Bol. Acad. peru. leng. 69. 2021 (265-295)

Criterios de evaluación en escritura académica en inglés: encuentros entre evaluadores y confiabilidad

Assessment Criteria in Academic Writing in English: Rater Sessions and reliability

Critères d'évaluation de l'écriture académique en anglais : rencontres entre évaluateurs et fiabilité

Andrea de los Ángeles Canavosio
Facultad de Lenguas, Universidad Nacional de Córdoba, Córdoba, Argentina
andrea.canavosio@unc.edu.ar
https://orcid.org/0000-0001-7301-2254

Ana Cecilia Cad
Facultad de Lenguas, Universidad Nacional de Córdoba, Córdoba, Argentina
anaceciliacad@unc.edu.ar
https://orcid.org/0000-0002-2894-5060

Julieta Salinas
Facultad de Lenguas, Universidad Nacional de Córdoba, Córdoba, Argentina
jusalinas@unc.edu.ar
https://orcid.org/0000-0002-5123-3782

https://doi.org/10.46744/bapl.202101.010

e-ISSN: 2708-2644

https://doi.org/10.46744/bapl.202101.010

Resumen:
La evaluación de la escritura en lengua extranjera es un proceso entretejido de complejos y múltiples aspectos, lo que conlleva a que este campo de la investigación sea abordado desde diferentes perspectivas. El trabajo que se presenta a continuación surge de la necesidad de indagar sobre los criterios que aplican los docentes cuando evalúan la escritura en lengua extranjera (inglés) en un contexto universitario. Esta investigación se encuentra enmarcada en un proyecto bianual de mayor envergadura denominado «Criterios de evaluación de la escritura en lengua extranjera (inglés) en el nivel superior», avalado por la Secretaría de Ciencia y Técnica de la Universidad Nacional de Córdoba (Argentina). En este trabajo en particular, se analizarán los datos recolectados en un encuentro durante el cual nueve docentes explicaron cómo llegaron a la calificación de cinco ensayos escritos en instancias de evaluación sumativa. Consideramos que este análisis contribuirá a estimular la reflexión sobre la elección, ponderación y uniformidad en los criterios de evaluación utilizados a futuro.

Palabras clave: confiabilidad, escritura académica en ILE, criterios de evaluación, nivel universitario, percepciones de evaluadores.

Abstract:
The assessment of foreign language writing is a complex and manifold process, which entails that this field of research be approached from different perspectives. This paper arises from the need to research the teaching criteria when assessing writing in a foreign language (English) in a university context. This research is part of a larger biannual research project called «Criteria for the assessment of foreign language (English) writing at the higher level» (supported by the Secretariat of Science and Technology of the National University of Córdoba, Argentina). In this particular work, we will analyze the data collected in a meeting at which nine teachers explained how they scored five written essays under summative assessment. We believe that this analysis will foster reflection on the selection, weighting and uniformity of assessment criteria to be applied in the future.

Key words: reliability, academic writing in EFL, assessment criteria, undergraduate level, raters' perceptions.

https://doi.org/10.46744/bapl.202101.010

Résumé:
Le processus d'évaluation de l'écriture en langue étrangère est tissé par de complexes et multiples aspects, ce qui emmène à ce que ce domaine de recherche fasse l'objet de différentes approches. Le présent article obéit au besoin de découvrir les critères employés par les professeurs quand ils évaluent l'écriture en langue étrangère (anglais) dans un contexte universitaire. Cette étude ressort d'un projet de recherche biannuel de plus grande envergure dénommé «Critères d'évaluation de l'écriture en langue étrangère (anglais) dans l'enseignement supérieur», soutenu par le Secrétariat de Science et Technique de l'Université Nationale de Córdoba (Argentine). Dans cet article en particulier nous analyserons les données obtenues pendant une rencontre où neuf professeurs ont expliqué comment ils avaient attribué leur note à cinq essais écrits dans un contexte d'évaluation sommative. Nous pensons que cette analyse encouragera la réflexion sur la sélection, le poids et l'uniformité des critères d'évaluation à appliquer à l'avenir.

Mots clés: fiabilité, écriture académique en Anglais Langue Étrangère, critères d'évaluation, enseignement supérieur, perceptions des évaluateurs.

Recibido: 10/09/2020 Aceptado: 27/02/2021 Publicado: 30/06/2021

1. Introducción

Al evaluar composiciones académicas en lengua extranjera se debe tener en cuenta una variedad de aspectos a la vez, lo cual hace que este tipo de evaluación sea investigada desde diferentes ángulos para poder así entender su complejidad. Los resultados que surgen de dichos análisis revisten gran importancia, ya que son el punto de partida para nuevas prácticas docentes. Es por esto que la evaluación se convierte en foco de atención de investigaciones y análisis que permiten determinar o al menos

https://doi.org/10.46744/bapl.202101.010

sugerir prácticas docentes concretas a seguir. Por este motivo, el objetivo planteado por un grupo de investigación de la Facultad de Lenguas de la Universidad Nacional de Córdoba (Argentina) fue estudiar qué criterios de evaluación aplican los docentes a cargo del curso Lengua Inglesa II cuando evalúan ensayos académicos. Cabe destacar que la materia se dicta en el segundo año de las carreras de Traductorado, Profesorado y Licenciatura en Lengua y Literatura Inglesa de dicha institución.

Esta comunicación se desprende del proyecto bianual recién mencionado, el cual se llevó a cabo durante los años 2014-2015 y se titula «Criterios de evaluación de la escritura en lengua extranjera (inglés) en el nivel superior». El objetivo principal del proyecto marco fue indagar sobre cómo los docentes de este curso evalúan las composiciones en instancias de evaluación sumativa y qué creencias y actitudes tienen sobre qué es evaluar la escritura académica en inglés. En etapas previas, el equipo de investigación se dedicó a analizar la retroalimentación docente y el impacto que esta tiene en el desarrollo de la habilidad escrituraria de los estudiantes. En una etapa posterior, de la cual se desprende el objetivo principal del proyecto marco actual, surge el interés por averiguar acerca de los criterios que utilizan los docentes cuando evalúan ensayos en inglés y qué impacto tienen en el proceso de evaluación.

Como parte del mismo, se organizó una reunión con los docentes para discutir criterios de evaluación. Esta fue una de las técnicas utilizadas con el objetivo de indagar sobre los procedimientos, ponderaciones y estrategias a las que recurren los docentes de lengua extranjera al momento de calificar los ensayos académicos en inglés que fueron escritos por sus alumnos. Previo a dicha reunión, cada uno de los docentes evaluó individualmente una serie de composiciones, y luego informó, en el encuentro presencial, la calificación asignada a cada ensayo como así también las razones por las cuales tomó esa decisión. Luego se realizó una puesta en común e intercambio de ideas sobre el proceso de evaluación.

Así, como fruto de ese trabajo, en este artículo presentaremos datos recolectados durante dicho encuentro con el objetivo de intentar comprender si los procesos abordados para llegar a una determinada calificación de un

https://doi.org/10.46744/bapl.202101.010

escrito académico en inglés son similares entre colegas, y si este tipo de instancias de discusión pueden favorecer e incrementar la confiabilidad entre evaluadores.

