Revista de Investigación Educativa, 2003, Vol. 21, n.º 1, págs. 135-155

SISTEMAS DE PRESENTACIÓN DE LOS RESULTADOS DE LAS EVALUACIONES DEL RENDIMIENTO EDUCATIVO: APLICACIÓN AL ESTUDIO INTERNACIONAL DE LA LENGUA INGLESA EN LA EDUCACIÓN SECUNDARIA

*Guillermo Gil Escudero**

*Juan Carlos Suárez Falcón***

RESUMEN

La presentación de los resultados de rendimiento educativo de las evaluaciones globales de los sistemas educativos ha venido realizándose de modo fundamental mediante dos sistemas: la presentación de resultados en términos de las puntuaciones directas o los porcentajes que los estudiantes obtienen en las pruebas de rendimiento y la presentación de resultados en términos de escalas de rendimiento y niveles de rendimiento basados en la Teoría de Respuesta al Ítem (TRI). En este trabajo se presentan las características de ambos sistemas y se analizan sus ventajas e inconvenientes. Asimismo, se ejemplifica la aplicación del sistema basado en la TRI con los resultados obtenidos por la muestra de alumnos españoles que participaron en 1996 en el Estudio Internacional sobre la Enseñanza y el Aprendizaje de la Lengua Inglesa en la Educación Secundaria.

Palabras clave: *Evaluación Educativa, Rendimiento en Lengua Inglesa, Educación Secundaria, Teoría de la Respuesta al Ítem.*

---

* Instituto Nacional de Calidad y Evaluación (INCE).

** Universidad Nacional de Educación a Distancia (UNED). Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad Nacional de Educación a Distancia (UNED). Ciudad Universitaria, s/n. 28040 MADRID. E-mail: jcsuarez@psi.uned.es

ABSTRACT

In this paper a procedure, based on Item Response Theory (IRT), to display achievement outcomes from international and national educational assessments is presented. This procedure, which combines normative and criterial evaluation, is compared with the classical approach which uses direct scores or the percentages of correct responses. As an example, the procedure is applied to the secondary level student's outcomes from the International Study in Teaching and Learning of English Language. The obtained results and the advantages and limitations of classical and IRT procedures are described and discussed.

Key words: *Educational Assessment, English Language Performance, Secondary School, Item Response Theory.*

INTRODUCCIÓN

Tradicionalmente los resultados de rendimiento educativo obtenidos a través de pruebas estandarizadas se han presentado en términos del número de respuestas correctas proporcionadas por los alumnos y/o en términos de sus porcentajes de respuestas correctas, bien para cada una de las preguntas consideradas aisladamente o bien para el conjunto total o para diversos subconjuntos de las preguntas de las pruebas.

La presentación de los resultados pregunta por pregunta suele ser de interés para los especialistas en currículum y en construcción de pruebas de rendimiento, pero no suelen ser de mucha utilidad para los administradores de la educación y para el público en general, que prefieren una presentación más general y resumida de los resultados del conjunto de las pruebas obtenidos en los estudios.

Para la comprensión del significado de los resultados en términos del número de respuestas correctas, o puntuación directa, es necesario tener en cuenta el número de preguntas de las pruebas ya que estas puntuaciones directas, de carácter absoluto, sólo tienen significado con referencia al tamaño de las mismas, es decir, al número total de preguntas. Sin embargo, los resultados presentados en términos de porcentajes de respuestas correctas proporcionan una información, de carácter relativo, que es independiente del tamaño de la prueba.

La utilización de estos dos sistemas de presentación de resultados tiene tanto ventajas como inconvenientes. Estas ventajas e inconvenientes se han discutido y presentado en múltiples trabajos, por ejemplo en Beaton y Johnson (1992), Beaton y Zwick (1992), Bock, Mislevy y Woodson (1982), Hambleton y Cook (1977) y en Hambleton y Jones (1993).

Sin duda, es de fácil e intuitiva comprensión saber que los alumnos responden, como promedio, por ejemplo, a veinte preguntas en una prueba que consta de un total de treinta preguntas, o saber que los alumnos responden correctamente, por ejemplo, a ochenta de cada cien preguntas que se les plantean de una determinada materia y nivel. No obstante, esta comprensión intuitiva puede llevar a interpretaciones erróneas

sobre el significado de los resultados, especialmente, cuando se realizan comparaciones con base en los mismos.

El porcentaje medio de aciertos indica cuál es el tanto por ciento medio de aciertos de los alumnos examinados en una prueba determinada, aunque si los mismos alumnos, con el consiguiente mismo nivel de competencia, hubieran contestado a otra prueba, construida con preguntas bien más fáciles o bien más difíciles, habrían obtenido valores de porcentaje de aciertos más altos o más bajos de modo correspondiente. Ello se debe a que el porcentaje de aciertos no tiene en cuenta, como unidad de medida, la dificultad de los ítems que la componen, por lo que los resultados son directamente dependientes de cada selección específica de preguntas. Dicho de otro modo, la distribución de los porcentajes de respuestas correctas, y en consecuencia, su porcentaje promedio, depende de la distribución de las dificultades de las preguntas que forman la prueba, por lo que las puntuaciones en términos de porcentajes de preguntas correctas varían en función de las características de las preguntas de la prueba.

Además, teóricamente —aunque también muy improbablemente—, podría darse el caso de alumnos que poseen niveles de competencia claramente distintos y que, sin embargo, obtienen la misma puntuación, directa o en términos de porcentajes, en una prueba de rendimiento. Sería este el caso en el que un alumno ha contestado correctamente, por ejemplo, a las diez preguntas más fáciles de la prueba mientras que otro ha contestado correctamente a las diez preguntas más difíciles de la misma. De modo que se obtendría una puntuación idéntica a pesar de que los alumnos tuviesen diferentes niveles de habilidad.

Un problema adicional que presenta la utilización de puntuaciones directas y porcentajes consiste en la imposibilidad de comparar con garantías de validez entre resultados obtenidos con pruebas diferentes, pudiéndose solamente comparar los resultados entre diferentes grupos de sujetos o entre los mismos grupos a lo largo del tiempo si se utiliza la misma prueba o una forma paralela de la misma. Este hecho es de especial importancia en la evaluación educativa del rendimiento de los estudiantes a gran escala ya que gran parte de su capacidad para ofrecer información relevante para el mundo de la educación se fundamenta en las comparaciones entre grupos y en el estudio de la evolución de los resultados a través del tiempo.

Por otro lado, los porcentajes medios de respuestas correctas proporcionan información sobre los valores medios que obtiene una población, pero no aportan ninguna información sobre la variabilidad de los resultados dentro de la misma. Puede darse el caso de dos grupos de alumnos que tengan idéntico porcentaje de respuestas acertadas frente a una prueba, por ejemplo, el cincuenta por ciento, aunque en uno de ellos las respuestas correctas de los alumnos vayan del diez al noventa por ciento mientras que en el otro los resultados de los alumnos sean más homogéneos y los porcentajes de acierto individuales de los estudiantes varíen entre el cuarenta y el sesenta por ciento.

Asimismo, el porcentaje medio de aciertos no informa sobre lo que saben los alumnos de modo absoluto en relación con el currículum o con los objetivos de la materia, sino que proporciona información sobre el nivel de rendimiento en relación con la prueba específica que se utiliza en un estudio particular. Esto se debe a que, obviamente, no es posible evaluar todos los contenidos que pueden incluirse en una materia

en un determinado nivel educativo, por lo que habitualmente se selecciona una muestra de las posibles preguntas o ejercicios que corresponden a la misma con el objetivo de que representen al conjunto de los contenidos de dicha materia y nivel. La capacidad para extrapolar los resultados obtenidos con una prueba específica al conjunto de una materia y a sus objetivos depende del grado en el que dicha prueba represente adecuadamente los objetivos y el currículum de la materia en la correspondiente etapa educativa así como que tenga el nivel de dificultad adecuado a la misma.

