Aportes Científicos desde Humanidades
Número 13. Volumen 1. Junio 2018
ISSN 2545-8272
pp. 134-145

Lengua y Literatura

EVALUACIÓN DE LA ORALIDAD EN LE: ¿ESCALA HOLÍSTICA O ESCALA ANALÍTICA?

Lidia Soler
solerlidia@gmail.com

Florencia Giménez
fgimenezferrer@gmail.com

Griselda Bombelli
gbombelli@gmail.com

Facultad de Lenguas, Universidad Nacional de Córdoba

RESUMEN

Este trabajo presenta avances de actividades de investigación en el marco de un proyecto que comenzó en el año 2010 y que tiene como objetivo estudiar la confiabilidad de distintos métodos de evaluación para valorar el desempeño oral en una LE, más específicamente en la asignatura *Fonética y Fonología Inglesa II.* En un primer momento, se confeccionó una escala holística a fin de poder comparar los resultados que su aplicación arrojaba con aquellos alcanzados a través del método impresionista, que es el actualmente utilizado para la evaluación de la pronunciación en la Facultad de Lenguas, Universidad Nacional de Córdoba. Al no alcanzar ninguno de ellos estándares de confiabilidad significativos, se prosiguió con la confección de una escala analítica y se analizaron los índices de confiabilidad logrados a través de su aplicación. En esta presentación, describiremos los distintos procesos que culminaron en el diseño de los dos tipos de escalas empleadas, holística y analítica, y compararemos los resultados obtenidos al aplicarlas. Asimismo, nos referiremos a las ventajas y desventajas que se desprenden del uso de estos métodos de evaluación en el contexto de la asignatura en cuestión. Las conclusiones nos permiten diseñar futuras líneas de trabajo.

Palabras-clave:
Confiabilidad. Escalas. Evaluación. Pronunciación

Recibido: 30.nov.2016
Aprobado: 04.dic.2017

Lidia Soler
Florencia Giménez
Griselda Bombelli

ABSTRACT

This paper reports on the activities carried out as part of a research project that started in 2010 and whose objective is to study the reliability of different assessment methods used to evaluate university students' oral performance in English as a foreign language in the subject called Fonética y Fonología II. First, we designed a holistic scale and used it to assess students' oral performance. Next, we compared the results obtained with those that derived from the use of an impressionistic assessment method, which is the one currently used at the School of Languages, Universidad Nacional de Córdoba. As neither of the methods revealed statistically significant interrater reliability standards, we designed and applied an analytic scale and compared results with those obtained with the other two methods. In this paper, we describe the construction process of the two scales, holistic and analytic, and compare the results obtained after their application. Besides, we refer to the advantages and disadvantages of these assessment methods in the specific context in which they have been applied. Future research lines derive from the conclusions.

***Key Words**:*
reliability, scales, assessement, pronunciation

INTRODUCCIÓN

En contextos de enseñanza y aprendizaje, la evaluación cumple un rol fundamental. Los sistemas educativos requieren, por ejemplo, instancias de evaluación formal para determinar si los estudiantes pueden ser promovidos al próximo nivel del programa de estudios. Debido al rol central que cumple la evaluación y a las escasas investigaciones que se focalizan en esta instancia dentro del aprendizaje de una lengua extranjera (LE) es que decidimos comenzar a estudiar la confiabilidad de distintos métodos.

Aunque la evaluación puede adquirir diversas formas, usualmente involucra, por una parte, el uso de algún tipo de instrumento (prueba o examen) para recolectar muestras y, por otra, la evaluación de esas muestras con respecto a criterios de "corrección o nivel apropiado" establecidos con anterioridad (Bachman y Palmer, 1996:731); eventualmente, esta evaluación puede adquirir la forma de un número, como en el caso del contexto de este estudio.

