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 RESPUESTA AL ÍTEM

GUILLERMO GIL ESCUDERO (*)
JUAN CARLOS SUÁREZ FALCÓN (**)

RESUMEN: El presente trabajo de investigación describe la aplicación de un procedimiento para la construcción de puntuaciones de los alumnos, basado en la Teoría de la Respuesta al Ítem (TRI), para la prueba de rendimiento en lengua inglesa del Estudio Internacional sobre la Enseñanza y el Aprendizaje de la Lengua Inglesa. Los objetivos que se perseguían en este trabajo eran dos: en primer lugar, construir una escala TRI de rendimiento en lengua inglesa que se ajustara al currículum español; y en segundo lugar, derivar una puntuación porcentual de rendimiento, basada también en la TRI, que facilite una presentación clara y sencilla de los resultados de los alumnos. Con el fin de alcanzar ambos objetivos, se llevaron a cabo dos estudios en los que se aplicó el método TRI propuesto a la prueba de rendimiento de lengua inglesa que se había administrado a 4.320 alumnos españoles de educación secundaria.

Los resultados obtenidos señalan que la derivación de puntuaciones de la TRI y ajustadas al currículum español no implica una mejora sustancial en la precisión con la que se mide la competencia lingüística, sino más bien supone una mayor validez de dominio y el establecimiento de una escala que puede tener utilidad en el futuro. Asimismo, se pone de manifiesto la posibilidad de obtener puntuaciones porcentuales basadas en la TRI, las cuales superan las limitaciones métricas de los porcentajes clásicos a la vez que permite la presentación de los resultados de rendimiento para el público general.

INTRODUCCIÓN

En el presente trabajo se informa del método utilizado para construir una escala de referencia y puntuaciones globales de rendimiento de los alumnos para el Estudio Internacional sobre la Enseñanza y el Aprendizaje de la Lengua Inglesa en el nivel de la educación secundaria llevada a cabo por el Instituto Nacional de Calidad y

(*) Instituto Nacional de Calidad y Evaluación (INCE).
(**) Universidad Nacional de Educación a Distancia (UNED).

Fecha de entrada: 10-06-1999 Fecha de aceptación: 11-02-2000

Evaluación –INCE–. 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). Por otro lado, se publicaron las actas del coloquio internacional, celebrado en octubre de 1997 en París, en el que se hicieron públicos por primera vez los resultados comparativos generales del estudio (Bonnet, 1997), y en las que se incluyeron diversos trabajos sobre la educación en lenguas extranjeras, los currículos, las prácticas y métodos de enseñanza, la formación del profesorado y los mecanismos de evaluación nacionales en el área de la lengua inglesa en España, Finlandia, Francia, Portugal y Suecia.

La evaluación de carácter internacional se ha limitado a realizar una comparación centrada en los resultados de rendimiento junto con unas comparaciones limitadas de algunos aspectos adicionales. Este tipo de análisis de carácter muy general, en términos porcentuales, viene condicionado por las diferentes características de las tomas de datos e instrumentos utilizados en cada país que condicionan y limitan la capacidad de análisis de los datos y, por consiguiente, la posibilidad de realizar inferencias. Por ello, en dicho informe se presentaron somera y descriptivamente los resultados en términos de porcentajes concernientes al rendimiento de los alumnos en la prueba de lengua inglesa, junto con algún dato destacado de los obtenidos a través de los cuestionarios dirigidos a los alumnos y los profesores.

En relación con el análisis de los datos de rendimiento en el nivel nacional de este estudio, su presentación y su interpretación, se presentan tres problemas de carácter general.

Por un lado, el problema derivado de la generación de una escala y una puntuación basada en la Teoría de la Respuesta al Ítem (TRI) (Keeves, 1990; 1992) de modo que se superen los inconvenientes y limitaciones que implicaría la utilización de un enfoque exclusivamente basado en la Teoría Clásica de los Tests (TCT). Las ventajas e inconvenientes de la TCT y la TRI 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), Hambleton y Jones (1993), Hambleton y Swaminathan (1985), Hambleton, Swaminathan y Rogers (1991), Hulin, Drasgow y Parsons (1983), Martínez Arias (1995) y en Van der Linden y Hambleton (1997). Además, un objetivo complementario consistió en desarrollar una escala y un sistema de puntuaciones relacionado que pudiese ser de utilidad en el futuro, bien al aplicarse la misma prueba a otras muestras de alumnos o bien al aplicar otra prueba similar que midiese las mismas habilidades, utilizando algunos ítems puente o de anclaje y los consiguientes mecanismos de equiparación.

Por otro lado, el problema derivado del grado de ajuste de la prueba de rendimiento utilizada, tanto en cuanto a los contenidos del currículum español como en cuanto a la importancia que se concede a cada una de las capacidades medidas por la prueba, dado que una de las premisas base para la realización de este estudio consistió en utilizar como elemento para la medida del nivel de rendimiento de los alumnos una prueba de rendimiento en la lengua inglesa que ya había sido desarrollada con anterioridad por la *Direction de l'Évaluation et de la Prospective* (DEP) del Ministerio de Educación Nacional de Francia, y utilizada en anteriores evaluaciones en 1984 y 1990 en el ámbito del sistema educativo francés.

El tercer problema planteado es el relacionado con la fácil comprensibilidad de los resultados por un público general. Sin

duda, es de fácil e intuitiva comprensión la presentación de los resultados en términos de porcentajes de respuestas correctas que los sujetos han dado a una prueba determinada. Sin embargo, la presentación tradicional en términos de porcentajes derivados de la puntuación directa presentan diversos problemas y están sujetos a malinterpretaciones comunes. Por ello, parece conveniente generar puntuaciones porcentuales con base en la TRI que, a la vez que incluyan las ventajas aportadas por ésta, permitan una presentación de los resultados fácilmente comprensible.