Sin embargo, es difícil poder asegurar que una prueba represente adecuadamente el conjunto de contenidos de una materia, mida el logro de sus objetivos y posea el nivel de dificultad correspondiente a un determinado nivel educativo. Este hecho se debe a que, generalmente, los currícula definen un conjunto de contenidos, procedimientos y actitudes para un determinado nivel, pero no delimitan con precisión, ni probablemente deben hacerlo, las preguntas que los alumnos deberían saber responder ni el grado de dificultad de las mismas, de modo que pueda estimarse inequívocamente si los alumnos han alcanzado los objetivos propuestos en el currículum. Por ello, puede ser equívoco asociar cierto valor de la proporción de aciertos en una prueba determinada con el éxito o el fracaso en una materia concreta.

La definición de éxito o fracaso escolar en una materia se fundamenta, o debería fundamentarse, en criterios que son externos a las pruebas de rendimiento en sí mismas, ya que la atribución de éxito o fracaso no puede ser dependiente de la dificultad específica de las pruebas que se utilizan para estimar el rendimiento, ya que pruebas similares en cuanto a contenido pero con diferentes grados de dificultad proporcionarían resultados contradictorios para los mismos alumnos, es decir, con unas pruebas se consideraría que han alcanzado los objetivos mientras que con otras se consideraría que han fracasado.

Para la adecuada definición de éxito o fracaso en relación con una materia es necesario definir criterios operacionales que permitan delimitar qué conjunto de conocimientos, capacidades, habilidades, destrezas o aptitudes, son suficientes para considerar que se han alcanzado los objetivos de una etapa educativa. Estos criterios operacionales deben estar vinculados a preguntas específicas con objeto de que se pueda evaluar de modo inequívoco qué niveles de rendimiento alcanzan los alumnos y qué niveles de los mismos se consideran como éxito o fracaso. Este proceso de especificación de niveles de rendimiento se conoce como la formulación de estándares educativos (González y Beaton, 1994; Linn y Baker, 1995; Linn y Dunbar, 1992; Masters y Forster, 1996 a y b; Phillips, 1994).

En otro nivel de análisis, un hecho importante consiste en que las escalas de porcentajes, debido a que son escalas ordinales, no suelen ser lineales en relación con variables externas, de modo que una determinada diferencia entre porcentajes no es equivalente a lo largo de la escala, es decir, la diferencia entre dos resultados, por ejemplo del 10 y el 20 por ciento, no es equivalente a la misma diferencia en términos numéricos entre otros dos resultados, por ejemplo del 50 y 60 por ciento. Así, una variación de un punto porcentual en los extremos de la escala de porcentaje, por ejemplo entre el 98 y el 99 por ciento, representa un efecto más grande que un cambio, numéricamente equivalente, también de un punto, en la parte media de la escala, por ejemplo entre el

49 y el 50 por ciento. Una manera de afrontar el problema derivado de esta característica de no linealidad de las escalas porcentuales es someter a la escala de porcentajes a las trasformaciones necesarias para que sus relaciones con otras variables resulten lineales.

PRESENTACIÓN DE RESULTADOS EN TÉRMINOS DE ESCALAS DE RENDIMIENTO BASADAS EN LA TEORÍA DE RESPUESTA AL ÍTEM

Gran parte de las dificultades que se presentan con la utilización de las puntuaciones directas y los porcentajes, basadas en el marco teórico de la denominada Teoría Clásica de los Test, o TCT (Gulliksen, 1950; Lord y Novick, 1968), se han resuelto con la utilización de escalas de rendimiento derivadas de la aplicación de la Teoría de la Respuesta al Ítem, o TRI. En los trabajos de Hambleton (1989a; 1989b), Hambleton y Swaminathan (1985), Hambleton, Swaminathan y Rogers (1991), Hulin, Drasgow y Parsons (1983), Lord (1980), Lord y Stocking (1990), Martínez Arias (1995), Muñíz (1994), Van der Linden y Hambleton, (1997) o Wright y Stone (1979) se exponen con detalle los fundamentos y desarrollos de la TRI. Asimismo, varios estudios han comparado en profundidad las propiedades y características de la TRI con las de la TCT (Fan, 1998; Hambleton y Jones, 1993; Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Rogers, 1991; Martínez Arias, 1995).

La Teoría de Respuesta al Ítem postula que un conjunto de rasgos, capacidades o habilidades subyacen al rendimiento de un sujeto en una prueba y que puede especificarse matemáticamente la relación entre el rendimiento de los sujetos frente a los ítems y los rasgos subyacentes de habilidad, de modo que la TRI relaciona la capacidad o habilidad de un sujeto que se intenta medir con la probabilidad de que dicho sujeto responda correctamente a cada una de las preguntas de la prueba.

Así como en la TCT la unidad de referencia es la prueba en su conjunto y se fundamenta en un modelo lineal, en la TRI el elemento de referencia es el ítem, o pregunta individual, y se fundamenta en modelos no lineales. El modelo más general de la TRI, o modelo logístico de tres parámetros, tiene en cuenta la dificultad de cada pregunta, su capacidad de discriminación entre los sujetos que tienen un mayor o menor dominio de la habilidad medida y la pseudo adivinación o grado de acierto por azar.

Una primera ventaja de la TRI sobre la TCT consiste en que los valores obtenidos en la estimación de la habilidad de los sujetos son independientes de la dificultad de los ítems de la prueba, de su número y de las características del subconjunto específico de preguntas que la compongan, siendo también independiente del número de sujetos examinados, así como de su variabilidad. En paralelo, las estimaciones de los parámetros de discriminación y dificultad son, asimismo, independientes de los grupos examinados.

Una segunda ventaja consiste en que las escalas TRI tienden a tener relaciones lineales con variables externas, al no producirse los efectos suelo y techo de los porcentajes y puntuaciones directas. Técnicamente, es interesante el hecho de que las escalas TRI permiten una mayor precisión en el cálculo de determinadas propiedades (por ejemplo, fiabilidad, errores estándar) de los instrumentos de medida y de los resultados.

Además, las escalas desarrolladas con base en la TRI permiten medir las tendencias a lo largo del tiempo y comparar entre grupos con una métrica consistente al situar la TRI a todos los estudiantes en una escala común, incluso cuando el conjunto de ítems evoluciona de unas evaluaciones a otras y, en consecuencia, los alumnos no han respondido exactamente al mismo conjunto de preguntas. Este enlace entre escalas se realiza mediante la utilización de ítems puente, o ítems de anclaje, que comparten diferentes pruebas. Esta capacidad de la TRI permite que las pruebas evolucionen a lo largo del tiempo y se puedan adaptar a cambios en los niveles educativos o en sus currículos (Mislevy, 1993).

Las escalas basadas en la TRI permiten desarrollar una descripción del rendimiento con significado, a la vez que eluden los problemas de las estadísticas basadas en el porcentaje de respuestas correctas, al permitir estimar la distribución de destrezas entre los estudiantes y asignar interpretaciones con significado a los niveles de puntuación en términos del rendimiento predicho sobre ejercicios específicos y en términos de tareas educativamente relevantes, mediante la creación y definición de escalas y la delimitación de diversos niveles de rendimiento en dichas escalas.

