Nebrija Universidad
Revista Nebrija
de Lingüística Aplicada a la Enseñanza de las Lenguas

Escalas de descriptores y fiabilidad de la evaluación de la expresión e interacción orales del usuario competente

Francisco del Moral Manzanares
Universidad de Verona.
francisco.delmoral@univr.it

del Moral Manzanares, F. (2013). Escalas de descriptores y fiabilidad de la evaluación de la expresión e interacción orales del usuario competente. *Revista Nebrija de Lingüística Aplicada* (2013) 13.

RESUMEN

La presente investigación se centra en una experiencia docente desarrollada en el Centro Lingüístico de la Universidad de Verona, en el ámbito de la evaluación de la expresión e interacción orales de nivel C2. Más concretamente, se trató de averiguar si, teniendo en cuenta el propio contexto de enseñanza, podría ser útil para aumentar la fiabilidad de la prueba, el uso de escalas de descriptores, que hasta ese momento no se habían utilizado. Se crearon para ello una serie de escalas, partiendo de las analíticas que el Instituto Cervantes aplica en los DELE.

El estudio empírico se realizó en colaboración con otra profesora del centro, durante la sesión oficial de exámenes de julio de 2012. Una vez grabados, se seleccionó una muestra representativa y se procedió a calificarlos sucesivas veces con un lapso de tiempo de varios días, hasta conseguir cuatro calificaciones de cada examen. Dos se realizaron con ayuda de escalas de descriptores y las otras dos, sin ella.

Los resultados permitieron comprobar que el uso de las escalas de descriptores aumentó significativamente la fiabilidad externa de la prueba (interevaluadora, entre los dos evaluadores) y, en menor medida, la interna (intraevaluadora, de cada evaluador consigo mismo).

Palabras clave: ELE, escalas de descriptores, evaluación, expresión oral, fiabilidad, usuario competente

ABSTRACT

This piece of research was carried out by teaching staff at the University of Verona Language Centre and regards the evaluation of oral production and interaction skills at a C2 level. In particular, the research aimed to find out whether, within this teaching context, the use of a rating scale, which up until that point was not being used, could lead to an increase in test reliability.

The empirical study was carried out in collaboration with a fellow teacher from the Language Centre, and began during the July 2012 exam session. A representative sample was chosen from recordings of the test sessions and each test was rated a total of four times with an interval of a few days between each evaluation. Two of the evaluations were carried out using the rating scale and two without.

The results showed that the use of the rating scale increased the inter-rater reliability of the test and, to a lesser extent, the intra-rater reliability.

Keywords: Spanish as a Foreign Language, rating scales, assessment, spoken production, reliability, competent user

1. INTRODUCCIÓN

El presente artículo pretende dar testimonio de una experiencia docente del ámbito de la evaluación de ELE, basada en conceptos como la fiabilidad, la evaluación subjetiva y las escalas de descriptores.

Según el acuerdo general de los expertos, la fiabilidad es, junto a la validez y la viabilidad, uno de los requisitos que ha de cumplir todo instrumento de evaluación capaz de proporcionar información correcta. Es definida como consistencia de la medición (cfr. Bordón, 2006:62) o estabilidad en la obtención de resultados (cfr. Diccionario de términos clave de ELE), por lo que se considera fiable la herramienta que aporta siempre los mismos resultados, independientemente del contexto o el momento en que esta se utiliza, y de la consistencia de los evaluadores.

Por otra parte, las pruebas de corrección subjetiva son aquellas en las que esta no puede hacerse por medio de instrumentos tales como lectores ópticos o plantillas y, consecuentemente, necesitan la intervención de al menos una persona para la interpretación de los resultados. Se trata de pruebas de expresión e interacción que requieren respuestas abiertas. En el ámbito de este tipo de pruebas, se diferencia entre fiabilidad externa, interevaluadora o entre correctores, por una parte, y fiabilidad interna, intraevaluadora o de un evaluador consigo mismo, por otra (cfr. Alderson, Clapham y Wall, 1998:128).