Al respecto, se pueden mencionar algunos estudios que se han dedicado a explorar las diferencias entre evaluadores al calificar ensayos académicos, aunque es importante señalar que se observa una escasez de trabajos en un contexto como el que aquí se describe. Es importante destacar, además, que la mayoría de los trabajos recientes hacen hincapié en los diferentes criterios que ponderan los evaluadores y las razones por las cuales aplican diferentes criterios; sin embargo, no se estudian alternativas para disminuir diferencias y aumentar la confiabilidad a través de sesiones de discusión entre evaluadores. Por ejemplo, Trace, Janssen y Meier (2016) llevaron a cabo un estudio de metodología mixta en el que estudiaron el impacto de la discusión y la negociación entre evaluadores sobre la consistencia en las calificaciones de ensayos académicos escritos por estudiantes de inglés como segunda lengua. Se analizaron las secciones de escritura de exámenes que rindieron estudiantes de una universidad colombiana como requisito de ingreso para la carrera de doctorado. Los resultados del estudio revelan que la negociación no modificó la severidad de los evaluadores a la hora de calificar, pero sí redujo significativamente su sesgo individual. Asimismo, los participantes entrevistados expresaron que la negociación ayuda a generar consenso y comprensión de los criterios y categorías de evaluación, además tiene un impacto positivo en las prácticas de enseñanza.

En el trabajo de Kuiken y Vedder (2014) se analizaron las evaluaciones de composiciones escritas por hablantes de inglés como lengua materna y como lengua extranjera. El propósito fue establecer en qué se basan los evaluadores para decidir las calificaciones y para analizar la correlación entre la evaluación de la complejidad lingüística y la adecuación comunicativa. Los resultados muestran que, aunque hubo correlación entre los dos aspectos de la evaluación, los participantes asignaron mayor importancia a la adecuación comunicativa (contenido, uso de argumentos, organización retórica, estilo y comprensibilidad general) que a la complejidad lingüística (léxico, gramática, ortografía y precisión). Los evaluadores expresaron haber

https://doi.org/10.46744/bapl.202101.010

tenido dificultades para mantener el mismo criterio al evaluar exámenes de diferentes niveles y reconocieron haber tenido diferentes expectativas según el nivel de los alumnos.

En su trabajo, Schaefer (2008) indagó sobre los patrones de sesgo en evaluadores hablantes nativos de inglés, quienes califican composiciones escritas por estudiantes de inglés en Japón. El análisis de los datos identificó algunos patrones en grupos de evaluadores. En los casos de sesgo hacia las categorías, si el contenido y la organización fueron calificados de manera severa, el uso de la lengua y la mecánica se calificaron de forma indulgente y viceversa. En los casos de sesgo hacia los escritores, se identificó más sesgo ya sea hacia la severidad o la tolerancia en los casos de alumnos con mayor proficiencia. Algunos evaluadores fueron excesivamente severos con alumnos más proficientes y excesivamente tolerantes con alumnos de bajo desempeño.

Por su parte, Eckes (2008) sostiene que, luego de analizar investigaciones previas, los evaluadores experimentados difieren en la interpretación de los criterios de evaluación. Con este objetivo, aplica un enfoque cuantitativo de carácter clasificatorio para identificar los diferentes tipos de evaluadores. El contexto de su estudio es el de la sección de escritura del examen de Alemán como Lengua Extranjera (*Test Deutsch als Fremdsprache,* TestDaF), el cual rinden estudiantes extranjeros que solicitan admisión a estudios de nivel superior en Alemania. La sección de escritura evalúa la habilidad del aspirante para producir un texto coherente y estructurado sobre un tema determinado. Los resultados de su análisis concluyen que los evaluadores difieren significativamente sobre la importancia general que les dan a los criterios de evaluación, lo que da lugar a futuras investigaciones sobre el tema.

Asimismo, Eckes (2012) continúa analizando el rol de los evaluadores, y enfoca su estudio en cerrar la brecha entre el conocimiento que poseen los evaluadores y su comportamiento. Su investigación adopta un enfoque clasificatorio y, además de continuar analizando las diferentes percepciones que tienen los evaluadores sobre los criterios de evaluación, también analiza la forma en que evaluadores de diferentes tipos hacen uso de esos

https://doi.org/10.46744/bapl.202101.010

criterios en una sesión de evaluación. Su supuesto parte de la noción de que la percepción e interpretación de los criterios de evaluación juegan un rol importante a la hora de determinar el sesgo de los evaluadores. Las diferentes percepciones que cada evaluador atribuye a los criterios de evaluación equivalen a su inclinación ya sea por la severidad o la indulgencia. Así, establece que poder categorizar los tipos de evaluador permitiría alcanzar una percepción más equilibrada de la importancia de los criterios y, por ende, reducir las probabilidades de adoptar una actitud severa o más indulgente a la hora de corregir los escritos. También sugiere continuar la línea de investigación con una orientación hacia el proceso de toma de decisiones que caracteriza a cada tipo de evaluador.

Siguiendo una línea similar de investigación, Yen (2016) reconoce la importancia de la validez entre evaluadores para realizar un proceso evaluador justo. Por ello, se deben lograr niveles altos de consistencia entre evaluadores. Los participantes de este estudio fueron diez evaluadores de la Facultad de Inglés de la Universidad de Lenguas y Estudios Internacionales. Cada uno evaluó cinco tareas de escritura de dos tipos empleando la escala del examen internacional IELTS. Los evaluadores asignaron una nota a las tareas y también realizaron comentarios sobre las mismas. Yen concluye que las discrepancias encontradas entre los evaluadores se deben a tres motivos: primeramente, no todos leían todos los descriptores de una banda; en segundo lugar, cada uno asignaba una nota basada en aspectos no mencionados en la escala; y, finalmente, los docentes manifestaban que al corregir un ensayo se sentían influenciados por las producciones leídas con anterioridad. Por esto, el autor sugiere que, cada banda dentro de la escala, comience enunciando los elementos más determinantes de esta última. Los descriptores deben ser reescritos en caso de no presentar descriptores sobre uso específico de vocabulario o prolijidad. Por último, sugiere que palabras tales como *inadecuado* o *bien* deberían ser reemplazadas por vocabulario más claro y directo.

En su investigación, Kayapinar (2014) compara la confiabilidad de diferentes instrumentos de evaluación de escritura en inglés como lengua extranjera. En el estudio se reconoce la importancia de llevar a cabo un proceso evaluador de la escritura que sea consistente, ya que las decisiones

https://doi.org/10.46744/bapl.202101.010

de los evaluadores tienen un impacto sobre la vida personal y profesional de los alumnos. La investigación se abocó a analizar posibles variaciones o consistencias en el proceso de evaluación de tareas de escritura de 44 alumnos universitarios de lengua inglesa cuyos trabajos fueron analizados por 10 evaluadores empleando tres diferentes escalas: a) escala de evaluación basada en impresión general (GIM por sus siglas en inglés), b) escala de criterios de ensayos (ECC) y c) la escala de evaluación de ensayos (ESAS). La escala de evaluación basada en la impresión general demostró ser muy poco efectiva para generar validez entre evaluadores. ECC y ESAS son escalas que lograron generar mayor fiabilidad, pero siempre hubo casos de discrepancia entre evaluadores. Según el autor, esta discrepancia se suscita o emerge del tiempo dedicado por cada evaluador a la corrección de cada trabajo. También, sostiene que el entrenamiento grupal y los encuentros de evaluadores para aunar criterios puede ser una buena opción para incrementar la confiabilidad.

Otro antecedente importante brinda Azun (2019), quien se concentró en analizar los factores que afectan el proceso de evaluación de ensayos de inglés como lengua extranjera en una universidad de Ankara. Los datos fueron recolectados a través de cuestionarios, protocolos de pensamiento manifestado y entrevistas realizados a 15 docentes. En los resultados se destacan los usos de escalas, estilos de calificación, (des)conocimiento de criterios, experiencia docente, capacidad de adaptación al nivel y objetivos institucionales como factores que impactan negativamente en la confiabilidad entre pares. La autora sostiene que es necesario recurrir a una variedad de estrategias para mejorar la confiabilidad, entre las cuales se mencionan las escalas de evaluación con categorías y descriptores claramente definidos, los encuentros de estandarización, los talleres docentes de evaluación y las sesiones de consulta y retroalimentación.

