*Aportes Científicos desde Humanidades*
Volumen 12. Número 2. Diciembre 2017
ISSN 2545-8272
pp. 122-128

Lingüística y Literatura

# LA ESTANDARIZACIÓN: POTENCIAL HERRAMIENTA PARA MEJORAR LA CONFIABILIDAD INTEREVALUADOR

*Florencia Giménez*
fgimenezferrer@hotmail.com

*M. Josefina Díaz*
diazm.josefina@gmail.com

*María Garay*
bopgff@gmail.com

**RESUMEN**

En este trabajo se presentan parte de las actividades realizadas hasta el momento en el marco de un proyecto de investigación cuyo objetivo principal es estudiar la confiabilidad de distintos métodos de evaluación para valorar el desempeño oral en una lengua extranjera, más específicamente, en la asignatura *Fonética y Fonología Inglesa II*, sección inglés, de la Facultad de Lenguas, U.N.C. En esta oportunidad, nos focalizamos en los resultados obtenidos en torno a la utilización de una escala analítica de evaluación y a la necesidad inminente de llevar a cabo un proceso de estandarización de evaluadores que, creemos, puede contribuir a mejorar los índices de confiabilidad interevaluador. Como lo señala Shaw (2001), la variación entre evaluadores "es una seria debilidad al momento de valorar el desempeño en una lengua", por lo que "la estandarización de evaluadores ha sido altamente recomendada a fin de mantener la variación dentro de límites aceptables" (p. 2) y de "alcanzar la mayor uniformidad posible entre las valoraciones de distintos evaluadores" (Nguyen, 2015:41). Para cumplir con este objetivo, nos proponemos referirnos al marco teórico que sustenta el proyecto y esta propuesta en particular. Asimismo, realizamos una sinopsis de lo realizado hasta el momento para explicar cómo surge la necesidad del proceso de estandarización al que hiciéramos referencia. De esta manera, podremos detallar el proceso de confección del protocolo a seguir y las decisiones que se tomaron en torno a la concreción de este.

**Palabras-clave**:
Confiabilidad interevaluador. Estandarización. Evaluación. Nivel superior. Pronunciación del inglés.

*Recibido: 30.nov.2016*
*Aprobado: 05.jul.2017*

*Florencia Giménez*
*M. Josefina Díaz*
*María Garay*

## INTRODUCCIÓN

La evaluación, entendida como la valoración del desempeño de los aprendices, constituye un elemento esencial en el proceso de enseñanza-aprendizaje en distintos contextos educativos. En este sentido, su importancia es tal que los métodos por medio de los cuales se vehiculiza continúan siendo objeto de estudio en trabajos de investigación. Si bien la evaluación puede llevarse a cabo de maneras diferentes, generalmente, se elabora un instrumento para cada instancia de evaluación y en base a criterios de corrección[1] establecidos con anterioridad (Bachman y Palmer, 1996). A la vez, de acuerdo con estos criterios preestablecidos, se puede arribar a un número que indique, dentro de un continuo, el grado de concreción de los objetivos planteados para una determinada instancia. Éste es el caso en el presente trabajo, dado el contexto en el que se inscribe.

Específicamente, en el área de la enseñanza del inglés como lengua extranjera (ILE), la evaluación del desempeño lingüístico se encuentra presente en reiterados momentos durante el proceso. La modalidad y rol que cumplen los actores involucrados –docentes, estudiantes, pares, evaluadores externos, por ejemplo– varía, o puede variar, en cada oportunidad. La teoría de la competencia comunicativa ha ejercido influencia sobre la forma de concebir la evaluación de la habilidad lingüística en el ámbito de ILE. En consonancia con estos cambios, Bachman (1991, citado en Cohen, 1994) puntualiza que los métodos de evaluación han recibido particular reconocimiento y que se evidencia la necesidad de diseñar instrumentos de evaluación "más sofisticados" (p. 3). Si bien parte de la bibliografía sobre la evaluación del uso de la lengua relevada para este estudio trata acerca de la valoración del desempeño oral, aquella que se centra en la pronunciación resulta extremadamente escasa. En otras palabras, este aspecto se considera uno más de los varios componentes que intervienen, y que deben tenerse presentes, al evaluar la producción oral de los estudiantes, pero no se hace foco en esta área en sí. Creemos que la causa de esta realidad se relaciona con el hecho de que, siguiendo a Celce-Murcia, Brinton y Goodwin (1996), "existen rasgos propios de la pronunciación que afectan la manera en que se realiza la evaluación" y se conjetura que es por esta misma razón que, según estos autores, "en la literatura sobre pronunciación, se presta poca atención a la evaluación" (p. 341). Nuestra experiencia docente nos demuestra que los procesos de evaluación conllevan cierto grado de complejidad en sí mismos, y que estos se tornan aún más complicados en el caso particular de la pronunciación debido a la naturaleza efímera del discurso oral, lo que requiere de un alto grado de experticia por parte del evaluador.

