Evaluación y calificación de resúmenes de textos expositivos en el aula de ILE/IFE: la guía "BABAR"

Irina Argüelles Álvarez
Universidad Politécnica de Madrid

Resumen

Los estudiantes de Inglés para Fines Específicos (IFE) practican la técnica del resumen en el aula cuando trabajan con textos expositivos relacionados con sus estudios. Si se lleva el resumen al aula dentro de una perspectiva pedagógica en la que enseñanza y evaluación se entienden como ligadas en el sistema de enseñanza-aprendizaje, se llega a la cuestión de cómo evaluar el resumen escrito por un estudiante. Para que sea formativa, la evaluación debe responder a distintas preguntas: ¿Qué está bien? ¿Qué está peor? ¿Qué puede hacerse para mejorar? En el caso concreto del resumen, no es sólo importante la calidad de la expresión escrita, sino que debe tenerse en cuenta también la calidad con la que el escritor recoge el mensaje en general y, en particular, las ideas básicas de un texto origen; además, en el contexto de IFE intervienen factores lingüísticos fundamentales. Aunque se han propuesto baremos para evaluar la expresión escrita, es difícil encontrar guías para evaluar resúmenes y, por esta razón, este trabajo repasa algunas de las propuestas de evaluación de la expresión escrita más conocidas y aporta un baremo para la calificación de resúmenes de textos expositivos en el aula de IFE.

Key words: resumen, evaluación, expresión escrita, inglés como lengua extranjera (ILE), IFE

Abstract

Students of English for Specific Purposes (ESP) need to summarize and to write summaries in the classroom when they work with expository texts related to their studies. If we adopt a pedagogical perspective where teaching and assessment are understood to be linked to the teaching-learning system, we will have to address the matter of how summaries written by students are assessed. If the required assessment is to be formative, it must provide answers to questions such as: What is good? What is poor? What can the student do to improve? In this case, not only the quality of the writing should be assessed but also the quality with which the writer transmits the message in general and, in particular, the main ideas from the original text. In the context of ESP, linguistic factors have also to be considered owing to their importance. Guides for the assessment of writing have been produced but it is difficult to find guides directed towards the assessment of summary writing. In this article, a number of well-known writing assessment proposals are reviewed and a new guide for the assessment of summaries of expository texts is presented.

Palabras clave: summary, assessment, writing, English as a Foreign Language (EFL), ESP

Introducción

En la actualidad, los trabajos de investigación que se aproximan a los procesos de evaluación de la expresión escrita lo hacen desde una perspectiva formativa en la que el papel preponderante es el de retroacción o retroalimentación (Cassany et al., 1994; White, 1994; Grabe & Kaplan, 1996). Dentro de esta perspectiva formativa de la evaluación, uno de los principales objetivos es el aprendizaje y, en lo que se refiere a la evaluación de la expresión escrita, se han propuesto distintas alternativas para tratar los errores que el estudiante comete al escribir. Pero, como recuerdan Grabe y Kaplan (1996: 395), la evaluación de la expresión escrita no supone sólo la respuesta del profesor, sino que implica mecanismos más formales que suponen calificaciones.

Ciertamente, en muchos casos dentro del contexto académico se hace imprescindible contar con dichos mecanismos que permitan ofrecer una calificación "justa"[1] al trabajo de un estudiante. Estos mecanismos de evaluación, más concretamente, los métodos de calificación de pruebas directas de evaluación de la expresión escrita, son el tema central de la introducción al trabajo empírico que se presenta en este estudio. Uno de los principales objetivos de este trabajo es promover la reflexión sobre cuestiones de validez y fiabilidad de las pruebas directas de expresión escrita, dado que la corrección de éstas supone una preocupación constante por la "justicia" de la nota. Para ello, se parte del proceso de elaboración de las pruebas directas de expresión escrita y se repasan distintas propuestas de calificación para, finalmente, aportar una nueva guía de calificación para resúmenes de textos expositivos en inglés como lengua extranjera.

Las pruebas directas de evaluación de la expresión escrita

Las pruebas directas de evaluación de la expresión escrita son aquéllas que evalúan la aptitud del estudiante para escribir. Esta evaluación se basa en un texto escrito por el estudiante que suele tener más de cien palabras y que responde a una actividad especialmente formulada para que el escritor pueda producir discurso. Cuatro características definen una prueba directa de evaluación (Grabe & Kaplan, 1996: 229; Ferris & Hedgcock, 1998: 233):

(a) claridad: el estudiante no tiene que perder tiempo en entender qué debe hacer,

(b) validez: una prueba válida permite la producción de escritos cuyas notas están en consonancia con las obtenidas durante el curso. Como en prácticamente cualquier contexto de evaluación, las pruebas directas necesitan cumplir con cuatro tipos de validez (Hamp-Lyons, 1991b: 70-73; Ferris & Hedgcock, 1998: 230-232):

- Validez superficial (*face validity*) supone que tanto alumnos como profesores entienden que el instrumento es adecuado para lo que se quiere evaluar. Según Harris (1969: 21) y Oller (1979: 52), este tipo de validez sólo es importante en la medida en que pueda afectar a la realización de la misma por parte del estudiante.
- Validez de criterio (*criterion validity*)[2]. Si el instrumento es válido, producirá calificaciones similares a las de otro instrumento ya validado y usado en condiciones similares.
- Validez de contenido (*content validity*). Tiene relación con la eficacia del método para obligar al estudiante a demostrar sus habilidades y estrategias en esa área específica.
- Validez de constructo[3] (*construct validity*). Esencial en la evaluación directa de la expresión escrita, se refiere a si la prueba evalúa la capacidad o habilidad de escribir.