Pero quien sí pone el foco en otro aspecto es Mnur Karadenizli-Çilingir (2019), ya que el autor describe el efecto de las sesiones de estandarización en la confiabilidad. Los 24 docentes de inglés participantes pertenecen a una escuela secundaria de Ankara. En una primera etapa, participaron de sesiones de estandarización, y luego evaluaron una serie de composiciones. Ocho meses después, los mismos docentes evaluaron

https://doi.org/10.46744/bapl.202101.010

las mismas composiciones sin una sesión de estandarización previa. Los resultados indican diferencias significativas entre las evaluaciones hechas por los diferentes evaluadores en una y otra etapa de la investigación. El investigador concluye que las sesiones de estandarización contribuyen a mejorar la confiabilidad entre pares y que deberían llevarse a cabo previamente a cualquier evaluación cualitativa que involucre la subjetividad de los evaluadores.

De este modo, la mayoría de los estudios recientes sobre la evaluación de la escritura académica en inglés indagan sobre las diferencias de criterios de evaluación y sobre la incidencia de los instrumentos de evaluación en el proceso. Mientras que algunos sugieren el entrenamiento de los docentes en el uso de instrumentos de evaluación, pocos estudian el impacto que las sesiones de discusión pueden tener en la confiabilidad entre evaluadores. Es por ello que, en este trabajo, proponemos analizar los datos recolectados en un encuentro durante el cual nueve docentes de inglés explicaron individualmente cómo llegaron a la calificación de cinco ensayos escritos en instancias de evaluación sumativa en una universidad argentina. La explicación individual fue seguida de una puesta en común grupal. Los comentarios analizados tenían que ver con el uso de la lengua, como por ejemplo cuestiones de sintaxis, concordancia, y vocabulario, y también con el contenido y la organización de los ensayos. Este análisis dará lugar a conclusiones relacionadas con el tipo y frecuencia de retroalimentación que brindan los docentes de la cátedra en los distintos ensayos analizados. Finalmente, el trabajo procura, mediante la búsqueda de patrones, realizar una contribución que ayude a que los evaluadores puedan seleccionar, ponderar y unificar los criterios de evaluación en escritura en lengua extranjera para generar un mayor nivel de confiabilidad en los procesos de calificación.

2. Marco conceptual

2.1. La evaluación

El proceso de evaluación en el nivel superior juega un papel crucial, ya que a través de diferentes herramientas se colecta información que luego

https://doi.org/10.46744/bapl.202101.010

se emplea para tomar decisiones relacionadas con el planeamiento, diseño y estrategias de enseñanza. Distintos autores presentan definiciones del concepto de evaluación. Collins y O'Brien (2003) definen a la evaluación como «cualquier método que se use para averiguar qué conocimiento poseen los estudiantes en un momento dado» (p. 29). Crooks (2001) se refiere a la evaluación como «cualquier proceso que nos brinde información sobre los logros y el progreso de los estudiantes». Por su parte, Allen (2004), Erwin (1991) y Huba y Freed (2000) se refieren al acto de evaluar como un proceso sistemático que propone colectar, seleccionar, describir, analizar e interpretar información recolectada por medio de diversas herramientas de evaluación para inferir qué es lo que los estudiantes saben, comprenden y pueden realizar con su conocimiento luego de haber transitado una experiencia de aprendizaje. La información obtenida sirve para luego tomar decisiones que pueden afectar diferentes aspectos del proceso de enseñanza y aprendizaje, tales como la selección de materiales didácticos y metodología de enseñanza, y la confección de programas y cronogramas, entre otros. Generalmente, este es un proceso que suele estar a cargo de un plantel docente que regularmente emplea diferentes instrumentos evaluativos como exámenes, tareas, reportes, proyectos, presentaciones orales u otras actividades.

Hyland (2002, 2003) sostiene que los puntajes y la retroalimentación de las distintas situaciones evaluativas impactan sobre el proceso de aprendizaje de los estudiantes y el potencial desarrollo de la habilidad escrituraria. Por esto, se postula que es importante desarrollar procedimientos de evaluación que sean efectivos para asegurarse que el proceso de enseñanza tenga el impacto deseado y que el alumnado sea evaluado de manera justa. Diversos instrumentos de evaluación permiten obtener información sobre el conocimiento alcanzado por los estudiantes y pautar objetivos de aprendizaje. Es posible aseverar que la evaluación puede motivar a los alumnos cuando se sienten orgullosos de sus logros. La evaluación también actúa como una guía orientadora sobre la selección de contenidos de enseñanza mientras que permite evaluar la eficacia de los métodos, tareas y materiales que se emplean. Se puede ver que el proceso de evaluación tiene objetivos pedagógicos, ya que sus resultados tienen un impacto directo sobre el proceso de enseñanza.

https://doi.org/10.46744/bapl.202101.010

Según los objetivos que persigamos, podemos distinguir entre dos tipos de evaluación: sumativa (evaluación *del* aprendizaje) y formativa (evaluación *para* el aprendizaje). La evaluación sumativa se materializa al cierre de un ciclo o curso y se emplea para recabar información sobre el contenido aprendido por los estudiantes y la efectividad del proceso de enseñanza y aprendizaje con respecto de los objetivos planteados. Este tipo de evaluación suele usarse para determinar si el alumno ha aprobado o no el curso y suele hacerse a través de exámenes, pruebas, proyectos o tareas. Su resultado se expresa generalmente en una nota (Bloom *et al*., 1971; Black y Wiliam, 2009; Crooks, 2001; Shepard, 2005).

A su vez, la evaluación formativa se lleva a cabo durante el periodo de cursado y su objetivo es fundamentalmente obtener retroalimentación sobre el proceso de enseñanza y aprendizaje para poder identificar sus fortalezas y debilidades (Collins y O'Brien, 2003; O'Malley y Pierce, 1996). Este es un proceso continuo a través del cual se recolecta información sobre el desempeño de los alumnos empleando múltiples instrumentos de evaluación para intervenir sobre la base de los resultados. Además de medir el grado de proficiencia de los estudiantes, evalúa también el progreso del estudiante, la efectividad del curso, o permite identificar problemas para sugerir soluciones. La evaluación formativa se alinea con los ciertos preceptos del aprendizaje constructivista (Vygotsky, 1962; Bruner, 1986, 1990), ya que promueve la interacción entre docentes y alumnos al mismo tiempo que estimula a los alumnos a adoptar un rol activo en la construcción de significado dentro de su contexto educativo. La información obtenida por medio de este tipo de evaluación permite diagnosticar y guiar dicho proceso, en el que es frecuente el uso de listas de verificación, entrevistas, auto-evaluación, escalas de evaluación, observación participante, entre otros (Black, 2013; Gipps, 1994; Guskey, 2003).