DESCRIPCIÓN DE LA PRUEBA

La prueba de rendimiento en lengua inglesa está constituida por cinco partes dirigidas a evaluar diferentes aspectos del dominio de dicha lengua: los conocimientos lingüísticos, la comprensión oral, la comprensión escrita, la expresión escrita y los conocimientos culturales relacionados con la lengua inglesa. Los ítems utilizados en la prueba se construyeron tomando como base una matriz de especificaciones de carácter curricular, es decir, su adscripción a cada escala estaba basada exclusivamente en criterios teóricos.

La parte correspondiente a los conocimientos lingüísticos consta de 56 ítems, la de comprensión oral de 12, la de comprensión escrita de 11, la de expresión escrita de 9 y la de conocimientos culturales de 3, lo que constituye un total de 91 ítems. La prueba no pretende estudiar la expresión oral debido a la complejidad y el coste que supondría una evaluación individualizada de esta capacidad para un diseño muestral tan amplio como el que se utiliza en este estudio. Este diseño, a grandes rasgos y a excepción de la expresión oral, coincide con la importancia que se concede a cada una de estas dimensiones en el currículum francés para la enseñanza de la lengua inglesa en la educación secundaria.

Un primer paso para conocer las posibilidades de aplicación de la prueba en el contexto del sistema educativo español fue analizar si los contenidos evaluados por la prueba formaban parte del currículum establecido por el decreto de enseñanzas mínimas para la educación secundaria obligatoria (Real Decreto 1007/1991, de 14 de junio). El análisis detallado de cada una de las preguntas que forman la prueba mostró que todas ellas estaban enmarcadas en algún apartado del currículum común de la educación secundaria. Este hecho puso de manifiesto la adecuación de la prueba desde el punto de vista de los contenidos.

Sin embargo, el tamaño concedido a cada una de las partes de la prueba, en términos del número de ítems, no resulta coincidente con las intenciones curriculares actualmente vigentes en España. La prueba original francesa concede una gran importancia a los conocimientos lingüísticos (aproximadamente, un 62% de los ítems) mientras que otorga menor importancia a las capacidades comunicativas (aproximadamente, un 35%) de comprensión oral (13%), comprensión escrita (12%) y de expresión escrita (10%). Como resultado de esta estructura de la prueba, la puntuación directa que se deriva de la misma refleja en gran medida esta distribución del número de ítems para cada una de las partes de la prueba, es decir, los resultados en la parte correspondiente a los conocimientos lingüísticos condiciona la mayor parte de la puntuación directa mientras que la capacidad de influencia sobre la misma es notoriamente menor para las otras partes de la prueba. Por otro lado, el actual currículum para la educación secundaria en España enfatiza los aspectos comunicativos del aprendizaje de las lenguas extranjeras haciendo especial hincapié en la comprensión y en la expresión en las mismas.

Se plantea pues el problema de construir, a partir de una prueba diseñada con unas especificaciones preestablecidas ajustadas al currículum francés, una puntuación que refleje no solamente el conocimiento de contenidos considerados en el currículum español, sino que refleje la adquisición de dominio de la lengua inglesa de modo acorde a las intenciones curriculares del mismo, es decir, otorgando una mayor importancia e incidencia a los aspectos comunicativos del aprendizaje de la lengua inglesa.

SUJETOS

La población utilizada en los dos estudios que se presentan en este trabajo estuvo constituida por los alumnos que en 1996 cursaban el 4.º año de la Enseñanza Secundaria Obligatoria (ESO) y los que cursaban el 2.º año del Bachillerato Unificado y Polivalente (BUP), cuyas edades estaban comprendidas, en su mayoría, entre los 15 y 16 años. En consecuencia, la población objeto de estudio se definió en función del curso y nivel. Estos alumnos habían cursado 4 años de aprendizaje en lenguas extranjeras, aunque bajo organizaciones escolares diferentes derivadas de distintas leyes de enseñanza: la Ley General de Educación de 1970 –LGE– y la Ley Ordenación General del Sistema Educativo de 1990 –LOGSE.

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 en todo el territorio nacional, de los que 3.352 estaban cursando sus estudios en 2.º curso de BUP y 1.210 lo hacían en 4.º curso de ESO. La prueba de rendimiento fue administrada a 4.320 alumnos, 3.209 de centros públicos (un 74,3%) y 1.111 de centros privados (un 25,7%).

ESTUDIO 1

LA CONSTRUCCIÓN DE UNA ESCALA DE PUNTUACIONES TRI

La Teoría de Respuesta al Ítem (TRI) proporciona el entramado teórico necesario para poder resolver estos problemas en la práctica. Las propiedades de las puntuaciones calculadas con base en los procedimientos y supuestos de la TRI hacen posible la creación de escalas de habilidad independientes de los ítems específicos utilizados para su estimación e independientes de las muestras utilizadas inicialmente para la calibración de los elementos que componen dichas escalas, siempre que se cumplan los supuestos básicos que requiere la aplicación de la teoría.

En consecuencia, los objetivos de la metodología desarrollada para la construcción de una escala de rendimiento TRI pueden formularse del siguiente modo:

- Equilibrar el peso de la contribución de cada una de las partes de la prueba original a la puntuación total de modo que dicha contribución refleje la importancia de cada una de las capacidades (conocimiento lingüístico, comprensión oral y escrita, expresión escrita y conocimientos culturales) en el currículum español para las lenguas extranjeras en la educación secundaria. Aunque Lord (1980)

presentó procedimientos de ponderación basados en la función de información, en este estudio sólo se han utilizado para la valoración del efecto de diversas ponderaciones y comparaciones basadas en la fiabilidad de la TCT.

- Construir una escala de habilidad para el dominio de la lengua inglesa en el marco de la Teoría de la Respuesta al Ítem de modo que pueda ser utilizada en trabajos futuros, bien con la misma prueba y muestras diferentes, bien con una prueba similar que incluya el número adecuado de ítems puente o de anclaje para llevar a cabo procedimientos de equiparación, así como similar en cuanto a nivel y destrezas evaluadas y muestras comparables de modo relevante.