La mayor parte de la bibliografía sobre la evaluación del uso de una lengua, inglés como LE en nuestro estudio (por ejemplo, Alderson, Clapham y Wall, 1995; Bachman, 1990; Bachman y Palmer, 1996 y 2010; Cohen, 1994, entre otros) no presta particular atención a la pronunciación y la considera solamente como uno de los varios componentes a tener en cuenta al evaluar la producción oral de un estudiante. En consonancia con lo dicho anteriormente, Derwing y Munro (2005) afirman que:

> se han realizado muchas menos investigaciones sobre la pronunciación de una segunda lengua que sobre otras habilidades como la gramática y el vocabulario. En consecuencia, el sentido común o la intuición influyen al momento de confeccionar material para la enseñanza y en las prácticas de enseñanza y aprendizaje habituales (p. 380).

Esta falta de estudios es todavía más evidente a nivel de carreras universitarias de grado en inglés como LE. Como sostienen Celce-Murcia, Brinton y Goodwin (1996), "existen rasgos propios de la pronunciación que afectan la manera en que se realiza la evaluación" (p. 341). Es decir, la evaluación es un proceso complejo y parece ser aún más complejo en el caso de la pronunciación debido a la naturaleza escurridiza del discurso oral y al alto grado de percepción individual que se requiere del evaluador. En coincidencia con esta afirmación, Luoma (2004) manifiesta que es muy difícil evaluar el desempeño oral de los estudiantes ya que los evaluadores deben "juzgar de manera instantánea una multiplicidad de aspectos" (p. 10), entre los que menciona los sonidos individuales, los tonos, el volumen, la velocidad, las pausas y la acentuación. Asimismo, la autora se pregunta "si todos estos aspectos pueden abarcarse en un solo criterio de evaluación" (p. 11).

La complejidad del proceso de evaluación, la escasez de trabajos de investigación sobre la pronunciación en contextos similares al nuestro y el deseo de lograr resultados más justos nos llevaron a iniciar el estudio de la evaluación de la pronunciación en el marco de la asignatura *Fonética y Fonología Inglesa II* del profesorado, traductorado y licenciatura de inglés de la Facultad de Lenguas, Universidad Nacional de Córdoba, con el propósito de mejorar el proceso y lograr un método más confiable y válido. Desde el año 2010 hemos indagado acerca de la confiabilidad de distintos métodos a fin de hacer de la evaluación una instancia lo más objetiva posible.

PROCEDIMIENTOS DE CALIFICACIÓN

Antes de presentar los avances de nuestra investigación debemos realizar algunas consideraciones generales sobre distintos procedimientos que pueden emplearse para calificar el desempeño de un estudiante en un examen.

Los procedimientos de calificación pueden abordarse desde distintas perspectivas. Se puede, por ejemplo, diferenciar entre calificación objetiva y subjetiva y distinguir entre procedimientos impresionistas (o generales), holísticos (o globales) y analíticos. Cohen (1994) establece un continuo que va desde la calificación objetiva a la subjetiva y manifiesta que la naturaleza de las actividades o ítems a calificar probablemente determine la ubicación del método de calificación hacia el extremo objetivo o subjetivo del continuo. De esta manera, las pruebas de opciones múltiples se calificarán más objetivamente que las entrevistas, por ejemplo. Los procedimientos impresionistas se ubican "en el extremo más subjetivo del espectro" (p.95), los procedimientos holísticos se encuentran a continuación y finalmente, los procedimientos analíticos están ubicados hacia el extremo más objetivo del continuo.

Bachman y Palmer (1996) establecen una diferencia entre calificación global (u holística) y calificación analítica. El empleo de escalas de calificación globales u holísticas se basa en la concepción que la habilidad lingüística constituye un todo unificado y, por lo tanto, la pericia lingüística puede ser cuantificada en una sola nota o calificación global. El criterio de evaluación consiste en juzgar la calidad total de la producción lingüística con especial énfasis "en lo que está bien hecho" (Cohen 1994:315). Según Bachman y Palmer (2010:339), una de las desventajas del uso de las escalas globales es que "es difícil saber lo que la nota

refleja" con respecto a los múltiples componentes inherentes a la habilidad lingüística. Otra desventaja, según estos autores, es que, aunque las escalas son descriptas como globales, incluyen en realidad múltiples componentes, los cuales son probablemente considerados en forma diferente por diferentes evaluadores.

