Criterios de evaluación: construyendo fiabilidad y objetividad

SCHANDER, Claudia Elizabeth; MASSA, Agustín Abel
Universidad Nacional de Córdoba, Facultad de Lenguas

El presente trabajo se propone mostrar los resultados y conclusiones de un tramo del proyecto de investigación bianual llevado a cabo en el período 2016/17 en la Facultad de Lenguas, UNC que se tituló: Hacia la elaboración de criterios uniformados de evaluación que aseguren fiabilidad entre los calificadores de las pruebas escritas en las cátedras de Lengua Inglesa I, Práctica Gramatical del Inglés, Gramática Inglesa I, Práctica de la Pronunciación del Inglés, Lengua Castellana I y Lectocomprensión en Lengua Extranjera IV (Inglés). El mismo estuvo avalado por Secyt y dirigido por Mgtr. Fabián Negrelli y co dirigido por la Mgtr. Cecilia Ferreras. El proyecto plantea como objetivo general abordar la elaboración de criterios de evaluación uniformados con el objetivo de lograr fiabilidad inter evaluadora en las Cátedras de Lengua Inglesa I, Práctica Gramatical del inglés, Gramática Inglesa I, Práctica de la Pronunciación del inglés, Lengua Castellana I y Lecto comprensión en Lengua Extranjera IV (inglés) en la Facultad de Lenguas de la Universidad Nacional de Córdoba (UNC).

Educación superior – Evaluación - Fiabilidad

Introducción

El sistema educativo universitario requiere más y mejores niveles de educación y enseñanza, y a ello deben contribuir nuestras prácticas evaluativas. Numerosos autores (González, 2005; Anijovich, 2010; Mottier López, 2010; Matute Vázquez & Muriel Gómez, 2014; entre otros) consideran que evaluar es el tema más difícil de la pedagogía, y también es la instancia más difícil de un docente comprometido que debe calificar a un estudiante. Jorba y Sanmartí (1993) sostienen que "Cada vez más se considera que si se quiere cambiar la práctica educativa es necesario cambiar la práctica de evaluación, es decir, su finalidad y el qué y cómo evaluar" (p. 36). Comprender que la evaluación es intrínseca a los procesos de enseñanza y de aprendizaje tiene directa relación con la comprensión de la función pedagógica de la evaluación. En este sentido, es preciso puntualizar que la función pedagógica de la evaluación es actuar como un dispositivo pedagógico que debe regular continuamente los aprendizajes. Para que ello suceda es fundamental que el docente considere de vital importancia dicha regulación, comenzando por la planificación de la clase y culminando con la evaluación de los contenidos. Es precisamente en la etapa de la evaluación cuando surgen preguntas como: ¿cuál es la función pedagógica de la evaluación?; ¿qué tipo de decisiones se toman a partir de los resultados?; ¿cuáles son las consecuencias de la calificación otorgada?; ¿están aseguradas la validez, fiabilidad y objetividad de los instrumentos que se aplican y de las calificaciones que se otorgan?; ¿están los docentes suficientemente capacitados y/o entrenados para corregir las evaluaciones?

Referentes teórico-conceptuales

Numerosos estudios de investigación han argumentado en forma extensa acerca de la fiabilidad de las calificaciones y de la probada influencia negativa que ejerce la subjetividad en la percepción y actitud de los que corrigen (Bachman, 1990; Gass, 1994; Alderson, Clapham & Wall, 1995; Bachman & Palmer, 1996; Bachman & Cohen, 1998; Bachman, 2002; Hughes, 2003; Brown, 2004, Fulcher & Davidson, 2007, Bachman & Palmer, 2010; entre otros); es por ello, que

más allá de las discusiones acerca de las concepciones y de las formas que puedan considerarse más pertinentes o apropiadas para llevar a cabo la acción de evaluar por parte del docente, esa acción tiene una importancia fundamental, ya que en ella se produce el encuentro entre los criterios sostenidos por la institución educativa y por el que "enseña" por un lado, y lo que le es posible mostrar al alumno como adquisición durante un periodo preestablecido, por el otro.

No podemos ignorar el hecho de que es, precisamente, como resultado del proceso de evaluación, que el alumno es promovido a un nivel siguiente en la sucesión en que han sido ordenadas esas adquisiciones, o bien sufre el impedimento para continuar y, en consecuencia, debe volver sobre la etapa anterior para revisar, corregir, practicar aquellas competencias y/o conocimientos que aún no han sido adquiridos para superar la presente etapa y ser promovidos a la siguiente. Es relativamente simple concluir que, al menos parcialmente, el hecho de permanecer, detenerse, avanzar en la carrera al ritmo previsto o a un ritmo menor depende del juicio que el profesor elabore, de acuerdo con determinados criterios, sobre lo que demuestra el alumno en la situación diseñada a los efectos de la evaluación.