Método y procedimiento

Para alcanzar estos objetivos, se llevaron a cabo los siguientes pasos:

- Se analizó la unidimensionalidad de la prueba, como un requisito previo a la aplicación de los procedimientos de la TRI, y se estudió el ajuste de los modelos de uno, dos y tres parámetros de la TRI a los datos para aplicar los cálculos adecuados a la mejor estimación posible de los parámetros a, b y c, y de θ.
- Se estimó una puntuación TRI global para cada alumno, basándose en el modelo que proporciona un mejor ajuste a las características de los datos, analizándose la similitud entre la puntuación directa original (PD) y la puntuación global TRI estimada (PTRI). Asimismo, se generaron puntuaciones TRI por subescalas para cada alumno.
- Se estableció teóricamente la importancia de las subescalas, desarrollando diversas fórmulas para la asignación de pesos diferentes a cada subescala, y se analizó la incidencia de las diferentes combinaciones lineales de pesos propuestas sobre la fiabilidad de la puntuación final de rendimiento de los sujetos.
- Se calculó la nueva puntuación global a partir de las puntuaciones tipificadas de las subescalas según la fórmula elegida y se reescalaron las puntuaciones obtenidas con base en una escala de uso internacional, analizándose la similitud entre la puntuación directa original (PD), la puntuación global TRI (PTRI) y la puntuación global TRI ponderada (PTRIP) calculada con los pesos establecidos por la combinación lineal elegida.

Resultados

Análisis de la unidimensionalidad de la prueba de rendimiento en lengua inglesa y del ajuste de los modelos logísticos de la TRI a los datos

Como un requisito previo a la aplicación de los procedimientos de la TRI, se analizó el cumplimiento del supuesto de unidimensionalidad de la prueba bajo estudio, para lo que se aplicó un análisis factorial para variables dicotómicas tomando como base la matriz de correlaciones tetracóricas entre todos los ítems (Bock y Aitkin, 1981), utilizando el programa TESTFACT (Wilson, Wood, Kandola y Gibbons, 1991). Se verificó la existencia de un único factor predominante, lo que indica que se trata de una prueba de rendimiento unidimensional.

En un segundo paso, se analizó el ajuste de los modelos de uno, dos y tres parámetros de la TRI a los datos utilizando el programa BILOG 3 (Mislevy y

Bock, 1990). Los análisis indicaron que el modelo con mejor ajuste resultó ser el de tres parámetros, seguido por el de dos parámetros y siendo el inferior en esta comparación el modelo de un parámetro. Asimismo, se calcularon los contrastes correspondientes para estimar la significatividad de las diferencias entre los ajustes de los diferentes modelos. Los datos indican que hay diferencia significativa entre el modelo de dos parámetros y el de un parámetro ($\chi^2_{91,\ 0,99}$ = 1071,76; p<.01) aunque no se encontró diferencia significativa entre los modelos de dos y tres parámetros ($\chi^2_{91,\ 0,99}$ = 94,86). No obstante, se calculó asimismo el número de ítems que se ajustaban a cada uno de los modelos como un criterio adicional para estimar la bondad del ajuste de los mismos. Se encontró que se ajustaban un total de 62 ítems bajo el modelo de un parámetro, 87 ítems bajo el modelo de dos parámetros y 91 ítems, es decir, la totalidad de los ítems que componen la prueba, bajo el modelo de tres parámetros. Este procedimiento es similar al utilizado en otros trabajos sobre la selección de modelos logísticos para la construcción de pruebas de rendimiento (Gil, Suárez y Martínez Arias, 1999). Los resultados obtenidos se presentan en la tabla I.

TABLA I

Resultados de los análisis del ajuste de los datos a los tres modelos logísticos de la TRI

MODELO LOGÍSTICO	Grados de Libertad	$-2 \log \lambda$
L1 (modelo de un parámetro)	4.228	85781,8614
L2 (modelo de dos parámetros)	4.137	84710,1008
L3 (modelo de tres parámetros)	4.046	84615,2374

Contrastes: L1-L2 $\chi^2_{91,\ 0,99}$ = 1.071,76 (p<.01)

L2-L3 $\chi^2_{91,\ 0,99}$ = 94,86 (p>.05)

A partir de estos resultados, se consideró que el modelo de tres parámetros era el que debía ser considerado para la estimación de los parámetros y la asignación de puntuaciones TRI para los sujetos al ser la mejor opción dado su mayor ajuste a la estructura de los datos, ya que 1) su grado de ajuste se diferenciaba clara y significativamente del modelo de un parámetro; 2) tenía un mejor ajuste global que el modelo de dos parámetros, a pesar de que la diferencia en el ajuste entre este modelo y el de dos parámetros, en términos de la diferencia en χ^2, no fuese significativa; 3) el número de ítems que se ajustaban era mayor, la totalidad de los ítems, que bajo el modelo de uno o dos parámetros; y 4) este modelo posee una estructura teórica más flexible, general y mejor adaptada a los ítems de elección múltiple, siendo los modelos de uno y dos parámetros simplificaciones de éste. Por ello, los cálculos subsiguientes en el trabajo se realizaron utilizando el modelo de tres parámetros con el programa BILOG y utilizando para la estimación de los parámetros el procedimiento de máxima verosimilitud marginal.

Análisis de la similitud entre puntuaciones

Tras generar una puntuación TRI global para cada alumno (PTRI), se analizó la

similitud entre la puntuación directa (PD) y la puntuación global TRI. Para ello, se calculó la correlación existente entre ambas puntuaciones resultando una correlación extremadamente alta y significativa (r=.9928, p<.0001). Este resultado indica que, independientemente de la aportación teórica que incorpora la TRI y las propiedades que poseen las puntuaciones derivadas de la misma, los resultados que se obtienen, en cuanto a la estimación del nivel de habilidad de los sujetos con esta prueba, con la aplicación de la Teoría Clásica de los Tests (TCT) y con la aplicación de la TRI son equivalentes en un altísimo grado.

De modo similar se generaron puntuaciones directas y puntuaciones TRI por subescalas para cada alumno. La tabla II presenta la matriz de intercorrelaciones tanto de las puntuaciones directas como de las puntuaciones TRI, que resultaron ser todas ellas estadísticamente significativas.