Uno de los instrumentos utilizados con el fin de obtener mayor objetividad en la evaluación de este tipo de pruebas, lo que redunda en una una mayor fiabilidad, son las escalas de descriptores. Se entiende por descriptor cada una de las explicaciones que se añade a las diferentes puntuaciones (numéricas o verbales) para aclarar el significado de un determinado juicio. La serie completa de todas ellas, de la más baja a la más alta, constituye una escala de calificación o evaluación (cfr. Luoma, 2004:59). A modo de ejemplo, se incluye a continuación la que hemos utilizado en nuestro proyecto para evaluar el criterio “Capacidad crítica”[1].

95	Expresa opiniones personales o cuenta experiencias propias con naturalidad, de manera que**enriquece notablemente el contenido de las fuentes**.**Es capaz de contradecir, apoyar, convencer, matizar inteligentemente con aportaciones nuevas,demostrando capacidad de escucha y elaboración de las opiniones del interlocutor, siendo flexible y coherente**.
85	Expresa opiniones personales o cuenta experiencias propias que**enriquecen el contenido de las fuentes y la conversación.** Argumenta sus puntos de vista correctamente ysus aportaciones van siempre encaminadas a sostener su idea inicial.
75	Tiene capacidad para opinar sobre los contenidos de las fuentes e ilustrar sus opiniones con experiencias propias, **de un modo pertinente y correcto,** aunqueno hace aportaciones nuevas que enriquezcan el tema tratado.
65	**Expresa algunas opiniones personales pertinentes**, perono las argumenta o no lo hace convincentemente.Puede dejarse llevar por el interlocutor en ocasiones, pero**es coherente con su opinión**, aunquemuestra dificultades para convencer al interlocutor.
50	Puede expresar opiniones personales o relatar experiencias propias, peroa veces estánvagamente relacionadas con el tema tratado, que se limita a tocar superficialmente,y repite casi al pie de la letra los contenidos de las fuentes.

La unión de las diferentes escalas de calificación referidas a todos los criterios empleados en la evaluación de la prueba (por ejemplo, coherencia, fluidez, corrección y alcance) constituye la escala de descriptores completa.

El objetivo principal de nuestra investigación fue determinar si la utilización de una escala de descriptores ayudaría a conseguir una mayor fiabilidad de la prueba oral de nivel C2 de nuestro centro, a pesar de que la creación de ambas no se haya producido paralelamente[2].

2. CONTEXTO DE LA INVESTIGACIÓN

El*Centro Linguistico di Ateneo* (en lo sucesivo, CLA) es un centro interdepartamental de servicios de la Universidad de Verona (Italia), encargado de ofrecer formación lingüística a todos los estudiantes que la necesiten, en virtud de los planes de estudio de sus respectivas facultades.

Los informantes fueron alumnos que han terminado de cursar el nivel C2 (en cuarto o quinto año de diferentes especialidades de la licenciatura de Lenguas Extranjeras), lo que significa 440 horas de formación de ELE en el CLA, más las realizadas de forma complementaria en la Facultad de Lenguas Extranjeras, referidas sobre todo a reflexión metalingüística (fonética, morfosintaxis, traducción). De los 37 exámenes que constituyeron la sesión oficial de julio de 2012 se seleccionaron dieciocho por "muestreo aleatorio estratificado" (León y Montero, 1997:101).

La examinadora es la profesora oficial de la asignatura durante el año académico 2011? 2012 y los evaluadores son tanto ella (evaluador 1, E1) como el propio investigador (evaluador 2, E2).

3. ESTUDIO EMPÍRICO

3.1 HIPÓTESIS

Hipótesis 1: La evaluación holística de la expresión oral sin ayuda de descriptores de evaluación previamente establecidos, produce una baja fiabilidad tanto externa como interna de la prueba de evaluación de expresión e interacción orales de nivel C2 del CLA de la Universidad de Verona.

Hipótesis 2: El empleo de unos criterios de evaluación y una escala de descriptores de calificación, establecidos previamente y conocidos por los evaluadores, aumentará la consistencia de la evaluación y la fiabilidad de la prueba de evaluación de expresión e interacción orales de nivel C2 del CLA de la Universidad de Verona.

Nuestra investigación constó de dos partes, denominadas Estudio 1 y Estudio 2, para poner a prueba sendas hipótesis.