Este lugar central que ocupan los momentos de evaluación en relación con la carrera académica de los alumnos justifica un análisis profundo de la situación y un estudio descriptivo que permita garantizar objetividad y consistencia por parte de los evaluadores a la hora de calificar lo que denominaremos genéricamente la "prueba" (remitiéndonos más a una categoría descriptiva de uso común en contextos educativos que a algo conceptualmente definido), y que, al mismo tiempo, garantice imparcialidad en la representación de los intereses de los alumnos. Así, este proyecto madre se encaminó a elaborar criterios uniformados de evaluación que aseguraran la fiabilidad de las calificaciones de las pruebas escritas en las cátedras de Lengua Inglesa I, Práctica Gramatical del Inglés, Gramática Inglesa I, Práctica de la Pronunciación del Inglés, Lengua Castellana I, y Lecto - comprensión en Lengua Extranjera IV (Inglés).

En este sentido, creemos que los criterios de evaluación – entendidos como indicadores para la evaluación de los aprendizajes de los alumnos en los diferentes niveles de concreción curricular - constituyen el referente fundamental para determinar el grado de consecución de los objetivos generales de la asignatura que hayan alcanzado los alumnos. Dichos criterios deben ser considerados, entonces, como puntos de referencia que hacen posible la calificación de lo que nos proponemos evaluar; en otras palabras, deben ser referentes de valor argumentados que nos ayuden a conocer en qué medida un sujeto alcanza el dominio de cada área.

Claro está que en el entramado de todo lo que se ha expuesto anteriormente se puede entrever un conjunto de preguntas subyacentes a las que básicamente se quiere dar respuesta durante la realización de este proyecto de investigación: ¿Qué sustento tienen las inquietudes planteadas por algunos alumnos de la Facultad de Lenguas con respecto a una aparente divergencia con respecto a los criterios de corrección y calificación y, en consecuencia, entre las puntuaciones asignadas en los exámenes escritos por los distintos evaluadores?; ¿cuál es el grado de convergencia entre las puntuaciones asignadas por los distintos evaluadores en las pruebas escritas de las cátedras objeto de estudio en este proyecto?; ¿en caso de ausencia significativa de consistencia entre los distintos profesores-calificadores en cuanto a la puntuación otorgada a una misma prueba escrita, es posible determinar las causas más comunes que la provocan?; ¿es posible elaborar criterios claros, uniformados y precisos de calificación que ayuden, por una parte, a los profesores evaluadores a lograr coincidencia en la puntuación y/o calificación de las pruebas escritas y, por otra, a los alumnos a confiar en la calificación obtenida?

En este contexto, evaluar un aprendizaje es, pues, una acción encaminada a estimar, apreciar o juzgar el valor o mérito que tiene el cambio en el conocimiento, capacidades o actitudes de los estudiantes. Cuando se aplica la evaluación a la enseñanza universitaria se

amplía el campo de ideas, términos y significados relacionados y derivados de la evaluación. Así también se habla de. (a) *medir*, como la asignación de un número a un objeto, según una regla aceptable, o (b) de *codificar,* como la atribución de un valor a una actuación en una prueba. La medición en la enseñanza es la comparación de una tarea de aprendizaje (tipos de percepciones, comprensiones, conocimientos declarativos y procedimentales, o capacidades de respuesta que un estudiante debe poseer para tener éxito en un aprendizaje) con su respectiva unidad (tal como las puntuaciones de una prueba educativa), con el fin de averiguar cuántas veces la segunda (unidad) está contenida en la primera (experiencia).

Así, el concepto *medición* se refiere a un amplio rango de tareas de aprendizaje (destrezas y competencias específicas) que tienen distintas valoraciones para los profesores, incluso de una misma asignatura o cátedra. En consecuencia, la medición requiere un análisis sistemático y una reflexión crítica acerca del rasgo, habilidad o tarea que está midiendo el ítem de una prueba.

El concepto de fiabilidad es aplicable a cualquier instrumento y, por lo tanto, es también aplicable al instrumento que emplea el docente para evaluar. De este modo, la fiabilidad es una cualidad esencial que debe estar presente en todos los exámenes de carácter académico-científico, y que deciden, como en el caso que nos convoca, la promoción al curso siguiente. Por definición, podemos decir que una prueba es fiable cuando es estable, equivalente o muestra consistencia interna. Una prueba alcanza un elevado coeficiente de fiabilidad si los errores de medida quedan reducidos al mínimo (Uebersax, 1988; Hayes & Hatch, 1999; Stembler, 2004; Johnson, Penny & Gordon, 2009; Gwet, 2014; entre otros).

Las pruebas se consideran fiables cuando, midan lo que midan, proporcionan puntuaciones comparables cuando se repite su aplicación, o se compara con otra equivalente. La fiabilidad debe entenderse como el término que describe la consistencia que existe entre las medidas, la ausencia de error. Así, se dice que un *test* o *prueba* es fiable cuando mide con la misma precisión, da los mismos resultados, independientemente del sujeto calificador. Podemos decir que, a más fiabilidad, más estables y consistentes son los resultados de las pruebas entre una aplicación y otra.