## CONTEXTO

Este trabajo se enmarca en una investigación cuyos comienzos datan del año 2010 y que surge del interés por investigar, mejorar y hacer lo más confiable y objetiva posible la evaluación en la asignatura *Fonética y Fonología Inglesa II*, de las carreras de Profesorado, Traductorado y Licenciatura de la Facultad de Lenguas, Universidad Nacional de Córdoba.

En una primera etapa estudiamos dos métodos de evaluación (impresionista y holístico) a fin de determinar cuál era más confiable. Debido a que la variación entre las notas asignadas por los evaluadores a cada una de las muestras utilizadas no se mantuvo

[1] Todas las citas de fuentes en inglés son traducciones de las autoras de este trabajo.

dentro de límites aceptables (Shaw, 2001), los números que arrojaron los análisis estadísticos no alcanzaron estándares de confiabilidad significativos.

En la etapa inmediatamente anterior a la presente, los objetivos centrales fueron confeccionar y poner en funcionamiento una escala analítica y comparar el nivel de confiabilidad de este método con los otros métodos ya estudiados. Para alcanzar este objetivo, se realizó búsqueda y lectura de bibliografía específica, al igual que consultas a expertos en el área de la evaluación. Asimismo, se definió el constructo[2] sobre el cual diseñar la nueva escala. Luego, se realizó el diseño de la escala analítica siguiendo protocolos específicos (Bachman y Palmer 1996, 2010; Cohen 1994, entre otros). Una vez confeccionada, 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. De esta disparidad entre las medias que cada investigador otorgó a la misma muestra y de otros problemas que surgieron con la utilización de la escala derivó la necesidad de realizar ajustes al instrumento. Se optó por continuar mejorando esta escala y no otra, ya que, no solo los evaluadores en la etapa 2014-2015[3] reportaron que el instrumento les había resultado más práctico que aquél diseñado en etapas anteriores (escala holística), sino que, a pesar de existir diferencias significativas entre las valoraciones medias asignadas por los distintos evaluadores, los análisis estadísticos realizados reportaron una mayor confiabilidad interevaluador. Una vez realizados los cambios, los evaluadores procedieron a evaluar el corpus de 32 muestras orales. Los valores obtenidos nuevamente indicaron diferencias significativas entre evaluadores. Es por esta razón que surge la inminente necesidad de realizar nuevos ajustes al instrumento y de llevar adelante un proceso de estandarización, para luego proceder, pasado un lapso de tiempo no menor a 150 días, a evaluar las 32 muestras nuevamente.

## LA ESTANDARIZACIÓN DE EVALUADORES: UNA ACCIÓN EN POS DE LA CONFIABILIDAD

Diferentes investigaciones que se centran en la evaluación en LE han podido comprobar que distintos evaluadores son capaces de alcanzar una cierta uniformidad al momento de evaluar (Alderson et al., 1995; Bachman 1990, 2010; Cohen, 1994; Eckes, 2011; McNamara, 1996, entre otros).

Weigle (1994) presenta el análisis de protocolos verbales previos y posteriores a sesiones de entrenamiento de 4 evaluadores no experimentados al momento de valorar la producción escrita de alumnos que rinden exámenes de ubicación. Dicho análisis parece reflejar que el proceso de estandarización contribuyó a esclarecer aspectos inherentes a los criterios de evaluación y modificó las expectativas con las que los evaluadores se enfrentaron a las producciones escritas de los alumnos. En consecuencia, todo esto redundó en una menor disparidad entre las valoraciones de los evaluadores.

Fahim y Bijani (2011) investigaron la relación que existe entre el entrenamiento y la severidad al momento de evaluar producciones escritas de estudiantes de ILE. Los

[2] El constructo resulta un complemento fundamental de la escala, tanto al momento de su diseño como de su implementación, ya que sitúa al evaluador respecto de los parámetros a los que debe prestar particular atención.

[3] La devolución es muy importante en los procesos de enseñanza y aprendizaje y una escala analítica permite contar con información detallada y con los aspectos a evaluar claramente diferenciados.

resultados revelaron que la mayoría de los evaluadores pudieron modificar sus criterios de evaluación, lo que redundó en una mayor uniformidad en las valoraciones.

