EL CONCEPTO DE LA UTILIDAD EN EL DISEÑO DE UNA PRUEBA DE COMPRENSIÓN DE LECTURA EN LA LENGUA EXTRANJERA

Pauline Moore Hanna[1]
Hugo Andrade Mayer[2]
Miriam Matamoros Sánchez[3]

RESUMEN

Este artículo representa un diagnóstico completo de la fase de diseño de la prueba de comprensión de lectura para ingreso a estudios de posgrado de la Universidad Autónoma del Estado de México. Para realizar el diagnóstico se determinó medir el alcance del instrumento en relación a las cualidades de utilidad de Bachman y Palmer (1996) siendo éstas: confiabilidad, validez de constructo, autenticidad, interactividad, impacto y practicidad. Como la valoración implícita en la medición de utilidad hace referencia al contexto de la evaluación se describió de manera detallada dicho contexto empleando el modelo de Alderson (2000). La conclusión principal que emana de dicho análisis es que la prueba es adecuada para su contexto en cuanto al diseño pero que la operatividad de la misma se beneficiaría del desarrollo de algunas medidas de difusión.

---

[1] Doctora en Lingüística por la UNAM (2009). Actualmente es Profesor de Tiempo Completo en la Facultad de Lenguas de la UAEM donde investiga cuestiones de comprensión de lectura en lengua extranjera.
[2] Profesor de lengua inglesa, estudiante de la Maestría en Lingüística Aplicada en la Facultad de Lenguas de la UAEM y Subdirector Administrativo desde el 2006.
[3] Maestra en Traducción e Interpretación Inglés-Español por la UAG (2005). Es Profesora de Tiempo Completo en la Facultad de Lenguas UAEM. Estudiante del Doctorado en Letras Modernas en la Universidad Iberoamericana.

Palabras clave: diseño de exámenes, comprensión de lectura en lengua extranjera, utilidad, confiabilidad, validez.

ABSTRACT

This article presents a full diagnostic of the design phase of a reading comprehension test designed as part of the admissions process to postgraduate studies in the *Universidad Autónoma del Estado de México*. The diagnostic is couched in terms of Bachman y Palmer's model of test usefulness (1996) and the extent to which the test achieves the dimensions of usefulness —reliability, construct validity, authenticity, interactiveness, impact and practicity— is evaluated. Since the assessment of usefulness is not possible without reference to the testing situation an analysis of the context of the test is also presented. The main conclusions of the analysis are that, while the test is appropriately designed with regard to context, its operation would benefit from the development of additional materials such as user guides and candidate practice materials.

Key words: design tests, reading comprehension, foreign language, utility, reliability, calidity.

INTRODUCCIÓN

En el presente artículo se efectúa el diagnóstico de la fase del diseño de la prueba, utilizando el modelo analítico de la utilidad, de acuerdo a Bachman y Palmer (1990, 1996). Por fase de diseño se entiende la planeación de la prueba, su especificación y el piloteo de los resultados preliminares (Bachman, 1990; Hughes, 2003). Es decir, todas aquellas acciones necesarias para permitir el lanzamiento de un instrumento de evaluación nueva.

En el campo del diseño de instrumentos de evaluación ha habido un movimiento gradual pero inexorable mayor hacia la rendición de cuentas por parte de los cuerpos evaluadores. Donde quizá antes el buen nombre del cuerpo evaluador era suficiente para convencer a los usuarios de las pruebas que los resultados serían confiables, hoy en día se tiene un concepto más sofisticado acerca de lo que la experiencia evaluadora debería ofrecer (Shohamy, 2001; Weideman,

2006). Entendemos como usuarios de una prueba a: las personas que toman el examen, aquellos que utilizan los resultados para tomar decisiones, los maestros e instructores que preparan a los candidatos para tomar la prueba, los padres de familia, etcétera. La creciente importancia de los exámenes y la certificación de competencias crean en los usuarios la necesidad de informarse mejor sobre lo que deben demandar en las pruebas que se usan para emitir juicios sobre sus habilidades y conocimientos, y al cuerpo evaluador se le exige ofrecer dicha información de manera accesible a los usuarios.

A pesar de que se ha incrementado la sofisticación del usuario, todavía persisten algunas creencias cuestionables. Por ejemplo, aun es común ver en las solicitudes de empleo la exigencia de que el candidato cuente con cierto porcentaje de dominio de la lengua extranjera. Para un experto en la evaluación de lenguas extranjeras resultaría imposible calcular competencia lingüística en términos de un porcentaje fijo. El lenguaje humano tiene posibilidades infinitas de expresión, por lo que no se puede hablar de un porcentaje. En cualquier caso, para poder estimar la adecuación del dominio de la lengua en un candidato es esencial saber para qué propósitos el candidato tendrá que utilizar la lengua extranjera. Una evaluación sin parámetros otorgaría un juicio carente de valor real.

De la misma manera, es imposible calcular con exactitud la competencia lingüística de una persona, toda vez que es imposible indagar sobre todos los posibles campos de expresión. Por ejemplo, es posible que una persona domine con excelencia el uso formal del lenguaje pero en cuanto tenga que participar en una conversación casual y espontánea no pueda expresarse de manera aceptable. En este mismo tenor, es posible que una persona lea muy bien en su campo de especialización, pero al confrontar textos de otra índole tenga problemas de comprensión. Esto último es cierto tanto para lengua nativa como para lengua extranjera. Dada la imposibilidad de diseñar una prueba que cubra todos los potenciales usos del lenguaje —que tardaría una vida entera en aplicarse y todavía no alcanzaría más de una mínima parte de los usos potenciales— los diseñadores de exámenes se ven en la necesidad de muestrear una pequeña parte

de los dominios posibles y extrapolar una decisión imprecisa pero práctica en base a dicho muestreo. Por lo que es importante determinar cuáles son los dominios o ámbitos del uso del lenguaje más relevantes para el propósito de la situación de evaluación para especificar cuáles se incluirán en la prueba.

El propósito del presente artículo consiste en valorar el diseño de la prueba utilizada para certificar la comprensión de textos en inglés. La mayor exigencia de los candidatos a la prueba ha resultado en modificaciones en la forma en que aplicamos nuestros instrumentos de evaluación, particularmente en la necesidad de incluir un mayor número de criterios, sobre todo en una situación de prueba cuyos resultados son tan significativos para todos los usuarios. Este incremento en los criterios de evaluación de las pruebas, al mismo tiempo ha ampliado nuestro concepto del instrumento de evaluación para tomar en consideración aspectos que van más allá de cómo se ve el instrumento en papel o las técnicas de medición que se ocupan. Nuestras inquietudes tienen que rebasar las meras consideraciones sobre qué tipo de reactivos estamos utilizando para incorporar todo el ámbito que rodea la prueba.

El trabajo del investigador Lyle Bachman (1990, 1996) ha sido fundamental en este cambio en la perspectiva sobre los exámenes. Una de sus aportaciones más importantes es el concepto de utilidad como medida de control de la calidad de los exámenes. Este concepto reorienta el interés de la evaluación del instrumento de los criterios objetivamente medibles de validez y confiabilidad, para incluir otros criterios que, aun siendo menos susceptibles a la cuantificación, nos permiten considerar el lado más humano de las pruebas que diseñamos. Esencialmente, se trata de dar una nueva dirección al enfoque tradicional de la evaluación de las pruebas, basado en las características técnicas de confiabilidad y validez, hacia cualidades más contextualizadas, lo que obliga al diseñador de exámenes a pensar más claramente sobre las acciones que ha realizado para asegurarse de que el examen se adapta a las condiciones de la situación de prueba y no viceversa.

1. UTILIDAD

Bachman y Palmer (1996: 17) plantean que el factor más importante cuando se diseña una prueba es el uso para el que está destinado, por lo que el criterio de evaluación más relevante es que la prueba sirva su propósito correctamente. Los autores llaman **utilidad** a esta cualidad, no obstante, cabe señalar que no se trata de una concepción utilitaria en el sentido negativo, sino que más bien indica una tendencia más humanista en el diseño de pruebas. Consta de seis componentes: confiabilidad, validez del constructo, autenticidad, interactividad, impacto y practicidad. Las cualidades de confiabilidad, validez y practicidad tienen una trayectoria reconocida en la teoría de evaluación; sin embargo, la autenticidad, interactividad y el impacto son menos conocidos por lo que ameritan un tratamiento más profundo. No son las cualidades por separado las que constituyen la aportación original de Bachman y Palmer sino su trabajo conjunto y el análisis de la interacción entre ellas. Dado que la propuesta de Bachman y Palmer es poco utilizada todavía en el campo de la evaluación en México consideramos importante ofrecer una breve exposición de los conceptos básicos del modelo. Es este modelo el que ocuparemos para aportar las categorías de análisis en el diagnóstico de la fase de diseño de la prueba. En este apartado solamente se explican los conceptos como tales, en la tercera sección se retomarán las cualidades descritas en términos específicos de la prueba que se está evaluando.