En este sentido, fiabilidad, confiabilidad o precisión denotan la cualidad de un instrumento que permite que cualquier docente-calificador asigne la misma puntuación bajo las mismas condiciones. Por lo expuesto, se esperaba que este estudio nos permita a los docentes que integramos cada una de las cátedras involucradas en este proyecto, trabajar en pos de unificar u homogeneizar criterios para calificar a los alumnos, ya que una calificación sólo es fiable si se asienta sobre un constructo informado de validación (Bailey, 1998; Mc Namara, 2000; Celce-Murcia, 2001; Purpura, 2004; Bordón, 2006; entre otros).

Otro aspecto a considerar para lograr prácticas equitativas de evaluación, son los *criterios de evaluación*. Creemos que las escalas o criterios de evaluación – entendidos como indicadores para la evaluación de los aprendizajes de los alumnos en los diferentes niveles de concreción curricular - constituyen el referente fundamental para determinar el grado de consecución de los objetivos generales de la asignatura que hayan alcanzado los alumnos. Dichos criterios deben ser considerados, entonces, como puntos de referencia que hacen posible la calificación de lo que nos proponemos evaluar; en otras palabras, deben ser referentes de valor argumentados que nos ayuden a conocer en qué medida un sujeto alcanza el dominio de cada área.

Como hemos señalado en el párrafo anterior, las escalas o criterios de corrección se diseñan para lograr valoraciones sistemáticas. Sin embargo, a menudo fallan en su propósito, ya que, al interpretar la escala, los jueces pueden diferir en su interpretación. En este sentido, diversos autores (Watts & García Carbonell, 2005; Robb Singer & Lemahieu, 2011) recomiendan la adopción de criterios con descriptores de los diversos niveles o grados de corrección de las respuestas. Según estos autores, este tipo de criterio combina las virtudes de dos clases de

criterios: por un lado, la rapidez de los criterios globales, holísticos o de impresión que utilizan los jueces muy experimentados y, por otro lado, la obligación de considerar aspectos específicos que conllevan los criterios analíticos, que benefician a los jueces con menos diplomacia y entrenamiento.

El presente proyecto se plantea como objetivo general abordar la elaboración de criterios de evaluación uniformados con el objetivo de lograr fiabilidad inter evaluadora en las Cátedras de Lengua Inglesa I, Práctica Gramatical del inglés, Gramática Inglesa I, Práctica de la Pronunciación del inglés, Lengua Castellana I y Lecto comprensión en Lengua Extranjera IV (inglés) en la Facultad de Lenguas de la Universidad Nacional de Córdoba (UNC).

Como objetivos específicos se proyecta:

Describir y analizar los objetivos, contenidos programáticos, metodología de enseñanza, metodología y criterios de evaluación, tal cual se explicitan en el programa vigente correspondiente a cada una de las asignaturas objeto de análisis en este proyecto.

Verificar el grado de convergencia entre las puntuaciones asignadas por los distintos profesores-calificadores a cada una de las pruebas escritas.

Analizar, interpretar y/o describir las posibles causas que conduzcan a la ausencia de criterios unificados de corrección, según el tipo de actividad de que se trate y el tipo de error detectado.

Elaborar criterios de corrección uniformados referenciales de suficiencia de la competencia lingüística esperada en forma de escalas de estimación (de niveles de habilidad), que reflejen los requerimientos de suficiencia en términos de ausencia de errores tipificados y/o presencia de errores tipificados cuantificables.

Verificar la eficacia y utilidad de dichos criterios uniformados de suficiencia de la competencia lingüística esperada, en términos de uniformidad lograda en las calificaciones, independientemente de la identidad del profesor-calificador.

El presente trabajo se propone mostrar los resultados y conclusiones de un tramo del proyecto de investigación bianual llevado a cabo en el período 2016/17 en la Facultad de Lenguas, UNC que se tituló: *Hacia la elaboración de criterios uniformados de evaluación que aseguren fiabilidad entre los calificadores de las pruebas escritas en las cátedras de Lengua Inglesa I, Práctica Gramatical del Inglés, Gramática Inglesa I, Práctica de la Pronunciación del Inglés, Lengua Castellana I y Lectocomprensión en Lengua Extranjera IV (Inglés).* El mismo estuvo avalado por Secyt y dirigido por Mgtr. Fabián Negrelli y co dirigido por la Mgtr. Cecilia Ferreras.

Aspectos metodológicos

Sujetos:

Los sujetos estudiados en este tramo del proyecto fueron:

a) Un grupo de 40 alumnos inscriptos en la asignatura objeto de análisis.

b) Los docentes a cargo tanto del dictado de los contenidos teórico-prácticos como de su posterior evaluación en la asignatura involucrada en este estudio.

Materiales

Programa vigente de la asignatura objeto de estudio.