Se observa en estas tablas de intercorrelaciones que la subescala de conocimientos lingüísticos presenta, en conjunto, el nivel más alto de correlación con el resto de las subescalas, especialmente con la de comprensión escrita. Una hipótesis que surge del análisis de estos datos puede concretarse en la idea de que el nivel de conocimiento lingüístico (reglas y convenciones gramaticales, léxico, morfología, sintaxis, etc.) es un requisito previo, o al menos un elemento facilitador, del desarrollo de los aspectos comunicativos y, en especial, de la expresión escrita.

Otro resultado a destacar de estas tablas es el hecho de que la subescala de conocimientos culturales presente como término medio las correlaciones más bajas con el resto de las escalas. Esto puede deberse, probablemente, al hecho de que la subescala de conocimientos culturales está formada únicamente por 3 ítems, lo que proporciona poca estabilidad en la medida, aparte de que los conocimientos

TABLA II

Tablas de intercorrelaciones entre las puntuaciones directas de las subescalas y entre las puntuaciones TRI de las subescalas

Puntuaciones directas	PD-CE	PD-CO	PD-CC	PD-EE
PD-CL	.6976	.5871	.3819	.7067
PD-CE		.5232	.4128	.5926
PD-CO			.3770	.5330
PD-CC				.3438
Puntuaciones TRI	**TRI-CE**	**TRI-CO**	**TRI-CC**	**TRI-EE**
TRI-CL	.6605	.5851	.3825	.7290
TRI-CE		.4969	.4005	.5982
TRI-CO			.3805	.5334
TRI-CC				.3585

CL = Conocimientos lingüísticos
CE = Comprensión escrita
CO = Comprensión oral
CC = Conocimientos culturales
Nota: todas estas correlaciones p<.001

culturales relacionados con una lengua extranjera no necesariamente están relacionados con el aprendizaje de la misma, dado que es obvio que puede tenerse un amplio conocimiento de la cultura, costumbres, situación actual e historia de países de lengua extranjera sin tener un dominio de dicha lengua.

Adicionalmente, el hecho de que en esta tabla todas las subescalas presenten unas intercorrelaciones positivas y significativas refleja la propiedad de unidimensionalidad de la prueba antes mencionada.

Se calcularon, asimismo, las correlaciones entre las puntuaciones directas originales y las puntuaciones derivadas de la TRI para cada subescala. La correlación entre estas puntuaciones para la subescala de conocimientos lingüísticos fue igual a .9863, para la de comprensión escrita de .9635, para la de comprensión oral de .9953, para la de conocimientos culturales de .9943 y para la de expresión escrita de .9870, todas ellas con una probabilidad menor que .001.

La conclusión que se obtiene del análisis de estas correlaciones es similar a la comentada en cuanto a la correlación entre la puntuación directa global y la puntuación global TRI para el conjunto de la prueba, volviendo a mostrar estos resultados, como era de esperar, la equivalencia desde un punto de vista práctico de la asignación de puntuaciones individuales a los alumnos mediante el cálculo de la puntuación directa derivada de la TCT y asignación de puntuaciones θ derivadas de la TRI.

Asignación de pesos a cada subescala y análisis de su incidencia sobre la fiabilidad

Se estableció teóricamente, con base en la opinión de expertos en la enseñanza del inglés, la importancia de las subescalas según se deriva del análisis del currículum de la educación secundaria. La opinión de los expertos se resume en la idea de que la enseñanza de la lengua inglesa debe estar orientada de modo que se enfaticen los aspectos comunicativos y, de modo especial, los aspectos de comprensión, siendo clara, en este sentido, la inadecuación del diseño global de la prueba para el currículum español. Partiendo de la opinión experta, se establecieron diversas fórmulas para la asignación de pesos diferentes a cada subescala.

Teniendo en cuenta las fiabilidades originales de cada subescala considerada independientemente y estimadas mediante el α de Cronbach (1951) (α=.8781 para la subescala de conocimientos lingüísticos, α=.7015 para la de comprensión escrita, α=.6602 para la de comprensión oral, α=.5085 para la de conocimientos culturales y α=.7567 para la de comprensión escrita), es de esperar que las fórmulas que otorguen un mayor peso a las escalas con mayor fiabilidad en origen y un menor peso a las subescalas con una menor fiabilidad en origen proporcionen una escala ponderada con una mayor fiabilidad de conjunto.

Se consideró conveniente no considerar la escala de conocimientos culturales para el cálculo de una puntuación final TRI por dos razones: en primer lugar, debido a su escaso número de ítems y, por lo tanto, a su baja fiabilidad como escala independiente y, en segundo lugar, debido a su no necesaria relación teórica con el conjunto de las subescalas que constituyen la prueba.

La tabla III muestra los pesos originales de cada parte de la prueba según su construcción original y según las varias propuestas derivadas de la opinión de los expertos. Asimismo, presenta la fiabilidad global resultante para cada una de las fórmulas de ponderación.

TABLA III

Pesos originales de cada parte de la prueba según su construcción original y según las fórmulas derivadas de la opinión de los expertos (sobre 100). Fiabilidad resultante para cada una de las fórmulas de ponderación

Fórmula	Subescalas				Fiabilidad α
	CL	CE	CO	EE	
1	63,6364	12,5	13,6364	10,2273	.8995
2	50	16,6667	16,6667	16,6667	.8992
3	33,3333	22,2222	22,2222	22,2222	.8831
4	25	25	25	25	.8687

CL = Conocimientos lingüísticos
CE = Comprensión escrita
CO = Comprensión oral
EE = Expresión escrita

El mecanismo utilizado para la estimación de la fiabilidad resultante de la aplicación de las fórmulas de ponderación se basa en la idea de considerar al conjunto de la prueba como una batería de tests y considerar a cada subescala como un test independiente. Para la estimación del coeficiente de fiabilidad global a partir de los coeficientes de fiabilidad, varianzas, covarianzas y pesos de las subescalas, se utilizó la siguiente fórmula (Muñiz, 1994).