La primera cualidad es la **confiabilidad**, una de las dimensiones más tradicionales en la evaluación de pruebas. Entre los autores que incluyen discusiones sobre las cualidades deseables en el diseño de un examen se encuentran Hughes (2003), Alderson y Wall (1993). Bachman y Palmer definen a esta cualidad como *consistency of measurement* (1996: 19), es decir, la uniformidad con que el mismo examen arroja resultados a pesar de que se aplique en momentos o lugares distintos y/o se evalúe por diferentes examinadores. Si se trata de un examen de versiones múltiples, será confiable en caso de que las diferentes versiones arrojen resultados parecidos. Si bien el resultado de un examen necesariamente se ve afectado por las circunstancias espacio-

temporales y hasta psicológicas del sustentante, la variación causada por este tipo de factores tendría que reflejar un mínimo error de medición para decir entonces, que el examen es confiable. Rudner (1994: 3) abunda sobre las posibles fuentes de error de medición, que en su definición "incluyen fatiga, nerviosismo, muestreo del contenido, errores al transcribir la respuesta, falta de comprensión de las instrucciones y adivinar respuestas, dado que estos errores aumentan la calificación del sujeto evaluado y disminuyen la confiabilidad de la prueba". En cuanto a las circunstancias espacio-temporales, posibles fuentes de variación en los resultados pueden ser: la hora del examen (muchas personas logran mejores niveles de concentración por la mañana, por ejemplo), el nivel del ruido exterior, el mobiliario, la temperatura del ambiente, entre otras. Con respecto a las circunstancias psicológicas del sustentante, su desempeño puede verse afectado por el trato del examinador o administrador de la prueba, su estado anímico, problemas personales o de trabajo, y su capacidad de respuesta ante la presión que implica la evaluación. Para Bachman y Palmer (1996), es importante reconocer que será imposible eliminar todas las posibles fuentes de error en la medición. La meta del diseñador de la prueba es procurar minimizar los efectos de las fuentes de error, para que los resultados del examen reflejen con mayor fidelidad la habilidad del examinado.

El segundo elemento del modelo de Bachman y Palmer (1996) es la **validez del constructo**. Validez se ha considerado una cualidad importante en el diseño de pruebas desde hace mucho tiempo y se entiende como la medida en que la prueba mide lo que se supone que debería de medir (Hughes, 2003). Generalmente se consideran tres dimensiones de validez: predictiva o de criterio, de contenido y de constructo (Moskal y Leydens, 2000). *Validez predictiva o de criterio* se refiere a la precisión con la que un resultado positivo en la prueba puede predecir éxito en las actividades que la prueba pretende evaluar. Por ejemplo, los exámenes de selección que se utilizan para determinar el ingreso a distintos niveles educativos idealmente predicen quienes serán estudiantes exitosos.

La *validez de contenido* es la medida en la que los conocimientos y competencias muestreados por la prueba reflejan el contenido que deberá o pudiera saber el sustentante. Esta dimensión de la validez es de particular importancia para los exámenes de logro, no es justo para el estudiante que se le evalúe algún conocimiento que no se le ha ofrecido durante el curso.

Algunos autores, como Hughes (2003), agregan a este listado de tipos de validez como la visual. Se trata de la medida en que una prueba parece superficialmente medir algo. A pesar de no tratarse de un concepto muy científico es de gran importancia para el diseñador de la prueba, ya que si carece de validez visual, aun cuando reúna todos los otros tipos de validez los usuarios no creerán en su eficacia.

Es notorio que para Bachman y Palmer (1996) el único de todos estos tipos de validez que amerita consideración es la *validez de constructo*. Estos autores la definen como la medida en que el resultado del examen se constituye en un parámetro significativo y digno de ser tomado en cuenta para tomar decisiones sobre la presencia o ausencia de alguna habilidad o conocimiento en el sustentante, por ejemplo, para determinar si una persona cuenta con la competencia para redactar oficios o para leer cierto tipo de textos; lo que nos interesa como evaluadores es poder decir que es así o no. Aquellos que diseñan un examen para medir un conocimiento o habilidad, tienen también que dar cuenta de cómo se pueden interpretar los resultados y de cuál es el sustento que apoya la interpretación de los resultados. Un examen es diseñado con una finalidad específica; pues bien, la validez es el indicador de que se está midiendo únicamente lo que se desea medir (Bachman y Palmer, 1996). En el caso de los exámenes de comprensión de textos en una lengua extranjera para estudios de posgrado, se desea conocer, a partir de los textos del examen, hasta qué punto el sustentante puede realizar una lectura significativa y útil para sus estudios. Bachman y Palmer (1996) indican que la validez del constructo en un examen de lengua permite hacer, a partir del resultado, una generalización que se proyecta hacia el uso de la lengua en circunstancias ajenas a la prueba.

Para entender mejor de lo que se trata la validez del constructo es importante definir qué es un **constructo**. Según Alderson (2000: 118) es un concepto psicológico que se deriva de una teoría de la habilidad que se evalúa. Los constructos son los componentes principales de la teoría y la relación que existe entre ellos también se especifica en la teoría. De una revisión de las teorías actuales de la comprensión de lectura queda claro que se involucran una amplia gama de estrategias y habilidades de niveles bajos de procesamiento cognitivo, como la identificación de letras y clases gramaticales de palabras; pero también una gama de procesos de alto nivel cognitivo, como la asimilación del contenido textual al conocimiento previo. Inevitablemente para representar adecuadamente la teoría será importante muestrear ampliamente estas estrategias y habilidades. El nivel logrado de validez del constructo depende de la solidez de la teoría sobre la que la prueba está construida, es decir, no se puede medir correctamente lo que no se entiende. Una teoría es una idealización de lo que se espera observar en un individuo que posee cierta competencia, la validez del constructo es una medición del grado en que las conductas esperadas coinciden con las conductas reales de un sujeto que cuenta con una competencia lectora.

Es precisamente esta definición lo que le permite aportar mayor importancia a la validez de constructo que a cualquier otro tipo. Una vez que la prueba refleje de la manera más fiel posible los procesos cognitivos que forman una habilidad es menos probable que no se logre el éxito en la aplicación de dicha habilidad. Para Bachman y Palmer (1996) la validez de contenido tiene menor trascendencia, ya que los contenidos de importancia no son los contenidos de enseñanza del curso que pudiera anteceder al examen, sino los de los procesos cognitivos que subyacen la capacidad del aspirante para leer textos. De la misma manera, validez de contenido es irrelevante en las situaciones de prueba que no presuponen un curso específico anterior, como estos exámenes de competencia general para ingreso a un nivel superior educativo.

La tercera categoría de análisis en la utilidad es la **autenticidad**. De acuerdo a Bachman y Palmer (1996: 23) la podemos entender como el grado de relación entre la tarea a realizar en el examen y el dominio en un contexto real del uso de la lengua objetivo. En otras palabras, la tarea que pone a prueba la competencia del sustentante debe reflejar o asimilarse a una tarea en la vida real, en la que tendrá que hacer uso de la lengua objetivo. Por ejemplo, en algunos países, las personas que solicitan una licencia de manejo son evaluadas mediante una prueba práctica de manejo en compañía de un evaluador y en dicha práctica se ponen a prueba las habilidades necesarias para llevar a cabo esta tarea en la vida real. Este examen tiene un alto grado de autenticidad. En otros países, la prueba de manejo es teórica y sobre la comprensión de la señalización utilizada o el reglamento de tránsito, lo cual tiene un grado menor de autenticidad. En nuestra área de interés, si alguien solicitara un trabajo como personal de limpieza y se le pidiera hacer un examen de idioma, esta prueba tendría un bajo grado de autenticidad, puesto que poco tendría que ver el conocimiento de una lengua extranjera con las labores que desempeñaría en caso de ser contratado.

Por esto es que una de las metas del evaluador consiste en medir el nivel de lengua alcanzado a través de un instrumento de cierta relevancia en el objetivo del uso de la lengua. Bachman y Palmer (1996: 23) resaltan la importancia de la autenticidad en el examen de idioma; el resultado de la evaluación debe reflejar de manera fiel el desempeño del sustentante al hacer uso de la lengua en una situación específica. El grado de autenticidad es de especial importancia para el sustentante, puesto que un examen de bajo grado en esta dimensión podría afectar su desempeño durante la prueba, al causar confusión o desconcierto, ya que la relación entre el uso de la lengua y los aspectos a evaluar (autenticidad) no son congruentes uno con el otro.

En cuarto lugar, para determinar la utilidad de una prueba se estima necesario determinar su **interactividad**. Ésta es definida como el grado de identificación entre las características individuales de la persona a ser evaluada y la tarea de una prueba (Bachman y Palmer, 1996: 25). Un examen será alto o bajo en interactividad de acuerdo al grado

con que las características individuales contribuyan a la resolución del mismo. En el contexto de exámenes de comprensión de lectura tres características tendrán mayor influencia sobre la interactividad: conocimiento del tema, competencia lingüística y afectividad (Bachman y Palmer, 1996: 65-78).