A diferencia de los métodos globales, las escalas analíticas constan de una serie de componentes explícitos que son evaluados separadamente. Cuando se requiere una sola nota, esta se obtiene del total de estas calificaciones parciales. La ventaja de este método, sostienen Bachman y Palmer (1996, 2010), es que la calificación refleja el nivel de habilidad para cada componente o área incluida en la escala. Además, "las escalas analíticas tienden a reflejar lo que los evaluadores realmente hacen cuando califican muestras del uso de la lengua" (1996:211), puesto que aún los evaluadores globales admiten que consideran diferentes áreas al evaluar. No obstante, una de las desventajas de la evaluación analítica es que el desempeño es considerado una suma de partes en lugar de un todo integrado. Otro inconveniente es que el uso de escalas analíticas insume mucho tiempo y, en consecuencia, no son muy prácticas cuando hay un gran número de estudiantes para calificar, pocos evaluadores y tiempo limitado, como en el caso del contexto del presente estudio.

CONTEXTO DEL ESTUDIO

En la Facultad de Lenguas de la Universidad Nacional de Córdoba, la enseñanza de la pronunciación del inglés como LE a futuros licenciados, profesores y traductores está organizada en un programa de tres materias instrumentales obligatorias: Práctica de la Pronunciación en primer año y Fonética y Fonología en segundo y tercer año. Aunque los cursos son dictados por diferentes profesores, los contenidos, objetivos y material didáctico son comunes a todos los grupos del mismo nivel.

En el marco de las materias mencionadas, la evaluación de la pronunciación aborda tres aspectos: conocimiento teórico de la materia, percepción y producción lingüística. Con respecto a la producción oral, existe consenso general sobre los rasgos segmentales y suprasegmentales a considerar para evaluar y calificar el desempeño de los estudiantes. Se presupone que el "constructo a medir" (Bachman y Palmer, 1996:115) está orientado a la comunicación y se define por los objetivos específicos del curso. Sobre estas bases, los profesores del área de fonética inglesa evalúan y califican la producción oral a través de la 'valoración guiada' (Consejo de Europa, 2002) o, siguiendo a Weigle (2002), empleando un procedimiento impresionista, es decir, sin utilizar una "escala explícita" (p. 149). Obviamente, entonces, el método utilizado es aquel que, en el continuo de Cohen (1994), se ubica en el extremo más subjetivo.

BREVE RESEÑA DE LO REALIZADO

Etapa I

En una primera etapa estudiamos la confiabilidad del método impresionista que, como ya lo dijéramos, es el que se emplea en las asignaturas del área de fonética inglesa. En primer lugar, se recolectaron 32 muestras, grabando la producción oral de 32 alumnos cursantes de *Fonética y Fonología Inglesas II*. Los estudiantes fueron seleccionados al azar y se les

solicitó que realizaran dos actividades similares en tipo y duración a las actividades que realizan en los exámenes finales, a saber: la lectura de un cuento y una exposición oral de alrededor de un minuto. Antes de realizar las grabaciones, los estudiantes tuvieron tiempo para ensayar, como lo hacen en los exámenes.
Una vez obtenidas las muestras, se pidió a 5 profesores de fonética que las evaluaran en forma impresionista y que asignaran, a cada producción, una nota del 0 al 10, como es norma en la Universidad Nacional de Córdoba. La evaluación se realizó en una sesión y, al igual que en los exámenes finales, los profesores escucharon las muestras solamente una vez. Los resultados obtenidos indicaron que la variación entre las notas asignadas por los distintos evaluadores a cada una de las muestras no se mantuvo dentro de 'límites aceptables' (Shaw, 2001); dicho de otra manera, los análisis estadísticos señalaron que no se alcanzaron estándares de confiabilidad significativos para el método impresionista.