$$\rho_{XX'} = \frac{\sum_{j=1}^{n} a_j^2 \sigma_j^2 \rho_{jj'} + \sum_{j=1}^{n} \sum_{k=1, j \neq k}^{n} a_j a_k \sigma_{jk}}{\sum_{j=1}^{n} a_j^2 \sigma_j^2 \rho_{jj'} + \sum_{j=1}^{n} \sum_{k=1, j \neq k}^{n} a_j a_k \sigma_{jk}}$$

donde:

n = número de subescalas
σ_j^2 = varianzas de las subescalas
$\rho_{jj'}$ = coeficientes de fiabilidad de las subescalas
σ_{jk} = covarianzas entre las subescalas
a_j y a_k = ponderaciones de las subescalas

La combinación lineal 1 es el resultado de considerar la prueba original en sí misma como una fórmula de ponderación, por lo que los pesos en este caso son simplemente la traducción en porcentajes del número de ítems relativo al número de ítems de la prueba, una vez excluidos los tres ítems de conocimientos culturales. Lógicamente, la fiabilidad resultante es la propia fiabilidad de la prueba original. La fórmula 2 constituye una ligera modificación en relación con la primera al otorgar un peso del 50% a la subescala de conocimientos lingüísticos y repartir el 50% restante equitativamente entre las subescalas dirigidas a la medida de las capacidades comunicativas.

La combinación lineal 3 se basa en la idea de otorgar un mayor peso a las subescalas de competencia comunicativa (un 66,67% en su conjunto) distribuido de modo equitativo entre ellas, lo que implica conceder un mayor peso a los aspectos de comprensión sobre los de expresión escrita, y disminuyendo el peso del componente de conocimientos lingüísticos. Por último, la fórmula 4 plantea la idea de conceder a cada una de las subescalas de la prueba un peso equivalente, otorgando por tanto un peso del 75% a las capacidades comunicativas y el 25% restante a los conocimientos lingüísticos.

Se observa en la tabla 3 que la incidencia de la aplicación de las fórmulas estudiadas sobre la fiabilidad del conjunto de las subescalas estimada mediante el α de Cronbach es pequeña (dándose la máxima diferencia entre la fiabilidad original y la fiabilidad resultante en el caso de la fórmula 4 y siendo ésta igual a .0308). La escasa variación de la fiabilidad global, a pesar de la variación significativa en los porcentajes asignados por las fórmulas, se explica al existir un considerable grado de intercorrelación entre las subescalas.

A la vista de los resultados en los que se mantiene un nivel muy alto de fiabilidad para el conjunto de las subescalas parece posible, pues, utilizar para la asignación de puntuaciones a los alumnos la fórmula que presente teóricamente un mejor ajuste a la importancia que el currículum español concede a cada una de las habilidades medidas por la prueba. Por ello, se decidió seleccionar la combinación lineal 3, PTRI = (33,33 × CL) + (22,22 × CE) + (22,22 × CO) + (22,22 × EE), dado que desde el punto de vista teórico su estructura parece la más adecuada y desde el punto de vista práctico su incidencia negativa sobre la fiabilidad se considera mínima al ser igual a .0164.

Análisis de la similitud entre la puntuación directa original, la puntuación global TRI y la puntuación global TRI calculada con los pesos establecidos por la fórmula elegida

Subsiguientemente, se calculó la nueva puntuación global a partir de las puntuaciones TRI tipificadas para cada subescala según la combinación lineal seleccionada, resultando por tanto una nueva puntuación TRI ponderada (PTRIP), y se reescalaron las puntuaciones obtenidas con base en una escala de uso internacional de media 500 y desviación típica 100 (Keeves, 1990, 1992). Se eligió este reescalamiento al ser el habitualmente utilizado por la *International Association for the Evaluation of Educational Achievement* (IEA) en los estudios internacionales que coordina (Binkley y Rust, 1994; Elley, 1994; IEA Secretariat, 1998; Martin y Kelly, 1996).

Las correlaciones entre las nuevas puntuaciones TRI ponderadas con base en la combinación lineal y las puntuaciones directas originales y la puntuación TRI sin ponderar son extremadamente altas (.9828 con la puntuación directa original y.9816 con la puntuación TRI sin ponderar) y, por supuesto, notoriamente significativas ($p<.001$). Este resultado pone de manifiesto que el procedimiento de ponderación seguido tiene un escaso efecto distorsionador de las puntuaciones originales debido al importante grado de intercorrelación entre las escalas.

ESTUDIO 2

LA CONSTRUCCIÓN DE UNA PUNTUACIÓN PORCENTUAL DE RENDIMIENTO BASADA EN LA TEORÍA DE RESPUESTA AL ÍTEM

Como se señaló en la introducción, al tratar del sistema de presentación de los resultados, la presentación en términos de porcentajes tiene la ventaja sobre las puntuaciones en términos de una escala de rendimiento basada en la Teoría de Respuesta al Ítem (TRI) de ser de más fácil e intuitiva comprensión, además de relacionar la puntuación de los sujetos con el número de cuestiones de la prueba de un modo relativo. Por ello, se ha desarrollado una puntuación porcentual de rendimiento como complemento de la puntuación ponderada basada en la TRI descrita en los apartados anteriores, estimándose ésta también desde el modelo de tres parámetros de la TRI. Asimismo, se trata de una puntuación ponderada que se ha calculado a partir de las puntuaciones de los sujetos en las cuatro subescalas de la

prueba, siendo la ponderación aplicada la misma que la que se utilizó en la puntuación TRI y que responde a la importancia atribuida al contenido de cada subtest en el currículum español para la enseñanza de la lengua inglesa.

La puntuación TRI expresada en una escala de media 500 y desviación típica 100 permite ordenar a los sujetos con respecto a la media y poner en relación el rendimiento del alumno con variables del contexto a través de análisis estadísticos inferenciales. Sin embargo, esta puntuación TRI no informa sobre el porcentaje de ítems de la prueba que los alumnos han respondido correctamente. Esta limitación puede complementarse con una puntuación porcentual que cumpla dos requisitos que preserven las características impuestas en la puntuación TRI:

- la estimación de la puntuación debe realizarse desde el modelo de tres parámetros de la TRI.
- la importancia de cada subescala en el porcentaje global debe ser ponderada de modo que refleje los pesos atribuidos a cada subárea en el currículum español para esta materia.