El conocimiento del tema, o esquemas de conocimiento, se refiere a las estructuras del conocimiento previamente almacenadas en la memoria a largo plazo (Bachman y Palmer, 1996: 65); es decir, aquello que el individuo conoce del mundo y que le ayuda a tener un entendimiento del mismo. **La relevancia** de esta primera característica radica en que el individuo usa este conocimiento para comprender los textos incluidos en la prueba. Si tal conocimiento no existiera, el resultado cambiaría al medir el conocimiento del lenguaje del aspirante puesto que se manifestaría su *desconocimiento* del tema tratado. Por ello, el individuo obtendrá un mejor resultado en la prueba cuando los materiales tratan temas con los que está familiarizado y asimismo tendrá una mayor interactividad. *Ceteris paribus*, el logro de la comprensión de un texto depende en gran medida del tema, así por ejemplo, puede esperarse que un texto que habla de la inflación será mejor comprendido por un economista que por un médico. Para el economista, la tarea es de alta interactividad, mientras que para el médico, puede resultar de baja interactividad.

La segunda característica es **la competencia lingüística**, la cual es mencionada junto con el conocimiento del mundo y las competencias estratégicas como un componente del concepto de habilidad comunicativa lingüística propuesto por Bachman (1990: 84). El autor divide la competencia lingüística en competencia organizacional, la cual se divide en: gramatical y textual, por otro lado la competencia pragmática, conformada por la competencia elocutiva y la competencia sociolingüística. Así Bachman (1990: 86) agrupa dentro de la competencia organizacional los elementos de morfología, sintaxis, cohesión y organización, y dentro de la competencia pragmática los elementos sociolingüísticos y las habilidades relacionadas con el uso del lenguaje. De mismo modo, Bachman habla de competencia estratégica (1990: 98), como otro elemento del marco de referencia, y

describe la naturaleza interactiva de dicha competencia a través de la mención de diversos estudios al respecto (Faerch y Kasper, 1984; Canale y Swain, 1980; Canale, 1983). La *competencia estratégica* de acuerdo a Canale (1983) es el uso de estrategias verbales y no-verbales para sortear problemas de comunicación causados por una competencia insuficiente o un pobre desempeño, y Bachman (1990: 100) la secciona en tres componentes: valoración, planeación y ejecución. En el caso de este examen se tiene que especificar que la parte de la competencia lingüística que nos interesa es la lectora. Nuestra aseveración en este sentido implica una decisión teórica en el sentido de que es posible subdividir la competencia lingüística de esta manera, es decir, que se puede de alguna manera hablar de evaluar la comprensión de textos sin necesariamente evaluar todas las habilidades lingüísticas. Dicha postura se considera normal dentro de la lingüística aplicada al menos desde el trabajo de Widdowson (1979).

Más tarde, Bachman y Palmer (1996: 25) retoman el concepto de **competencia lingüística**, compuesta por el conocimiento lingüístico y la competencia estratégica, o estrategias metacognitivas, como una parte importante de las características individuales que tienen un efecto en la interactividad de un examen. De acuerdo con esto, el sustentante tratará de compensar la falta de conocimiento usando su competencia estratégica. Es decir, es posible que en algún momento de la prueba el sustentante no conozca algún significado, o no comprenda una idea del texto, es entonces cuando buscará alguna manera de entender, ya sea infiriendo, o haciendo uso de su conocimiento del tema, o de sus experiencias para llevar a cabo la tarea de manera exitosa. No obstante, pensamos que es la competencia, o habilidad lingüística, la que tiene mayor peso en el momento de la resolución de la prueba.

Finalmente, la tercera característica individual, **los esquemas afectivos**, también podrán influir en la interactividad de la prueba. La relación afectiva o emocional que tenga el individuo con el conocimiento del tema, es decir, que represente algo vivido por el sustentante para bien o para mal. La experiencia activaría, de alguna manera, una respuesta positiva o negativa a la tarea, y podría tanto facilitarla como hacerla más difícil.

Así una prueba será alta en interactividad de acuerdo con el grado en que le permite al individuo hacer uso de sus características personales para su resolución. Un texto acerca de un tema ajeno, con vocabulario desconocido, estructuras demasiado complejas, o sin un significado emocional, será de baja interactividad y representará problemas para el individuo. Es un objetivo para el examinador lograr que la prueba tenga un grado aceptable de interactividad, que el tema sea familiar, o al menos no desconocido por completo y que el texto sea accesible. Todo esto es factible, pero aún queda sobre la mesa la cuestión de los esquemas afectivos, ya que por tratarse de factores muy personales, es prácticamente imposible que el evaluador tenga control sobre esta característica.

La quinta categoría es el **impacto**. La palabra más significativa que Bachman y Palmer (1996: 30) usan para hablar de esta cualidad en los exámenes es **consecuencias**. El resultado de un examen trae consecuencias, positivas o negativas, significativas o poco trascendentes; y no solamente a nivel micro, para el sustentante, sino para los que toman decisiones basándose en los resultados. Yendo aun más lejos, puede decirse que en ocasiones se genera un efecto dominó que llega hasta la sociedad en general en un impacto macro. Se puede decir que el impacto de la prueba es la manera en que las cualidades obtenidas por su diseño realimentan a las otras partes del sistema en el que se encuentra inmersa.

En los contextos en los que una prueba se emplea para determinar quiénes serán seleccionados para participar en una actividad que se percibe como deseable (como el ingreso a estudios de posgrados) los aspirantes se esforzarán para adquirir las habilidades y competencias necesarias para aprobar la prueba. El impacto o *washback* se puede definir como la relación que existe entre la preparación necesaria para lograr un resultado exitoso en la meta inmediata de la prueba y la preparación necesaria para lograr un resultado exitoso en la meta más a largo plazo de éxito en la vida rea: los estudios de posgrado.

Para que se optimice el impacto de una prueba sobre el contexto es necesario que los aspirantes cuenten con información útil y accesible sobre lo que se evaluará. Por ejemplo, no hace mucho se dio a conocer la nota de que México obtuvo pobrísimos resultados en las pruebas de nivel[4] en educación primaria. Ahora son notorios los esfuerzos en el sistema educativo (incluyendo los docentes, los directivos y los padres de familia) para mejorar esta situación. En la medida en que las pruebas de ENLACE midan habilidades y competencias reales y útiles el impacto de la prueba será positivo. Sin embargo, si para obtener buenos resultados en la prueba solamente es necesario desarrollar la habilidad de responder a exámenes de este tipo aunado a un mínimo repertorio de conocimientos, se reducirán los esfuerzos de los profesores a trabajar sobre esta gama reducida de conocimientos y el impacto de la prueba será negativo. La confiabilidad y validez todavía no son conocidas por el público en general, sin embargo, el impacto de los resultados, que finalmente se determina en función de la interacción de dichos factores, alcanza a todos.

Los exámenes de comprensión de lectura para posgrado tienen un impacto directo en los interesados en continuar con su preparación profesional, y también en las instituciones que toman los resultados como un factor, de mayor o menor peso, para aceptar a los aspirantes. Más allá de los usuarios directos de la prueba, el examen tendrá un impacto sobre la planeación de las actividades formativas en la preparación de los candidatos para la prueba. Si el examen no alcanza cierto nivel de validez de constructo, es decir, si en realidad las habilidades y competencias que muestrean la prueba no constituyen la comprensión de lectura, sino otra dimensión del conocimiento del aspirante, como conocimiento previo del tema, vocabulario técnico o la habilidad de adivinar correctamente, entonces el impacto de la prueba sobre la enseñanza será más bien negativo. En la medida en que

[4] ENLACE es una prueba general de conocimientos que se aplica en toda la República Mexicana. Se administra por la Secretaría de Educación Pública. Las siglas de ENLACE son el acrónimo de "Evaluación Nacional del Logro Académico en Centros Escolares".

para lograr una calificación aprobatoria en la prueba el candidato en realidad deberá poder comprender lo que lee en inglés, más benéfico será el impacto de la prueba.

Finalmente, la sexta cualidad del modelo de la utilidad es la **practicidad**. Se trata de la relación entre los recursos necesarios para llevar a cabo la prueba y los recursos disponibles (Bachman y Palmer, 1996: 35). Este aspecto de la utilidad indica que tan factible es su realización. Hemos de considerar los distintos tipos de recursos: humanos (examinadores, realizadores del examen, administradores, etcétera), materiales (espacios, equipos, y otros materiales como de papelería) y por último, el tiempo (para desarrollar, aplicar, administrar y evaluar la prueba). Es indispensable tomar en cuenta el costo de los recursos mencionados ya que la utilidad de la prueba se vería afectada seriamente si los recursos necesarios exceden los disponibles; lo anterior indica una baja practicidad, y por ende poca factibilidad.