En este trabajo nos focalizaremos en la evaluación sumativa dado que se trabajaran con instancias de evaluación de exámenes parciales y finales. En ambos casos los alumnos son calificados al final de un proceso de enseñanza y aprendizaje para determinar si han cumplido con los

https://doi.org/10.46744/bapl.202101.010

objetivos de cada etapa. Ambas instancias evaluativas se consideran directas de la escritura porque se evalúa la producción textual individual de cada alumno, y no se emplea una evaluación indirecta que recurre a ejercicios de componentes lingüísticos tales como vocabulario y gramática (Williams, 2005). La evaluación directa suele ser considerada como la forma más efectiva para lograr inferir el nivel de habilidad escrituraria alcanzada por un estudiante (Hamp-Lyons, 2001; Williams, 2005), principalmente por tratarse de una instancia real de escritura, en la que interactúan escritores y evaluadores; es decir, el componente humano es central en este tipo de evaluación (Hamp-Lyons, 2001).

2.2. La retroalimentación

La enseñanza de la escritura destaca la importancia de generar, formular y reformular las ideas (Bromley, 2003; Hayes, 2004), por lo tanto, la retroalimentación es una instancia fundamental en el desarrollo de la escritura (Hyland, 2003). Según Hyland (2003), el estudiante logra adquirir fluidez en la escritura cuando transita las diferentes etapas del proceso al escribir un borrador, editarlo, y seguir haciendo cambios basados en la retroalimentación recibida ya sea por sus docentes o inclusive por sus pares. Es así que la retroalimentación se convierte en una herramienta de gran importancia para el docente. La habilidad de escritura, por lo tanto, es considerada como un proceso en el que existe una constante construcción de significado, y el cual incluye dichas etapas de planificación, escritura, revisión y reescritura.

De esta manera, en cada etapa de este proceso se pueden dar distintas instancias de retroalimentación para que así los escritores puedan reflexionar sobre sus trabajos (Hyland, 2003; Williams, 2005), por lo que ciertas etapas —como la revisión y reelaboración del borrador— resultan de vital importancia. Es así que la retroalimentación cobra un rol preponderante en el aprendizaje (Cohen y Cavalcanti, 1990; Ferris, 2002) y, en especial, durante todas las fases del proceso escriturario. De este modo, la retroalimentación resulta formativa por la posibilidad de trabajar y modificar los distintos borradores, y no solamente sumativa que considera la versión final del texto escrito.

https://doi.org/10.46744/bapl.202101.010

Este tipo de retroalimentación apunta a justificar el puntaje que se asigna a un texto determinado así como también a proponer puntos de mejora en futuras producciones. Autores como Mandel *et al.* (2003) consideran que la retroalimentación desde un enfoque formativo e interactivo es más recomendable y efectiva. En este caso, el rol del docente es más activo que cuando se trata de una retroalimentación que solo apunta a la corrección y evaluación. De acuerdo a Hyland y Hyland (2006), este tipo de retroalimentación formativa sobre aspectos de forma y contenido, intenta «promover el desarrollo de la escritura y se juzga crucial en mejorar y consolidar el aprendizaje» (p. 177). Un enfoque formativo apunta también a estimular el desarrollo de la habilidad de escritura de los estudiantes en sus producciones futuras. El interrogante que se plantean los investigadores es quién debe proporcionar dicha retroalimentación —si docentes, pares, u otros— y cuál debería ser el foco de atención —si la lengua o el contenido y la organización.

Fathman y Whalley (1990) demuestran en su investigación que los textos exhibieron mejoras cuando los estudiantes recibieron retroalimentación sobre aspectos relacionados al uso de la lengua y también a la organización y contenido de los textos. Resulta necesario que el docente atienda todos los aspectos del texto para así no enfocarse solo en uno y descuidar otros igual de relevantes. Asimismo, los docentes deben lograr un equilibrio entre comentarios negativos y positivos en la retroalimentación que proporcionan para poder motivar a los estudiantes a mejorar su producción escrita, no solamente deben indicar errores. Según Hyland y Hyland (2006), los estudiantes se benefician de los comentarios positivos, pero también necesitan críticas, siempre que sean constructivas, para poder entender la naturaleza de los errores y así poder corregirlos. Por lo tanto, la retroalimentación no debería implicar la mera detección de errores, sino el puntapié para poder corregir el error y mejorar el texto escrito.

En cuanto a las formas de ofrecer retroalimentación, existen distintas alternativas para ello. Ellis (2009) establece modalidades tales como la retroalimentación directa, indirecta y metalingüística o indirecta explícita, que utiliza un código que enumera y describe el tipo de error a corregir, estableciendo categorías que hacen referencia a cada tipo de error, ya sean

https://doi.org/10.46744/bapl.202101.010

de lengua o de contenido y organización. Autores como Ferris (2003) indican que la retroalimentación puede ser también iniciada desde los pares o también se puede proporcionar retroalimentación electrónica. La retroalimentación puede brindarse a través de comentarios al margen o comentarios al final del texto. En este último caso, los comentarios tienden a considerarse más una respuesta al trabajo de los estudiantes que un tipo de evaluación. Si bien ambas metodologías —tanto comentarios al margen como al final del texto— tienen ventajas y desventajas, autores como Ferris enfatizan la utilidad de los comentarios al final. Esto permite al docente desarrollar comentarios más completos, detallados, y también ofrecer una mirada más holística del texto. Sin embargo, los comentarios marginales también tienen la ventaja de proporcionar un *feedback* más inmediato. Es así que un equilibrio entre ambas metodologías proporciona una retroalimentación más completa para que el estudiante pueda desarrollar la habilidad escrituraria. A través de este proceso de escritura y reescritura, los estudiantes pueden editar sus textos teniendo en cuenta criterios tales como estructura, estilo, vocabulario, etc., los cuales están plasmados en los comentarios.

Dentro de nuestro contexto educativo, es decir, desde la cátedra de Lengua Inglesa II de las carreras de Profesorado, Traductorado y Licenciatura de la Facultad de Lenguas de la Universidad Nacional de Córdoba, la metodología utilizada para brindar retroalimentación es indirecta explícita, por medio del uso de un código que describe el tipo de error, que puede ser de lengua, de contenido o de organización, o cualquier otro aspecto relacionado a la redacción. Durante el dictado de la materia, y como parte del proceso de aprendizaje de la habilidad escrituraria, los estudiantes reciben retroalimentación de manera electrónica, por medio de la función comentario de Microsoft Word. En una primera instancia, se presenta a los estudiantes el código que será utilizado para brindar retroalimentación. Dicho código se divide en aspectos de uso de la lengua (categoría que incluye concordancia, preposiciones, ortografía, sintaxis, entre otros) y aspectos relacionados con el contenido y la organización, por ejemplo, ideas irrelevantes o innecesarias, problemas en el desarrollo de ideas, ideas poco claras, etc. Los alumnos realizan las actividades de escritura a través del aula virtual en la plataforma *Moodle*. Si bien estas son

https://doi.org/10.46744/bapl.202101.010

instancias de evaluación informal, ya que no llevan una calificación, sirven para que los estudiantes puedan practicar y, de este modo, desarrollar la habilidad de escritura previamente a la evaluación formal. Esta consiste en dos parciales y un examen final donde se asigna gran valor a la sección de redacción, particularmente en el examen final, ya que es de carácter eliminatorio.

3. Metodología

3.1. Participantes

Los participantes de este trabajo fueron nueve docentes de la cátedra de Lengua Inglesa II que se encargan de la evaluación de ensayos durante las instancias de evaluación formativa escrita, presentes al final de cada semestre, y la instancia de evaluación sumativa final, que se necesita pasar para aprobar el curso.

3.2. Procedimientos y materiales

Los nueve docentes que participaron en esta investigación calificaron individualmente 5 ensayos que alumnos de la cátedra habían escrito durante un proceso de evaluación sumativa final, los que fueron seleccionados por una persona no participante del comité evaluador. Esta persona escogió aleatoriamente tres ensayos que habían sido aprobados y dos desaprobados, luego cada docente recibió una copia de los cinco ensayos, pero sin revelar la nota final que se les había otorgado. Se les pidió a los docentes que corrigiesen los ensayos de manera individual, simulando las condiciones de una situación de examen, es decir, dedicando alrededor de quince minutos a la corrección de cada trabajo.