Exámenes finales escritos para los alumnos “regulares”, correspondientes a los turnos febrero-marzo de 2016 y febrero-marzo de 2017.

Taxonomías descriptivas y explicativas de errores que permitan consensuar criterios uniformados de corrección.

Si bien existen otros instrumentos para realizar la evaluación de los aprendizajes de los estudiantes, las pruebas escritas y/u orales constituyen actualmente el único instrumento que

contempla el Reglamento de Exámenes vigente en la Facultad de Lenguas de la Universidad Nacional de Córdoba (en adelante FL, UNC) para evaluar a los alumnos que han obtenido la regularidad en una asignatura y determinar si están en condiciones de ser promocionados al nivel superior. El tipo de "prueba" objeto de análisis en la presente investigación puede considerarse, siguiendo a Hyland (2003), como una *evaluación sumativa*, dado que se califica formalmente y se lleva a cabo al final de los procesos de enseñanza y de aprendizaje con el fin de determinar si se han logrado los objetivos planteados para todo el curso. Así, en las asignaturas que fueron objeto de estudio en este proyecto*, la* "prueba" se constituye en una herramienta de poder y de control y en una suma de puntos con el objetivo de generar una nota o calificación que impacta definitivamente en la decisión del docente de promover a los estudiantes a niveles superiores, lo cual, se torna, al menos, cuestionable. Recordemos el pensamiento de Díaz Barriga (1990) sobre esta cuestión:

> El examen es un espacio donde se realiza una multitud de inversiones de las realidades sociales y de las pedagógicas. Es un espacio que invierte las relaciones de saber y de poder. De tal manera que presenta como si fueran relaciones de saber las que fundamentalmente son de poder. Los problemas de orden social: posibilidad de acceso a la educación, justicia social, estratos de empleo, estructura de la inversión para el desarrollo industrial, etc.; son trasladados a problemas de orden técnico, objetividad, validez, confiabilidad. (p. 189)

Dada la naturaleza de este proyecto, se formaron seis subgrupos de trabajo; es decir, los integrantes del equipo fueron distribuidos según la asignatura que dictan. Cabe destacar que todas las actividades que desempeñó cada uno de estos subgrupos estuvieron supervisado por el director y / o Codirectora de este proyecto. Por otra parte, estos subgrupos trabajaron, de manera alternada, en forma independiente o mancomunada, según las fases o etapas por las que transitó la investigación.

El proyecto se dividió en dos fases:

Fase I

En un primer momento, y tal como sucede habitualmente, los Profesores Titulares de cada una de las asignaturas involucradas en este estudio (debemos señalar que todos ellos forman parte de este proyecto) procedió a diseñar el examen final ("la prueba") correspondiente al turno de examen febrero-marzo de 2016. Dicho examen fue elaborado en un todo de acuerdo con los contenidos disciplinares y formato de evaluación correspondientes al ciclo lectivo 2015. Una vez administrado y resuelto por los alumnos, se seleccionaron aleatoriamente 20 exámenes de alumnos "Regulares" de cada una de las asignaturas objeto de estudio en esta investigación. A los fines de preservar la versión original de la muestra, se procedió a fotocopiar cada uno de los exámenes seleccionados antes de que comience el proceso de corrección de los mismos.

Cada subgrupo de trabajo se avocó luego a realizar la descripción y el análisis de los objetivos, contenidos disciplinares, metodología de trabajo, formas y criterios de evaluación de la asignatura objeto de estudio en cada caso*,* según el programa vigente de la asignatura. Seguidamente, los docentes a cargo tanto del dictado de la asignatura como del diseño, administración y corrección de las pruebas escritas procedieron a corregir los exámenes que fueran fotocopiados y preservados oportunamente.

Para asegurar mayor fiabilidad con respecto a los resultados de la investigación, los exámenes fueron corregidos por al menos tres docentes diferentes; por lo tanto, en aquellos casos en que el número de docentes que formó parte de un subgrupo fue menor al número señalado, se solicitó la colaboración de otros docentes externos a este proyecto. En este sentido, deseamos destacar que todos los Profesores Adjuntos y Asistentes que forman parte

de las cátedras objeto de estudio en este proyecto, pero que no integran este equipo de investigación, ya han sido consultados al respecto; la mayoría de ellos se han mostrado sumamente interesados tanto en colaborar con la corrección de los exámenes como en discutir y elaborar, de manera conjunta, los criterios de calificación.

De esta manera, los docentes de cada subgrupo, con la colaboración de algún docente externo al proyecto en caso de haber sido necesario, corrigieron en forma individual (sin ningún tipo de influencia o injerencia de los otros docentes) cada uno de los 20 exámenes que formaron parte de la muestra recolectada. Posteriormente, se procedió a verificar el grado de convergencia entre las puntuaciones asignadas por los distintos docentes-calificadores a cada una de las pruebas escritas que formaron parte de la muestra.