La practicidad está estrechamente relacionada con otras cualidades, notoriamente con la autenticidad. Por ejemplo, para evaluar la comprensión auditiva en el campo de las lenguas tradicionalmente se utilizan grabaciones en audio con sus correspondientes preguntas de comprensión. Los CD son relativamente baratos y su reproducción en la sala de aplicación de la prueba es sencilla y práctica; el equipo es fácil de conseguir. Sin embargo, en términos de autenticidad, la práctica de escuchar un CD para obtener información es bastante remota de las realidades del uso de una lengua extranjera y por lo tanto resulta poco auténtica. Ur (1984: 4,26) señala varias diferencias entre la comprensión auditiva en la vida real y en las evaluaciones. El oyente en la vida real tiene acceso a pistas contextuales sobre el significado, normalmente se escuchan intervenciones cortas, como es el caso de una conversación normal en la cual los interlocutores tienen turnos de duración relativamente cortos. De la misma manera en una interacción cotidiana hay cierta reciprocidad, por ejemplo, aunque se esté escuchando una clase magistral (clásico monólogo) la persona que da la conferencia puede tomar en cuenta las reacciones lingüísticas y no-lingüísticas de los que están escuchando y existen oportunidades para la negociación del significado. Evaluar la comprensión auditiva con un casete es práctico pero no es auténtico.

Tradicionalmente en el campo de la evaluación de lenguas extranjeras, se ha dicho que un buen examen deberá tener tres cualidades: confiabilidad, validez y practicidad (Morrow, 1979; Underhill, 1987; Heaton, 1988). El mensaje ha sido que existe una tensión entre las cualidades y que optimizar una de ellas llevará a una reducción casi total en otra. Es decir, que los evaluadores tienen que determinar si les importa más la confiabilidad, la validez o la practicidad ya que una prueba con altos niveles de confiabilidad necesariamente será una prueba poco válida y que una con altos niveles de validez tendrá que sacrificar confiabilidad. Morrow (1979) señala que es imposible diseñar pruebas que sean auténticas y confiables a la vez. Esta postura es bastante radical y, de acuerdo con Bachman y Palmer, insostenible. Hughes (2003) ofrece una perspectiva más razonada comentando que aunque siempre existirá cierta tensión entre la validez y la confiabilidad en la construcción de pruebas, la meta del evaluador debe ser buscar un equilibrio entre dichas cualidades. Bachman y Palmer (1996) observan que en lugar de hacer hincapié en la tensión que existe entre las cualidades es importante considerarlas como complementarias.

Las seis cualidades descritas en conjunto forman el modelo de la utilidad. Sin embargo, no es suficiente medir cada una de las cualidades de manera aislada, sino que hay que considerar la forma en que se complementan en relación al instrumento evaluado y su propósito. El equilibrio perfecto entre las seis cualidades varía con respecto a la situación de la prueba. Para determinar si la prueba en realidad cumple con los propósitos para los que fue diseñada es importante seguir tres principios operativos (Bachman y Palmer, 1996: 18).

- **Principio 1**: la meta es maximizar la utilidad global de la prueba en vez de enfocarse a las cualidades de manera individual.

- **Principio 2**: las cualidades individuales de la prueba no pueden medirse de manera independiente, pero habrán de evaluarse en términos de su efecto combinado sobre la utilidad global de la prueba.

- **Principio 3**: la utilidad de la prueba y el equilibrio apropiado entre las cualidades no puede recetarse de manera general, sino que se determina específicamente para cada situación evaluativa.

En este sentido, debemos considerar, por ejemplo, que una prueba que tendrá un alto impacto sobre la vida del aspirante deberá ser altamente confiable. No es justo que una prueba que limitará las posibilidades del aspirante a entrar a la educación superior tenga un error de medición alto. En el caso de la prueba de manejo anteriormente descrita, es posible que omitir la necesidad de una prueba práctica redunde en otorgar licencias de manejo a personas que no saben manejar por su baja autenticidad pero indudablemente la omisión de la prueba práctica reduce sustancialmente los costos de su aplicación. Finalmente, cada situación evaluativa genera sus propios parámetros de utilidad que se tendrán que establecer claramente para optimizar la utilidad de la prueba.

2. DESCRIPCIÓN DEL CONTEXTO DE LA PRUEBA

Bachman y Palmer (1996) establecen claramente que no es posible determinar la utilidad de una prueba sin referencia detallada al contexto de diseño de la prueba y los propósitos para los que se diseñó. Está implícito en esta aseveración que entre más se aproximan las tareas evaluativas incluidas en la prueba y las tareas de la vida real que el candidato tendrá que llevar a cabo, mejor será el ajuste alcanzado con la situación de evaluación. La prueba de comprensión de lectura a que nos referimos en este artículo es, de hecho, una serie de instrumentos diseñados para evaluar comprensión de lectura dentro de un conjunto de áreas disciplinares que reflejan una clasificación relativamente suelta de las disciplinas estudiadas en los posgrados de la UAEM. Por motivos de seguridad se cuenta con más de una versión de cada instrumento por disciplina para que se pueda rotar su aplicación, evitando así las probabilidades de que un aspirante presente más de una vez la misma versión del instrumento que le corresponde. Por lo anterior, cuando hacemos referencia a la prueba nos referimos a todo el conjunto de instrumentos incluidos en su operatividad.

Siguiendo a Alderson (2000: 167-201) describiremos la situación evaluativa de la prueba que nos interesa en cuanto a cinco aspectos: el contexto de la evaluación, las características del escenario, las características de la rúbrica de la prueba, las características de los materiales utilizados y las características de las respuestas esperadas.

1.1. Contexto de evaluación

La prueba objeto de este estudio es de alto impacto para los usuarios. El resultado que obtiene el individuo es uno de los factores que los comités de admisión a estudios de posgrado toman para determinar la aceptabilidad de un candidato. Para el candidato, reprobar la prueba significa no ser aceptado al programa de posgrado deseado. La legislación universitaria en el Reglamento de Facultades y Escuelas Profesionales de la UAEM en el artículo 166, inciso V, establece que los posgrados de la UAEM deberán incluir entre sus criterios de selección de aspirantes la acreditación de la comprensión de lectura de un idioma extranjero. En la práctica, los aspirantes eligen generalmente la lengua inglesa.

La prueba se ubica en una situación académica. Su propósito es identificar si los aspirantes a ingresar a estudios de posgrado en la UAEM podrán leer con suficiente competencia en inglés para responder a las exigencias de sus estudios. Potencialmente, y dependiendo del área de especialidad que elige, a lo largo de sus estudios podrá tener que leer desde un texto en inglés hasta tres o cuatro artículos por semana. Como estas variaciones dependen de muchos factores, incluyendo los maestros asignados y sus actitudes hacia la lengua inglesa, es difícil de predecir. En el diseño del examen de comprensión de lectura, se ha optado por suponer que los estudiantes tendrán que leer regularmente en inglés y que esa habilidad es esencial para su buena formación como profesional o investigador. Por lo anterior, es importante que la prueba diseñada evalúe su competencia como lector en inglés en relación a los textos académicos, y que el examen no sea un ejercicio vacío de medición para cumplir con un requisito legislativo.

1.2. Características del escenario

Los examinados son aspirantes a estudios de posgrado en la UAEM. La mayoría de la población tendrá entre 24 y 45 años, aunque algunos aspirantes pueden ser mayores a los 45. Casi sin excepción serán mexicanos. Suelen ser de ambos sexos, aunque dentro de disciplinas específicas puede haber una tendencia hacia un género u otro, por ejemplo, en ingeniería hay más hombres, mientras que en lenguas hay más mujeres. Sin excepción contarán con estudios completos de licenciatura, la mayoría ya tendrá su título profesional, otros aspirantes estarán a punto de obtenerlo, por lo que, todos habrán ya desarrollado cierta competencia en la lectura de textos, sobre todo de su área disciplinaria. En algunos casos, un aspirante podrá estar entrando a estudiar un campo que no fue su especialidad en la licenciatura, por ejemplo, el estudiante de ingeniería que planea ingresar a una maestría en administración de empresas.

Si aprueban el examen, se puede esperar que los estudiantes tengan que presentar algún texto escrito que demuestre su comprensión en relación a la lectura de los textos de sus estudios de posgrado, como un resumen, fichas o mapa conceptual, o evidencia oral en discusiones o presentaciones. Aun si no fuera así, se esperaría que el estudiante de posgrado fuera al menos capaz de incorporar los significados reconstruidos de los textos que lee a los artículos, proyectos de fin de semestre o tesis que se encuentre redactando. Casi sin excepciones, esta evidencia de comprensión escrita u oral al texto se realizará en español. En resumen, la finalidad primordial de la lectura es el procesamiento y reconstrucción del significado.

En el contexto para el que los sustentantes requieren la lectura, puede haber limitaciones considerables del tiempo que disponen para leer los textos. Sobre todo en el contexto de lecturas académicas para artículos y tesis se espera que el sustentante pueda autorregular su uso de estrategias, por ejemplo, que pueda elegir entre la lectura superficial y la necesidad de una lectura más profunda.