(c) interés: la prueba tiene interés suficiente para motivar la redacción por parte del estudiante y la lectura por parte del corrector y

(d) fiabilidad: tiene relación con la consistencia de todo el procedimiento de calificación. Calificar una prueba directa suele requerir más esfuerzo que calificar pruebas indirectas, dado que, si la intención es que las notas sean fiables, el corrector debe recibir una formación adecuada (Cooper & Odell, 1977: 21; Hamp-Lyons, 1991a: 245; Hamp-Lyons, 1991b: 73-82; Kroll, 1998: 225). Aunque se ha investigado sobre la posibilidad de prescindir de la laboriosa parte de la formación de los correctores dentro de los programas de evaluación de la expresión escrita (Hamp-Lyons & Henning, 1991: 366), se ha llegado a la conclusión de que tal entrenamiento supone un aspecto clave y, por lo tanto, necesario.

Además de cumplir con estas cuatro características, las pruebas han de ser prácticas –pueden aplicarse dentro de los límites de un presupuesto– y tener valor pedagógico –en algunos casos la prueba se convierte en un medio para enseñar (Harris, 1969: 21; Oller, 1979: 52).

A pesar del esfuerzo que supone trabajar con pruebas directas de evaluación de expresión escrita, existen ventajas que hacen de este tipo de evaluación una herramienta muy útil. En opinión de Ferris y Hedgcock (1998: 230), "there is no question that direct methods are the most appropriate and potentially valid form of assessment in the writing classroom". Jacobs et al. (1981: 3), por ejemplo, enumeran las siguientes ventajas:

(a) enfatiza la importancia del aspecto comunicativo de la lengua,
(b) redunda en una conexión más clara entre lo que se enseña y lo que se aprende,
(c) es más válida que una prueba de aspectos concretos para ofrecer información sobre la competencia comunicativa,
(d) es más fácil de preparar que una actividad de elección múltiple, que requiere mayor preparación en aspectos técnicos para conseguir validez y fiabilidad,
(e) es un método que ofrece resultados significativos y adecuados para ser interpretados,
(f) puede indicar con precisión el nivel de aptitud, además de puntos fuertes y puntos débiles dentro de las habilidades de composición,
(g) ofrece una alta fiabilidad si se aplica adecuadamente,
(h) tiene en cuenta a otros participantes en el proceso de comunicación –los lectores– y sus opiniones intuitivas y subjetivas.

La importancia del género

Además de las características hasta aquí expuestas, una variable considerada igualmente crítica con respecto a la elaboración de pruebas directas de expresión escrita es el producto escrito que será evaluado y distintas investigaciones relacionadas con el producto han centrado su atención en las peculiaridades retóricas de cada género (Kroll, 1998: 223-224).

El género, según Downing (1996: 11), tiene relación con la lengua en un contexto y, como dice van Dijk (1977: 153-154), dos categorías intervienen en el reconocimiento

de lo que él denomina "tipos de discurso": por un lado, las categorías estructurales relacionadas con el orden lineal y jerárquico de las macroestructuras y por otro, las categorías conceptuales que permiten reconocer el tema del que trata el texto. Los géneros son modelos y según Swales (1990: 8-9), la comunidad discursiva reconoce e identifica distintos géneros, aunque sea de una forma intuitiva, gracias a su estructura esquemática y a patrones lingüísticos u opciones léxico-gramaticales específicos de un tipo concreto de discurso.

Según Purves (1992: 116), el producto escrito de un estudiante y su evaluación por parte de un profesor están condicionados por el género. Los productos escritos por los alumnos son susceptibles de variación en relación con el género y condicionan la evaluación de los correctores. Sin embargo, las guías de calificación no suelen elaborarse teniendo en cuenta el género del producto escrito (Fulcher, 1997: 91), aunque el enfoque basado en el género es especialmente útil en aquellas situaciones en las que el tipo de escrito que se espera del estudiante puede ser definido con cierta precisión. Para Landa (1993: 50, citado en Fulcher, 1997: 93), la mayor ventaja de esta aproximación a la evaluación de la expresión escrita es que los alumnos han aprendido cuáles son las reglas que dominan el género que se les pide y que se tendrán en cuenta para la calificación.

El diseño de pruebas de expresión escrita en L2

El análisis del género es también importante en L2 y especialmente, según Dudley-Evans (1994: 219), en el campo del inglés para fines específicos (IFE). La investigación sobre el género en IFE se centra en la enseñanza de expresión escrita a los estudiantes de lenguas no nativas para su aplicación en las distintas materias de sus áreas de conocimiento (Swales & Feak, 1999).

Si se entiende la escritura como una actividad comunicativa y, por lo tanto, guiada por una serie de principios que se refieren al uso de la lengua en un acto comunicativo, al evaluar el producto escrito, interesa especialmente lo que el escritor quiere decir y cómo lo dice (Flower & Hayes, 1981; Widdowson, 1983). Ha de tenerse en cuenta, además, la excepcionalidad del contexto en el que se aplican las pruebas directas de evaluación –el escritor se ve forzado a escribir, el tema es impuesto desde fuera y el producto escrito es evaluado (Reid & Kroll, 1995: 17-19)– aunque, como dice White (1995: 35): "Even as we deplore the lack of revision opportunities for essay test writers, we should note that intense focus on writing for an hour or so can lead to

valuable work" y para que la prueba de escritura sea efectiva, ésta debe, según Kroll y Reid (1994: 234-241), ser clara, apropiada y pedagógica.

Dos factores distintos intervienen en la redacción en una lengua extranjera: los lingüísticos y los retóricos y, por esto, los escritores tienen en ILE unas necesidades especiales (Hamp-Lyons, 1991a: 241). Como puntualizan Grabe y Kaplan (1996: 113), el contexto de L2 no sólo introduce la variable de la lengua, sino también variables sociales y culturales. Las diferencias entre escribir en L1 y en L2 pueden y deben suponer prácticas distintas dentro del aula y, como consecuencia, instrumentos de evaluación distintos.