Seguidamente, previa tabulación de los datos recabados en una tabla matriz, se analizaron las posibles causas que hayan conducido a divergencias entre los distintos examinadores, para poder elaborar, sobre la base de lo analizado en el paso anterior, criterios consensuados y uniformados de corrección que permitan reducir los errores de medición y/o puntuación por parte de los distintos docentes-calificadores. De este modo, se trató de lograr mayor consistencia en las calificaciones.

Finalmente, se realizó una puesta en común, con el fin de socializar, por un lado, las distintas experiencias vividas durante esta primera fase de la investigación y, por otro, los resultados obtenidos por cada uno de los subgrupos de trabajo. En esta instancia, la técnica empleada fue el *grupo de discusión*; en este caso, se llevó a cabo una reunión general en la que el director y la codirectora propiciaron, por un lado, el intercambio de distintos puntos de vista, opiniones e interpretaciones sobre el tema objeto de estudio y, por otro lado, la evaluación de la consecución de los objetivos de este proyecto y del cumplimiento de las actividades planeadas para esta fase. Creemos que esta instancia fue sumamente enriquecedora no solo para los integrantes del equipo de investigación sino para todos los docentes que forman parte de cada una de las cátedras objeto de estudio en esta investigación.

Fase II

A los fines de verificar la funcionalidad y/o eficacia de los criterios que se hayan consensuado, los mismos docentes que corrigieron los exámenes de la primera muestra, procedieron individualmente a corregir una segunda muestra de pruebas escritas que correspondió a los exámenes finales escritos correspondientes al turno febrero-marzo de 2017 para alumnos “Regulares”. Luego, el grupo de docentes que conformó cada subgrupo de trabajo tabuló nuevamente en una tabla matriz los datos recabados y analizó los resultados en términos de grado de uniformidad en la calificación que ha sido asignada a cada una de las veinte pruebas escritas corregidas por cada profesor-calificador. El objetivo central, en este caso, estuvo centrado en comprobar si la variación en las puntuaciones y calificaciones propuestas por los distintos examinadores han disminuido en valor, es decir, si dichas puntuaciones y calificaciones muestraron un mayor grado de consistencia.

Finalmente, se realizó una puesta en común, con el fin de socializar, por un lado, las distintas experiencias vividas durante esta segunda fase de la investigación y, por otro, los resultados obtenidos por cada uno de los subgrupos de trabajo. En esta instancia, la técnica a emplear fue la de el *grupo de discusión*; en este caso, se llevó a cabo una reunión general en la que el director y la codirectora propiciaron, por un lado, el intercambio de distintos puntos de vista, opiniones e interpretaciones sobre el tema objeto de estudio y, por otro lado, la evaluación de la consecución de los objetivos de este proyecto y del cumplimiento de las actividades planeadas para esta fase.

Resultados alcanzados y/o esperados
Acerca de los resultados en la asignatura Lengua Inglesa I

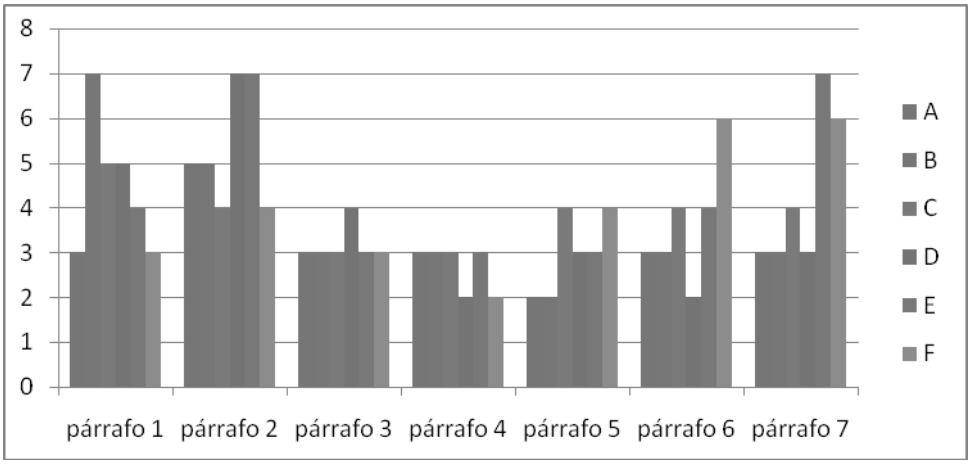
Lengua Inglesa I es una asignatura de primer año del Profesorado de Lengua Inglesa, Licenciatura en Lengua Inglesa y Traductorado Público Nacional de Inglés de la Facultad de Lenguas, UNC. Esta asignatura tiene como objetivo lograr que los alumnos desarrollen las cuatro macro habilidades de la lengua (lectura, escucha, escritura y habla) a fin de alcanzar un nivel intermedio en su conocimiento de inglés. *Lengua Inglesa I,* que tiene una carga horaria de 10 horas cátedra semanales, está divida en dos cátedras: la cátedra "A" cuenta con 6 comisiones en el turno mañana y la cátedra B está, a su vez, subdividida en dos comisiones del turno tarde. Cada ciclo lectivo, cursan la asignatura aproximadamente 600 alumnos. Como se desprende del dato anterior, cada comisión cuenta con un número elevado de estudiantes y, dado que el cuerpo docente a cargo del dictado suele fluctuar dependiendo de variables tales como licencias por salud, estudios, embarazos, entre otros factores, se consideró necesario acordar criterios de corrección uniformados que aseguraran que las correcciones se lograran de la manera más efectiva y objetiva posible. Por este motivo, profesores del equipo docente de la Cátedra formamos parte de este subgrupo de trabajo, la Profesora titular de la Cátedra "B" actuó como evaluadora externa y una Ayudante-Alumna de la Cátedra participó aportando su experiencia desde otra perspectiva.