Las escalas de rendimiento son el resultado de un proceso por el que el conjunto de datos que contiene información compleja se reduce a un número limitado de estadísticos, de modo que las respuestas obtenidas de un alumno se resumen en una única puntuación que representa su rendimiento general, con el fin de que los resultados puedan ser manejables e interpretables (Keeves, 1990, 1992; Masters y Forster, 1996 a).

Las escalas de rendimiento posibilitan la situación en la misma escala no solamente de los alumnos sino también de las preguntas, de modo que el significado de las puntuaciones de la escala puede determinarse mediante la inspección del contenido de las preguntas que resuelven correctamente los sujetos situados en los diferentes tramos de las escalas, sin necesidad de hacer referencia a una norma —es decir, a la situación que ocupa el rendimiento de un sujeto en comparación con el del resto de los sujetos de la muestra—, al asignarse cada pregunta de una prueba a un nivel determinado de competencia y de ese modo definir niveles de habilidad de los alumnos que alcanzan una determinada puntuación, en términos de los conocimientos, capacidades o tareas que los alumnos deben dominar para resolver adecuadamente las preguntas.

Esta posibilidad brinda la oportunidad de combinar la evaluación referida a una norma con evaluación referida a un criterio ya que, además de proporcionar una descripción del rendimiento en relación con los resultados del resto de los alumnos en el grupo evaluado, proporciona una descripción del rendimiento en términos de los conocimientos, contenidos y habilidades que los alumnos poseen.

La situación de las preguntas de una prueba en la escala representa en cierto grado un mapa de las habilidades y conocimientos de los alumnos en el que se sitúan los mismos en el orden que son adquiridos según los alumnos progresan e incrementan su rendimiento en un área determinada (Masters y Forster, 1996a y b).

Por todo ello, y dadas las ventajas de la TRI, se debería utilizar de modo fundamental este enfoque en la presentación de resultados de las evaluaciones del rendimiento educativo. No obstante, parece evidente que la presentación en términos de porcentajes de respuestas correctas resulta claro, especialmente en lo que se refiere a los

resultados de ítems o de grupos de categorías de ítems. Para ello, se pueden derivar porcentajes para la presentación de los resultados de promedios de acierto para las distintas categorías de ítems, siguiendo un procedimiento que permite su estimación a partir de la TRI. Este tipo de presentación combina las ventajas de la precisión de la TRI y la comprensibilidad de la presentación de resultados en porcentajes. Una metodología adecuada para la derivación de los porcentajes de acierto a partir de la TRI se presenta en el trabajo de Gil y Suárez (2000).

EJEMPLIFICACIÓN: ESCALAS DE RENDIMIENTO Y NIVELES DE RENDIMIENTO

Para ejemplificar este enfoque se van a presentar los resultados obtenidos en el Estudio Internacional sobre la Enseñanza y el Aprendizaje de la Lengua Inglesa en la Educación Secundaria, en el que participaron tres instituciones dedicadas a la evaluación de los sistemas educativos, la *Direction de l´Évaluation et de la Prospective* (DEP) del *Ministère de l´Éducation Nationale* de Francia, la *Skolverket* (Agencia Nacional de Educación de Suecia) y el Instituto Nacional de Calidad y Evaluación (INCE) del Ministerio de Educación y Cultura de España. Los resultados comparativos internacionales se presentaron en la publicación *Evaluación Comparada de la Enseñanza y el Aprendizaje de la Lengua Inglesa: España, Francia, Suecia* (Gil y Alabau, 1997), desde una perspectiva española, y en la publicación correspondiente a cargo de la *Direction de l´Évaluation et de la Prospective* (DEP, 1997).

La prueba de rendimiento en lengua inglesa utilizada medía la competencia lingüística del alumno en tres destrezas básicas. La prueba incluía ejercicios dirigidos a estimar el nivel de rendimiento en cuanto a los conocimientos lingüísticos (incluyendo el léxico y la gramática), la comprensión escrita y oral y la expresión escrita.

La población objeto de esta evaluación se definió en función del curso y estuvo constituida por los alumnos que en 1996 cursaban el 4º año de la Enseñanza Secundaria Obligatoria (ESO) y los que cursaban el 2º año del Bachillerato Unificado y Polivalente (BUP), cuyas edades estaban comprendidas, en su mayoría, entre los 15 y 16 años.

El diseño muestral utilizado en esta evaluación fue el de un muestreo estratificado, utilizando la técnica de probabilidad proporcional al tamaño, bietápico, tomando como primer nivel de muestreo al alumno y, como segundo, al centro. El diseño y procedimientos de muestreo se basaron en las especificaciones técnicas utilizadas por la *International Association for the Evaluation of Educational Achievement* —IEA— (Rosier y Ross, 1992; Ross, 1991). El número total de alumnos evaluados fue de 4.562 de los que 4.320 respondieron a la prueba de rendimiento.

ESCALAS DE RENDIMIENTO

Por un lado, se va a utilizar una escala de rendimiento que representa la dimensión de conocimientos, capacidades o competencias en lengua inglesa. El rango de las puntuaciones en esta escala se fija a través de la definición arbitraria de su media y su desviación típica, con el objeto de facilitar la comprensión de los resultados.

La escala que va utilizarse en la presentación de los resultados de esta evaluación es similar a las escalas utilizadas por la Asociación Internacional para la Evaluación del Rendimiento Educativo *(International Association for the Evaluation of Educational Achievement)* —IEA— (IEA Secretariat, 1998) con una media de 500 puntos y una desviación típica de 100 puntos. Un modo de interpretar estos resultados consiste en considerar las puntuaciones como el resultado obtenido por los alumnos en una escala hipotética de 1000 preguntas, de modo que, por ejemplo, una puntuación de 600 puntos para un alumno indicaría que este alumno sería capaz de responder correctamente a 600 preguntas de una prueba similar compuesta por 1000 preguntas. No obstante, las puntuaciones en dicha prueba hipotética se distribuyen habitualmente entre los 200 y los 800 puntos. La razón subyacente a la utilización de estos grandes números en las escalas consiste en que de ese modo se puede obviar la utilización de decimales simplificándose de ese modo la lectura de los datos.

Con esta escala normalizada, de media 500 puntos y desviación típica de 100 puntos, representada en el Gráfico 1, aproximadamente el 2 por ciento de los sujetos con habilidad más baja tendrá puntuaciones inferiores a 300 puntos; el 16 por ciento de los sujetos con baja habilidad tendrá una puntuación menor de 400 puntos y el 50 por ciento de los sujetos con menor habilidad tendrá puntuaciones menores de 500 puntos; aproximadamente un 84 por ciento de los sujetos tendrá una puntuación menor de 600 puntos; un 98 por ciento menor de 700 y un 2 por ciento mayor que 700 puntos.

De ese modo, la normalización de las puntuaciones implica que aproximadamente un 14 por ciento tendrá puntuaciones entre 300 y 400 puntos y otro 14 por ciento entre 600 y 700 puntos; y, por último, un 34 por ciento aproximadamente tendrá puntuaciones entre 400 y 500 puntos y otro 34 por ciento de los alumnos evaluados tendrá puntuaciones entre 500 y 600 puntos.