Con respecto a la calificación, lo más corriente hoy en día es el uso de guías de puntuación especialmente diseñadas para que los correctores den una puntuación rápida basada en su impresión general. El establecimiento de criterios y el empleo de guías facilitan la fiabilidad entre correctores y, si son presentados al estudiante en el aula, le ayudan a centrar su atención en aquellos aspectos que se consideran importantes. Además, especificar criterios que se orientan hacia aspectos retóricos de un género facilita la evaluación de la competencia en ese género determinado (Murphy, 1999: 119).

Métodos de calificación de la expresión escrita

Algunos investigadores como Cooper (1977: 4) o Lloyd-Jones (1977: 33) entienden la evaluación holística –no calificación– como opuesta a aquélla más usada en investigación, que exige el recuento de unidades o características especiales asociadas a un estilo concreto y que Cooper denomina *frequency-count marking* y Lloyd-Jones llama *atomistic*[4]. Dentro del enfoque holístico, existen básicamente dos posibles formas de calificar un producto escrito: la primera, denominada analítica, trata de evaluar por separado los distintos componentes de un escrito; la segunda, holística, tiene como objetivo primordial la calidad global de la comunicación (Madsen, 1983: 120-122).

En los últimos años se ha propuesto la siguiente clasificación de métodos de calificación: holística, analítica, basada en rasgos primarios y basada en rasgos múltiples (Grabe & Kaplan, 1996: 404-409; Ferris & Hedgcock, 1998: 234-245). Cooper (1977: 8-9) incluye también la escala de dicotomía (*dichotomous scale*), que consiste en una serie de proposiciones o preguntas que pueden ser contestadas afirmativa o negativamente (sí–no). No parecen existir por el momento estudios de

fiabilidad de este método y tampoco se recoge con frecuencia en clasificaciones que proponen otras investigaciones[5].

Calificación holística

El método de "impresión general" es el método más básico dentro de lo que se entiende por calificación holística. El evaluador lee la composición sin escribir nada en ella y decide una calificación general y subjetiva. Normalmente, para aplicar este método de calificación participan varios correctores, con el objeto de compensar la falta de fiabilidad de uno sólo y las calificaciones se apoyan con frecuencia en una escala numérica (*focused holistic scoring* en Hamp-Lyons, 1991a: 244). El significado de los números de la escala suele estar descrito en una guía de calificación, como es el caso del *Test of Written English (TWE) Scoring Guide*, quizás la más conocida de este tipo en el contexto de inglés como lengua extranjera.

La calificación holística es, según Reid (1993: 241), poco útil en el contexto del aula de ILE, pues es difícilmente explicable por parte del corrector a otros correctores con los que supuestamente tenga que ponerse de acuerdo o a cualquiera otra de las partes interesadas en el resultado de la misma –estudiantes, familiares u otras instituciones.

Calificación por rasgos primarios

El principal objetivo de la calificación mediante rasgos primarios es proponer criterios que se adapten a una tarea en particular sobre un tema determinado, en un género determinado. La escala más conocida de este tipo –de hecho la única citada por otros investigadores que se han dedicado a la evaluación de la expresión escrita (Cooper, 1977: 11; Jacobs et al., 1981: 29, en nota al pie; Hamp-Lyons, 1991a: 246; o Kroll, 1998: 228)– es la de Lloyd-Jones (1977), que considera que la ventaja de este método es la cantidad de información que puede extraerse de un solo escrito.

Según el propio Lloyd-Jones (1977: 45), el principal inconveniente de este sistema de evaluación es quizás que se trata de un método poco económico, porque las guías se redactan para un tipo de actividad muy particular y su confección requiere mucha dedicación.

Calificación por rasgos múltiples

La calificación mediante guías de rasgos múltiples consiste en centrar la atención del corrector en distintos aspectos de la redacción (Hamp-Lyons, 1991a: 247). El peligro potencial de este sistema, según Grabe y Kaplan (1996: 405), son los fallos en el diseño y la interpretación apropiada de los descriptores que determinan las posibles distintas calidades de los productos.

Hamp-Lyons (1991a: 253) prefiere éste a otros métodos de calificación en el contexto de ILE por dos razones: la primera es que "the judgments made by assessment readers can be translated into information which can be shared with the writers, their academic advisors and other concerned parties", y la segunda es su nivel de validez y la facilidad con la que los correctores pueden usarlo (Hamp-Lyons, 1991a: 247). Distintas investigaciones han demostrado que, tras un periodo de formación, los correctores pueden llegar a dar notas estadísticamente válidas y fiables (Coffman, 1971; Diederich et al., 1971; Cooper, 1977: 18, 19; Jacobs et al., 1981; Dunbar et al., 1991; Hamp-Lyons, 1991a y 1991b).

Cada vez más, se tiende a desarrollar este tipo de instrumentos para su aplicación en distintos contextos de escritura, de forma que sea más fácil reconocer aquellos aspectos que el escritor debe dominar. Desde la perspectiva del género, si diferentes géneros exigen distintas habilidades lingüísticas, operaciones, destrezas y estrategias, es lógico pensar como Murphy (1999: 120) que las guías que sirven para dirigir la composición de un estudiante en un género específico son más útiles que las holísticas.

Calificación por escalas analíticas

La calificación mediante escalas analíticas se basa en una guía de calificación que ofrece un valor a priori para cada componente textual. Diseñar una escala analítica es un trabajo que requiere un estudio teórico-práctico previo; la lista de rasgos debe ser completa, es decir, debe incluir todos aquéllos que caracterizan el género con el que se va a trabajar y según Cooper (1977: 14-15), dada la naturaleza de este tipo de guías, sería conveniente la colaboración de un estadístico en su elaboración.

Siguiendo al mismo autor, una vez creada la lista de rasgos, se describe qué se considera un nivel de calidad alto, medio y bajo para cada rasgo. Un valor numérico máximo y uno mínimo –la característica distintiva de este tipo de guías– son preasignados a

componentes como la organización, el contenido, la cohesión, el estilo, el registro, el vocabulario y, entre estos valores límite, se establece una escala o banda descendente para cada uno de ellos. Esta escala numérica que se asigna a cada uno de los aspectos es lo que diferencia fundamentalmente este tipo de guías de aquéllas de rasgos múltiples.