Luego que los docentes corrigieron, se realizó la reunión donde explicaron a sus pares cómo cada uno había arribado a la calificación de cada ensayo y se dispuso una puesta en común grupal. Toda esa interacción fue grabada, y luego transcrita para su posterior análisis. Sobre cada uno de los cinco ensayos se registró lo siguiente: el puntaje asignado por cada evaluador, número de docentes que lo consideraron aprobado, número

https://doi.org/10.46744/bapl.202101.010

de docentes que lo consideraron desaprobado y justificación del puntaje (comentarios positivos y comentarios negativos).

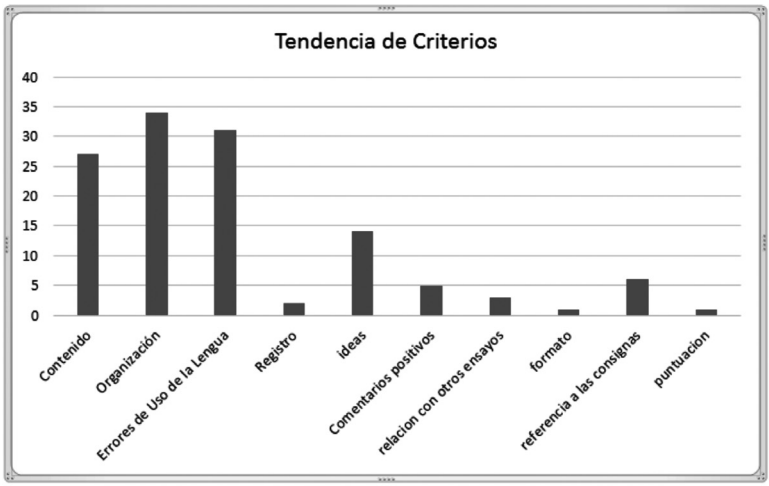
4. Resultados

En esta sección se analizarán los datos recogidos durante la sesión en la que nueve docentes de inglés como lengua extranjera a nivel universitario expusieron las razones por las que otorgarían una calificación aprobatoria o desaprobatoria a un trabajo. La dinámica que se empleó fue de turno por turno, por lo que cada docente presentó una explicación y su punto de vista de manera ininterrumpida. En el proceso de recolección y análisis de datos se registraron 124 comentarios a través de los cuales las docentes describieron qué ensayos se encuentran desaprobados o aprobados y su justificación. Estos datos fueron volcados en Excel para facilitar su análisis.

El análisis de los comentarios realizados nos permite conocer en mayor detalle cuáles son los criterios que prioriza cada docente para ponderar los ensayos. Sin bien se presupone que el criterio de evaluación de escritura académica es compartido por todos los docentes de una misma cátedra, el análisis de los datos revela que no todos los docentes le otorgan el mismo peso a cada uno de los criterios. En la tabla 1 se evidencia que mientras hay criterios tales como organización (34), contenido (27) y desarrollo de ideas (31) que son tenidos en cuenta por muchos docentes a la hora de fundamentar su evaluación, hay otros tales como registro (2) y puntación (1) que tan solo algunos docentes traen a colación. Además, de esta reunión, también se desprende que hay circunstancias contextuales que también tienen incidencia en la calificación final. En tal sentido, algunos docentes manifestaron que ciertas notas habían sido determinadas por los criterios de evaluación y al mismo tiempo habían sido influenciadas por el o los ensayos que habían sido leídos y evaluados previamente. Es decir, los docentes manifestaron que comparaban la calidad (en términos de uso de la lengua, estructura y organización) de los ensayos corregidos con anterioridad en una misma instancia de evaluación para determinar la calificación final de algunas composiciones.

https://doi.org/10.46744/bapl.202101.010

Tabla 1
Tendencia de criterios



De acuerdo a lo que se observa en este cuadro, existe un sesgo entre los docentes sobre ciertos aspectos de los ensayos evaluados. Un análisis más detallado de lo que cada docente mencionó como justificación nos permite ratificar que el contenido y la organización son los aspectos más mencionados en casi todos los ensayos y por la mayoría de los docentes. Al mismo tiempo, los datos en la tabla 2 parecen indicar que hay docentes que realizan sus ponderaciones basándose sobre todo en tres criterios: organización, contenido y errores de lengua (evaluador 2, 9), mientras que hay otros docentes que consideran una mayor gama de aspectos: organización, contenido, ideas, errores de lengua. Este docente también manifiesta que la nota de uno de los ensayos puede haber estado influenciada por un ensayo corregido con anterioridad (evaluador 1), lo que puede dar evidencia de una mayor capacidad de reflexión sobre los múltiples factores que se consideran durante el complejo proceso de evaluación de la escritura.

Por otra parte, los datos revelan que la subjetividad es inherente a todo proceso de evaluación. Esto se evidencia en las palabras de la evaluadora

https://doi.org/10.46744/bapl.202101.010

4 que compara sus ponderaciones con la de sus colegas y dice «hay algunas coincidencias, pero quizá fui más exigente». Otra lectura que nos permite realizar el análisis de los comentarios de los docentes es que, de los nueve docentes participantes, solo cuatro realizaron comentarios positivos sobre algunos de los ensayos evaluados. Por ejemplo, la evaluadora 2, al referirse al ensayo 1, el cual era una composición cuyas características abrían la posibilidad a dudas respecto de si debía ser aprobada o reprobada, expresó «está bien enfocado, se podría premiar eso». En futuras investigaciones sería necesario indagar en profundidad la razón por la cual más de la mitad de los docentes no realizó comentarios positivos, y si esto es una tendencia que se repite, sobre todo en instancias de evaluación sumativa. Una de las razones mencionadas por una de las participantes en la reunión fue que durante las instancias evaluativas cada docente debe corregir varios ensayos en un corto tiempo, por lo que el foco está puesto en los resultados. Los comentarios positivos parecen estar más asociados solo a instancias de evaluación formativa, lo que demuestra también que hay una disociación entre el proceso de desarrollo de la habilidad escrituraria y las instancias de evaluación. Las evaluaciones parecen no ser consideradas como parte del proceso de aprendizaje.

Tabla 2

Frecuencia de criterios observados por cada docente en los cinco ensayos evaluados

Evaluador	Contenido	Organización	Uso de la lengua	Registro	Ideas	Comentarios positivos	Relación con otros ensayos	Formato	Referencia a las consignas	Puntuación
1	3	3	3		2	1	1	1		
2		3	3		2					
3	3	5	5		2	2			2	
4	4	4	1		4				2	
5	4	3	4	2			1		1	
6	2	4	5		2	1			1	
7	1	4	4			1				1
8	5	3	2		1					
9	5	5	4		1		1			

https://doi.org/10.46744/bapl.202101.010

Estas diferencias en las ponderaciones de los diferentes criterios de evaluación están reflejadas en comentarios docentes como «si bien el ensayo presentaba errores, no eran graves». Idealmente, el proceso de evaluación debería minimizar niveles de subjetividad por parte de los docentes, a través de la búsqueda de diversos mecanismos, como el uso de escalas de evaluación y sesiones de estandarización para reducir el impacto de la subjetividad y así lograr una mayor confiabilidad entre los docentes responsables. Por esta falta de uniformidad en la aplicación de criterios, en la tabla 3 se puede advertir que hubo docentes con más tendencia a desaprobar (evaluador 3) o a aprobar (evaluador 7) algunos ensayos más que otros.