Método y procedimiento

Para la construcción de la puntuación porcentual con las características especificadas, se utilizó el siguiente procedimiento:

- Estimación de los parámetros de los ítems de discriminación, dificultad y adivinación al azar en cada una de las cuatro subescalas por separado.
- Estimación del nivel de aptitud de cada alumno en cada una de las subescalas de la prueba.
- Cálculo de la puntuación verdadera de los alumnos en cada ítem a través de una transformación no lineal del nivel de habilidad estimado en probabilidad de acertar al ítem correctamente, la cual corresponde con la puntuación verdadera en el ítem. Esta transformación del nivel de aptitud a una escala que se corresponde con la de la prueba facilita la comparación de los sujetos y permite la obtención de una puntuación porcentual desde la TRI. Para la transformación no lineal de la habilidad en puntuación verdadera se utilizó el modelo de tres parámetros con las estimaciones paramétricas de los ítems obtenidas en la calibración de las subescalas.
- Cálculo de la puntuación verdadera en cada subescala y en el test global. Para ello, se calcularon las puntuaciones verdaderas obtenidas por cada alumno en cada subescala y, a continuación, se sumaron las puntuaciones de las cuatro subescalas obteniéndose así la puntuación verdadera de cada sujeto en la prueba.
- Conversión de las puntuaciones verdaderas de cada subescala en porcentajes y obtención del porcentaje global ponderado con los pesos seleccionados anteriormente en función del currículum español. Este porcentaje se calculó mediante la siguiente fórmula:

$$\begin{aligned} PVP = {} & (0{,}333 \times PorCL) + \\ & + (0{,}222 \times PorCE) + \\ & + (0{,}222 \times PorCO) + \\ & + (0{,}222 \times PorEE) \end{aligned}$$

Siendo PorCL, PorCE, PorCO y PorEE los porcentajes de respuestas correctas obtenidos a partir de las puntuaciones verdaderas en las subescalas de conocimientos lingüísticos, comprensión escrita, comprensión oral y expresión escrita, respectivamente.

Según este porcentaje global ponderado (PVP), el porcentaje promedio

de ítems acertados por los sujetos se sitúa en un 50,73%, siendo su desviación típica $S_x = 16,11$.

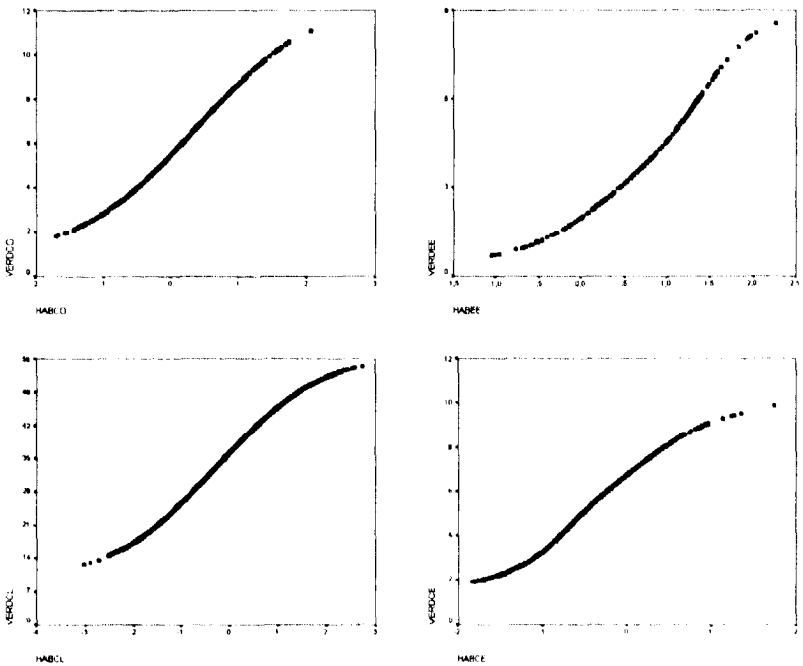
- Comparación de los porcentajes ponderados y sin ponderar tanto de la TRI como de la TCT, calculados estos últimos a partir de la puntuación directa, con la puntuación global TRI ponderada. Teniendo en cuenta que la única diferencia entre la puntuación TRI ponderada y el porcentaje global ponderado consiste en una transformación de escala, parece lógico esperar que la correlación de la puntuación TRI con el porcentaje global ponderado sea algo superior que con el resto de las puntuaciones. Asimismo, se esperaba que el porcentaje global ponderado obtenido con la puntuación directa no se diferenciara significativamente del porcentaje global ponderado obtenido con la TRI.

RESULTADOS

En la figura I se muestran los diagramas de dispersión que relacionan las aptitudes

FIGURA 1

Curvas características de las subescalas de Conocimientos Lingüísticos, Comprensión Escrita, Comprensión Oral y Expresión Escrita



VERDCL, VERDCE, VERDCO, VERDEE = Puntuaciones verdaderas en Conocimientos Lingüísticos, Comprensión Escrita, Comprensión Oral y Expresión Escrita.
HABCL, HABCE, HABCO, HABEE = Niveles de habilidad estimados en Conocimientos Lingüísticos, Comprensión Escrita, Comprensión Oral y Expresión Escrita.

estimadas en cada subescala con la correspondiente puntuación verdadera producto de la transformación no lineal con el modelo de tres parámetros. La curva que relaciona las dos puntuaciones de cada subescala, situando en el eje de abscisas la puntuación verdadera del test y en el eje de coordenadas el nivel de aptitud estimado, se denomina Curva Característica del Test, en este caso de la subescala. Las correlaciones lineales de Pearson entre las dos puntuaciones de las escalas CL, CE,CO, y EE son r=0,9961 (p=0,000), r= 0,9897 (p=0,000), r=0,9946 (p=0,000) y r=0,9696 (p=0,000), respectivamente.