Davies (2016) nos interesa sobremanera, ya que es unos de los pocos estudios que se focaliza en la evaluación de la oralidad, creemos debido a que, como mencionáramos anteriormente, "existen rasgos propios de la pronunciación que afectan la manera en que se realiza la evaluación" (Celce-Murcia et al., 1996:341). Davis investiga en qué medida la estandarización y la experiencia para evaluar de 20 experimentados profesores de inglés influye en la evaluación de la sección *Speaking* del examen internacional TOEFL IBT. Los resultados posteriores a las sesiones de entrenamiento a las que los evaluadores fueron sometidos revelan una mejoría en la correlación interevaluador, al igual que un acercamiento a los puntajes de referencia.

Los resultados de los estudios que mencionamos parecen indicar que la estandarización de evaluadores es recomendable, dado que contribuye a que la variación interevaluador se mantenga dentro de límites aceptables (Nguyen, 2015). Esto nos ha llevado a preguntarnos si es posible que surjan cambios luego de someter a evaluadores experimentados a sucesivas sesiones de entrenamiento en el uso de la escala analítica para valorar el desempeño oral de alumnos de *Fonética y Fonología Inglesa II*. Si los hubiera, ¿cuál sería la naturaleza de estos cambios?, ¿contribuirían a mejorar la variación entre evaluadores que valoran una misma muestra a fin de alcanzar un nivel de confiabilidad satisfactorio? Por último, otro interrogante del proyecto marco del que deriva este trabajo es si, luego del proceso de estandarización de evaluadores, el uso de la escala analítica resulta más confiable en términos estadísticos que los otros métodos ya valorados. Si bien las respuestas a estos interrogantes no se expondrán en esta oportunidad, resulta de importancia exponer qué nos impulsa a llevar adelante las acciones que describimos a continuación.

## CONFECCIÓN DEL PROTOCOLO

Previo al entrenamiento de los evaluadores, consideramos necesario confeccionar un protocolo de acción que constituya un registro pormenorizado de los procedimientos que se aconseja llevar adelante. Esto sin lugar a dudas facilitará, agilizará y eficientizará las sesiones de estandarización. Asimismo, resultará beneficioso para implementar en momentos en los que se produzcan incorporaciones docentes a la cátedra. Por otro lado, como dice Wang (2010), "el entrenamiento de evaluadores no es un evento aislado y único, sino que es una tarea continua" (p. 110). Los resultados del entrenamiento de evaluadores pueden no perdurar mucho más allá del entrenamiento en sí. Es por esto que es necesario mantener sesiones de entrenamiento periódicas antes de evaluaciones a fin de refrescar criterios de evaluación, para lo que el protocolo resulta de vital importancia.

Luego de relevar bibliografía específica sobre procedimientos para la estandarización de evaluadores y teniendo en cuenta el contexto específico en el que este proceso se llevará adelante, consideramos que el procedimiento debe constar de dos sesiones de entrenamiento en las cuales haya momentos de trabajo grupal e individual. Asimismo, debe involucrar de manera activa a los evaluadores. Todo esto resulta en un protocolo de acción que incluye los siguientes pasos:

1. Lectura y discusión de la escala analítica. Este primer paso resulta crucial para que los evaluadores se familiaricen con el instrumento y la terminología allí utilizada. Resulta esencial que, previo a la evaluación de muestras, se logre un consenso respecto de qué aspecto evaluar en cada categoría. Esto resulta especialmente

dificultoso en el área de fonética y fonología, en particular, debido a la dificultad de aislar rasgos que pueden ser evaluados en distintas categorías. Se hará un registro pormenorizado de todo lo consensuado a fin de que los evaluadores puedan recurrir a esta información cada vez que sea necesario. Este instrumento y la detallada descripción de cada una de sus partes servirá, además, para entrenar nuevos evaluadores y realizar sesiones de seguimiento en el que se retomen todos o algunos de los descriptores para trabajar en mayor profundidad y continuar así con la estandarización de criterios una vez terminado este proyecto.

2. Análisis y discusión de muestras ya evaluadas por expertos. En este contexto en particular, las directoras y co-directora del proyecto actuarán como los evaluadores expertos, quienes, a su vez, son las profesoras titulares de *Fonética y Fonología II*. La importancia de este paso reside en que los evaluadores en plena etapa de entrenamiento tienen la posibilidad de analizar, discutir y comprender los criterios de evaluación que adoptaron experimentados profesores de fonética y fonología inglesa. En esta etapa se prevé trabajar con 5 muestras que reflejen distintos grados de experticia, incluyendo un mayor número de aquellas de alumnos con un nivel de desempeño medio (Alderson et al., 1995) que, de acuerdo con nuestra experiencia como evaluadores, resultan sumamente difíciles de estimar.