Tabla 3

Ponderaciones por ensayo según evaluador

	Ensayo 1	Ensayo 2	Ensayo 3	Ensayo 4	Ensayo 5
Evaluador 1	Reprueba	Reprueba	Aprueba	Aprueba	Duda
Evaluador 2	Aprueba	Reprueba	Aprueba	Aprueba	Reprueba
Evaluador 3	Aprueba	Reprueba	Aprueba	Reprueba	Reprueba
Evaluador 4	Reprueba	Reprueba	Aprueba	Reprueba	Reprueba
Evaluador 5	Reprueba	Duda	Aprueba	Reprueba	Aprueba
Evaluador 6	Duda	Reprueba	Aprueba	Reprueba	Aprueba
Evaluador 7	Aprueba	Reprueba	Aprueba	Aprueba	Aprueba
Evaluador 8	Reprueba	Reprueba	Duda	Reprueba	Duda
Evaluador 9	Reprueba	Reprueba	Aprueba	Duda	Reprueba

4.1. Análisis individual de los ensayos evaluados

A continuación, se presenta una breve descripción de la evaluación realizada por los docentes en cada ensayo, la cual se encuentra ordenada teniendo en cuenta los criterios de evaluación sobre los que los docentes lograron de un mayor a un menor grado de acuerdo. Al analizar la evaluación de cada uno se puede concluir que, aparentemente, se logró un alto grado de confiabilidad en aquellos casos en los que aspectos clave como organización, desarrollo de ideas y el uso de la lengua fueron evaluados positivamente. Por ejemplo, en el ensayo 3 se advierte un alto grado de coincidencia entre

https://doi.org/10.46744/bapl.202101.010

las docentes, ya que todas aprobaron el ensayo (100 %) y la mayoría de sus comentarios hacen referencia a aspectos relacionados con el contenido y la organización, por lo que se desprende que estos dos aspectos tienen un alto nivel de incidencia al momento de decidir la calificación de un alumno. Sin embargo, existen algunas inconsistencias o falta de coincidencia cuando se analiza en detalle qué llevó a cada evaluadora a tomar esa decisión: por un lado, algunas docentes expresaron que el ensayo presenta una organización clara, con una oración tópica bien planteada, y con buen desarrollo de ideas; por otro lado, algunas docentes comentaron que ciertas ideas no se entendían completamente, que el desarrollo de las ideas era pobre y vago, o que varias de las soluciones propuestas no eran claras. En uno de los casos, una de las docentes mencionó que tal vez sus comentarios se debían a que se sentía condicionada por el ensayo que corrigió anteriormente, el cual no había sido aprobado. De este modo, se puede apreciar que el concepto de claridad está fuertemente influenciado por factores subjetivos, tales como la lectura de otros trabajos similares.

Por su parte, en relación con los errores de lengua, el análisis muestra que las docentes mencionaron la presencia de algunos problemas —como concordancia o sintaxis—, pero también destacaron el uso de vocabulario específico. De lo expresado por las docentes, los errores relacionados con el uso de la gramática y la sintaxis poseen un peso mayor que los errores de léxico.

El ensayo 2 tampoco presentó discrepancias, ya que todas las docentes (100 %) le asignaron un puntaje inferior a 24, nota mínima con la que se aprueba esta sección. Ocho docentes desaprobaron el ensayo en una primera instancia (89 %), mientras que una le asignó inicialmente un signo de pregunta (11 %) que luego definió como desaprobado. Entre las razones expresadas para no aprobar el ensayo, todas las docentes hicieron referencia a aspectos relacionados con el contenido y la organización, tales como la falta de profundidad o relación en el desarrollo de ideas, el enfoque incorrecto del contenido, problemas con la oración tópica, entre otros. En cuanto a errores de lengua, solo algunas de las docentes mencionaron la presencia de errores elementales. Esta tendencia de las docentes a focalizar su atención principalmente en el desarrollo de ideas y organización de

https://doi.org/10.46744/bapl.202101.010

contenido evidencia la presencia de un sesgo hacia dichas categorías, las cuales son analizadas con mayor minuciosidad y severidad.

El ensayo 5 fue aprobado por 7 evaluadoras (78 %) y desaprobado en dos casos (22 %). Los comentarios de quienes lo consideraron desaprobado son en su mayoría relacionados a contenido y organización (problemas de foco, desarrollo de ideas y de distribución). Asimismo, otro comentario que se repitió fue que los párrafos eran muy cortos y la conclusión presentaba una idea nueva. En cuanto al uso de la lengua, solo se indicaron algunos errores. Es importante mencionar que los comentarios positivos frente a este ensayo desaprobado son escasos. Como se ha mencionado, para futuras investigaciones, sería interesante explorar qué factores influyen en este comportamiento. Los evaluadores que lo consideraron aprobado realizaron comentarios sobre contenido y organización, la falta de ideas, las generalizaciones, problemas de foco y de contenido pobre. Asimismo, en algunas oportunidades, las docentes hicieron comentarios positivos sobre el buen desarrollo de ideas, buena organización y buena introducción. La clara diferencia de criterio que parece ser de suma importancia para los miembros de esta cátedra puede deberse a diferentes factores tales como la antigüedad del docente en la cátedra o su severidad para corregir un determinado aspecto. De este hecho, se desprende la necesidad de pensar en reuniones de estandarización para disminuir diferencias entre evaluadores, aumentar la confiabilidad entre los mismos y garantizar un proceso evaluativo justo. En cuanto al uso de la lengua, se destacaron algunos errores de sintaxis y preposiciones, pero no representaron errores serios. Esto parece indicar que, si bien los errores relacionados con el uso de la lengua son parte de la retroalimentación que el alumno recibe del docente, no parecen ser ponderados con tanta severidad como para inclinar la balanza de un resultado aprobado a uno desaprobado.

El ensayo 1 fue aprobado por tres docentes (33 %), desaprobado por cinco (56 %) y en un caso presenta signo de pregunta (11 %). En cuanto a las razones para desaprobar el ensayo, en su mayoría hacen referencia al contenido y organización: se menciona la falta de justificación y detalle, el desarrollo pobre de ideas y de contenido, la generalidad y poca claridad de las ideas, e inclusive la falta de foco. Las docentes también indicaron

https://doi.org/10.46744/bapl.202101.010

la presencia de errores de lengua y de registro. En este caso, es la primera vez que se menciona el registro de escritura empleado por el alumno entre los criterios para desaprobar un ensayo. Por su parte, los comentarios en el caso de las evaluadoras que lo aprobaron (33 %) fueron en su mayoría positivos y hacen referencia a la buena organización y foco, así como también a la buena introducción y desarrollo (*body paragraphs)*; mientras que en el caso de la docente que le asignó un signo de pregunta (11 %), los comentarios indican que las ideas no eran muy claras, que la conclusión debería invertir el orden y que la tesis era muy general. Asimismo, esta última docente destacó errores de lengua.