1.3. Características de la rúbrica de la prueba

Tradicionalmente en el estudio de algunas disciplinas se requiere de manera poco frecuente de la habilidad de leer, en el sentido común de extraer significado, por ejemplo, en matemáticas y computación. Otras disciplinas requieren de la lectura normal pero de forma limitada, como química, biología o ingeniería, mientras que para otras la lectura es una actividad central que se espera se lleve a cabo con profunda comprensión y análisis crítico, como las humanidades, Lingüística e Historia (Alderson, 2000: 180). No obstante, estas diferencias no se han comprobado con estudios empíricos, mucho menos en el contexto de la UAEM. Las especificaciones de la prueba identifican un núcleo básico de estrategias y habilidades que caracterizan a un buen lector y este núcleo se evalúa a pesar de posibles sospechas de que lo que se evalúa mediante el examen va más allá de lo que en realidad necesitan unos, y queda corta de lo que requieren otros.

El núcleo básico de habilidades de comprensión de lectura se fundamentó en un análisis de las necesidades percibidas de una muestra de coordinadores y profesores de posgrado, con lo que se determinó que los estudiantes tienen que lograr más que una comprensión superficial de lo leído: necesitan identificar el significado global del texto, el propósito de secciones menores al texto, como párrafos o grupos de párrafos, los argumentos principales y secundarios y, en cierta medida, la postura que evidencia el autor hacia el tema del texto. Para realizar dichos procesos el lector requiere de destreza con técnicas más mecánicas, como la identificación de sinónimos, antónimos y de los referentes de elementos deícticos. Sin embargo, su nivel de comprensión no necesariamente se ubica en el nivel de la lectura crítica de los textos. Basándose en esta definición se construyeron las especificaciones del examen en la siguiente manera: comprensión general, 20%; comprensión de detalles, 70%; inferencia, 10%. A continuación se presenta una muestra de las estrategias que componen cada habilidad y que conforman las unidades de medición en cada versión de la prueba.

Comprensión general

- Identificar la idea general del texto o de una sección del texto
- Identificar el título de una gama de opciones

Comprensión de detalles

- Identificar la información contenida en una oración o un párrafo
- Identificar las ideas secundarias
- Seguir argumentos en texto escrito
- Identificar los pasos en un procedimiento escrito
- Identificar los referentes en un texto
- Identificar los sinónimos en un texto

Inferencia

- Identificar tenor positivo o negativo en argumentación
- Identificar el propósito del autor
- Comprender la relación entre partes del texto
- Realizar inferencias a partir de unidades de información discretas en el texto

Por cuestiones tanto de fatiga de los sustentantes como operativas, se restringe el tiempo disponible para realizar la prueba a dos horas, lo cual a su vez limita la extensión posible de los textos a leer, la cantidad de preguntas que se pueden realizar y el tipo de respuesta que se espera del aspirante. Esta limitación al tiempo es una fuente importante de desviaciones entre el formato final de la prueba y las actividades de lectura en los estudios de posgrado.

Finalmente, en cuanto a las rúbricas de la prueba, solo el texto a leer se presenta en inglés. Tanto las preguntas sobre el texto como las respuestas de los aspirantes se presentan en la lengua española. Esta medida de diseño procura asegurar que lo que se mide en la prueba sea la competencia lectora del aspirante en cuanto al texto. Aún cuando se cuenta con la capacidad de redactar preguntas en la lengua inglesa esta práctica incorpora la medición de otras competencias en los

aspirantes que no necesariamente están relacionadas con el constructo de comprensión de lectura. Adicionalmente, solicitar al aspirante que conteste en la lengua inglesa representa la evaluación de su competencia en la producción escrita del inglés: una habilidad que no se desea medir.

1.4. Características de los materiales utilizados

Para empezar a determinar el tipo de materiales que se pueden seleccionar para el examen, es obligado hacer referencia a las tareas de la vida real que se requieren de los estudiantes. Típicamente, durante los estudios de posgrado, los textos que requieren leer son relativamente largos, un artículo generalmente tiene más de 20 cuartillas y los libros más aún. Los estudiantes de un programa de maestría o especialidad también tienen que leer textos breves en forma de resúmenes para evaluar la necesidad de leer el texto completo. El registro lingüístico de dichos textos es bastante elevado, los artículos y libros se escribieron para ser leídos por nativos hablantes del inglés, especialistas en el área. El léxico será técnico y poco frecuente y la sintaxis compleja y elaborada.

Se puede decir que el estilo de redacción es propio de cada disciplina. Gran parte de los conocimientos que se imparten en cada una de ellas serán inaccesibles sin el conocimiento del léxico técnico adecuado. La valoración crítica de un texto también depende del dominio del lector de los supuestos básicos de la disciplina. Por lo anterior, queda claro que un individuo leerá mejor en un área que domina. Como resultado lógico, un candidato tiene mayores probabilidades de aprobar la prueba si el texto a leer viene de su disciplina. Sin embargo, garantizar a todos los sustentantes el derecho a leer dentro de su propia disciplina obligaría a diseñar una prueba especial para cada una de las 33 maestrías en la UAEM. Aún con este grado de especialización es posible que un aspirante a la Maestría en Administración tenga mayor afinidad con la Contabilidad Gerencial que con las Tecnologías de la Información. Elegir un tema de una u otra de estas disciplinas podría sesgar la prueba a favor o en contra de algún sustentante.

Otra alternativa que utilizan los grandes exámenes internacionales, como el TOEFL[5] y el IELTS[6] es ocupar materiales de lectura menos académicos, como entradas de enciclopedia, libros o revistas de divulgación y periódicos. Esta postura es algo extrema, ya que no se podrá diseñar una prueba especializada al gusto de cada sustentante, se diseña una prueba tan generalizada que ningún sustentante podría tener una ventaja sobre los otros, porque el texto del examen es de su área de conocimiento. Es sumamente práctica, pero sacrifica un aspecto importante de autenticidad, los textos que se ocupan ya no son textos académicos.

Otra alternativa para utilizar textos académicos sin tener que diseñar un examen para cada candidato, es realizar agrupaciones por disciplina, como ciencias naturales, o ciencias sociales. Como egresado de una licenciatura es probable que el aspirante a posgrado tenga un conocimiento relativamente profundo del área a la que se dedica, por lo que, como se ha dicho puede convertirse en una ventaja leer un texto de su área de especialidad.

Considerando lo anterior, cada versión de la prueba se realiza con dos textos académicos que pertenecen al género académico y de fuentes auténticas, es decir que se publicaron anteriormente para una audiencia lectora de hablantes nativos. Los textos de cada versión deben proceder de áreas temáticas distintas para minimizar el efecto del conocimiento previo. Los textos deben tener entre 500 y 1 000 palabras cada uno, sin exceder las 1 500 palabras en cualquier versión de la prueba. Las únicas modificaciones que se permiten a los textos

[5] TOEFL (*Test of English for Speakers of Other Languages*) es una prueba que desarrolla el *Educational Testing Service* que pertenece a la Universidad de Princeton. Se considera la medida por excelencia de dominio de inglés para propósitos académicos en los Estados Unidos.

[6] IELTS (*International English Language Testing System*) es una prueba que desarrolla Cambidge ESOL, una dependencia de la Universidad de Cambridge. Evalúa dominio del inglés para propósitos académicos.

son: la omisión de parte o partes del texto para respetar los límites superiores sin afectar a la integridad del texto, el reemplazo de verbos compuestos en inglés por su contraparte de raíz grecolatina y la inclusión de un glosario para algún elemento, léxico irreemplazable que se estime desconocido por los candidatos.

1.5. Características de las respuestas esperadas

En el contexto real de los estudios de posgrado las respuestas a la lectura en muchos de los casos podrán ser invisibles, es decir, en forma de significado reconstruido a partir del texto. En ocasiones las respuestas tendrán que ser escritas u orales, en forma de resúmenes o presentaciones. Claramente, no será posible medir la precisión con la que el lector ha interpretado el texto si queda en silencio. Sin embargo, tampoco sería práctico pedir a los estudiantes que produzcan respuestas tan largas como las que se les exige en clase, tanto por el tiempo que se tiene disponible para la prueba, como por la dificultad de calificar de manera confiable estas respuestas.

En la prueba, la amplitud de las respuestas de los aspirantes será mínima, limitada a la identificación de la solución correcta en los reactivos de opción múltiple. Se justifica esta elección en el proceso de diseño por varios motivos, en primer lugar, para nuestros propósitos lo importante es que el aspirante sepa leer en la lengua inglesa pero no se ha indicado que necesariamente tenga que saber escribir. En segundo lugar, las limitaciones al tiempo de aplicación del examen descartan la posibilidad de solicitar una respuesta más larga. Finalmente, las respuestas más complejas implicarían un sacrificio muy alto de la confiabilidad de la prueba.

Como mínimo cada versión de la prueba debe tener 25 reactivos entre los dos textos para garantizar cierto nivel de confiabilidad. En cuanto al tipo de reactivo se aceptan opción múltiple, respuesta breve y verdadero/falso. Los reactivos tipo verdadero/falso se limitan a 5 reactivos por versión, por el sacrificio que representan en cuanto la confiabilidad de la prueba.