Ev1 I	*Ev2 I*	*Ev3 I*	*Ev4 I*	*Ev5 I*	*T^2*	*p*
3,55	*2,91*	*3,28*	*3,36*	*1,91*	*8,77*	*<0,0001*

Tabla 1. *Prueba de Friedman*
***Ev** seguido de un número designa a un evaluador; **I**: método impresionista*
(Fuente: elaboración propia)

Etapa II

Dado el resultado de los análisis estadísticos y teniendo en cuenta que la mayor parte de la bibliografía específica (Alderson et al., 1995; Bachman y Palmer 1996, 2010; Cohen, 1994; Luoma, 2005, entre otros) sugiere la utilización de escalas de evaluación como forma de lograr una mayor confiabilidad entre evaluadores, nos propusimos diseñar y utilizar dos tipos diferentes de escalas: una escala holística, en una primera etapa, y una escala analítica, en una etapa posterior. El objetivo era comparar el nivel de confiabilidad de estos métodos. Se comenzó con el diseño de la escala holística a partir de los criterios que se desprenden de los objetivos del programa de la asignatura *Fonética y Fonología Inglesas II* y de los resultados obtenidos de una entrevista semi-estructurada que se administró a cada uno de los evaluadores al finalizar la evaluación impresionista. La escala fue validada por dos jueces externos al proyecto, quienes también realizaron observaciones que resultaron de suma utilidad ya que guiaron el proceso de elaboración de la escala definitiva. Esta fue el resultado de varias pruebas y reuniones en las que los miembros del equipo pudimos identificar problemas, intentar soluciones y realizar cambios hasta que, finalmente, llegamos a una versión consensuada.

Se evaluaron las 32 grabaciones utilizando esta última versión de la escala holística y se aplicaron análisis estadísticos a los resultados obtenidos. Esto nos permitió descubrir que aún existían diferencias significativas entre las valoraciones medias asignadas por los distintos evaluadores a las mismas muestras empleando la escala holística.

Ev1 H	Ev2 H	Ev3H	Ev4H	Ev5H	T^2	p
3,42	3,67	2,44	2,83	2,64	5,47	0,0004

Tabla 2. *Prueba de Friedman*
***H:** método holístico*
(Fuente: elaboración propia)

Para profundizar el análisis, se indagó sobre dónde estaban las diferencias; con este objetivo se realizó la Prueba a Posteriori que determina entre cuáles evaluadores hay diferencias y entre cuáles no las hay. La Tabla 3 presenta los resultados de esta prueba, la cual revela entre qué evaluadores se encuentran las diferencias significativas.

Mínima diferencia significativa entre suma de rangos = 20,117

Tratamiento	Suma(Ranks)	Media(Ranks)	n				
Ev3H	78	2,44	32	A			
Ev5H	84,5	2,64	32	A	B		
Ev4H	90,5	2,83	32	A	B	C	
Ev1H	109,5	3,42	32			C	D
Ev2H	117,5	3,67	32				D

Medias con una letra común no son significativamente diferentes (p > 0,050)

TABLA 3. *Prueba a posteriori*
(Fuente: elaboración propia)

Etapa III

Siempre en pos de alcanzar niveles de confiabilidad interevaluador estadísticamente significativos, luego de la escala holística, diseñamos una escala analítica. Vale aclarar que, aunque la confiabilidad de un instrumento de esta naturaleza parece resultar superior a la de las escalas holísticas (Barkaoui y Knouzi, 2011; Weigle, 2002), en su momento se tomó la decisión de comenzar con una holística por su practicidad.