El ensayo 4, al igual que el ensayo 1, fue aprobado en tres oportunidades (33 %), desaprobado en 5 (56 %) y un caso presenta signo de pregunta (11 %). Las razones para desaprobar el ensayo hacen hincapié en el contenido y la organización, como así también se menciona aspectos relacionados al desarrollo de ideas, la falta de mención de temas indicados en la consigna, fragmentos largos, ideas fuera de foco, la copia textual de las instrucciones y la falta de detalles en el desarrollo. En cuanto al uso de la lengua, solo se mencionaron algunos errores. Los evaluadores que lo consideraron aprobado hicieron comentarios similares a los de los ensayos desaprobados. Señalaron errores de foco y desarrollo, así como también errores de formato, por ejemplo, el número de palabras al final del ensayo. En este caso también surgen dudas de aspectos no explícitos en el criterio de evaluación. Esto se evidencia en la siguiente manifestación «bajé puntos por una conclusión poco desarrollada. ¿Corresponde un desaprobado? Pregunto porque no está especificado en los criterios. Estrictamente no es una conclusión bien hecha, pero lo colocaría como error de desarrollo». Como en el análisis de la corrección de los ensayos previos, puede observarse que tanto el desarrollo como el foco juegan un papel preponderante en la asignación de una calificación definitiva. Al mismo tiempo, se destaca que por primera vez se menciona el criterio de formato. Esto puede darse por dos motivos. Por un lado, puede deberse a que es el único de los cinco ensayos que presenta este problema Por otro lado, puede ser a causa de que cuando los docentes deciden otorgar una nota que implica un aplazo, tienden a brindar retroalimentación más extensa y detallada para argumentar la nota desaprobada e indicar que el proceso evaluativo ha

https://doi.org/10.46744/bapl.202101.010

sido justo, por lo cual también destacan errores de menor relevancia, como las cuestiones de formato. En cuanto al uso de la lengua, los comentarios hacen referencia a errores de expresión, aunque no se registraron muchos problemas en esta área. La evaluadora que presenta signo de pregunta también tuvo comentarios, en su mayoría relacionados con el contenido y la organización (principalmente problemas con la tesis), o la necesidad de un desarrollo más profundo y específico del tema en cuestión. Con respecto al uso de la lengua, no presentó inconvenientes.

De esta manera, y a modo de síntesis, puede establecerse que, de lo discutido en la reunión, resulta claro que hay tres aspectos que son los más determinantes en el momento de ponderar un ensayo: la organización, el desarrollo de ideas y el uso de la lengua. Sin embargo, a partir de esto último se abrió un rico debate donde una docente puntualizó que cada integrante de la cátedra le da «un peso diferente al contenido y la organización» (evaluador 2), lo que se materializa en diferentes ponderaciones, y esto a su vez impacta en el nivel de confiabilidad entre evaluadores. Mientras tanto, otra docente estableció que «es importante determinar qué peso va a tener el contenido para saber si la falta o la pobreza del mismo implica un desaprobado o descontar puntos» (evaluador 1) y, a su vez, la evaluadora 9 manifestó su interés en considerar el contenido como un criterio de aprobación o desaprobación para no focalizar solamente en aquellos respecto de la forma. Su argumento se basa en que uno de los objetivos de la materia es desarrollar el pensamiento crítico, por lo que es necesario ser precisos en la evaluación del contenido, ya que «es en donde ellos muestran si han hecho la conexiones» (evaluador 9). Por otro lado, se halló un alto grado de coincidencia en la valoración de aspectos formales, como la concordancia de *sujeto-verbo*.

Al cierre de la reunión, y a modo de procurar mayor claridad en los criterios, una de las evaluadoras planteó la necesidad de «ver qué valor le damos a cada criterio como en bandas» (evaluadora 5). De esta manera, avanzando en esa dirección, se reduciría el nivel de subjetividad que incide en el proceso evaluativo y se ganaría mayor confiabilidad entre evaluadores. Asimismo, se planteó la posibilidad de elaborar un instrumento de evaluación particular al contexto educativo de la cátedra, como una

https://doi.org/10.46744/bapl.202101.010

escala analítica, en el que se establezca cuáles son los estándares aceptables en cada aspecto de la escritura a evaluar para que un alumno apruebe una instancia de evaluación sumativa.

5. Discusión y conclusiones

En el presente trabajo, se analizaron los comentarios realizados por docentes universitarios de inglés durante una sesión de intercambio sobre los criterios de evaluación utilizados para calificar cinco ensayos académicos que habían sido evaluados individualmente por cada uno de ellos simulando una instancia de evaluación sumativa. Durante el encuentro, cada participante socializó la calificación decidida para cada composición y esgrimió las razones por las cuáles se tomó la decisión. La exposición individual fue seguida de una puesta en común.

En las explicaciones brindadas sobre todos los ensayos, tanto aprobados como desaprobados, se observa una preponderancia significativa de comentarios sobre la organización y el contenido. Estos datos coinciden con los resultados del análisis realizado por Kuiken y Vedder (2014), ya que los participantes de su estudio reconocieron asignar mayor valor a los aspectos relacionados con la adecuación comunicativa (el contenido, los argumentos y la organización retórica del ensayo) que el uso de la lengua (léxico, precisión, ortografía y gramática). Al mismo tiempo, se observa una gran cantidad de comentarios negativos y escasez o incluso ausencia, en ciertos casos, de comentarios positivos. Esto indicaría una falta de equilibro entre comentarios positivos y negativos que puede tener un impacto dañino en la motivación de los estudiantes por mejorar su habilidad escrituraria. Aunque, como bien apunta Hyland y Hyland (2006), los estudiantes necesitan críticas constructivas para poder entender la naturaleza de los errores y poder corregirlos, también necesitan y se benefician de los comentarios positivos sobre sus producciones. Esto resalta el rol formativo e interactivo de la retroalimentación, incluso en instancias de evaluación sumativa (Morrow y otros, 2003).

Es importante mencionar que, en las justificaciones presentadas para aprobar ciertos ensayos, no hay comentarios positivos sobre el uso

https://doi.org/10.46744/bapl.202101.010

de la lengua de los alumnos. Se puede observar un cierto alineamiento entre estos resultados y los arrojados por el estudio de Schaefer (2008), ya que en este último se detectó una tendencia de los evaluadores a ser excesivamente estrictos con alumnos que tienen un mayor dominio de la lengua; además, en su estudio de 2016, también aseveró que algunos evaluadores ubicaron el ensayo de un alumno en una banda baja simplemente por su uso del lenguaje, sin considerar todos los descriptores de coherencia, o contenido, en una banda. Sin embargo, como se mencionó anteriormente, es fundamental incluir valoraciones sobre los aspectos positivos de las producciones de los estudiantes para estimular la motivación y el deseo de progreso. Eckes (2012), citando los resultados de Schaefer (2008), establece que la variable a considerar es la importancia de los criterios de evaluación. De esta manera, aquellos criterios a los que se asigna mayor importancia están asociados a un sesgo más severo; mientras que aquellos a los que se otorga menor importancia se vinculan con un punto de vista más indulgente.