1.6. Relación entre los materiales y las respuestas esperadas

Durante los estudios de posgrado, la selección de textos será adaptativa. En parte, la reacción del estudiante a un texto determinará el siguiente texto que lee. Si no entiende un texto, puede elegir uno más simple para su siguiente lectura. Si durante la lectura de un texto surge algún concepto interesante, podrá seleccionar textos futuros que complementan el tema anterior. En el contexto de la prueba esto no será posible, porque todos los lectores tendrán que leer los mismos textos. La relación entre los materiales y la respuesta es compleja, en ocasiones lo que se necesita es una respuesta que demuestre una comprensión detallada de alguna sección reducida del texto, en otras partes se requerirá de una respuesta más general a algún segmento relativamente largo de texto o quizá el texto entero. Sin embargo, la relación entre los contenidos de los textos y las respuestas deseadas es relativamente directa, no es necesario que el lector transforme el significado o critique el estilo del autor.

Al terminar un primer borrador de la prueba el diseñador especifica que habilidad se muestrea con cada pregunta, y compara este resultado con las ponderaciones determinadas. Si encuentra que su prueba no incluye los porcentajes correctos de preguntas sobre las diferentes habilidades debe adecuar su diseño a dichas especificaciones. Después del diseño inicial, la nueva versión se lee por dos diseñadores para asegurar la adecuación en su contenido, así como para eliminar cualquier error de estilo. Una vez que se ha establecido de esta manera la validez visual de la prueba se formatea y se aplica durante un periodo de tres meses, para posteriormente realizar las medidas de confiabilidad y validez de la prueba.

1.7. Discusión

Por las razones mencionadas, no todos los aspirantes toman la misma versión de la prueba. Por un lado, se usan versiones diferentes para tomar en consideración su especialidad en diversas áreas del conocimiento; por otro lado, para asegurar la confiabilidad de la prueba. A pesar de los riesgos de seguridad del examen, se han

diseñado varias versiones, las cuales cambian después de cierto volumen de aplicaciones. Por ello, es importante asegurar equivalencia de versiones, para garantizar que una calificación aprobatoria en una versión de la prueba signifique lo mismo en otra.

La interpretación de los resultados para todas las maestrías es la misma. La calificación mínima aprobatoria es 7.0 y no se reporta a menos que sea mayor que ésta. No tenemos conocimiento acerca de que haya alguna maestría que utilice los resultados de la prueba para tomar decisiones finales entre candidatos, es decir, si se toma en cuenta que el candidato obtenga 7.0 o 9.0, por ejemplo. Para evitar el uso de esta información se podría optar por reportar las calificaciones solamente en términos de "acreditada" o "no acreditada". Sin embargo, esta decisión resulta en menor satisfacción para el aspirante, quien se ha educado para esperar cierto detalle en sus resultados.

Se puede esperar algún efecto colateral de la selección de textos. Si algunos candidatos conocen mejor el tema de los textos utilizados que otros, eso les dará una ventaja sobre los otros, sin embargo, ese es un aspecto sobre el que el diseñador de la prueba no tiene control alguno.

3. EVALUACIÓN DE LA UTILIDAD DE LA PRUEBA DE COMPRENSIÓN DE LECTURA PARA SELECCIÓN DE CANDIDATOS A ESTUDIOS DE POSGRADO EN LA UAEM

Una vez descritas las características de la situación de la prueba, se les puede someter a un análisis más adecuado que aporte información acerca del logro de las diferentes dimensiones del concepto de utilidad. En las siguientes secciones se presenta dicho diagnóstico. Los datos sobre el comportamiento de la prueba representan un piloteo de 90 sujetos en el periodo comprendido entre noviembre 2007-marzo 2008. Cincuenta de las pruebas pertenecen al campo de Ciencias Médicas y 40 al área de Administración.

1.1. Confiabilidad

La confiabilidad de los exámenes de comprensión de textos para posgrado se apoya en la elección del nivel de dificultad de los textos, la extensión, el área de especialidad, el número, tipo y grado de dificultad de la preguntas, entre otros factores. La trascendencia de las decisiones tomadas con base en esta prueba de comprensión de lectura hace que su confiabilidad sea de suma importancia, lo que se ha buscado lograr a través de versiones múltiples elaboradas de acuerdo a criterios paralelos. Los textos deben ser de fuentes auténticas, de preferencia, provenientes de libros o revistas que no estén disponibles en Internet, por cuestiones de seguridad. La extensión de los textos tiene un mínimo y un máximo que son fijos. El número y tipo de reactivos también son fijos y existen estándares predeterminados que deben cumplirse. No es raro el caso de candidatos que presentan varias veces el examen y que, habiendo obtenido una calificación no aprobatoria, vuelven a presentarlo (en otra versión) sin apoyarse en cursos que los capaciten, lo que repercute nuevamente en un resultado reprobatorio.

Los exámenes se aplican en salones amplios y ventilados, con iluminación suficiente, en horarios matutinos o vespertinos. Se ha procurado evitar al máximo el ruido exterior que está en nuestras manos controlar, como pueden ser personas en el pasillo. Aunque se considera que las condiciones físicas son aceptables, se busca su mejora constante. Se sabe que es imposible aseverar que los exámenes de comprensión de lectura están libres de error en la medición, sin embargo, a través de un diseño cuidadoso y el constante monitoreo de los resultados obtenidos por los sustentantes, se trabaja para brindar resultados fidedignos.

En términos de esta prueba, la confiabilidad entre calificadores tendría que ser de 1.0, ya que los resultados se obtienen mediante una plantilla donde una sola respuesta es la correcta y siempre que un candidato haya puesto esta respuesta el calificador se lo pondrá como correcto. Esto siempre es cierto en cuanto a las pruebas de opción múltiple, tal vez la pregunta que en realidad es más

interesante es, si estos conocimientos y habilidades que se están midiendo con tanta confiabilidad tienen algo que ver con las habilidades que se desean medir; en otras palabras, esta cantidad de confiabilidad se logró a costo de la validez de constructo, lo que a su vez tendría un efecto negativo en el impacto de la prueba.

Para determinar la consistencia interna de la prueba se aplicó el coeficiente de Guttman sobre mitades de la prueba obteniendo: Ciencias Médicas (r = .84); Administración (r = .81). Esto representa un nivel aceptable de consistencia interna, es decir, que los reactivos de la prueba muestrean de manera relativamente consistente la misma habilidad. No es tan alto como se desearía en una prueba de este tipo sobre todo dada la importancia de las decisiones que se toman con base en sus resultados. No obstante, es importante reconocer la probabilidad de las fuentes de error que se comentaron en la primera sección de este artículo, sobre todo, en el caso de reactivos de opción múltiple, el efecto sobre la calificación de que el sujeto adivine la respuesta correcta o que conteste correctamente por saber que los distractores son incorrectos.

Sobre los mismos datos del piloteo se realizó un análisis de reactivos para determinar si los reactivos funcionaban correctamente. Se calculó el índice de dificultad de los reactivos, obteniendo buenos resultados. La interpretación de los resultados del índice de dificultad se realizó de acuerdo a los parámetros ofrecidos por MacNamara (1996). Reactivos que presentan índices de 1.00-0.85 son muy fáciles; de 0.84-0.70, fáciles; de 0.69-0.30, óptimos; de 0.29-0.15, difíciles y menores de 0.14 son muy difíciles. Una prueba bien diseñada deberá presentar una mayoría de reactivos en condiciones óptimas. A continuación se presentan los datos de ambas pruebas.

Tabla 1
Proporción de reactivos de acuerdo al índice de dificultad

	Muy fácil	Fácil	Óptimo	Difícil	Muy difícil
Ciencias Médicas	8	36	48	0	8
Administración	4	28	52	8	8

Los resultados no son idóneos. Aun cuando se ha logrado una mayoría de reactivos en condiciones óptimas en ambas versiones de la prueba hay demasiados reactivos fáciles. Posteriormente, se calcularon los índices de discriminación sobre los mismos datos del piloteo para determinar su capacidad de discriminar entre buenos y malos lectores. En este caso los resultados no fueron tan alentadores. Solamente un porcentaje menor de los reactivos se pueden considerar buenos en este aspecto, mientras la mayoría son dudosos o pobres. Los resultados se presentan a continuación:

Tabla 2
Proporción de reactivos de acuerdo al índice de discriminación

	Pobre	Dudoso	Bien
Ciencias Médicas	16	56	28
Administración	08	66	26

Se revisaron uno por uno los reactivos, para identificar cualquier problema que sea causa tanto del alto porcentaje de reactivos fáciles, como la falta de discriminación entre los reactivos de la prueba. Este tipo de problemas incluyen errores ortográficos o la existencia de dos respuestas correctas. Era improbable que se encontrara algún problema de esta naturaleza, dado que el instrumento ya se había revisado por dos otros diseñadores y no surgió ningún problema de este tipo. En el caso de un reactivo que se había estimado demasiado fácil se identificó un posible problema, la respuesta correcta era un poco más larga que los distractores, lo cual se corrigió y se están esperando resultados del piloteo posterior a este cambio.