Una de las principales ventajas de las guías analíticas es que, si sus descriptores son suficientemente explícitos, el sistema de valoración facilita la formación de los correctores. Además, este tipo de guías sirve para cualquiera de los propósitos para los que normalmente se requiere una calificación en el contexto educativo: diagnóstico (*diagnosis*), consecución (*achievement*), progreso (*progress*), adscripción (*placement*) y competencia (*admission* o *proficiency*) (Cooper, 1977: 16-18)[6]. Desde una perspectiva más pedagógica, también pueden ser útiles en la evaluación formativa para ofrecer retroalimentación a los alumnos y comentar sus trabajos escritos, para que las usen ellos mismos en sus revisiones o para favorecer la auto evaluación o la evaluación entre compañeros.

La evaluación del resumen como producto

El creciente interés por el resumen y su enseñanza dentro del aula han dado lugar a algunas investigaciones que se han orientado hacia la evaluación del producto de tal actividad cognitiva y textual. Algunos investigadores (Garner, 1982: 277; Brown & Day, 1983: 12; Alonso Tapia et al., 1991: 24) han mencionado algunos de los criterios hasta aquí expuestos para describir sistemas usados por ellos mismos o por otros investigadores (Sherrard, 1989: 6-7) para analizar la calidad de un resumen; otros han propuesto adaptaciones de dichos métodos para su uso dentro del aula (Sherrard, 1989: 6). Sin embargo, estos sistemas de calificación son en general creados *ad hoc* para obtener datos en una investigación particular y suelen ser de carácter cuantitativo más que cualitativo.

Aunque existen algunas propuestas para la evaluación del resumen en inglés como lengua extranjera, como la guía de Lucisano y Kádár-Fülöp –para pruebas a gran escala– y la de Sarig –elaborada explícitamente para un experimento–, no parece haber ejemplos de guías de demostrada validez y fiabilidad especialmente diseñadas para su aplicación dentro del aula de ILE o, al menos, no han tenido suficiente repercusión dentro del mundo académico.

El contexto del aula de IFE es muy particular en este sentido pues muchas de las actividades que se proponen en un curso de inglés para fines específicos tienen como

base textos que se leen dentro o fuera del aula y, tanto en el caso de resúmenes como en el de respuestas a preguntas abiertas, los estudiantes tienen que reproducir parte del texto que previamente han leído y contarlo con sus propias palabras. La cuestión es cómo se corrigen y se puntúan esas respuestas basadas en textos expositivos.

Baremo de bandas analíticas para resúmenes: BABAR[7]

Como otros baremos, el BABAR (ver apéndice) se aproxima al resumen de un texto desde un enfoque holístico, que juzga el texto como un todo y admite como parte de la evaluación la subjetividad del lector. Tal subjetividad, sin embargo, debe ser controlada para que distintos correctores centren su atención en los aspectos comunicativos del escrito desde puntos de vista cuanto menos parecidos, sin olvidar que el profesor–corrector se ve afectado por condicionantes o estímulos exteriores que pueden reducirse e incluso evitarse. Para conseguirlo, se exige la aplicación de ciertas "normas" o "procedimientos" que sistemáticamente han demostrado ser útiles a la hora de moderar inconsistencias o discrepancias, además de para aumentar el escaso grado de fiabilidad que suele ofrecer un solo corrector.

El BABAR es una guía subdividida en cinco componentes, cada uno de los cuales tiene un peso distinto con respecto al resumen como un todo: 25 % el contenido, 20 % la organización, 17.5 % el vocabulario, 22.5 % el uso de la lengua y 5 % la presentación. Además, cada banda se divide en cuatro niveles de destreza (flojo, regular, bien y muy bien), que se representan con una calificación numérica distinta según el aspecto del que se trate y, por lo tanto, según su peso con respecto al todo. Tales niveles de destreza se describen en el apartado "criterios" y son éstos el primer punto de estudio y discusión en una sesión de formación de correctores.

Igual que sucede en el caso de cualquier otro trabajo de expresión escrita (Diederich, 1967: 576), se estima que el tiempo que un corrector debe emplear para dar una calificación al resumen de un texto original es de dos o tres minutos. Normalmente los correctores tienden a hacer dos lecturas rápidas en este tiempo y generalmente se les pide que no vuelvan a revisar el escrito después de dar una calificación.

Antes de comenzar una sesión de corrección de pruebas con el baremo, los profesores participantes deben seleccionar una cantidad de escritos que sirvan de ejemplo para "calentar" –es suficiente con unos cinco. El organizador de las sesiones, o ellos mismos

si tal figura no existe, los distribuye en grupos de tres y, una vez finalizada la calificación de estas pruebas, compara con ellos sus notas globales para asegurarse de que están aplicando los criterios con homogeneidad. Si los correctores se mueven dentro de un margen no superior a un punto de diferencia se entiende que están aplicando correctamente los criterios de evaluación (Jacobs et al., 1981: 103) y la calificación del resto de las pruebas sigue adelante. Es útil supervisar las puntuaciones con cierta frecuencia –cada 25 o 30 trabajos aproximadamente– para detectar a tiempo cualquier tendencia marcada de un corrector a puntuar más alto o más bajo que el resto de sus compañeros. Si se apreciaran divergencias dentro del grupo de correctores, éstos deben poner en común su interpretación de los criterios que les ha llevado a determinado resultado. Los correctores no siempre estarán completamente de acuerdo en las calificaciones parciales, es decir, aquéllas que se dan a cada uno de los aspectos; a pesar de ello pueden llegar, y de hecho alcanzan calificaciones globales similares.