Para la presentación de los resultados relativos a los factores que afectan al rendimiento y que implican la comparación entre dos o más grupos o categorías se pueden utilizar tablas en las que se presente la puntuación media para cada grupo o categoría, su error típico (E.T.), su desviación típica (D.T.) y el número de sujetos a partir del cual se han realizado los cálculos para dicho grupo. Además, las tablas pueden incluir una representación gráfica de la puntuación media estimada y del intervalo en el que (por ejemplo, con un grado de confianza del 95 por ciento) se encuentra la media verdadera del grupo, en relación con la escala de rendimiento de media 500 puntos y desviación típica de 100 puntos. Asimismo, la representación gráfica de los resultados puede incluir una representación de las características de la dispersión de los resultados en cada grupo mediante la representación de los cuartiles, con un sistema similar al de diagramas de cajas, o *box-plots,* en el que se representan los cuatro cuartiles de izquierda a derecha, cubriendo el primer cuartil, o cuartil inferior, el rango de puntuaciones del 25 por ciento de los sujetos con puntuaciones más bajas. El segundo cuartil representa el rango de puntuaciones del 25 por ciento de los sujetos con puntuaciones situadas entre el límite superior del primer cuartil y la mediana mientras que el tercer cuartil cubre el rango de puntuaciones del 25 por ciento de sujetos que obtienen un resultado superior a la mediana e inferior al límite más bajo del cuarto

cuartil. Por último, el cuarto cuartil cubre el rango de puntuaciones de los sujetos con mejores resultados. Como es tradicional en este tipo de representación mediante diagramas de caja, las representaciones del límite inferior del primer cuartil y del superior del cuarto cuartil se restringen a un valor mínimo y máximo equivalentes a una vez y media la distancia intercuartil, es decir, la distancia entre los límites que separan el primer y segundo cuartil y el tercer y cuarto cuartil, respectivamente, y que cubre al 50 por ciento de los sujetos de cada grupo o categoría representada con valores medios. Esta representación se aplica con el objeto de que los sujetos con valores atípicos, o *outliers*, no distorsionen el significado global de los resultados y, en consecuencia, no se representan en la misma dichos casos extremos. En el Gráfico 2 se muestra un ejemplo del significado de cada uno de los elementos de la representación gráfica de los resultados.

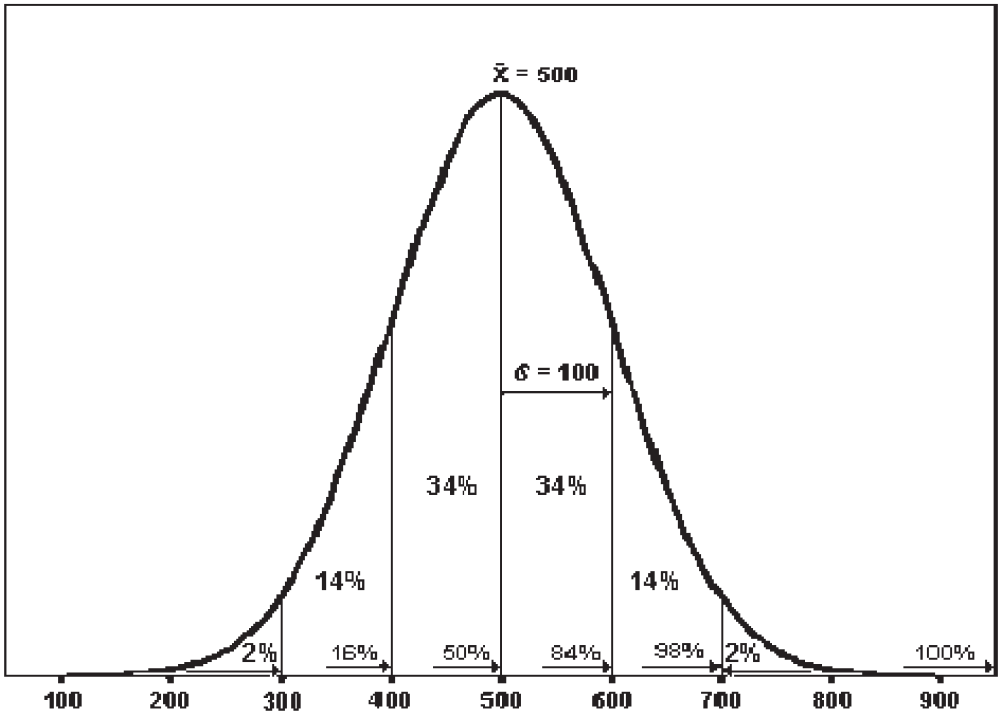


Gráfico 1

Escala de rendimiento con media igual a 500 y desviación típica igual a 100, con su curva normal y los porcentajes de sujetos correspondientes a cada tramo de la escala por unidad de desviación típica.

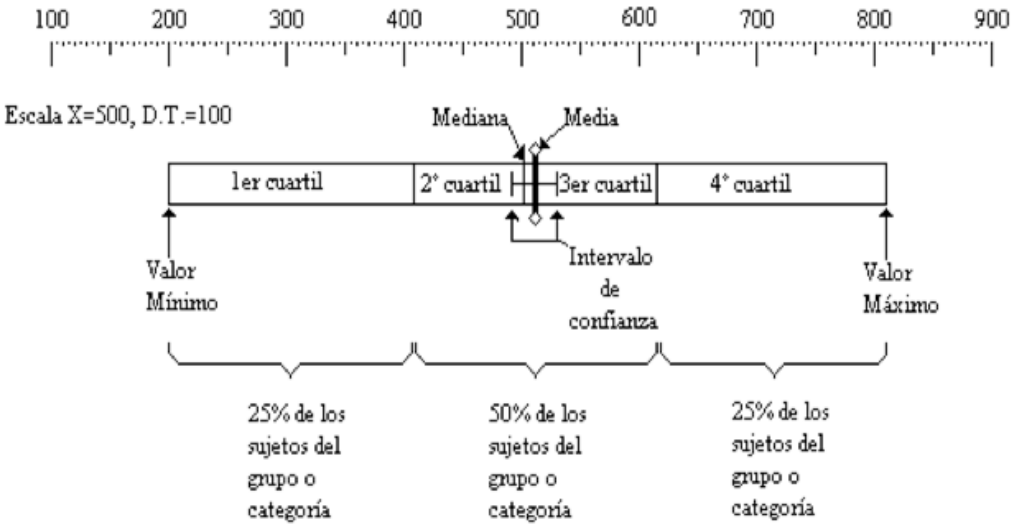


Gráfico 2

Ejemplo de representación gráfica utilizada para la presentación de los resultados comparativos que incluye la puntuación media estimada de cada grupo, o categoría analizada, y del intervalo en el que, con un grado de confianza del 95 por ciento, se encuentra su media verdadera, en relación con la escala de rendimiento de media 500 puntos y desviación típica de 100, así como las características de la dispersión de los resultados en cada grupo mediante la representación de los cuartiles.

Dos ejemplos de presentación de los resultados en términos de escalas de rendimiento: sexo y edad de inicio al estudio del inglés

Sexo

De los 4.146 alumnos con datos completos sobre el sexo 1.735 eran varones (el 41,85 por ciento) y 2.411 mujeres (el 58,15 por ciento).

En el gráfico 3 se presentan las medias de la puntuación de rendimiento para los alumnos y las alumnas evaluados, tal como se explicitó en el apartado sobre el sistema de presentación de los resultados, junto con su error típico y desviación típica, y el número de alumnos de los que se dispone de datos para el análisis.