En el caso de los ensayos que recibieron puntajes aprobados por algunos evaluadores y desaprobados por otros, hubo cierto grado de coincidencia en los comentarios hechos tanto sobre el uso de la lengua como sobre el contenido y la organización, por lo que se infiere que las diferencias en puntaje se deben a diferencias de criterio en la ponderación de los diversos errores. Una instancia de discusión y negociación entre evaluadores, en donde se analicen los criterios de evaluación, se los conceptualice y se acuerde una interpretación común de los mismo, podría contribuir a mejorar la confiabilidad entre pares, como sugieren *Trace et al.* (2016). Este debate debe darse en relación a un instrumento de evaluación claramente definido, como una escala analítica, puesto que, como Kayapinar (2014) indica las meras impresiones de los evaluadores no conducen a generar un proceso evaluador de alta validez. En este sentido, se puede sugerir la adopción del uso de una escala para calificar los ensayos académicos y la organización de sesiones de estandarización previas a las instancias de evaluación para incrementar la confiabilidad entre evaluadores, tal como propone Nur Karadenïzlï-çïlïngïr (2019) a partir de los resultados de su estudio.

https://doi.org/10.46744/bapl.202101.010

Por otra parte, lo que parece determinar que algunos ensayos tengan un signo de pregunta son cuestiones relacionadas con el contenido y la organización, mas no con el nivel de lengua. La preponderancia de comentarios sobre contenido y organización coincide con lo expresado por docentes —tanto en cuestionarios como en entrevistas— a la hora de señalar los aspectos más importantes durante la evaluación de la escritura. La tendencia a enfocarse en el error o cuestiones a mejorar y casi no hacer comentarios positivos coincide con lo expresado por los docentes y con las notas que se observó al margen de los ensayos evaluados. Como bien lo indica Azun (2019), la planificación sistemática de sesiones de estandarización y los encuentros entre colegas para intercambiar experiencias sobre las prácticas de evaluación pueden contribuir a aunar criterios, a interpretar de la misma manera las escalas utilizadas y a revalorizar el error como elemento natural y necesario en los procesos de aprendizaje de una habilidad, para nuestro caso, la escritura en inglés como lengua extranjera.

Finalmente, en este artículo se analizaron los comentarios realizados por evaluadores durante un encuentro en el cual se debatió cómo cada uno de ellos llegó a la calificación final de cada ensayo. Esta puesta en común contribuyó a entender qué criterios de evaluación fueron tenidos en cuenta por cada participante a la hora de brindar retroalimentación sobre una composición académica en inglés. Este proceso constituyó un primer paso en la desafiante tarea que implica la unificación de criterios que permitan lograr una mayor confiabilidad entre evaluadores. Los resultados de esta etapa de la investigación nos permiten proponer implicancias pedagógicas que en etapas subsiguientes deberán ser estudiadas de manera sistematizada para poder determinar su impacto. A partir de los resultados, se sugiere la incorporación periódica de sesiones de discusión, entre los docentes de una misma cátedra, sobre los criterios de evaluación de la habilidad escrituraria para consensuar prácticas. El uso de un instrumento de evaluación, ya sea una escala de evaluación analítica o holística, puede también ser incorporado. La sesiones periódicas de estandarización de criterios pueden tener lugar previo a instancias de evaluación sumativa. Estas prácticas podrían contribuir a reducir la subjetividad a la hora de asignar una calificación a un ensayo y, en consecuencia, a aumentar la confiabilidad entre pares, lo cual ayudará a mejorar la transparencia y la validez del proceso de

https://doi.org/10.46744/bapl.202101.010

evaluación. El objetivo de la confiabilidad sigue siendo un norte en el área de la evaluación de una segunda lengua y adquiere aún más relevancia en el contexto universitario, en donde los resultados de las evaluaciones tienen un impacto directo en la vida profesional y personal de los estudiantes.

https://doi.org/10.46744/bapl.202101.010

REFERENCIAS BIBLIOGRÁFICAS

Allen, M. J. (2004). *Assessing academic programs in higher education.* Anker Publishing Company, Inc.

Azun, F. M. (2019). *An Analytic Approach to English Language Instructors' Scoring Differences Of Writing Exams* [Tesis de maestría no publicada]. Graduate School of Educational Sciences. Hacettepe University.

Black, P. (2013). Pedagogy in theory and in practice: Formative and summative assessments in classrooms and in systems. En D. Corrigan, R. Gunstone, y A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy* (pp. 207-229). Springer.

Black, P., y Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Bloom, B. S., Hastings J. T., y Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. McGraw-Hill Book Co.

Bromley, K. (2003). Building a sound writing program. En L. Mandel Morrow, L. Grambell, y M. Pressley (Comps.), *Best Practices in Literacy Instruction*. The Guilford Press.

Bruner, J. (1986). *Actual Minds, Possible Worlds.* Harvard University Press.

Bruner, J. (1990). *Acts of Meaning*. Harvard University Press.

Cohen, A. D., y Cavalcanti, M. C. (1990). Feedback on compositions: Teacher and student verbal reports. En B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 155-177). Cambridge University Press.

https://doi.org/10.46744/bapl.202101.010

Collins, J., y O'Brien, N. (2003). *The Greenwood dictionary of education*. Greenwood Press.

Crooks, T. (13-15 de setiembre de 2001). *The validity of formative assessment*. Paper presented to The British Educational Research Association Annual Conference, University of Leeds. http://www.leeds.ac.uk/educol/documents/00001862.htm.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25* (2), 155-185. http://doi.org/10.1177/0265532207086780

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9*(3), 270-292. http://doi.org/10.1080/15434303.2011.649381

Ellis, R (2009). A typology of written corrective feedback types. *ELT Journal*, (63), 97-107. https://doi.org/10.1093/elt/ccn023

Erwin, T. D. (1991). *Assessing student learning and development: A guide to the principles, goals, and methods of determining college outcomes.* Jossey-Bass.

Fathman, A., y Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. En B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178-190). Cambridge University Press.

Ferris, D. R. (2002). *Treatment of Error in Second Language Student Writing.* University of Michigan Press.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment.* Falmer Press.

https://doi.org/10.46744/bapl.202101.010

Guskey, T. (2003). How classroom assessments improve learning. *Educational Leadership, 60* (5), 6-11. https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1008&context=edp_facpub

Hamp-Lyons, L. (2001). Fourth generation writing assessment. En T. Silva, y P. Matsuda (Eds.), *On second language writing.* Lawrence Erlbaum Associates.

Hayes, J. (2004). A new framework for understanding cognition and affect in writing. En R. B. Ruddellm, y N. J. Unrau (Comps.), *Theoretical Models and Processes of Reading* (5ª ed.). International Reading Association.

Huba, M. E., y Freed, J. E. (2000). *Learner-Centered Assessment on College Campuses - Shifting the Focus from Teaching to Learning*. Allyn and Bacon.

Hyland, K. (2002). *Teaching and Researching Writing.* Longman.

Hyland, K. (2003). *Second Language Writing*. Cambridge University Press.

Hyland, K., y Hyland, F. (2006a). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83-101.

Kayapinar, U. (2014). Measuring Essay Assessment: Intra-rater and Inter-rater Reliability. *Eurasian Journal of Educational Research*, *57*, 113- 136.

Kuiken, F., y Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, *31*(3), 279-284.

Mnur Karadenïzlï-çïlïngïr, M. (2019). *The Effect of Standardisation Sessions Conducted before English Language Writing Exams on Inter-rater and Intra-rater Reliability* [Tesis de maestría no publicada]. Graduate School of Social Sciences.

https://doi.org/10.46744/bapl.202101.010

Mandel Morrow, L., Grambell, L., y Pressley, M. (Comps.). (2003). *Best Practices in Literacy Instruction*. The Guilford Press.

O'Malley, J, y Pierce, L. (1996). *Authentic Assessment for English Language Learners: Practical Approaches for Instructions.* Addison-Wesley Publishing Company.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*, 465-493.

Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership*, *63*(3), 66-70. http://www.ascd.org/publications/educational-leadership/nov05/vol63/num03/Linking-Formative-Assessment-to-Scaffolding.aspx

Trace, J., Janssen, G., y Meier, V. (2016). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing*, *34*(1), 3-22.

Vygotsky, L. S. (1962). *Thought and language*. MIT Press.

Williams, J. (2005). *Teaching writing in second and foreign language classrooms*. McGraw-Hill.

Yen, N. T. Q. (2016). Rater Consistency in Rating L2 Learners´ Writing Task. *VNU Journal of Science: Foreing Studies*, *32*(2), 75-84.