En la tabla IV figuran los estadísticos descriptivos de las cuatro puntuaciones porcentuales: el porcentaje basado en la puntuación directa del test (PD), el porcentaje basado en la puntuación verdadera del test (PV), el porcentaje basado en las puntuaciones directas de las subescalas ponderadas (PDP) y el porcentaje basado en las puntuaciones verdaderas de las subescalas ponderadas (PVP).

TABLA IV

Estadísticos descriptivos de las cuatro puntuaciones porcentuales

Porcentaje	Media	Desviación Típica	Rango	Mínimo	Máximo	N
PD	57,48	17,60	84,09	14,77	98,86	4.320
PV	57,33	15,53	73,36	21,27	94,63	4.320
PDP	52,15	18,43	91,07	8,33	99,40	4.143
PVP	50,73	16,11	75,38	18,27	93,64	4.320

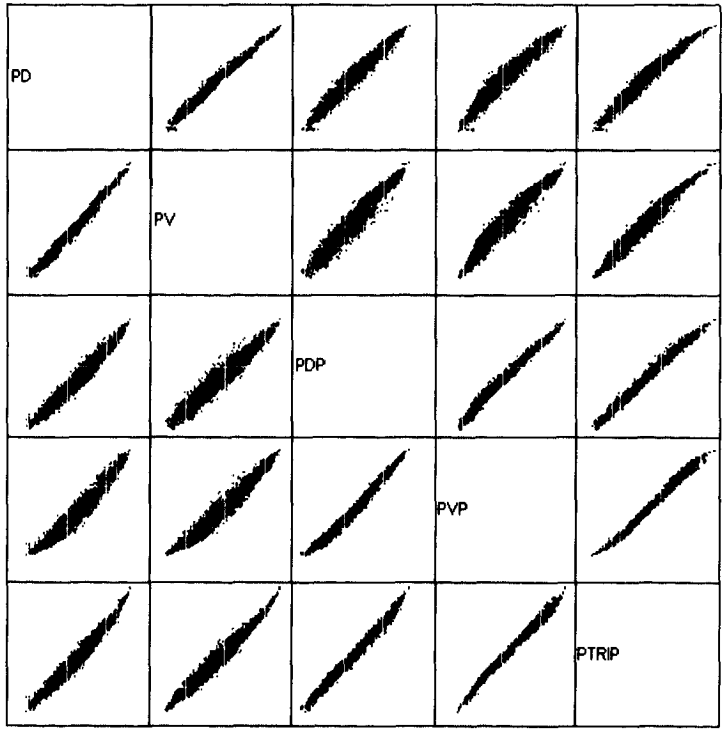
Como puede apreciarse, la ponderación de los porcentajes supone una ampliación del rango, sobre todo en el caso de la puntuación directa, lo que conlleva a un aumento de la variabilidad en las distribuciones ponderadas. Asimismo se aprecia que el porcentaje promedio de ítems acertados por los alumnos disminuye significativamente cuando se ajusta la puntuación al currículum español.

Si se comparan los dos porcentajes ponderados, se observa que el porcentaje basado en la TRI disminuye ligeramente el porcentaje promedio de aciertos y la variabilidad del porcentaje clásico, disminución que no llega a ser significativa desde un punto de vista práctico. La diferencia entre ambos porcentajes radica en el estrechamiento del rango en el caso de la puntuación basada en la TRI de un 16% aproximadamente, lo que lleva a mejorar el porcentaje promedio de la zona baja de la distribución, y a empeorar el promedio en los valores extremos de la zona alta.

En la figura II se representan a través de una matriz de diagramas de dispersión las relaciones entre los porcentajes obtenidos entre sí así como con la puntuación global ponderada TRI que sirve de criterio. Todas las relaciones son lineales y con coeficientes de correlación por encima del 0,97 en todos los casos. Las correlaciones más altas se obtienen cuando se analiza la concordancia entre las puntuaciones no ponderadas entre sí (r=0,9933), o los porcentajes ponderados entre sí (r=0,9924). Según este resultado, parece ser que la importancia atribuida a cada subescala a

FIGURA II

Matriz de diagramas de dispersión de las intercorrelaciones entre la puntuación TRI (PTRI), la puntuación directa (PD), la puntuación verdadera estimada (PV), el porcentaje ponderado basado en la puntuación directa (PDP) y el porcentaje ponderado basado en la puntuación verdadera (PVP)



PD = Puntuación directa
PV = Puntuación verdadera
PDP = Porcentaje ponderado basado en la puntuación directa
PVP = Porcentaje ponderado basado en la puntuación verdadera

través de los pesos determina más el grado de relación entre las variables que la puntuación original, puntuación directa o puntuación verdadera, en la que se basa el porcentaje.

En cuanto a la relación con la variable criterio, como era de esperar, con las variables ponderadas se obtienen unos coeficientes de correlación superiores a los correspondientes a las puntuaciones no ponderadas, siendo ligeramente superior la correlación de la puntuación porcentual basada en la TRI al porcentaje ponderado clásico.

Una vez que se dispone de una puntuación porcentual de cada alumno, y sabiendo que dicha puntuación ofrecería resultados equivalentes a los obtenidos utilizando la puntuación global TRI, pero en una escala más intuitivamente

interpretable, se podría estudiar con esta variable dependiente la influencia de diversas variables independientes, por ejemplo, contextuales, socioeconómicas, culturales, etc., sobre el rendimiento de los alumnos.

CONCLUSIONES

El procedimiento seguido para la construcción de una puntuación a partir de una prueba que no se ajusta en su diseño global a las especificaciones que serían deseables en cuanto a la importancia de cada una de las partes que la componen pone de manifiesto la posibilidad de considerar a cada una de estas partes como una subescala o subtest que forma parte una batería de test y ponderar su peso en una combinación lineal que se utiliza para el cálculo de una nueva puntuación. Por otro lado, el caso específico con el que se ejemplifica este procedimiento, la puntuación de rendimiento en el estudio sobre la enseñanza y el aprendizaje de la lengua inglesa revela que la variación en cuanto a la fiabilidad global de la prueba utilizando las puntuaciones originales, puntuaciones TRI o puntuaciones ponderadas con las combinaciones lineales analizadas es mínima. Asimismo, se pone de manifiesto la equivalencia, en términos prácticos, de la asignación de las puntuaciones a los sujetos utilizando cualquiera de las tres puntuaciones antes mencionadas, aunque la mayor capacidad explicativa, la mayor flexibilidad y generalidad del modelo TRI hace que se considere deseable utilizar las puntuaciones derivadas del mismo, tanto para la asignación de puntuaciones a los sujetos como para establecer una escala que pueda tener utilidad en el futuro.