A fin de pautar criterios de evaluación uniformados y alcanzar así el objetivo principal del presente proyecto, cada docente-evaluador corrigió la sección de composición de párrafos académicos de siete exámenes finales escritos correspondientes al turno febrero de 2017 para alumnos "Regulares". Los exámenes fueron elegidos aleatoriamente por el director del proyecto. Para la primera y segunda sesión se utilizaron los mismos párrafos mientras que para última sesión se corrigieron párrafos distintos con el objeto de poner a prueba la grilla final de evaluación elaborada por los integrantes de este subgrupo. Para comparar y aunar criterios, se realizaron 3 sesiones (mayo 2017, octubre 2017 y febrero 2018) en las que participaron todos los integrantes del subgrupo 1 junto con el director y la codirectora. Para la recolección de datos cualitativos, se organizaron grupos de discusión estructurados siguiendo la dinámica de una conversación entre el grupo de docentes participantes. En dichas instancias, el director del proyecto fue registrando los intercambios y, posteriormente, comparando los puntajes obtenidos por cada docente y señalando las discrepancias como también las similitudes que se observaban en cada corrección.

Los resultados de la primera sesión fueron un poco desalentadores. Cada docente-evaluador se guió por los criterios de corrección consensuados por todos los miembros de la asignatura (cantidad de puntos que se descuentan en los casos en los cuales los alumnos escriben más de un párrafo, incluyen ideas no relevantes, cometen errores básicos de lengua). Se detectaron algunas diferencias en los puntajes otorgados por los distintos evaluadores; en este sentido, las discrepancias más notables fueron identificadas en relación a problemas con el título otorgado al párrafo, con la oración que presenta la idea principal con la temática que se desarrollará en el párrafo (*Topic Sentence*) y con el uso incorrecto de ciertas estructuras gramaticales y el uso (no) relevante del léxico. Dichas diferencias resultaron muy significativas ya que, en algunos casos; las calificaciones oscilaban entre un 4 ó superior (aprobado) a un 2 ó 3 (desaprobado). Este fenómeno se observó en el caso de los párrafos 1, 3, 5, 6 y 7.

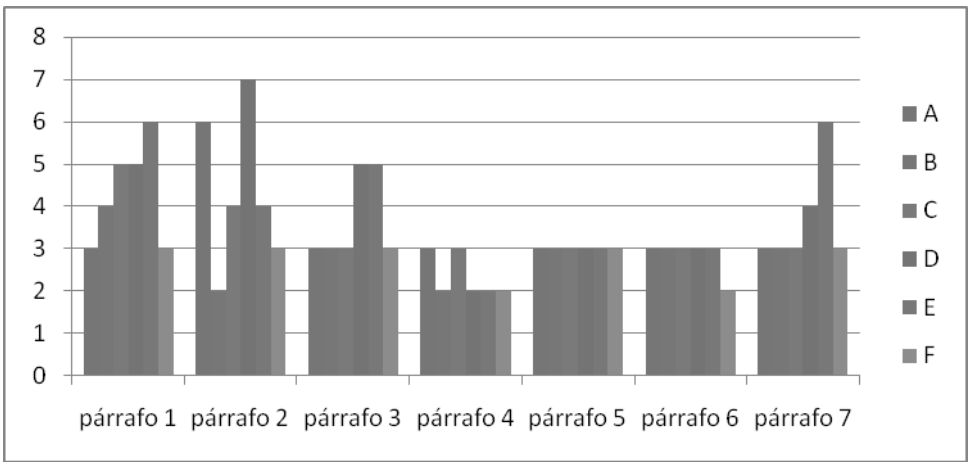
Gracias al debate generado entre los docentes-evaluadores en esta primera sesión, se confeccionó una grilla analítica en donde se especificó el porcentaje de la nota que se restaría por cada uno de los distintos tipos de errores (Ver Apéndice 1). Esta grilla se utilizó para evaluar los mismos párrafos en la segunda sesión.