3. Evaluación de nuevas muestras; discusión y comparación de resultados. Para esta etapa se seleccionarán 10 muestras de un corpus recolectado para el entrenamiento. Las muestras elegidas, al igual que en el paso anterior, deberán reflejar distintos grados de experticia, incluyendo nuevamente un mayor número de producciones de alumnos con un nivel de desempeño medio (Alderson et al., 1995). Durante el proceso de evaluación, los evaluadores en proceso de entrenamiento podrán socializar sus valoraciones y, de esta manera, llegar a una calificación de manera conjunta para así avanzar en la unificación de criterios y la discriminación de los componentes que integran las categorías del instrumento.

4. Evaluación y discusión de nuevas muestras. En esta fase del proceso, los evaluadores siguen el mismo procedimiento del paso anterior con otras 10 nuevas muestras, pero ya provistos de un instrumento que materializa criterios unificados y categorías claramente delimitadas, producto de la conformidad que manifestaron los integrantes del equipo evaluador en la etapa que precede a esta. En esta etapa los evaluadores en proceso de entrenamiento valorarán las muestras de manera individual, sin la posibilidad de discutir ni socializar durante el transcurso de la valoración que cada miembro del equipo realice. El objetivo de este paso es, por consiguiente, ejercitar el proceso de evaluación una vez más, siguiendo los criterios preestablecidos. Podrán, si los evaluadores consideraran necesario, socializar los resultados obtenidos con el fin de identificar y, así, resolver alguna discrepancia que surgiera una vez finalizada la evaluación.

5. Evaluación en forma independiente e individual por parte de cada evaluador del corpus de 32 muestras. En esta parte del proceso, cada evaluador escuchará las 32 muestras que forman parte del corpus en orden diferente al que fueron presentadas antes de comenzar el entrenamiento. Las notas que cada evaluador asigne a cada muestra se registrarán en una grilla para luego poder tabularlas y someterlas a análisis estadísticos. Al igual que al momento de evaluar las muestras previo al entrenamiento, también en esta oportunidad se llevará un registro del tiempo que los evaluadores toman para valorar cada muestra ya que esta instancia también tiene como objetivo agilizar la utilización del instrumento y, por consiguiente, el proceso de evaluación.

Una vez concluida la etapa de estandarización, se prevé la aplicación de pruebas estadísticas con el fin de determinar la confiabilidad de los resultados obtenidos. El objetivo que se persigue es observar si los índices de confiabilidad interevaluador aumentan luego del entrenamiento y, en caso de ser así, si dicha confiabilidad alcanza límites aceptables, es decir si los valores son estadísticamente significativos. Es de esperar, como ha sido el caso de las investigaciones mencionadas, que este proceso redunde en una mejora en la confiabilidad interevaluador y, por consiguiente, de la instancia de evaluación en sí y de los resultados que de ésta se obtengan.

## CONCLUSIONES

Si bien, como ya lo mencionáramos, la estandarización de evaluadores parece ser una práctica más comúnmente utilizada para mejorar la confiabilidad interevaluador al momento de valorar producciones escritas y, no así orales. Sin embargo, consideramos ésta una posible alternativa en pos de la obtención de resultados más confiables al momento de evaluar el desempeño oral de alumnos en el área de fonética y fonología inglesa.

De ser positiva la experiencia, y de mantenerse el grupo de trabajo medianamente estable, podría pensarse también en sesiones de entrenamiento periódicas, en miras a exámenes, en las que de manera alternativa se discutan aspectos aislados en la escala, siguiendo también los pasos explicitados en el protocolo que se ha confeccionado.

En sucesivas etapas puede considerarse, también, entrenar a los evaluadores en el uso de la escala holística confeccionada anteriormente y así comparar los índices de confiabilidad obtenidos con cada instrumento. Esto, sin embargo, excede los objetivos planteados en el proyecto marco del cual deriva este trabajo.

Es de esperar que este proceso que comenzamos a transitar hace ya varios años, y en el que ha habido avances y también retrocesos, contribuya a hacer de la evaluación del desempeño oral en el área de pronunciación una instancia lo más objetiva y confiable posible.

## BIBLIOGRAFÍA

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

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.

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.

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

Eckes, T. (2011). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. Frankfurt, Germany: Lang.

Fahim, M. y Bijami, H. (2011). The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment. *Iranian Journal of Language Testing*, *1*(1), 1-16.

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

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: 05/08/2016)

Shaw, S. D. (2001). Issues in the assessment of second language writing. En *Research Notes 6* (pp. 2-5). Cambridge: CUP.

Weigle, S. (1994). Effects of training on raters o ESL compositions. *Language Testing, 11*(2), 197-223.