Validez y fiabilidad del BABAR

El BABAR ha sido objeto de un estudio casi experimental que se llevó a la práctica para investigar, bajo condiciones controladas, el grado de validez y fiabilidad del baremo para evaluar resúmenes de textos expositivos en inglés como lengua extranjera.[8]

Método

Participaron en el estudio 127 alumnos de la Facultad de Lenguas Aplicadas de la Universidad Alfonso X el Sabio que cursaban estudios de Traducción e Interpretación; 42 estudiantes de primer curso (1º), 38 estudiantes de segundo curso (2º), y 47 estudiantes de tercer curso (3º), a los que se les pidió que escribiesen un resumen, de no más de cien palabras, de un texto expositivo relacionado con sus estudios. Trabajar con alumnos de tres cursos permitió contar con escritos de calidades distintas pues, según Johns y Mayes (1990: 265), aquellos alumnos con un nivel de competencia alto en lengua extranjera producen escritos de mejor calidad que aquéllos que tienen un nivel de competencia menor.

Los profesores que participaron como correctores trabajan en la sección de idiomas –Idioma Inglés– de la Universidad Alfonso X el Sabio, Facultad de Lenguas Aplicadas; algunos son hablantes nativos, aunque la mayoría son españoles que han terminado estudios de Filología.

La corrección de pruebas se llevó a cabo en dos fases. En la primera, se aplicó un método de impresión general y los correctores no fueron previamente formados. Para la segunda, que tuvo lugar cinco meses después, se diseñó el baremo de bandas analíticas (BABAR) que los correctores aplicaron a los mismos trabajos, después de someterse a un periodo de formación. Como resultado de ambas correcciones, la primera de impresión general (IG) y la segunda producto de la aplicación del baremo (BABAR), se recogieron tres calificaciones por cada uno de los dos métodos y para cada trabajo escrito, a través de las cuales se analizó el grado de validez y el grado de fiabilidad del baremo para calificar este tipo de pruebas en ILE/IFE.

Fiabilidad entre correctores mediante IG y BABAR

Como se ha avanzado, para analizar la fiabilidad del baremo se contó con las notas que dieron tres correctores a cada resumen escrito aplicando ambos métodos –impresión general y BABAR. Para el análisis de los datos, se eliminaron siete de los 127 resúmenes porque no respetaban el número de palabras exigido para el estudio, por lo tanto, para el análisis final, se partió de un total de 120 resúmenes.

En la tabla 1 se muestran los coeficientes de correlación entre correctores mediante los dos métodos que se aplicaron.

Número de correctores	1	2	3	Intervalo de fiabilidad para 3 correctores
Método IG	0.476	0.645	0.732	.6363-.8051
Método BABAR	0.571	0.727	0.799	.7282-.8544

Tabla 1. Coeficientes de correlación

Como se aprecia en la tabla, mediante la aplicación del BABAR, los coeficientes de fiabilidad para uno, dos y tres correctores son más altos que en el caso del método de impresión general; lógicamente la media entre tres correctores produce en ambos casos una nota más fiable que la nota de un solo corrector o la media de dos. La fiabilidad que se alcanza mediante la aplicación del BABAR para tres correctores es .799, muy cercana al .80 considerado aceptable (Diederich, 1974: 33; Cooper, 1977: 18; Jacobs, 1981: 69; Hamp-Lyons, 1991b: 69).

El intervalo de fiabilidad para tres correctores en cada uno de los métodos muestra que, mediante la aplicación del BABAR, el coeficiente de correlación mínimo y

máximo son más altos con respecto al método de impresión general. Además, los valores son más cercanos –el intervalo es más pequeño– lo cual evidencia una mayor fiabilidad del método.

Notas mínimas, máximas, medias y desviación típica

Al analizar las notas medias, se observa que el BABAR es más severo o restrictivo que el método de impresión general; las notas medias son más bajas. Las notas mínimas mediante IG van de 1 a 3 puntos, mientras que al aplicar el BABAR tienden al 2 y las notas máximas son más bajas mediante la aplicación del BABAR –máximas de 8 puntos frente a máximas de 9 mediante IG. Se aprecia también una desviación típica más baja en el segundo método (BABAR). En la tabla 2 se muestran los estadísticos descriptivos totales de ambos métodos.

	N	Min	Máx	Media	DT
Método IG	120	1.67	8.00	4.94	1.38
Método BABAR	120	2.00	7.58	4.13	1.10

Tabla 2. Estadísticos descriptivos por método (DT = Desviación Típica)

Consistencia del método

La *consistencia interna* se refiere al grado en que las bandas o las escalas y, más concretamente, las notas de cada una de ellas son homogéneas o tienen correlación con el resto de las notas parciales y con la nota global. En el BABAR hay cinco bandas de análisis que evalúan el resumen como un producto desde una perspectiva global en la que todos los aspectos recogidos en las bandas interaccionan. Dado que unos aspectos no pueden separarse de otros en un resumen escrito por un estudiante, se entiende que debe haber una correlación alta entre los componentes y que una nota alta en una de las áreas presupone notas altas en el resto y viceversa. Pero, como se ha demostrado en estudios anteriores (Hamp-Lyons, 1991a; Grabe & Kaplan, 1996), los correctores de pruebas escritas en inglés como lengua extranjera tienden además a evaluar aspectos cognitivos por un lado y otros más lingüísticos por otro.

Siguiendo la idea de que unos aspectos se ligan a los otros, debería esperarse entonces una correlación fuerte entre contenido y organización y también entre uso de la lengua y vocabulario; la correlación podría ser menor entre otros aspectos

–contenido y uso de la lengua u organización y vocabulario– si se tiene en cuenta que los resúmenes están escritos en ILE. La tabla 3 muestra la correlación entre los distintos aspectos evaluables.