Como primer paso, se definió el constructo sobre el cual diseñar la escala analítica. El constructo, esencial para el diseño del instrumento y la interpretación de los resultados (Kim, 2006), implica una precisa delimitación de aquello a lo que se debe prestar particular atención en el "desempeño en una evaluación" (Bachman y Palmer, 2010:43). En el marco del modelo de Habilidad Lingüística Comunicativa (Bachman, 1990) y el Enfoque de Doble Foco de Atención (Morley, 1994), para el diseño de la escala analítica, al igual que para la holística, se tuvieron en cuenta los objetivos y contenidos propios de la asignatura *Fonética y Fonología Inglesas II*, así como también contenidos recurrentes que se abordan en forma espiralada desde el primer año de estudios. De esta manera, se consensuó y elaboró el constructo que se presenta en la Tabla 4.

Dominio de la habilidad lingüística hablada con un propósito específico y con atención a los siguientes componentes: estructuración y cohesión discursiva, rasgos paralingüísticos, entonación y ritmo, fluidez, sonidos consonantales y vocálicos en la cadena hablada, vocabulario y sintaxis.

Tabla 4. *Constructo a medir*
(Fuente: elaboración propia)

Una vez definido el constructo, se realizó el diseño de la escala analítica siguiendo protocolos específicos (Bachman y Palmer, 1996, 2010; Cohen, 1994; entre otros) que se adaptaron al contexto de uso específico. Se trabajó en pos de que el instrumento diseñado constara de categorías independientes, cada una definida por un número variable de componentes. El evaluador debía asignar una nota a cada categoría, según la valoración del desempeño del individuo con relación a cada uno de esos componentes. Es decir, la escala tiene la forma de una grilla de componentes a evaluar que conforman "un conjunto de escalas paralelas" (Consejo de Europa, 2002:189).

En cuanto a los descriptores de los niveles de dominio de las distintas categorías, se decidió presentar las bandas numéricas en forma detallada y separada de las categorías. Se establecieron cinco bandas que definen cinco niveles de desempeño (dominio total: 10, dominio amplio: 9-8, dominio moderado: 7-6, dominio limitado: 5-4 y dominio insuficiente: 3-2-1). Cada banda tiene una breve descripción holística del nivel de desempeño. La nota final que representa el desempeño del evaluado resulta del promedio de las notas asignadas a cada categoría.

En esta etapa, se realizaron dos pruebas con la escala analítica. Para la primera, se solicitó a 3 jueces externos al proyecto que evaluaran, en forma independiente, las grabaciones de 10 producciones orales con las que contábamos de etapas anteriores y que realizaran comentarios sobre el uso de la escala. Los resultados obtenidos indicaron la existencia de diferencias significativas entre las notas que los jueces asignaron a cada grabación. Teniendo en cuenta estos resultados y los comentarios de los jueces, realizamos ajustes al instrumento. A continuación, aplicamos la escala modificada para evaluar las 32 grabaciones que habían sido usadas para estudiar los métodos impresionista y holístico. Una vez implementada la evaluación, los resultados obtenidos se analizaron siguiendo la misma metodología utilizada para los métodos anteriores. Es decir, se utilizó una prueba no paramétrica para comparar los resultados entre evaluadores. Sus resultados se presentan en la Tabla 5.

Ev1 A	*Ev2 A*	*Ev3 A*	*Ev4 A*	*Ev5 A*	*T^2*	*p*
2,90	*3,31*	*1,90*	*3,29*	*3,60*	*6,37*	*<0,0001*

Tabla 5. *Prueba de Friedman*
A: *método analítico*
(Fuente: elaboración propia)

Los valores obtenidos permiten afirmar que, también en este caso, existen diferencias significativas entre los evaluadores. Al indagar entre qué evaluadores se producían estas diferencias, encontramos que un solo evaluador se diferenciaba del resto. Es decir, como

muestra la Prueba a Posteriori de la Tabla 6, de los cinco evaluadores, cuatro no mostraron diferencias significativas en su forma de evaluar por el método analítico.