Asimismo, el gráfico muestra, además de la media y el intervalo en el que ésta se encuentra con un 95 por ciento de confianza, la distribución por cuartiles de los resultados de los alumnos y las alumnas. Tal como se mencionó en el apartado dedicado a la presentación de los resultados, el primer tramo de cada barra gráfica horizontal indica en sus extremos la puntuación mínima obtenida por los alumnos y las alum-

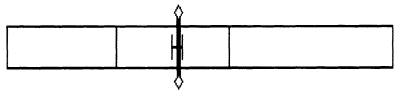
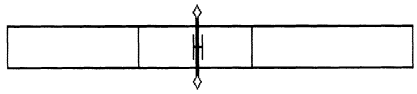
	Media	E.T.	D.T.	N	Escala de rendimiento 200 - 300 - 400 - 500 - 600 - 700 - 800
Alumnos	486	2,32	97	1735	
Alumnas	512	2,05	101	2411	

Gráfico 3
Puntuación media de rendimiento por sexo.

nas y la puntuación por debajo de la cual se encuentra el 25 por ciento de los alumnos o alumnas, es decir, el primer cuartil. El segundo tramo de la barra muestra el rango de puntuaciones que comprende desde el primer cuartil hasta la puntuación correspondiente a la mediana, es decir, la puntuación en la que el 50 por de los alumnos queda por debajo y el restante 50 por ciento queda por encima de la misma. El tercer tramo indica el rango entre la mediana y el tercer cuartil, es decir, la puntuación por debajo de la cual se encuentra el 75 por ciento de los sujetos. Y, por último, el cuarto tramo muestra el rango entre el tercer cuartil y la puntuación máxima para los alumnos y las alumnas.

Tal como se indica en la tabla las alumnas obtuvieron un resultado (512 puntos) significativamente mejor ($F=67,70$; g.l.=1, 4144; $p\leq.0000$) que los alumnos (486 puntos) (recuérdese que la media para el conjunto de los alumnos examinados es de 500 puntos).

Este resultado concuerda con el resultado habitual de los estudios de rendimiento escolar que muestra que, para esta edad y para las áreas de conocimiento relacionadas con el lenguaje, las alumnas muestran un mejor rendimiento académico que los alumnos. La varianza de las puntuaciones en estos grupos resultó homogénea (Levene=3,18; g.l.=1, 4144; $p<.074$) lo que implica que la diversidad de rendimiento en los dos grupos, alumnos y alumnas, es equivalente.

En el Gráfico 4 se muestra la distribución del alumnado por los cinco niveles de rendimiento y por el sexo. Se observa como los alumnos se distribuyen en un mayor porcentaje en los niveles de rendimiento inferiores, 300 y 400, con unos porcentajes del 9 y del 27 por ciento respectivamente frente a los del 5 y 24 correspondientes a las alumnas. De modo complementario, la presencia de las alumnas es mayor que la de los alumnos en los dos niveles de rendimiento más altos, 600 y 700, con unos porcentajes del 25 y 10 frente a los del 22 y 4 por ciento respectivamente correspondientes a los alumnos.

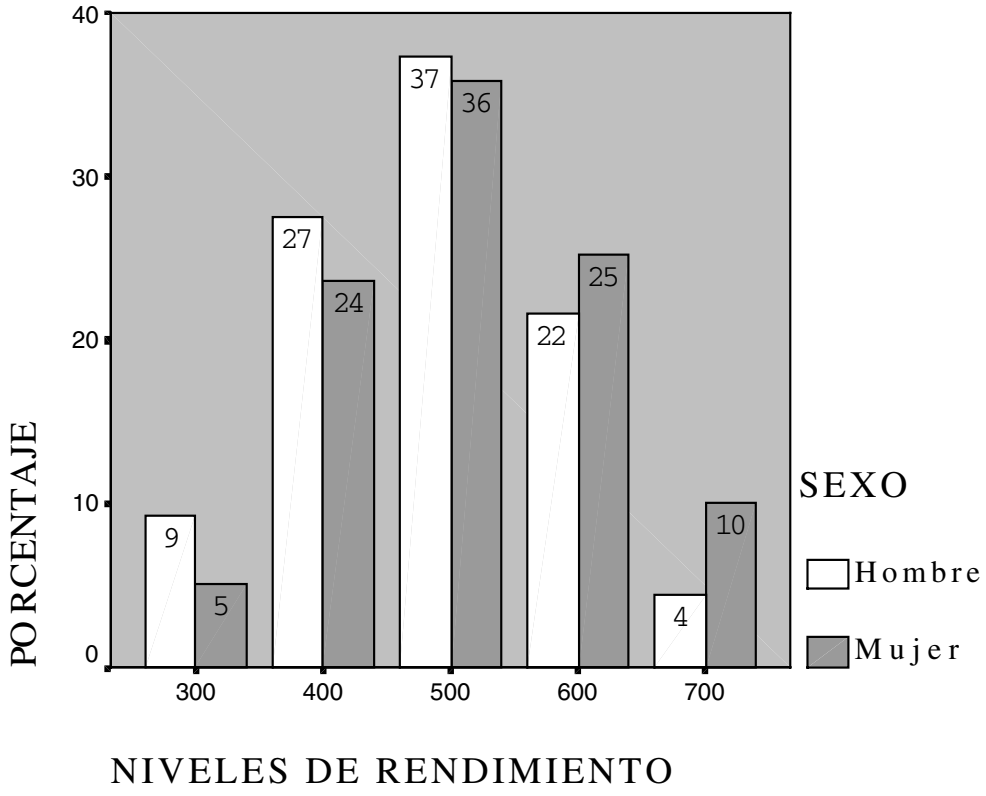


Gráfico 4
Distribución porcentual del alumnado en los cinco niveles de rendimiento por sexo.

Edad de inicio de estudio del inglés

El Gráfico 5 presenta las puntuaciones de rendimiento en lengua inglesa obtenidas por los alumnos de la muestra examinada en función de la edad de inicio al estudio del inglés.

Puede observarse en el gráfico cómo el rendimiento de los alumnos es mejor cuando estos han iniciado el estudio del inglés a los 6 años o antes, y cómo desciende sistemáticamente según su edad de inicio al aprendizaje del inglés es mayor, correspondiendo el nivel de rendimiento más bajo a los alumnos que iniciaron el aprendizaje del inglés a los 12 años o más tarde. La influencia de la edad de inicio al estudio del inglés sobre el rendimiento fue significativa ($F=75{,}98$; g.l.=6, 3.998; $p<.0001$) observándose diferente variabilidad en las puntuaciones de los alumnos según su edad de inicio, dado que hay más variabilidad en los grupos que iniciaron el estudio de la lengua extranjera a los nueve años o antes y menos en los grupos de alumnos que empezaron a aprender

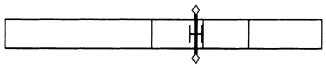
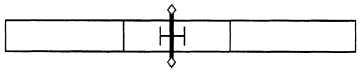
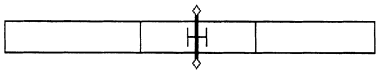
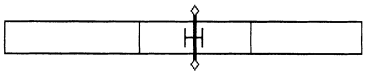
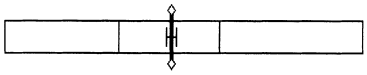
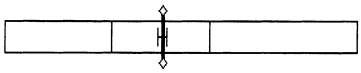
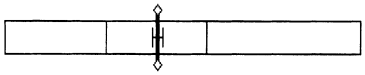
Edad de inicio de estudio del inglés	Media	E.T.	D.T.	N	Escala de rendimiento 200 - 300 - 400 - 500 - 600 - 700 - 800
6 años o menos	566	4,43	96	472	
7 años	544	7,70	108	197	
8 años	531	6,70	105	244	
9 años	524	6,02	100	278	
10 años	496	3,46	92	716	
11 años	484	2,56	91	1258	
12 años o más	467	3,16	92	840	

Gráfico 5
Puntuación media de rendimiento en función de la edad de inicio de estudio del inglés.

inglés a los 10 años o más tarde, que son los que tienen un menor rendimiento (Levene=3,67; g.l.=6, 3.998; p<.001). No se presentan en estos resultados los datos desagregados para los cinco años o menos debido al escaso número de alumnos de la muestra que habían iniciado su aprendizaje del inglés a estas edades.