Aspectos	Contenido	Organización	Vocabulario	Uso de la lengua
Organización	.841			
Vocabulario	.726	.763		
Uso de la lengua	.798	.809	.852	
Presentación	.403	.501	.483	.517

Tabla 3. Correlaciones entre áreas

En la tabla 3 se aprecia que el uso de la lengua y el vocabulario tienen un grado alto de correlación (.852); son los aspectos más lingüísticos de los resúmenes. También el contenido y la organización se entienden como áreas directamente relacionadas con una correlación de .841. En general, se aprecian correlaciones más altas entre las cuatro primeras áreas –contenido, organización, vocabulario y uso de la lengua– y más bajas entre éstas y la presentación, lo cual tiene sentido por ser la presentación un aspecto más mecánico que lingüístico o retórico.

Validez de criterio

Un método corriente para evaluar la validez de criterio de una prueba es analizar hasta qué punto los resultados de la misma son similares a los resultados obtenidos mediante la aplicación de otras pruebas que evalúan habilidades similares. Lo normal, según Cronbach (1970:135), es no obtener coeficientes superiores a .60 ya que los coeficientes de validez se ven limitados por la fiabilidad –o falta de fiabilidad– de cada uno de los métodos que se compara. En la tabla 4, se muestra la correlación entre los dos métodos: IG y BABAR.

Correlaciones IG y BABAR	IG
BABAR	.592

Tabla 4. Correlaciones entre métodos

Aunque se evalúa el mismo producto, los métodos son distintos y además, para aplicar BABAR ha habido una formación y entrenamiento previo de correctores. Un coeficiente entre métodos de .592 implica que BABAR goza de validez de criterio.

Validez de constructo

La comparación de las notas obtenidas por estudiantes de distintos cursos es el método que se aplica en este estudio para analizar estadísticamente si las puntuaciones de los estudiantes están directamente relacionadas con otras medidas ya tomadas que evalúan la misma destreza o habilidades muy relacionadas con ésta.

Como ya se ha adelantado, en la fase de recogida de datos se trabajó con estudiantes de tres cursos distintos de la Facultad de Traducción e Interpretación. Para el análisis de los datos, se parte de que los estudiantes de tercer curso han recibido más instrucción relacionada con la producción escrita en ILE que los de primer curso y es prudente pensar que los estudiantes de tercero tendrán mejores notas, tanto por el método IG como por el BABAR, que los de primero. En la tabla 5 se muestran los resultados de los estudiantes por ambos métodos y cursos.

		IG		BABAR	
N	**CURSO**	**Media**	**Desv. Típica**	**Media**	**Desv. Típica**
39	1º	4.35	1.43	3.69	1.05
36	2º	4.99	1.21	4.21	1.07
45	3º	5.41	1.30	4.46	1.07
120	Total	4.94	1.38	4.13	1.10

Tabla 5. Estadísticos descriptivos por curso y método

Contrastes	Curso (F2,117)	*p*
IG	6.864	.002
BABAR	5.712	.004

Tabla 6. Significación estadística

La nota media de los estudiantes de primero por el método IG es de 4.35 y la de los estudiantes de tercero de 5.41 –los correctores dan a los estudiantes del curso superior notas más altas– y la diferencia es estadísticamente significativa (.002) $p<.05$. Con respecto al BABAR, la nota media de los estudiantes de primer curso es de 3.69, mientras que la media de los estudiantes de tercer curso es de 4.46. La diferencia de las notas de los estudiantes del curso superior con respecto a las de los estudiantes del curso inferior es también significativa (.004) $p<.05$. Los correctores dan a los estudiantes de cursos superiores notas significativamente más altas por ambos métodos.

Conclusiones

Este trabajo ha querido dar al resumen la relevancia que merece como actividad pedagógica. Desde tal perspectiva pedagógica, no solo interesa saber cómo llevarlo al aula teniendo en cuenta las variables que afectan a su producción, sino también y más concretamente cómo evaluarlo. Con respecto a la evaluación, cabe destacar que el resumen se ve condicionado por una serie de factores que afectan a su producción; evidentemente el hecho de resumir en una lengua extranjera afecta todavía más a su calidad. BABAR es un baremo que se ha demostrado válido y fiable para la calificación de resúmenes de textos expositivos en inglés como lengua extranjera.

NOTAS

1 Se emplea en adelante este término, tanto en el sentido de equidad –cada uno tiene lo que le corresponde por sus merecimientos–, como en el de precisión o adecuación.

2 También denominada *concurrent validity* en Oller (1979: 51) y Jacobs et al. (1981: 74).

3 Del inglés *construct* < lat. c*onstruo* (acumular). Entidad, a menudo denominada constructo teórico, elaborada por abstracción o postulada por necesidades inherentes al aparato teórico con que trabaja el lingüista (Cerdà Massó et al., 1986).

4 Uno de los ejemplos más recientes de este tipo de evaluación es el desarrollado en España por la Dra. María del Mar Duque, Directora Técnica del Proyecto Anestte–Analizador de Estilo para Textos Científico-Técnicos en inglés.

5 *Trinity College London* investiga sobre el uso de escalas de este tipo para la evaluación de la expresión escrita en ILE aunque, por el momento, no se han hecho públicos resultados concluyentes a partir de los datos obtenidos de su aplicación.

6 "Diagnóstico, aprovechamiento, control, clasificación y nivel de dominio" en Bordón (1999: 16).

7 Basado en *ESL Composition Profile* (Jacobs et al., 1981) originalmente diseñado para la calificación de textos expositivos (no resúmenes) escritos en ILE.

8 El diseño del trabajo empírico orientado a probar el baremo se presentó en las Primeras Jornadas de AELFE, Madrid, UPM, 2002. Los análisis estadísticos completos se recogen en la Tesis Doctoral inédita de Argüelles Álvarez (2002: 175-196).

REFERENCES

Aleksander, I. (1996). *Impossible Minds: My Neurons My Consciousness.* London: Imperial College Press.