Figura 1: Resultados de la primera sesión



Como se observa en la Figura 2, la grilla elaborada permitió consensuar los casos de los párrafos 4, 5, y 6, puesto que todos los docentes evaluadores les asignaron una nota inferior a 4. Sin embargo, la grilla no fue efectiva para el resto de los párrafos. Resulta llamativo el caso del párrafo 2 en donde el docente D le asignó la nota 7 mientras que dos docentes lo evaluaron con 7 y otros dos con 3 y 2 respectivamente. En el grupo de discusión se manifestó la dificultad de utilizar una grilla en donde se "descuentan" puntos y en donde no hay claridad en cuanto al puntaje a descontar en el caso de errores gramaticales y de vocabulario. Por estos motivos, se decidió elaborar una grilla en donde se procediera de manera inversa, es decir, que, en vez de restar, se sumaran puntos según cada sección. Para determinar cada sección se adaptó el cuadro propuesto por Brown (2007, p.413) teniendo en cuenta los objetivos de la asignatura *Lengua Inglesa I* y la grilla confeccionada previamente (VER Apéndice 2).

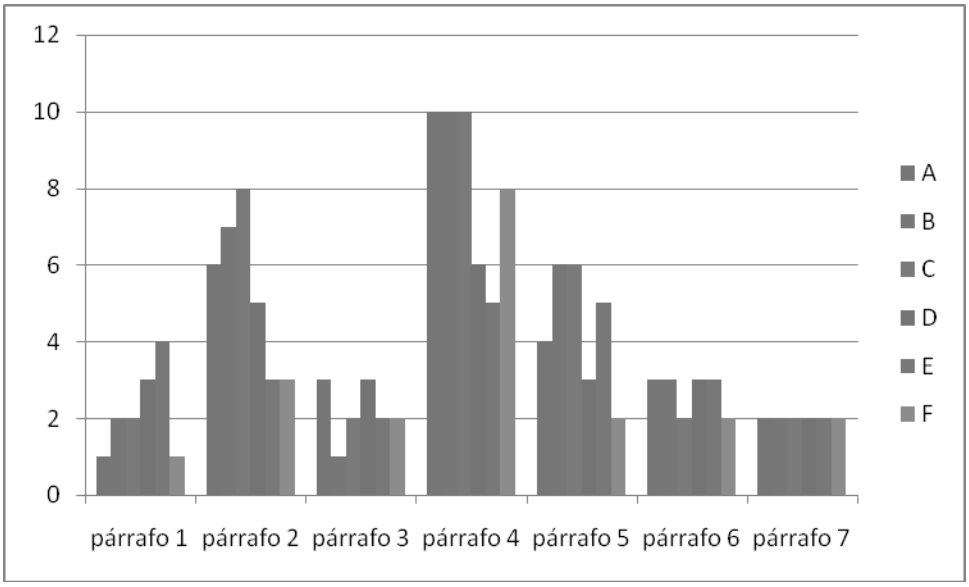
Figura 2: Resultados de la segunda sesión



Afortunadamente, las evaluaciones del nuevo set de párrafos con la nueva grilla tuvieron muchas coincidencias (Figura 3). Hubo coincidencias en cuanto a no aprobar al alumno en el

caso de los párrafos 1, 3, 6 y 7. También hubo coincidencias con respecto al párrafo 4 aunque las notas variaron entre una calificación de 6 (70%) a 10 (100%). Con respecto al párrafo 5, cuando la evaluadora D fundamentó las razones por las cuales el párrafo tenía pocos puntos en las secciones "contenido" y "organización y discurso" de la grilla, los evaluadores B, C y F estuvieron de acuerdo en que su puntaje en esa sección tendría que haber sido menor. El único párrafo con el que no hubo acuerdo fue en el caso del párrafo 2.

La grilla confeccionada para la última sesión, la cual indicaba que se deberían sumar los puntos, fue más efectiva ya que permitió a los profesores-evaluadores acordar criterios y poder determinar en qué aspectos había discrepancias en caso de haberlas. El subgrupo 1 va a seguir perfeccionando esta grilla para futuras instancias de evaluación y para colaborar en la formación de Profesores Adscriptos y de los nuevos profesores que se incorporen a la cátedra.



Apéndice 1: Grilla elaborada en la primera sesión

Problema	**% que se descuenta**
Hay más de 1 párrafo	74%
2 de las 3 ideas secundarias (*supports*) son irrelevantes	70%
1 de las 2 ideas secundarias (*supports*) son irrelevantes	70%
1 de las 3 ideas secundarias (*supports*) es irrelevante	20%
No hay título o el título es incorrecto (es una oración, es irrelevante, etc.)	10%
No hay TS o la TS es incorrecta	20%
No hay conclusión o la conclusión es incorrecta	15%
No hay uso específico del léxico relacionado con el tema (words, semi-fixed expressions, fixedexpressions, idioms, collocations)	30%
Hay errores en el uso del léxico (words, semi-fixed expressions, fixed expressions, idioms, collocations, prepositions, spelling, connectors)	30%
Hay errores en el uso de la gramática (syntax, agreement, WO, gerund/infinitive, plural of nouns,	30%

article, [C], [U], passive voice, modal verbs, verb tense, sentence fragment, run-on sentence)	
No incluye el WC o hay un problema con el WC	5%

