 27	55	70	85	80	95	60	65	73	auditiva
Pregunta 28	55	65	93	85	90	55	75	74	auditiva
Pregunta 29	50	75	85	85	90	50	80	74	auditiva
Pregunta 30	50	70	80	80	95	50	75	71	auditiva
Pregunta 31	75	65	85	80	95	60	80	77	auditiva
Pregunta 32	75	65	80	80	55	50	65	67	auditiva
Pregunta 33	70	75	80	70	50	65	65	68	grammar
Pregunta 34	75	80	90	90	75	50	65	75	grammar
Pregunta 35	80	75	75	90	90	50	60	74	grammar
Pregunta 36	75	80	90	70	60	60	70	72	grammar
Pregunta 37	70	70	80	65	50	50	60	64	grammar
Pregunta 38	80	70	80	70	60	45	50	65	grammar
Pregunta 39	80	80	95	80	95	65	85	83	grammar
Pregunta 40	70	80	80	80	80	50	80	74	grammar
Pregunta 41	75	85	80	75	90	55	85	78	grammar
Pregunta 42	85	85	100	80	90	65	90	85	grammar
Pregunta 43	90	75	70	70	50	50	50	65	grammar
Pregunta 44	75	75	50	70	50	50	40	59	grammar
Pregunta 45	70	80	85	80	90	50	85	77	grammar
Pregunta 46	90	80	90	75	90	50	90	81	grammar

Reactivos	Esp1	Esp2	Esp 3	Esp 4	Esp 5	Esp 6	Esp 7	MJUECEAL	treactivo
Pregunta 47	90	80	95	75	75	70	90	82	grammar
Pregunta 48	90	75	95	80	75	70	80	81	grammar
Pregunta 49	85	75	90	80	70	60	85	78	grammar
Pregunta 50	80	80	95	75	70	70	85	79	grammar
Pregunta 51	80	75	95	80	80	60	80	79	grammar
Pregunta 52	70	70	70	75	75	45	60	66	grammar
Pregunta 53	35	65	70	65	50	45	65	56	grammar
Pregunta 54	40	70	80	65	50	45	50	57	grammar
Pregunta 55	40	65	70	65	50	45	50	55	grammar
Pregunta 56	50	65	60	65	30	45	65	54	grammar
Pregunta 57	70	65	70	80	60	50	65	66	grammar
Pregunta 58	75	70	70	75	85	50	70	71	grammar
Pregunta 59	90	70	75	75	60	50	65	69	grammar
Pregunta 60	70	70	75	75	50	55	75	67	grammar
Pregunta 61	70	70	85	75	45	60	80	69	grammar
Pregunta 62	70	70	85	80	45	60	70	69	grammar
Pregunta 63	85	70	93	80	60	60	80	75	grammar
Pregunta 64	90	75	90	75	50	60	70	73	grammar
Pregunta 65	80	70	80	70	40	40	60	63	grammar
Pregunta 66	70	70	75	75	40	40	45	59	grammar
Pregunta 67	70	70	70	80	75	70	70	72	lecture
Pregunta 68	70	75	75	80	60	70	60	70	lecture
Pregunta 69	75	70	90	80	60	65	60	71	lecture
Pregunta 70	70	80	85	80	75	60	80	76	lecture
Pregunta 71	80	75	80	70	60	60	65	70	lecture
Pregunta 72	70	75	80	80	70	65	60	71	lecture
Pregunta 73	70	70	65	70	80	55	65	68	lecture
Pregunta 74	80	75	85	80	85	70	70	78	lecture
Pregunta 75	80	70	80	75	50	50	65	67	lecture
Pregunta 76	70	75	80	75	80	50	50	69	lecture
Pregunta 77	70	75	85	75	50	50	50	65	lecture
Pregunta 78	45	70	85	80	75	45	40	63	lecture
Pregunta 79	65	70	65	75	80	45	40	63	lecture
Pregunta 80	70	70	70	80	75	45	40	64	lecture
Pregunta 81	70	75	93	75	90	45	65	73	lecture
Pregunta 82	70	80	80	75	85	45	65	71	lecture
Pregunta 83	70	75	70	75	50	45	45	61	lecture
Pregunta 84	75	80	80	80	90	55	70	76	lecture
Pregunta 85	75	80	94	80	90	70	65	79	lecture
Pregunta 86	80	75	70	75	90	60	45	71	lecture
Pregunta 87	80	80	65	80	75	60	50	70	lecture
Pregunta 88	70	80	80	80	50	65	55	69	lecture
Pregunta 89	65	80	70	80	90	50	45	69	lecture
Pregunta 90	70	85	90	80	80	70	40	74	lecture
Pregunta 91	80	80	92	80	90	50	65	77	lecture
Pregunta 92	70	80	90	70	90	50	80	76	lecture
Pregunta 93	80	80	85	75	95	50	70	76	lecture
Pregunta 94	80	75	95	85	90	70	50	78	lecture
Pregunta 95	80	80	95	75	95	60	65	79	lecture

Reactivos	*Esp1*	*Esp2*	*Esp 3*	*Esp 4*	*Esp 5*	*Esp 6*	*Esp 7*	***MJUECEAL***	treactivo
Pregunta 96	80	70	85	75	65	50	40	66	lecture
Pregunta 97	65	80	90	90	55	60	50	70	lecture
Pregunta 98	80	80	100	90	90	75	60	82	lecture
Pregunta 99	70	85	95	75	90	70	70	79	lecture
Pregunta 100	70	70	80	80	90	70	50	73	lecture
	72	73	81	77	72	55	64	**71**	**CUT**

72