En el gráfico 5 se presentan las puntuaciones medias de rendimiento para los diferentes grupos de sujetos en función de la edad de inicio al estudio del inglés junto con la representación gráfica de la distribución de sus puntuaciones. Estos resultados parecen indicar que el inicio temprano de los alumnos al estudio del inglés es un factor positivo e importante de cara a su buen aprendizaje, siendo claro que el empezar, al menos, a los seis años proporciona mejores resultados que empezar el estudio de las lenguas extranjeras a edades mayores.

Niveles de rendimiento

La definición de niveles de rendimiento es uno de los modos de atribuir significado a la escala de medida mediante la descripción de lo que la vasta mayoría de los estudiantes situados en cada nivel de rendimiento saben y pueden hacer diferenciándoles esta descripción de las capacidades de los alumnos de niveles de rendimiento inferiores (Beaton y Allen, 1992; Beaton y Johnson, 1992; Beaton y Zwick, 1990; Bock, Mislevy y Woodson, 1982; González y Beaton, 1994).

Existen diversos métodos de asignación de estándares y niveles de rendimiento, diferenciándose fundamentalmente en si la definición de niveles se lleva a cabo *a priori* o *a posteriori* y en si es criterial o teórica (es decir, se fundamenta bien en el consenso de grupos de expertos o en una teoría subyacente de la evolución del aprendizaje) o bien es empírica o normativa (es decir, se basa en el análisis de los datos reales obtenidos con los estudiantes). El procedimiento utilizado en este caso es un enfoque empírico-normativo *a posteriori*.

Básicamente, el procedimiento general consiste en dividir el conjunto de la muestra en varios grupos con un criterio arbitrario y definido para, posteriormente, analizar las cuestiones que los distintos grupos de sujetos de diferente nivel de habilidad son capaces de resolver correctamente. También existen distintos procedimientos y criterios para fijar los puntos de corte que clasifican a los sujetos en cada nivel de habilidad en la dimensión que mide la escala (Beaton y Johnson, 1992; Bock, Mislevy y Woodson, 1982).

En este ejemplo, la dimensión de habilidad o escala de rendimiento se ha dividido, de forma arbitraria en cinco intervalos de puntuaciones o niveles de rendimiento. Dichos intervalos son equivalentes en términos de la escala de medida estableciéndose los puntos de corte para la asignación de sujetos y cuestiones a los diferentes niveles de rendimiento en función de su distancia a la media, siendo el criterio de división de la escala las distancias enteras en unidades de desviación típica de la media. Los niveles resultantes se han denominado Nivel 300, Nivel 400, Nivel 500, Nivel 600 y Nivel 700. La denominación de estos niveles se utiliza en el sentido tradicional de los nombres de intervalos de modo que el Nivel 300 incluye a los sujetos con puntuaciones comprendidas entre 250 y 350 puntos.

Esta división *a posteriori* de los sujetos en cinco niveles de rendimiento tiene como consecuencia que, si la distribución de puntuaciones se aproxima a la distribución normal, alrededor del 7 por ciento de los sujetos debería encontrarse situado en el Nivel 300 y otro 7 por ciento en el Nivel 700; el 24 por ciento de los sujetos debería situarse en el Nivel 400 y otro 24 por ciento en el Nivel 600; y, por último, aproximadamente un 38 por ciento debería situarse en el Nivel 500. Esta distribución teórica de los sujetos en términos de porcentajes en los cinco niveles de rendimiento se presenta en el Gráfico 6.

En cuanto a los criterios para asignar las cuestiones y, en consecuencia, los conocimientos, procesos o habilidades que dominan los sujetos a cada nivel como elementos que definen el nivel de habilidad para cada grupo se ha utilizado el criterio habitual en la práctica tradicional en la teoría de los tests mentales considerando una cuestión

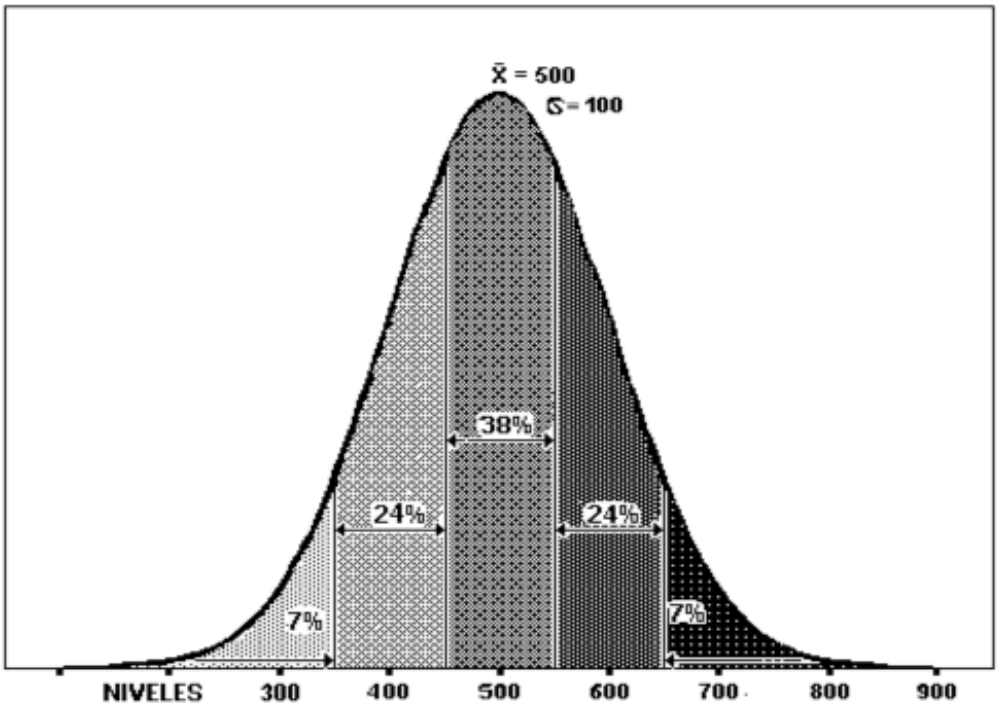


Gráfico 6
Distribución normal de los niveles de rendimiento en la escala de media igual a 500 y desviación típica igual a 100 y los porcentajes de sujetos asociados a cada nivel.

como representativa de un nivel de habilidad si el 50 por ciento de los sujetos de dicho nivel resuelven dicha cuestión correctamente (Bock, Mislevy y Woodson, 1982) frente a los niveles más exigentes del 65 y el 80 por ciento utilizados por Beaton y Johnson (1992).

El Cuadro 1 presenta la descripción de los niveles de rendimiento en términos de las competencias en lengua inglesa asociadas a cada uno de los mismos, así como el porcentaje de alumnos que se encuentran en cada nivel y el porcentaje acumulado.

Al analizar esta tabla con los niveles de rendimiento y sus correspondientes competencias, se aprecia claramente una progresión en los conocimientos y destrezas lingüísticas de los alumnos desde los niveles más bajos, 300 y 400, a los niveles de mejor rendimiento, 600 y 700.