Tratamiento	Suma(Ranks)	Media(Ranks)	n		
Ev3 A	59,00	1,90	31	A	
Ev1 A	90,00	2,90	31		B
Ev4 A	102,00	3,29	31		B
Ev2 A	102,50	3,31	31		B
Ev5 A	111,50	3,60	31		B

Medias con una letra común no son significativamente diferentes (p > 0,050)

TABLA 6. *Prueba a posteriori*
(Fuente: elaboración propia)

A continuación (Tablas 7 y 8) se compararon los resultados del método analítico con el impresionista y el holístico para cada evaluador; la prueba utilizada es la Prueba de Wilcoxon para muestras apareadas.

Evaluador	Método		N	Suma(R+)	Z	p(2 colas)
Ev1	Analítico	Impresionista	31	151,00	-1,90	0,0578
Ev2	Analítico	Impresionista	31	248,50	0,01	0,9744
Ev3	Analítico	Impresionista	31	86,50	-3,16	0,0010
Ev4	Analítico	Impresionista	31	226,00	-0,43	0,6802
Ev5	Analítico	Impresionista	31	415,00	3,27	0,0006

Tabla 7. *Prueba de Wilcoxon (muestras apareadas) (métodos analítico e impresionista)*
(Fuente: elaboración propia)

Evaluador	Método		N	Suma(R+)	Z	p(2 colas)
Ev1	Analítico	Holístico	31	38,00	-4,12	<0,0001
Ev2	Analítico	Holístico	31	47,50	-3,93	<0,0001
Ev3	Analítico	Holístico	31	61,00	-3,66	<0,0001
Ev4	Analítico	Holístico	31	157,50	-1,77	0,0798
Ev5	Analítico	Holístico	31	257,00	0,18	0,8494

Tabla 8. *Prueba de Wilcoxon (muestras apareadas) (métodos analítico y holístico)*
(Fuente: elaboración propia)

En la Tabla 7 se observa que dos evaluadores (Ev3 y Ev5) muestran diferencias significativas en las evaluaciones realizadas con el método analítico y el Impresionista. En la Tabla 8 tres evaluadores (Ev1, Ev2 y Ev3) muestran diferencias significativas en las evaluaciones realizadas con los métodos analítico y holístico. Ev3 presenta diferencias en

ambas comparaciones y Ev4 no presenta diferencias significativas a un nivel α=0,05 en ninguna de las comparaciones.

VENTAJAS Y DESVENTAJAS DEL USO DE LAS ESCALAS

Aunque no se alcanzó la confiabilidad buscada con ninguno de los tres métodos usados, es importante señalar las ventajas, desventajas y problemas en el uso de las escalas holística y analítica en este estudio.

Creemos que la escala holística es un instrumento práctico cuando se evalúa un número importante de alumnos, como es el caso en el contexto de este estudio. Sin embargo, en las pruebas piloto que se realizaron con evaluadores externos se presentaron diversos problemas, algunos de los cuales persistieron en la aplicación de la versión final a pesar de las revisiones realizadas. Entre ellos se destacan la delimitación de algunas de las bandas, pues la transición entre ellas no era lo suficientemente clara. En algunos casos, surgió la necesidad de hacer cambios respecto de la terminología utilizada con el fin de evitar palabras que denotaran alguna valoración subjetiva. Asimismo, se vio la conveniencia de agregar a la escala descriptores que hicieran referencia a la competencia gramatical y léxica.

Por otra parte, el rango de notas del 0 al 10 que se utilizó en la escala también desencadenó discrepancias. Se llegó a la conclusión de que el hecho de tener en cuatro de las seis bandas, dos posibles calificaciones (2-3, 4-5, 6-7, 8-9) favorecía las diferencias entre evaluadores. Como un instrumento de 11 bandas (del 0 al 10) es totalmente impráctico, se confeccionó una escala con bandas del 5 al 1 como suele usarse en algunos exámenes internacionales, discriminando los niveles micro y macro. De esta manera, la calificación asignada para cada aspecto es un número exacto y la nota final resulta del promedio de las calificaciones asignadas a cada aspecto. A raíz de estos cambios surgió el problema de convertir los resultados de la escala del 5 a 1 a la escala del 0-10 que necesariamente debe emplearse en las evaluaciones formales en la universidad.