3.2 DISEÑO DE INVESTIGACIÓN

Para la realización de nuestro proyecto se combinó metodología cuantitativa y cualitativa. Por una parte, una vez obtenidas las diferentes series de calificaciones (datos numéricos), se explicó la relación entre ellas por medio de un análisis de correlación[3], que permite evaluar la fuerza y dirección de la relación existente entre dos variables o el modo en que están vinculadas la una con la otra (cfr. Dörnyei, 2010:223). Para llevar a cabo dicho análisis, calculamos un coeficiente de correlación comprendido entre -1 y +1. Cuanto mayor sea el coeficiente, mayor será la relación entre ambas series de calificaciones. Se suele considerar alta una correlación mayor de 0,70 (cfr. Diccionario de términos clave de ELE y Bisquerra, 2004:212).

Por otro lado, han sido muy útiles para interpretar o matizar algunos de estos datos de tipo cuantitativo, ciertas informaciones que se desprenden del diario de investigación (datos de tipo cualitativo).

3.3 ESTUDIO 1 (CALIFICACIONES SIN DESCRIPTORES)

La primera parte del estudio empírico consistió en la emisión de cuatro calificaciones de las dieciocho pruebas orales grabadas, utilizando como herramienta una tabla en la que aparecen los nombres y datos de los alumnos, así como la nota obtenida en la prueba escrita (que se realiza previamente a la oral y cuya superación es necesaria para tener acceso a esta).

La primera calificación tenida en cuenta es la otorgada oficialmente en la sede de examen por la profesora examinadora. Se codifica como C1E1 (calificación 1 de la evaluadora 1). El investigador emite su primera calificación durante la misma sesión, al mismo tiempo que va grabando los exámenes (C1E2). Con un lapso de varios días, ambos evaluadores vuelven a oír las grabaciones en un orden diferente y aleatorio y emiten C2E1 y C2E2, respectivamente.

Una vez obtenidas las cuatro calificaciones, se lleva a cabo el siguiente procesamiento de datos:

Correlación entre C1E1 y C1E2, para hallar la primera fiabilidad interevaluadora, sin que ninguno de los evaluadores hayan hecho uso de descriptores. Resultado: 0,29 / 0,34.

Correlación entre C2E1 y C2E2, para hallar la segunda fiabilidad interevaluadora, sin que ninguno de los evaluadores hayan hecho uso de descriptores. Resultado: 0,52 / 0,53.

Correlación entre C1E1 y C2E1, para hallar la fiabilidad intraevaluadora de E1 sin que esta haya hecho uso de descriptores. Resultado: 0,44 / 0,48.

Correlación entre C1E2 y C2E2, para hallar la fiabilidad intraevaluadora de E2 sin que este haya hecho uso de descriptores. Resultado: 0,79 / 0,82.

Desviación típica de las cuatro calificaciones de cada examen (C1E1, C1E2, C2E1 y C2E2) y, una vez obtenidas las de los dieciocho exámenes de la muestra, se haya la media de todas ellas. Resultado: 4,67.

3.4 ESTUDIO 2 (CALIFICACIONES CON DESCRIPTORES)

La segunda parte del estudio empírico consistió en la emisión de otras cuatro calificaciones por parte de los mismos evaluadores, pero esta vez utilizando una escala de descriptores. Dicha escala fue creada teniendo en cuenta las necesidades de nuestro centro, a partir de las escalas analíticas empleadas por el Instituto Cervantes para los DELE C1 y C2. A los criterios de coherencia, fluidez, corrección y alcance se añadió el de capacidad crítica. La inclusión de este quinto criterio está justificada en los reglamentos didácticos de las licenciaturas cursadas por los estudiantes objeto de nuestro estudio.

Tras la creación de las escalas se realizó una sesión de pilotaje, cuyo objetivo era consensuar el significado de los descriptores con la E1. Posteriormente, del conjunto de los diecinueve exámenes que formaba parte de la población inicial pero que no fueron seleccionados para la muestra, se seleccionó un examen representativo de cada nivel de calificación y se proporcionó a la E1 el informe de calificación y la grabación correspondientes, para que pudieran servir de ejemplo.