Un 7% de los alumnos solamente domina las competencias más elementales del nivel inferior, nivel 300. Las capacidades asociadas a este nivel 300 son muy básicas, tales como clasificar palabras por campos semánticos o reconocer expresiones muy usuales, en lo que respecta a conocimientos léxicos, y conocer y utilizar con correc-

ción el presente simple y el continuo, en cuanto a conocimientos lingüísticos se refiere. En relación con la competencia lectora, los alumnos situados en el nivel 300 sólo reconocen frases de estructura y vocabulario muy simple. En cuanto a las capacidades evaluadas en compresión oral y expresión escrita, no consiguen superar ninguna de las capacidades evaluadas en la prueba de rendimiento. Los resultados de este 7% de los alumnos con puntuaciones muy bajas, por debajo de 350 puntos, resultan claramente insatisfactorios para lo que es esperable en esta etapa de la educación secundaria.

El Cuadro 1 muestra como el 93% de los alumnos ha adquirido las competencias correspondientes al nivel 400, aunque sólo un 32% de los alumnos no superan este nivel y, lógicamente, dominan también las capacidades asociadas al nivel inferior.

En este nivel 400 los alumnos son capaces de diferenciar algunos fenómenos de derivación y utilizar con corrección adjetivos y verbos para dar sentido a diversas frases, todo ello en lo que respecta a sus conocimientos léxicos. En cuanto a los conocimientos gramaticales, conocen y utilizan el verbo modal «can», además del presente simple y continuo de los verbos. En lo que respecta a la compresión de textos escritos además de comprender frases de estructura y vocabulario sencillos, son capaces de identificar algunos términos léxicos para completar frases, siempre que éstas sean de estructura simple.

En cuanto a la comprensión oral son capaces de identificar el tema de una noticia escuchada a través de una cassette de audio seleccionándolo de entre varias opciones. Sin embargo, no consiguen superar ninguna de las capacidades evaluadas referidas a la expresión escrita. Los resultados de estos alumnos, entre 350 y 450 puntos, pueden calificarse de pobres resultando insuficientes para esta etapa educativa.

En el nivel 500 se encuentra el 36% de los sujetos. En total un 68% de los alumnos han alcanzado al menos las competencias asociadas a este nivel.

En este nivel 500, los alumnos, aparte de lograr las capacidades de los niveles anteriores, conocen bien los fenómenos de derivación, en lo que se refiere a conocimientos léxicos, y conocen y utizan con corrección el pasado simple, los verbos modales «may», «must», «should» y «have to»; conocen y usan adecuadamente los artículos, al igual que los principales cuantificadores, dentro de los conocimientos gramaticales. En cuanto a lo que se refiere a la comprensión escrita, los alumnos de este nivel se desenvuelven bien, consiguiendo superar prácticamente todas las tareas evaluadas en la prueba en relación con esta destreza tales como la identificación de términos que completan una frase o que pertenecen a un mismo campo semántico, la realización de una tarea después de haber leído un texto y la compresión de frases en el contexto de un párrafo. En cuanto a la compresión oral, son capaces de identificar algunos detalles de la información recibida además de identificar el tema del que trata. Su nivel de competencia en la expresión escrita es inferior al resto de las destrezas evaluadas, desarrollando únicamente la capacidad de formular frases sencillas respondiendo a una información solicitada. Los resultados de los alumnos de este nivel resultan básicos, siendo suficientes, aunque no brillantes.

El 32% de los alumnos evaluados logra dominar las competencias correspondientes al nivel 600, obteniendo un 24% de los alumnos puntuaciones de este nivel, entre 550 y

CUADRO 1

NIVELES DE RENDIMIENTO Y COMPETENCIAS ASOCIADAS EN LENGUA INGLESA. PORCENTAJES DE ALUMNOS EN CADA NIVEL

NIVEL	Conocimientos Lingüísticos		Comprensión Escrita	Comprensión Oral	Expresión Escrita	PORCENTAJE DE ALUMNOS EN CADA NIVEL	PORCENTAJE ACUMULADO
	Léxico	Gramática					
700	❑Reconocer diferentes expresiones del lenguaje coloquial ❑ Utilizar los términos léxicos adecuados para dar sentido a frases	❑ Conocer y utilizar el presente perfecto y el presente simple con idea de futuro ❑ Transformar un texto de estilo directo a estilo indirecto	❑ Inferir el significado de términos por medio del contexto	❑ Resumir la información recibida	❑ Redactar un texto breve que resuma una situación	**8**	**8**
600		❑ Conocer y utilizar el futuro simple		❑ Identificar detalles concretos de la información recibida	❑ Completar un diálogo formulando frases en distintos tiempos verbales ❑ Describir un dibujo	**24**	**32**
500	❑ Distinguir la raíz de los afijos	❑ Conocer y utilizar el pasado simple ❑ Conocer y utilizar el artículo ❑ Conocer los verbos modales ❑ Utilizar los principales cuantificadores	❑ Comprender frases en el contexto de un párrafo ❑ Extraer información con el objeto de realizar una tarea ❑ Identificar las palabras de un campo semántico dado ❑ Identificar los términos que completan una frase	❑ Identificar el tema y algunos detalles de la información recibida	❑ Completar un diálogo formulando frases sencillas en pasado simple	**36**	**68**
400	❑ Distinguir los afijos más usuales ❑ Utilizar los términos léxicos (verbos y adjetivos) adecuados para dar sentido a frases	❑ Conocer el verbo "can"	❑ Identificar los términos léxicos que completan una frase con un vocabulario fácil	❑ Seleccionar de entre varios temas el referido a la información recibida	❑ No se alcanza el nivel mínimo evaluado en esta competencia	**25**	**93**
300	❑ Clasificar palabras por campos semánticos ❑ Reconocer expresiones muy usuales del lenguaje coloquial	❑ Conocer el presente simple y el presente continuo	❑ Comprender frases de estructura y vocabulario sencillos en el contexto de un párrafo	❑ No se alcanza el nivel mínimo evaluado en esta competencia		**7**	**100**

650 puntos. Estos alumnos superan todos los objetivos fijados en la prueba con objeto de evaluar los conocimientos léxicos del alumnado; por tanto, además de lograr las capacidades descritas en los niveles inferiores, son capaces de reconocer diferentes expresiones coloquiales y de utilizar los términos léxicos adecuados para dar sentido a diversas frases. En cuanto a los conocimientos gramaticales, dan un paso más al dominar el uso del futuro simple. En lo que respecta a la compresión escrita, también consiguen superar todas las tareas de la prueba relacionadas con esta destreza lingüística, al ser capaces de deducir el significado de términos por medio del contexto. En la comprensión oral, avanzan un paso más al captar detalles muy concretos de la información escuchada. Y, por último, en lo referente a la expresión escrita consiguen formular frases en diversos tiempos dando una opinión o una información no directamente solicitada, además de ser capaces de describir un dibujo. Se puede considerar que los resultados de los alumnos de este nivel 600, un 24 por ciento, muestran un buen rendimiento para esta etapa.

El nivel de máximo rendimiento, el nivel 700, se alcanza por un 8% de los alumnos, que conocen y utilizan correctamente el presente perfecto de los verbos y son capaces de transformar un texto de estilo directo a estilo indirecto, superando de esta forma todas las tareas de conocimientos lingüísticos incluidas en la prueba de rendimiento. En relación con la comprensión oral son capaces de resumir en la lengua materna la información recibida oralmente, tarea que presentaba la mayor dificultad de las utilizadas para evaluar esta destreza comunicativa. Son asimismo, capaces de redactar un texto, no sólo describiendo lo que un dibujo les sugiere, sino utilizando un vocabulario fluido y correcto ortográficamente y un lenguaje lógico y coherente. Sin duda, este 8% de los alumnos alcanza un nivel de rendimiento académico en lengua inglesa que, para esta etapa educativa, puede considerarse muy bueno.