Al seguir avanzando en nuestra investigación, descubrimos que nos encaminábamos hacia una escala más parecida a una analítica. Se decidió suprimir la banda del 0 ya que resulta muy improbable que un estudiante carezca de competencia fonética-fonológica.

Como ya mencionáramos, a pesar de todos los ajustes realizados hasta llegar a la versión que se consideró como la definitiva, creemos que algunas cuestiones no terminaron de resolverse. Por ejemplo, la inclusión de distintos componentes en cada banda descriptora plantea el problema de qué hacer cuando no todos los criterios incluidos "se satisfacen al mismo tiempo" (Bachman y Palmer, 2010:340) o se satisfacen en distinta medida. Es decir, es probable que los evaluadores decidan la asignación de nivel según el peso que cada uno otorgue a determinados criterios, lo que deriva en disparidades significativas. Cabe señalar, asimismo, la dificultad para brindar una adecuada devolución al alumno por la falta de descriptores más detallados.

Con respecto a la escala analítica, observamos que hubo mayor confiabilidad interevaluador ya que, con este instrumento, un solo evaluador se diferencia del resto. Es decir, de los cinco evaluadores, cuatro no muestran diferencias significativas en su forma de evaluar por el método analítico lo que parecería indicar que este método es más confiable.

Sin embargo, no podemos dejar de analizar e intentar solucionar algunos problemas que surgieron en el proceso de aplicación de la escala analítica, a pesar de haber discutido y aparentemente consensuado los criterios a seguir. En primer lugar, las categorías incluidas en la escala, estructuradas como un listado de componentes, no resultaron lo suficientemente precisas cuando hubo que aplicarlas en la evaluación de un número considerable de grabaciones en forma consecutiva y en una sola sesión. Es decir, a medida que avanzaba el proceso evaluativo, algunos evaluadores expresaron que comenzaron a valorar ciertos rasgos o componentes dentro de categorías en las que no estaban incluidos. En segundo lugar, el procedimiento para evaluar cada una de las 32 grabaciones tomó más tiempo del esperado, lo que le restó practicidad a la escala, condición indispensable en el contexto de uso. Esta demora puede haber sido consecuencia de la falta de claridad o precisión en la descripción de las categorías. Además, debido a la falta de definiciones precisas de los componentes, los evaluadores manifestaron que no se sintieron totalmente seguros con algunas de las decisiones tomadas, que es justamente lo que trata de evitarse con el uso de una escala.

Asimismo, cuando se optó por una escala analítica, se tuvo en cuenta el hecho de que estas escalas facilitan la devolución que se hace a los estudiantes sobre su desempeño, lo que consideramos de suma importancia para que las instancias de evaluación también contribuyan al éxito de los procesos de enseñanza y de aprendizaje. El no contar con categorías definidas con precisión constituye una dificultad también desde esa perspectiva. En suma, los resultados obtenidos parecen señalar la existencia de errores sistemáticos (Aiken, 2003), que tienen que ver con el diseño del instrumento de medición y la forma de aplicación de este. Si bien compartimos lo que la bibliografía especializada sostiene acerca de la mayor confiabilidad de la evaluación con el uso de una escala (Alderson et al., 1995; Bachman, 1990; Bachman y Palmer, 1996 y 2010; Cohen, 1994; entre otros), el instrumento no ha alcanzado aún los objetivos para los cuales fue diseñado.

CONCLUSIONES

Si bien los resultados obtenidos no han sido estadísticamente significativos, el camino recorrido hasta el momento no ha sido en vano. En primer término, nos ha permitido confirmar que el empleo de un instrumento de evaluación contribuye positivamente al proceso de evaluación. También, nos ha obligado a diseñar un plan de acción en pos de mejorar el diseño del instrumento elegido y su aplicación.