1.2. Validez del constructo

La finalidad de los exámenes de comprensión de lectura para posgrado es muy específica: valorar la capacidad lectora del candidato en otra lengua para realizar un posgrado que puede requerirle, en mayor o menor grado, de dicha habilidad. No se busca dar un veredicto sobre el dominio de la lengua en cuanto a otras habilidades; no es

un indicador de la capacidad para escuchar, hablar o escribir en ese idioma, únicamente se desea medir su habilidad de comprensión de lectura. Es por ello que, en términos muy simplistas, el examen contiene textos y preguntas con respecto al texto, esto es para conocer si se ha comprendido el contenido de éste. Las preguntas no requieren conocimientos previos del área, ya que no son exámenes de conocimientos generales, sino de lengua, y si bien es cierto que el conocimiento previo puede ayudar al sustentante, no son factor definitivo. Se utilizan textos del área del posgrado al que se aspira porque se presupone que esos son el tipo de texto que tendrán que leer durante sus estudios.

Los datos sobre confiabilidad solamente miden la consistencia interna de la prueba, es decir, el grado en que los reactivos estén midiendo el mismo constructo; no indica que este constructo sea el que se deseaba medir cuando se diseñó la prueba. Para determinar la validez lograda en el instrumento es necesario hacer referencia a una medida independiente. Para la mayoría de las pruebas de selección a un nivel educativo, la medida idónea de validez predictiva es el éxito del candidato seleccionado durante sus estudios. En este caso, aunque consideramos que la competencia lectora en inglés es importante para alcanzar cierta profundidad en los estudios en posgrado, también es una competencia secundaria que no necesariamente es una causa directa del éxito.

Por este motivo se realizó una comparación entre el desempeño en este examen, con el desempeño en otros exámenes estandarizados. Se decidió medir en qué grado el instrumento utilizado "aprueba" a los aspirantes que sepan leer en la lengua extranjera y "reprueba" a los que no. Como medida independiente se eligió la sección de comprensión de textos del TOEFL (ETS) por su prestigio en el campo de la evaluación de lenguas. Se aplicó esta prueba a 9 aspirantes y se comparó su resultado en TOEFL con su resultado en el examen de comprensión de lectura. Pruebas estadísticas (Wilcoxon) dieron un resultado significativo indicando que ambos instrumentos miden el mismo constructo.

1.3. Autenticidad

Para evaluar el grado en que se logra la autenticidad en la prueba, es necesario saber para qué les va a servir la comprensión de textos en inglés a los aspirantes una vez que ingresan a los estudios de posgrado en la UAEM. De lo anteriormente mencionado en la sección sobre el contexto de la prueba queda claro que no se sabe con certeza cuánto se exige leer en lengua inglesa al estudiante de posgrado. Seguramente varía considerablemente entre las diferentes disciplinas y probablemente puede variar de acuerdo al maestro de cada asignatura. Lo que se ha hecho a este respecto es establecer un nivel medianamente complejo de comprensión de textos académicos que pudiera servir de manera general como criterio a alcanzar. Este nivel es mayor a la comprensión meramente literal, pero no ahonda en asuntos de crítica de contenido y/o estilo que representarían el nivel más avanzado de comprensión de lectura.

Los materiales de lectura que forman la base de la prueba no son tan extensos como los artículos y libros que el aspirante tendrá que leer para sus estudios, pero por motivos de practicidad sería imposible incluir textos que fueran tan largos. En cualquier caso, una evaluación siempre debe proceder en relación de un muestreo parcial de la competencia a evaluar. Otros exámenes de la competencia lectora incorporan textos desde una oración, con la finalidad de variar la muestra de tipos de texto y temáticas. En este caso no es necesario muestrear muchos tipos de texto ya que las encuestas de los coordinadores de posgrado aclaran que el rango de textos que se tendrá que leer se reduce básicamente a artículos y libros de texto.

Otra dimensión de la autenticidad es el género de los materiales. En este caso, los textos utilizados siempre se redactaron originalmente para hablantes nativos de la lengua inglesa, pero se han modificado de su versión original de acuerdo a los criterios mencionados en el apartado 2.4. Estas modificaciones fueron necesarias porque en realidad la tarea solicitada en el examen no es similar a la tarea de leer en la vida real, ya que los aspirantes normalmente tendrían acceso a un diccionario. En este sentido, la autenticidad del examen es algo baja.

Finalmente, la autenticidad de la tarea que tienen que realizar los aspirantes durante el examen también tiene un nivel relativamente bajo. En lugar de producir apuntes para responder a sus propios problemas de investigación el candidato tiene que contestar a preguntas que han sido planteadas por el diseñador de la prueba. El formato del reactivo tiene poca relación con lo que se les exige a los estudiantes de posgrado cuando leen un texto. Morrow (1979) ha comentado que es imposible diseñar un *test* auténtico que sea también confiable y el análisis anterior parece apoyar esta postura.

Sin embargo, el modelo de Bachman y Palmer (1996) no trata de sacrificar una cualidad a costa de otra, sino de valorar la importancia relativa de cada una, considerando el contexto de la evaluación. En este caso, por la importancia de las decisiones que se toman con base en los resultados de la prueba la confiabilidad adquiere mayor peso que la autenticidad, sin embargo, se han intentado incorporar temáticas y tipos de textos que serían normales durante los estudios de posgrado.

1.4. Interactividad

Los exámenes de comprensión de texto que se aplican a los aspirantes a ingresar a algún posgrado de la UAEM se efectúan de acuerdo con el área específica en que se van a desenvolver; es así que se llevan a cabo exámenes en nueve diferentes áreas: Arquitectura-Diseño, Económico-Administrativas, Derecho, Educación, Lingüística, Medicina, Odontología, Ciencias Naturales y Ciencias Sociales. El grado de interactividad de la prueba, definido como la medida en que el instrumento interactúa con las características individuales de los aspirantes, tiene un peso importante en los resultados. En la primera sección de este artículo se identificaron tres categorías principales: conocimiento previo del tema, competencia lingüística y afectividad.

El primer aspecto del diseño que garantiza un buen nivel de interactividad es la selección de los temas de acuerdo con el interés de los aspirantes. Es probable que alguien que considere realizar estudios de posgrado en alguna área disciplinaria ya tenga cierto conocimiento del tema. El problema del diseñador es encontrar una

lectura que aproveche los esquemas de conocimiento anteriores del lector, pero que no haya leído antes. Se procura incluir temas que sean novedosos o poco comunes en el área pero que se abordan con un mínimo de terminología. Por supuesto que para disfrutar la lectura es esencial que el nivel de competencia en la lengua extranjera le permita leer con cierta fluidez. No obstante, es posible que el aspirante tenga un bajo nivel de competencia lingüística en inglés, aún así deberá hacer uso de otros recursos, como su conocimiento del tema y de sus esquemas afectivos, para lograr la comprensión del texto, lo que garantizaría que podrá desempeñarse con éxito durante su posgrado. Finalmente, en cuanto al aspecto afectivo de la interactividad se eligen textos sobre temas agradables que pueden captar la atención del aspirante para que su lectura no sea aburrida.

1.5. Impacto

Hemos mencionado que el impacto de la prueba puede ser a nivel macro, con un efecto a nivel de sistema educativo, y micro, aquel que afecta de manera directa al individuo. El examen de comprensión de lectura para aspirantes a posgrado tiene una inobjetable relevancia a nivel individual. Aprobar o no esta prueba significa para el aspirante continuar su vida profesional o académica, al menos en esta institución. Esto significa un alto impacto a nivel micro. A nivel macro, el análisis del impacto se puede realizar en distintas vertientes: nivel universitario o institucional, o nivel de organismo académico.

Para la UAEM, que los aspirantes a posgrado cuenten con la habilidad suficiente para aprobar este examen es imprescindible, tanto que es un requisito indispensable para ser admitido en cualquiera de los programas ofrecidos. Los resultados obtenidos por los sustentantes en general son bastante pobres. Más del 60% de los aspirantes reprueban, de lo que podemos concluir que hace falta mayor preparación a nivel de educación superior para obtener resultados satisfactorios al momento de acceder a estudios de posgrado. Quizá el aspecto más importante del impacto de la prueba es valorar la medida en que, para aprobar, el aspirante tenga que mejorar su nivel de comprensión de lectura en inglés, es decir, la validez de constructo de la prueba. Una prueba con alto nivel de validez, mejora

el ajuste entre la tarea de la vida real, en este caso la lectura de textos, y lo que el aspirante tiene que aprender para obtener un resultado aprobatorio. Si esto se logra, tendrá un efecto positivo para elevar los niveles de comprensión. Por ejemplo, si la prueba muestra una gama amplia de estrategias, no será posible aprobarla mediante el desarrollo de una sola estrategia. Siempre que la validez de constructo de una prueba sea alta, su impacto sobre la docencia tenderá a ser provechoso.