CONCLUSIÓN

La finalidad del presente trabajo era la descripción y aplicación de un procedimiento, basado en la Teoría de la Respuesta al Item, para la presentación de los resultados en términos de escalas y niveles de rendimiento educativo de las evaluaciones globales de los sistemas educativos. Las características de este procedimiento de presentación de resultados se han comparado con la aproximación clásica que utiliza las puntuaciones directas o los porcentajes que los estudiantes obtienen en las pruebas de rendimiento. Asimismo, se ha ejemplificado la aplicación de este procedimiento con los resultados obtenidos por la muestra de alumnos españoles que participaron en 1996 en el Estudio Internacional sobre la Enseñanza y el Aprendizaje de la Lengua Inglesa en la Educación Secundaria.

A la vista de los resultados obtenidos y de las ventajas teóricas descritas de la TRI sobre la TCT, parece aconsejable —para la presentación de resultados de las evaluaciones de rendimiento educativo en los estudios de evaluación global de los sistemas educativos— utilizar las escalas y los niveles de rendimiento derivadas de la Teoría de la Respuesta al Item. Este reciente marco teórico supera las limitaciones de la Teoría Clásica de los Tests, proporcionando medidas objetivas y estables tanto del

rendimiento de los alumnos como de las propiedades psicométricas de los ítems del test (la dificultad, el poder discriminativo y el grado de adivinación al azar de los ítems).

Por último, resulta conveniente presentar los resultados de tablas y representaciones gráficas de modo que incluyan la información mínima necesaria para una correcta interpretación de los resultados, lo que implica incluir tanto el número de sujetos de cada grupo sobre el que se estiman los estadísticos como como el error de medida de los mismos, junto con sus medidas básicas de tendencia central y dispersión.

REFERENCIAS

Beaton, A.E. y Allen, N. (1992). Interpreting Scales through Scale Anchoring. *Journal of Educational Statistics*, 17, 191-204.

Beaton, A.E. y Johnson, E.G. (1992). Overview of the Scaling Methodology Used in the National Assessment. *Journal of Educational Measurement*, 29, 2, 163-175.

Beaton, A.E. y Zwick, R. (1990). Overview of the National Assessment of Educational Progress, *Journal of Educational Statistics*, 17, 95-109.

Bock, R.D., Mislevy, R.J. y Woodson, C. (1982). The next Stage in Educational Assessment. *Educational Researcher*, 11, 3, 4-11.

Direction de l´Évaluation et de la Prospective (DEP) (1997). *Espagne, France, Suède: Évaluation des Connaissances et Compétences en Anglais des Élèves de 15-16 Ans.* Paris, Ministère de l´Éducation Nationale.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement*, 58, 3, 357-381.

Gil Escudero, G. y Alabau Balcells, I. (1997). *Evaluación Comparada de la Enseñanza y el Aprendizaje de la Lengua Inglesa: España, Francia, Suecia*. Madrid, Ministerio de Educación y Cultura.

Gil Escudero, G. y Suárez Falcón, J.C. (1997). Aplicación y Validación de un Modelo de Construcción de Pruebas de Rendimiento de Matemáticas, Ciencias y Lengua en la Educación Primaria. *Revista de Educación*, 312, 133-143, 1997.

Gil Escudero, G.A. y Suárez Falcón, J.C. (2000). Construcción de una escala y diversas puntuaciones de rendimiento en una prueba de lengua inglesa y derivación de puntuaciones porcentuales basadas en la Teoría de la Respuesta al Ítem. *Revista de Educación*, *322*, 325-340.

González, E.J. y Beaton, A.E. (1994). The Determination of Cut Scores for Standards en A.C. Tuijnman y T.N. Postlethwaite *Monitoring the Standards of Education*. London, Pergamon.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley

Hambleton, R. K. (1989a). (Ed.) Applications of item response theory. *International Journal of Educational Research*, 13.

Hambleton, R. K. (1989b). Principles and selected applications of item response theory. In R. I. Linn (Ed.). *Educational measurement* (3rd cd.) (pp. 147-200). New York Macmillan.

Hambleton, R.K. y Cook, L.L. (1977). Latent Trait Models and their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement*, 14, 75-96.

Hambleton, R.K. y Jones, R.W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12, 3, 38-47.

Hambleton, R.K. y Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, Kluwer-Nijhoff.

Hambleton, R.K.; Swaminathan, H. y Rogers, H.J. (1991). *Principles and Applications of Item Response Theory.* Beverly Hills, Sage.

Hulin, C.L., Drasgow, F. y Parsons, C.K. (1983). *Item Response Theory: Applications to Psychological Measurement*. Homewood, Dow Jones-Irwin.

IEA Secretariat (1998). *IEA Guidebook:1998: Activities, Institutions and People*. Amsterdam, The International Association for the Evaluation of Educational Achievement (IEA).

Keeves, J.P. (1990). Scaling Achievement Test Scores en H.J. Walberg y G.D. Haertel (Eds.) *The International Encyclopedia of Educational Evaluation*. Oxford, Pergamon.

Keeves, J.P. (1992). Scaling Achievement Test Scores, in J.P.Keeves (Ed.) *Methodology and Measurement in International Educational Surveys*. The Hague, The International Association for the Evaluation of Educational Achievement (IEA).

Linn, R.L. y Baker, E.L. (1995). What Do International Assessment Imply for World-Class Standards? *Educational Evaluation and Policy Analysis*, 17, 4, 405-418.

Linn, R.L. y Dunbar, S.B. (1992). Issues in the Design and Reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29, 2, 177-194.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, LEA.

Lord, F.M. y Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Addison-Wesley.

Lord, F.M. y Stocking, M.L. (1990). Item Response Theory en H.J. Walberg y G.D. Haertel (Eds.) *The International Encyclopedia of Educational Evaluation*. Oxford, Pergamon.

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

Masters, G. y Forster, M. (1996a). *Progress Maps*. Melbourne, ACER.

Masters, G. y Forster, M. (1996b) *Developmental Assessment*. Melbourne, ACER.

Mislevy, R.J. (1993). Foundations of a New Test Theory, in N.Frederiksen, R.J. Mislevy y I.I. Bejar, *Test Theory for a New Generation of Tests*. Hillsdale, LEA.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., González, E.J., Kelly, D.L. y Smith, T.A. (1996*). Mathematics Achievement in the Primary School Years: IEA´s Third International Mathematics and Science Study.* Chesnut Hill, Boston College.

Muñiz, J. (1994). *Teoría Clásica de los Tests.* Pirámide, Madrid.

Phillips, G.W. (1994). Methods and Issues in Setting Performance Standards en A.C. Tuijnman y T.N. Postlethwaite *Monitoring the Standards of Education.* London, Pergamon.

Rosier M.J. y Ross, K.N. (1992). Sampling and Administration en J.P. Keeves (Ed.) *The IEA Technical Handbook*. The Hague, The International Association for the Evaluation of Educational Achievement (IEA).

Ross, K.N. (1991). Sampling Manual for the IEA Reading Literacy Study. Hamburg, University of Hamburg.
Van der Linden, W.J. y Hambleton R.K. (1997). *Handbook of Modern Item Response Theory*. New York, Springer-Verlag.
Wright, B. D. y Stone, M. H.(1979). *Best test design*. Chicago: MESA Press University of Chicago.

Fecha de recepción: 8 de julio de 2001.
Fecha de aceptación: 30 de mayo de 2002.