Por último, se ha puesto de manifiesto la posibilidad de generar puntuaciones porcentuales basadas en la TRI derivadas de la estimación de la puntuación verdadera, pudiendo la utilización de dichas puntuaciones porcentuales ser de utilidad, tanto con finalidades descriptivas o inferenciales, para la presentación de los resultados de rendimiento para el público general.

BIBLIOGRAFÍA

BEATON, A. E.; JOHNSON, E. G. (1992): «Overview of the Scaling Methodology Used in the National Assessment», en *Journal of Educational Measurement,* 29, 2 (1992), pp. 163-175.

BEATON, A. E.; ZWICK, R.: «Overview of the National Assessment of Educational Progress», en *Journal of Educational Statistics,* 17 (1990), pp. 95-109.

BINKLEY, M.; RUST, K. (Eds.): *Reading Literacy in the United States: Technical Report.* Washington, National Center for Education Statistics: Office of Educational Research and Development: U.S. Department of Education. U.S. Government Printing Office, 1994.

BOCK, R. D.; AITKIN, M.: «Marginal maximum likelihood estimation of item parameters: application of an EM algorithm». *Psychometrika,* 46 (1981), pp. 443-459.

BOCK, R. D.; MISLEVY, R. J.; WOODSON, C. (1982): «The next Stage in Educational Assessment», en *Educational Researcher,* 11, 3 (1982), pp. 4-11.

BONNET, G. (Ed.): *The Effectiveness of the Teaching of English in the European Union.* Paris, Direction de l'Évaluation et de la Prospective (DEP), 1997.

CRONBACH, L. J.: «Coefficient Alpha and the Internal Structure of Tests», *Psychometrika,* 16 (1951), pp. 297-334.

DIRECTION DE L'ÉVALUATION ET DE LA PROSPECTIVE (DEP): *Espagne, France, Suède: Évaluation des Connaissances et Compétences en Anglais des Élèves de 15-16 Ans.* Paris, Ministère de l'Éducation Nationale, 1997.

ELLEY, W. B. (Ed.): *The IEA Study of Reading Literacy: Achievement and Instruction in*

Thirty-Two School Systems. Oxford, Pergamon, 1994.

GIL ESCUDERO, G.; ALABAU BALCELLS, I.: *Evaluación Comparada de la Enseñanza y el Aprendizaje de la Lengua Inglesa: España, Francia, Suecia.* Madrid, Ministerio de Educación y Cultura, 1997.

GIL ESCUDERO, G. A.; SUÁREZ FALCÓN, J. C. y MARTÍNEZ ARIAS, R.: «Aplicación de un procedimiento iterativo para la selección de modelos de la Teoría de la Respuesta al Item a una prueba de rendimiento lector», en *Revista de Educación,* 1999.

HAMBLETON, R. K.; COOK, L. L.: «Latent Trait Models and their Use in the Analysis of Educational Test Data», en *Journal of Educational Measurement,* 14 (1997), pp. 75-96.

HAMBLETON, R. K.; JONES, R. W.: «Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development», en *Educational Measurement: Issues and Practice,* 12, 3 (1993), pp. 38-47.

HAMBLETON, R. K.; SWAMINATHAN, H.: *Item Response Theory: Principles and Applications.* Boston, Kluwer-Nijhoff, 1985.

HAMBLETON, R. K.; SWAMINATHAN, H. y ROGERS, H. J.: *Principles and Applications of Item Response Theory.* Beverly Hills, Sage, 1991.

HULIN, C. L.; DRASGOW, F. y PARSONS, C. K.: *Item Response Theory: Applications to Psychological Measurement.* Homewood, Dow Jones-Irwin, 1983.

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

KEEVES, J. P.: «Scaling Achievement Test Scores», en H. J. WALBERG y G. D. HAERTEL (Eds.): *The International Encyclopedia of Educational Evaluation.* Oxford, Pergamon Press, 1990.

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

LORD, F. M.: *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, LEA, 1980.

MARTIN, M. O.; KELLY, D. L. (Eds.): Third International Mathematics and Science Study: Technical Report. Boston College, Chesnut Hill, 1996.

MARTÍNEZ ARIAS, R.: *Psicometría: Teoría de los Tests Psicológicos y Educativos.* Madrid, Síntesis, 1995.

MISLEVY, R. J.; BOCK, R. D.: *Bilog 3. Item Analysis and Test Scoring with Binary Logistic Models.* Mooresville, Scientific Software Inc., 1990.

MUÑIZ, J.: *Teoría Clásica de los Tests.* Madrid, Pirámide, 1994.

Real Decreto 1007/1991, de 14 de junio, por el que se establecen las enseñanzas mínimas correspondientes a la Educación Secundaria Obligatoria. ABOE@ número 152, de 26 de junio de 1991.

ROSIER, M. J.; ROSS, K. N.: «Sampling and Administration», en J. P. KEEVES (Ed.): *The IEA Technical Handbook.* The Hague, The International Association for the Evaluation of Educational Achievement (IEA), 1992.

ROSS, K. N.: *Sampling Manual for the IEA Reading Literacy Study.* Hamburg, University of Hamburg, 1991.

VAN DER LINDEN, W. J.; HAMBLETON, R. K., *Handbook of Modern Item Response Theory.* New York, Springer-Verlag, 1997.

WILSON, D. T.; WOOD, R.; KANDOLA, P.; GIBBONS, R.: *Testfact. Test Scoring. Item Statistics and Item Factor Analysis.* Mooresville, Scientific Software Inc., 1991.