Para lograr este objetivo debemos realizar nuevos ajustes al instrumento, para lo cual realizaremos consultas a expertos en metodologías para la toma de decisiones en equipo (Días y Clímaco, 2005). Con todos los insumos producto del trabajo ya realizado y el asesoramiento en toma de decisiones en equipo, realizaremos los cambios necesarios en la escala analítica que conllevarán ajustes de mayor o menor envergadura y procederemos a validarla a través del juicio de expertos.

Una vez hecho esto, consideramos fundamental que los evaluadores realicen un entrenamiento en el uso de la escala, en consonancia con gran parte de la bibliografía dedicada al estudio de la evaluación de una LE que señala que la estandarización de

evaluadores es recomendable y contribuye a que la variación interevaluador se mantenga dentro de límites aceptables (Davies, 2016; Nguyen, 2015; Weigle, 1994; entre otros).

La preocupación por hacer de la evaluación una instancia lo más objetiva posible y por lograr que las valoraciones de los docentes reflejen de manera fidedigna el desempeño oral de los alumnos en un contexto sumamente específico como es el de la enseñanza universitaria nos ha obligado a abrirnos camino en un área con muy poca trayectoria en investigación. Sin embargo, tenemos la firme convicción de que los resultados de este estudio nos señalan el camino para continuar profundizando los estudios sobre el empleo de escalas de evaluación, así como también para investigar otros aspectos relacionados con la valoración de la producción oral en la universidad.

BIBLIOGRAFÍA

Aiken, L. (2003) *Test Psicológicos y Evaluación* (11va Edición). México: Prentice Hall.

Alderson, J. Ch., Clapham, C., y Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (1990) *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, L. F., y Palmer, A. S. (1996). *Language testing in practice*. Oxford: OUP.

Bachman, L. F., y Palmer A. S. (2010). *Language assessment in practice*. Oxford: OUP.

Barkaoui, K.; y Knouzi, O. (2011). Rating scales as frameworks for assessing L2 writing: examining their impact on rater performance. Paper presented at ALTE 4th International Conference, Kraków, Poland.

Celce-Murcia, M., Brinton, D. M., y Goodwin, J. M. (1996). *Teaching pronunciation*. *A reference for teachers of English to speakers of other languages*. Cambridge: CUP.

Cohen, A. D. (1994). *Assessing language ability in the classroom*. 2nd edition. Boston: Heinle & Heinle Publishers.

Consejo de Europa (2002). *Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación.* Secretaría General Técnica del MECD-Subdirección General de Información y Publicaciones y Grupo ANAYA, S.A.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33,* 117-135.

Derwing, T., y Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39* (3), 379-397.

Días, L. y Clímaco, J. (2005). Dealing with Imprecise Information in Group Multicriteria Decisions: A Methodology and a GDSS Architecture ll. *European Journal of Operational Research, 160,* 291-307.

Kim, H. J. (2006). Issues in rating scales in speaking performance assessment. *Working papers in TESOL & Applied Linguistics, 6* (2), 1-3.

Luoma, S. (2004). *Assessing speaking*. Nueva York: CUP.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

Morley, J. (1994). *Pronunciation pedagogy and theory*. *New views, new directions*. Alexandria, VA: TESOL.

Nguyen, A. T. (2015). *Towards an Examiners Training Model for Standardized Oral Assessment Qualities in Vietnam*. Recuperado de http://www.melta.org.my/majer/vol11(1)/03NguyenAT-MaJER11-1-41-51-250315.pdf (última consulta 04/08/2016)

Shaw, S. D. (2001). Issues in the assessment of second language writing. En *Research Notes, 6,* 2-5.
Weigle, S. (1994). Effects of training on raters o ESL compositions. *Language Testing, 11* (2) 197-223.
Weigle, S. (2002). *Assessing writing.* Cambridge: CUP.