Alonso Tapia, J., N. Carriedo López y E. González Alonso (1991). "Evaluación de la comprensión lectora: ¿Cómo determinar si un lector distingue lo que es importante en un texto de lo que no lo es?" *Boletín del ICE* (UAM) 19: 7-43.

Argüelles Álvarez, I. (2002). "La evaluación de resúmenes de textos expositivos en el aula de ILE: Propuesta de un baremo de bandas analíticas". Tesis doctoral inédita, Universidad Autónoma de Madrid.

Argüelles Álvarez, I. (2003). "BABAR: baremo de bandas analíticas para la calificación de resúmenes". *Las lenguas para fines específicos y la sociedad del conocimiento*, 655-665. Madrid: DLACT, UPM.

Bordón, T. (1999). "La evaluación del español como lengua extranjera". *Boletín de ASELE* 20: 15-25.

Brown, A. L. & J. D. Day (1983). "Macrorules for summarizing texts: The development of expertise". *Journal of Verbal Learning and Verbal Behaviour* 22: 1-14.

Cassany, D., M. Luna & G. Sànz (1994). *Enseñar lengua*. Barcelona: Graó.

Cerdà Massó, R., M. C. Muñoz Olivares, J. L. Calero López de Ayala & J. Lloret Cantero (1986). *Diccionario de lingüística*. Madrid: Anaya.

Coffman, W. E. (1971a). "On the reliability of ratings of essay examinations in English". *Research in the Teaching of English* 5: 24-36.

Cooper, C. R. & L. Odell (eds.) (1977). *Evaluating Writing: Describing, Measuring, Judging*. Urbana, Illinois : NCTE.

Cooper, C. R. (1977). "Holistic evaluation of writing" en Cooper & Odell (eds.), 3-31.

Cronbach, L. J. (1970). *Essentials of Psychological Testing*. New Cork: Harper and Row.

Chaudron, C. (1988). *Second Language Classrooms: Research on Teaching and Learning*. Cambridge: Cambridge University Press.

Diedrich, P. (1967). "Cooperative preparation and rating of essay tests". *English Journal* 56: 573-584.

Diederich, P. (1974). *Measuring Growth in English*. Urbana, Illinois: National Council of Teachers of English.

Diederich, P. B., J. W. French & S. T. Carlton (1971). "Factors in judgments of writing ability". *ETS Research Bulletin* RB-61-15. Princeton: Educational Testing Service.

Downing, A. (1996). "Register and/or genre?" en I. Vázquez & A. Hornero (eds.), *Current Issues in Genre Theory*, 11-27. Zaragoza: Mira Editores.

Dudley-Evans, T. (1994). "Genre analysis: an approach to text analysis for ESP" en M. Coulthard (ed.), *Advances in Written Text Analysis*, 219-228. London & New York: Routledge.

Dunbar, S. B., D. M. Koretz & H. D. Hoover (1991). "Quality control in the development and use of performance assessments". *Applied Measurement in Education* 4,4: 289-303.

Duque, M. M. (2000). *Manual de estilo: El arte de escribir en ingles científico–técnico*. Madrid: Paraninfo.

Ferris, D. & J. S. Hedgcock (1998). *Teaching ESL Composition: Purpose, Process and Practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flower, L. & J. R. A. Hayes (1981). "Cognitive process theory of writing". *College Composition and Communication* 32,4: 365-387.

Fulcher, G. (1997). "Assessing writing" en G. Fulcher (ed.), *Writing in the English Language Classroom*, 91-116. London: Prentice Hall Europe ELT.

Garner, R. (1982). "Efficient text summarization: Costs and benefits". *Journal of Educational Research* 75: 275-279.

Grabe, W. & R. B. Kaplan (1996). *Theory & Practice of Writing*. London: Longman.

Hamp-Lyons, L. (1991a). "Scoring procedures for ESL contexts" en L. Hamp-Lyons (ed.), *Assessing Second Language Writing in Academic Contexts*, 241-276. Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (1991b). "Second language writing: assessment issues" en B. Kroll (ed.), *Second Language Writing: Research Insights for the Classroom*, 69-87. New Cork: Cambridge University Press.

Hamp-Lyons, L. & G. Henning (1991). "Communicative writing profiles: An investigation on the transferability of a multiple trait scoring instrument across ESL writing assessment contexts". *Language Learning* 41,3: 337-373.

Harris, D. P. (1969). *Testing English as a Second Language*. New York: Mc Graw Hill.

Jacobs, H. L., S. A. Zinkgraf, D. R. Wormuth, V. F. Hartfiel & J. B. Hugey (1981). *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Johns, A. y P. Mayes (1990). "An análisis of summary protocols of university ESL students". *Applied Linguistics* 11: 253-271.

Kroll, B. (1998). "Assessing writing abilities". *Annual Review of Applied Linguistics* 18: 219-240.

Kroll, B. & J. Reid (1994). "Guidelines for designing writing prompts: Clarifications, chavetas, and cautions". *Journal of Second Language Writing* 3,3: 231-255.

Landa, J. (1993). "Coping with "Advanced German composition": Teaching writing as genre writing". *Unterrichtspraxis*, 26,1: 50-55.

Lloyd-Jones, R. (1977). "Primary-trait scoring" en Cooper & Odell (eds.), 33-66.

Lucisano, P. & J. Kádár-Fülöp (1988). "The summary tasks" en T. P. Gorman, A. C. Purves & R. E. Degenhart, *The IEA Study of Written Composition I: The International Writing Tasks and Scoring scales*, vol. 5, 112-127. Oxford: Pergamon Press.

Madsen, H. S. (1983). "Writing tests". *Techniques in Testing*, 101-126. New York y Oxford: Oxford University Press.

Murphy, S. (1999). "Assessing portfolios" en Cooper & Odell (eds.), 114-135.

Oller, J. W. Jr. (1979). *Language Tests at School*. London: Longman.

Purves, A. (1992). "Reflections on research and assessment in written composition". *Research in the Teaching of English* 26,1: 108-122.