Por último, antes de proceder a recalificar los exámenes de la muestra, se realizó una sesión de entrenamiento con las grabaciones de los que no formaban parte de esta[4].

La emisión de calificaciones del estudio 2 (codificadas como C3 y C4 de cada evaluador) se hizo de modo análogo a la de la primera parte. El procesamiento de datos fue el siguiente:

Correlación entre C3E1 y C3E2, para hallar la primera fiabilidad interevaluadora, habiendo hecho uso los dos evaluadores de descriptores. Resultado: 0,79 / 0,71.

Correlación entre C4E1 y C4E2, para hallar la segunda fiabilidad interevaluadora, habiendo hecho uso los dos evaluadores de descriptores. Resultado: 0,71 / 0,68.

Correlación entre C3E1 y C4E1, para hallar la fiabilidad intraevaluadora de E1 con el uso de descriptores. Resultado: 0,83 / 0,90.

Correlación entre C3E2 y C4E2, para hallar la fiabilidad intraevaluadora de E2 con el uso de descriptores. Resultado: 0,80 / 0,87.

Desviación típica de las cuatro calificaciones de cada examen (C3E1, C3E2, C4E1 y C4E2) y, una vez obtenidas las de los dieciocho exámenes de la muestra, se haya la media de todas ellas. Resultado: 4,04.

3.5 INTERPRETACIÓN DE LOS RESULTADOS

3.5.1 FIABILIDAD INTERNA O INTRAEVALUADORA

La comparación de los resultados obtenidos en las dos partes del estudio, referidos tanto a la E1 (0,44-0,83), como al E2 (0,79-0,80), indican que el uso de descriptores aumenta la fiabilidad intraevaluadora (v. fig. 1). El hecho de que el E2 partiera ya de una fiabilidad alta sin haber usado los descriptores hace que, en su caso, el aumento sea mucho menor que en el de la E1.

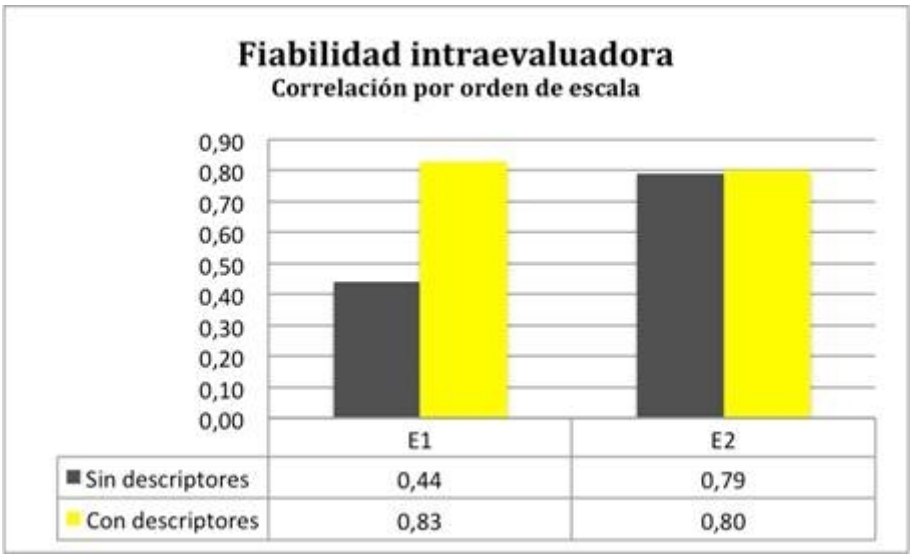

Figura 1. Fiabilidad intraevaluadora sin y con descriptores

3.5.2 FIABILIDAD EXTERNA O INTEREVALUADORA

La fiabilidad entre los dos evaluadores también aumenta cuando estos emiten sus calificaciones usando la escala de descriptores creada*ad hoc* (0,79 y 0,71 cuando se emplean, frente a 0,23 y 0,52 cuando no).