Mientras la prueba se encuentra en su fase de diseño hay poco que se puede indagar sobre el impacto de ésta. Para efectuar esta medición es importante que se haya aplicado la prueba durante cierto tiempo, además de que se cuenten con datos sobre el estado de las cosas antes de que la prueba se rediseñó.

1.6. Practicidad

La Facultad de Lenguas de la UAEM recibe alrededor de 1 000 aspirantes al año para tomar el examen de comprensión de lectura en inglés para ingreso a posgrado, lo cual implica una cuantiosa aplicación de recursos tanto humanos, como materiales. Esta dimensión de la utilidad de la prueba cobra, precisamente por lo anterior, una especial relevancia, para este caso. Una vertiente importante del desarrollo de la prueba ha sido la reducción de gastos de administración, tanto para los candidatos, como para la facultad. El examen tiene una duración máxima de dos horas y el Departamento de Evaluación ofrece lineamientos claros a solicitud del aspirante sobre la forma de pago y las fechas de aplicación. El proceso de comunicación de resultados es oportuno y eficiente. Los cuadernos de preguntas de la prueba que contienen los textos y las preguntas se vuelven a utilizar en aplicaciones consecuentes, siempre realizando una revisión pormenorizada para suprimir cualquier anotación que pudiera haber registrado algún candidato, reduciendo de esta manera los gastos en fotocopias. La homogeneidad en cuanto al número y tipo de reactivos permite el diseño de una hoja de respuestas única que puede valerse para cualquier versión de la prueba.

CONCLUSIONES

En este artículo se presenta un diagnóstico del grado de utilidad de la prueba de comprensión de lectura en lenguas extranjeras de la Facultad de Lenguas, según el concepto propuesto por Bachman y Palmer (1996). En general, se puede decir que la prueba es útil en relación a las dimensiones analizadas, sin embargo, resulta necesario difundir mayor información sobre su diseño. En este sentido, proponemos dos acciones específicas que consideramos aumentarán la transparencia de operación de la prueba. Ambas acciones se encuentran actualmente en vías de desarrollo:

- Una guía del usuario que servirá a dos audiencias principalmente: 1) Para los docentes que preparan a los aspirantes. Ellos deberán conocer puntualmente la estructura de la prueba para optimizar la eficiencia de los cursos que imparten. 2) Para las personas que toman decisiones con base en la prueba, en las Coordinaciones de Posgrado de cada Facultad, por ejemplo, para que puedan determinar el significado de los resultados obtenidos.

- Una guía para el aspirante, que incluya una breve descripción del propósito de la prueba y materiales de práctica para mejorar el índice de aprobación y el nivel de satisfacción del aspirante con la experiencia de la prueba.

Ahora bien, una vez propuesta la vía de comunicación entre los examinadores y los usuarios de la prueba en las Coordinaciones de Posgrado será importante hacer esta comunicación recíproca. En relación al uso de los resultados de la prueba, surgen las siguientes interrogantes: ¿es significativo para el organismo, o facultad, el resultado obtenido por el sustentante? o, ¿sólo es un requisito impuesto por la Universidad a nivel central?, ¿satisface el examen los requerimientos del organismo en particular?, ¿cuál es la necesidad real de cada organismo en cuanto al nivel de comprensión de textos en segunda lengua de sus aspirantes? Podemos tomar como un ejemplo a la Facultad de Lenguas, donde el resultado obtenido en este examen es de gran relevancia, ya que los dos programas de maestría ofrecidos

por esta facultad requieren de una habilidad de comprensión de textos en inglés altamente desarrollada, puesto que gran parte del material está escrito en esta lengua. Pero ¿sucede lo mismo en otras facultades? Debe extenderse la certeza sobre la pertinencia de estos resultados a otros planes de posgrado a través de estudios sobre las necesidades en materia de lectura en inglés que perciben las coordinaciones de posgrados en cada facultad.

Sin dudar de la importancia de las necesidades percibidas por las coordinaciones de posgrado al interior de cada Facultad es importante establecer cuáles son las necesidades reales. Para esto, se requiere hacer valoraciones dentro de las facultades, para que éstas nos puedan informar sobre las exigencias reales en cuanto a la comprensión de lectura en inglés. Dar un paso adelante para la obtención de un grado académico requiere también mayor capacidad de lectura, inferencia y análisis crítico, en este caso a partir de textos en una segunda lengua, los exámenes de comprensión de textos buscan dar razón de ésta habilidad para que el aspirante a estudios de posgrado cumpla con los objetivos educacionales. Una pregunta básica para investigaciones futuras sería: ¿un candidato que lee en inglés es un estudiante de posgrado más exitoso?

Incluso cuando los datos del análisis de la confiabilidad y validez de la prueba fueron aceptables podrían mejorarse. Sobre todo considerando la importancia de las decisiones que se toman sobre los resultados de la prueba. Para todos los usuarios valdría la pena aumentar el número de reactivos que se incluyen en cada versión, lo cual podría tener un impacto positivo. Sin embargo, tratándose de pruebas de comprensión de lectura es difícil aumentar la cantidad de texto que hay que leer, por lo que no es posible agregar otro texto al examen. Aún en su estado actual, los aspirantes generalmente toman las dos horas completas que se permiten para contestar al instrumento. Agregar otro texto podría aumentar el tiempo para responder el examen en una hora, lo cual sería contraproducente, considerando el cansancio como una fuente de error de medición. Entonces para aumentar el número de reactivos la única opción es explotar cada texto utilizado al máximo y sacar más preguntas.

Finalmente, el modelo de diagnóstico presentado en este artículo sirve de modelo para la evaluación de pruebas en general. Tomar en consideración aspectos más allá de la confiabilidad y la validez nos permiten desarrollar una perspectiva amplia sobre los instrumentos de evaluación que utilizamos. La amplitud del análisis del estudio de utilidad ofrece una visión de la prueba en relación a su comportamiento dentro de la sociedad vista como sistema, en lugar de la visión reducida y atomística ofrecida por métodos de análisis más cuantitativos como los estudios que solo miden confiabilidad. De la misma manera, un estudio de utilidad enfoca su análisis en los problemas de evaluación a los que da solución la prueba, en lugar de cuestiones meramente mecanicistas de su operación.

BIBLIOGRAFÍA

Alderson, J. C. (2000), *Assessing reading*, Cambridge University Press, Cambridge.

_____ (2003), *Language test construction and evaluation*, Cambridge University Press, Cambridge.

Alderson, J. C. y D. Wall, (1993), "Does washback exist?", en *Applied linguistics* 14(2), pp. 115-129.

Bachman, L. F. (1990), *Fundamental considerations in language testing*, Oxford University Press, Oxford.

Bachman, L. F. y A. S. Palmer (1996), *Language testing in practice: designing and developing useful tests*, Oxford University Press, Oxford.

Canale, M. y M. Swain (1980), "Theoretical bases of communicative approaches to second language teaching and testing", en *Applied linguistics* 1, 1, pp. 1-47.

Canale, M. (1983), "From communicative competence to communicative language pedagogy", en J. C. Richards y R. Schmidt (eds.) *Language and communication*, Longman, London, pp.2-27.

Cheng, L. (2005), "Changing language teaching through language testing: a washback study", en *Studies in language testing*, vol. 21, Cambridge, Cambridge University Press.

Faerch, C. y G. Kasper (1984), Two ways of defining communication strategies, en *Language learning: a journal of applied linguistics*, 5 (3), pp. 214-225.

Heaton, J. B. (1988), *Writing english language tests*, Longman, Londres.

Hughes, A. (2003), *Testing for language teachers*, Cambridge University Press, Cambridge.

MacNamara, T. (1996), *Measuring second language performance*, Longman, Londres.

Morrow, K. (1979), Communicative language testing: revolution or evolution?, en C. J. Brumfit y K. Johnson, *The communicative approach to language teaching*, Oxford University Press, Oxford.

Moskal, Barbara M. y Jon A. Leydens (2000), "Scoring rubric development: validity and reliability", en *Practical assessment, research & evaluation*, 7(10), retrieved april 20, 2008, from http://PAREonline.net/getvn.asp?v=7&n=10

Rudner, Lawrence M. (1994), "Questions to ask when evaluating tests", en *Practical assessment, research & evaluation*, 4(2), retrieved april 18, 2008, from http://PAREonline.net/getvn.asp?v=4&n=2

Shohamy, E. (2001), *The power of tests: a critical perspective on the uses of language tests*, Pearson Education, Harlow.

Underhill, N. (1987), *Testing spoken language: a handbook of oral testing techniques*, Cambridge University Press, Cambridge.

Ur, P. (1984), *Teaching listening comprehension*, Cambridge University Press, Cambridge.

Weideman, A. (2006), "Transparency and accountability in applied linguistics", en *South african linguistics and applied language studies*, 24/1, pp. 71-86.

Widdowson, H. G. (1979), *Teaching language as communication*, Oxford University Press, Oxford.

Fecha de recepción: 25/05/2008
Fecha de aprobación: 23/09/2009