Reid, J. M. (1993). *Teaching ESL writing*. Englewood Cliffs, NJ: Regents/Prentice Hall.

Reid, J. & B. Kroll (1995). "Designing and assessing effective classroom writing assignments for NES and ESL students". *Journal of Second Language Writing* 41: 17-41.

Sarig, G. (1993). "Composing a study-summary: A reading-writing

encounter" en J. G. Carson & I. Leki (eds.), *Reading in the Composition Classroom: Second Language Perspectives*, 161-182. Boston, MA: Heinle & Heinle.

Sherrard, C. (1989). "Teaching students to summarize: Applying text linguistics". *System* 17,1: 1-11.

Swales, J. M. (1990). *Genre Analysis*. Cambridge: Cambridge University Press.

Swales, J. M. & B. C. Feak (1999). *Academic Writing for Graduate Students*. Ann Arbor, Michigan: University of Michigan Press.

Test of Written English (TWE) Guide (1992). Princeton, NJ: Educational Testing Service.

Van Dijk, T. A. (1977). *Text and Context*. London. Longman.

White, E. (1994). *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating and Improving Student Performance*. San Francisco, CA: Jossey-Bass Publishers.

White, E. (1995). "An apologia for the timed impromptu essay test". *College Composition and Communication* 46: 30-45.

Widdowson, H. G. (1983). "New starts and different kinds of failure" en A. Freedman, I. Pringle & J. Yalden (eds.), *Learning to Write: First Language / Second Language*, 34-47. London: Longman.

Irina Argüelles Álvarez, Doctora por la Universidad Autónoma de Madrid, lleva más de trece años dedicada a la docencia de la lengua inglesa. En la actualidad es Profesora Titular Interina en la Escuela Universitaria de Ingeniería Técnica de Telecomunicación de la Universidad Politécnica de Madrid.

Apéndice: La guía (BABAR)

CONTENIDO	Calificación	Criterios
• Muy bien	2.5 / 2.25	**R**esalta la idea central y la conclusión. Incluye las ideas relevantes y las desarrolla con precisión y concisión. Omite información redundante o no importante.
• Bien	2 / 1.75	**E**xpone la idea central y la conclusión. Incluye muchas ideas relevantes y las desarrolla con bastante precisión y concisión. Poca información es redundante o no importante.
• Regular	1.5/ 1.25	**M**enciona la idea principal o conclusión además de algunas ideas relevantes. Desarrollo adecuado. Incluye información redundante o no importante.
• Flojo	1	**A**lgunas ideas son relevantes pero confunde otras o no las desarrolla con claridad. Puede omitir alguna(s) ideas importante(s). Incluye mucha información no importante o redundante.
ORGANIZACIÓN	**Calificación**	**Criterios**
• Muy bien	2 / 1.75	**S**e aprecia una clara estructuración expositiva. Evidente relación entre ideas y secuenciación lógica. Muy coherente.
• Bien	1.5 / 1.25	**S**e aprecia una estructura expositiva. Existe cierta conexión entre ideas. Secuenciación lógica, quizás algo incompleta. Coherente.
• Regular	1 / 0.75	**E**structura clara pero relación entre ideas poco evidente. Tiene cierta coherencia pero falla en la secuenciación de ideas.
• Flojo	0.5 / 0.25	**C**arece de estructuración o se aprecia con dificultad. No existe relación entre ideas y le falta coherencia. Secuenciación muy afectada.

VOCABULARIO	Calificación	Criterios
• Muy bien	1.75 / 1.5	**U**sa sus propias palabras y mantiene sólo aquéllas que son clave. Preciso. Registro y forma son apropiados.
• Bien	1.25 / 1	**E**xpresa las ideas con sus propias palabras aunque mantiene algunas prescindibles. Bastante preciso. Pocos errores de registro y forma.
• Regular	0.75 / 0.5	**I**ntenta no usar las mismas palabras pero mantiene muchas prescindibles. Algo /poco preciso. Errores de registro y forma.
• Flojo	0.25	**U**sa las mismas palabras o se aprecia desconocimiento de vocabulario apropiado. Muchos errores de registro o forma.
USO DE LA LENGUA	**Calificación**	**Criterios**
• Muy bien	2.25 / 2	**P**arafrasea correctamente. Usa recursos lingüísticos adecuados para combinar /enlazar ideas. Produce transformaciones con competencia. Apenas errores de tiempo, concordancia, preposiciones u orden.
• Bien	1.75 / 1.5	**P**arafrasea. Tiene recursos lingüísticos adecuados para combinar/ enlazar ideas. Produce transformaciones con cierta competencia. Pocos errores de tiempo, concordancia, preposiciones u orden.
• Regular	1.25 / 1	**P**arafrasea pero usa recursos lingüísticos básicos para combinar/ enlazar ideas. Produce pocas transformaciones o no son adecuadas. Bastantes errores de tiempo, concordancia, preposiciones u orden.
• Flojo	0.75 / 0.5	**P**rácticamente copia. No usa recursos lingüísticos para combinar/ enlazar ideas. No sabe producir transformaciones, o no lo demuestra. Muchos errores de tiempo, concordancia, preposiciones u orden.
PRESENTACIÓN	**Calificación**	**Criterios**
• Muy bien	0.5	**N**o tiene errores ortográficos ni de puntuación. Marca los párrafos. El número de palabras es exacto. Muy limpio.
• Normal	0.25	**T**iene algunos errores ortográficos o de puntuación. Párrafos poco claros. El número de palabras no es exacto. Limpio.
• Flojo	0	**T**iene varios errores ortográficos y de puntuación. Párrafos muy poco claros. El número de palabras no es exacto. Poco limpio.

NOTA: Aquellos trabajos que se consideren por encima de "muy bien" para cada uno de los aspectos o que alcancen tal calificación en cada uno de ellos optará a la calificación de 10.