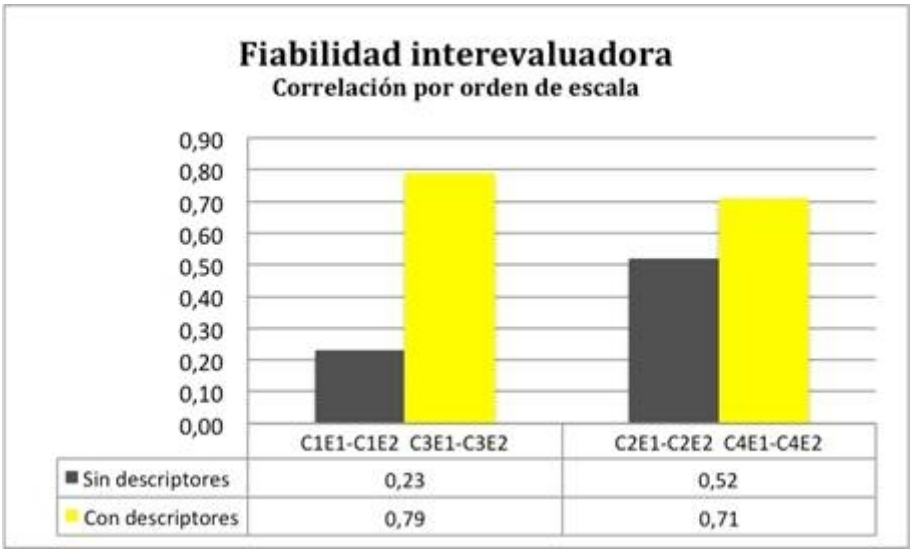

Figura 1. Fiabilidad interevaluadora sin y con descriptores

3.3.3 LA FIABILIDAD DE LA PRUEBA

Consideramos índice de una fiabilidad general de la prueba, sin vincularla particularmente a la consistencia de uno u otro evaluador, la media de las desviaciones típicas de las cuatro calificaciones otorgadas sin el uso de descriptores, por una parte (4,67), y con él, por otra (4,04). Tal diferencia, aunque indica que las calificaciones emitidas varían menos cuando se usan los descriptores, no es muy significativa. Esto se debe al bajo número de sujetos que componen la muestra. Suponemos que de haber trabajado con una muestra mayor, la diferencia entre estos valores también aumentaría.

4. CONCLUSIONES

Según los resultados expuestos, podemos concluir que la hipótesis 1 se ha cumplido parcialmente, puesto que no siempre los valores de fiabilidad obtenidos son bajos, cuando no se emiten las calificaciones con la ayuda de escalas de descriptores. En este sentido, los valores que más se alejan de la hipótesis inicial son los referidos a la fiabilidad interna del E2, que podemos considerar alta.

Sin embargo, la hipótesis 2 se cumple totalmente, pues en todos los casos el uso de las escalas de descriptores produce un aumento de la fiabilidad tanto interna como externa, aunque en el caso de la interna del E2 el aumento sea mínimo, debido al hecho de que ya partía de un valor alto, cuando no usó las escalas mencionadas.

En cualquier caso, parece ser que, en el contexto en que se ha desarrollado la investigación, la implantación oficial de escalas de descriptores para la calificación de las pruebas orales de nivel C2 podría suponer la mejora de la fiabilidad de la prueba. Para implantarlas oficialmente, sería necesario completar aún ulteriores fases de validación y pilotaje de nuevas versiones de las escalas, perfeccionadas a la luz de los resultados obtenidos.

REFERENCIAS BIBLIOGRÁFICAS

Alderson, J.C., Clapham, C., Wall, D. (1998). *Exámenes de idiomas*. Madrid: Cambridge University Press.

Bisquerra, R. (coord.) (2004). *Metodología de la investigación educativa. Madrid*: La Muralla.

Bordón, T. (2006). *La evaluación de la lengua en el marco de E/L2: Bases y procedimientos*. Madrid: Arco Libros.

Dörnyei, Z. (2010). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Instituto Cervantes (1997? 2012). *Diccionario de términos clave de ELE* [en línea] Disponible en http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/default.htm/

León, O., Montero, I. (1997). *Diseño de investigaciones. Introducción a la lógica de la investigación en Psicología y Educación* . Madrid: MacGraw-Hill.

Luoma, S. (2004) *Assessing Speaking*. Cambridge: Cambridge University Press.