Apéndice 2: Grilla elaborada en la segunda sesión basada en Brown (2007)

CATEGORIAS	PUNTAJE
Contenido	
La TS es efectiva Ideas principales relevantes Buenas ideas secundarias Buena conclusión	**10p**
El título es irrelevante	**-4**
TS incorrecta	**- 8**
1 de 3 "*support*" es irrelevante	**- 8**
1 de 2 "*support*" es irrelevante	**-18**
Irrelevant supporting details	**- 4**
Conclusión incorrecta	**- 6**
Organización yDiscurso	
Título TS 2 or 3 ideas principales Conclusión Longitud (180-220 palabras) Se incluye el número de palabras Cohesión (referencia, conector, cambio en el sujeto) Registro Punctuación	**10p**
No hay título o el título es una oración	**- 4**
Hay más de un párrafo	**- 30**
Uso del modo imperativo	**-1**
Uso de contracciones	**-1**
No se incluye el número de palabras	**- 2**
Syntax / Use of English	
Patrones verbales Concordancia Orden de las palabras Gerundio/infinitivo Formación del plural Uso del artículo [C], [U] Voz pasiva Verbos modales Tiempos verbales	**10p**
Para cada caso incorrecto	**-1**
Si el mismo caso se repite varias veces (e.j. "**people is*") se descuenta una sola vez	**-1**
Frase larga e inconexa (*run-on sentence*)	**-1**

Fragmento	**-1**
Vocabulario	
Palabras y frasesrelevantes Collocatciones Preposiciones Ortografía Mayúsculas	**10p**
No hay vocabulario relevante	**- 5**
Por cada pronombre indefinido (*something, anyone, anything, etc*)	**- 1**
Por cada caso incorrecto	**- 1**
TOTAL	**40p**

El sistema educativo universitario requiere más y mejores niveles de educación y enseñanza, y a ello debe contribuir la evaluación educativa. Lograr la excelencia académica es nuestro compromiso fundamental. Para ello es necesario que los docentes nos formemos, asumamos nuevas formas de hacer y modifiquemos nuestra actitud y mentalidad respecto de la acción evaluadora que, como docentes de un sistema educativo, nos corresponde. Esto supone un planteamiento cada vez más sistemático y científico de todos los elementos implicados: profesores, alumnos, metodología, contextos, evaluación, entre otros.

Se espera que los resultados del presente proyecto conduzcan a implicaciones relevantes para el mejoramiento de la calidad del proceso de evaluación de las asignaturas objeto de estudio en el presente proyecto, constituyendo una base de mayor solidez para perfilar líneas de acción en el trabajo docente. Asimismo, se espera que los resultados de esta investigación resulten de utilidad para mejorar la práctica educativa en otras asignaturas a cargo de los docentes integrantes de este proyecto. En términos más específicos, se pretende aportar criterios unificados para calificar a los alumnos, ya que, como se dijo anteriormente, una calificación sólo es fiable si se asienta sobre un constructo informado de validación (Bailey, 1998; McNamara, 2000; Celce-Murcia, 2001; Purpura, 2004; Bordón, 2006; entre otros).

En este sentido, este proyecto resulta significativo, ya que dará una respuesta a las constantes quejas planteadas por los alumnos respecto de la disparidad en la forma de calificar de los docentes de la Facultad de Lenguas, más específicamente, en las cátedras involucradas en este proyecto. En otras palabras, desde la perspectiva que trasciende al equipo mismo de investigación, alcanzar los objetivos propuestos constituiría un invalorable aporte para reflexionar sobre la propia práctica docente a los fines de concientizar a los docentes de la necesidad e importancia de mejorar la calidad del sistema de evaluación en nuestra Facultad y llevar a cabo acciones que conduzcan a tal fin.

Por otra parte, este proyecto contribuirá a iniciar la formación en el campo de la investigación educativa de un grupo importante de Profesores Asistentes, Adscriptos y Ayudantes-alumnos.

Bibliografía

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

Anijovich, R. (2010). La evaluación significativa. Buenos Aires: Paidós.

Bachman, L. &. Cohen A. (1998). *Interfaces Between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press.

Bachman, L. & Palmer A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L. & Palmer A. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Bachman, L. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. (2002). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21* (3), 5-19.

Bailey, K. (1998). *Learning about Language Assessment: Dilemmas, Decisions and Directions*. Pacific Gove, CA: Heinle & heinle.

Bordón, T. (2006). *La evaluación de la lengua en el marco de ELT: bases y procedimientos*. Madrid: Arco.

Brown, H. (2004). *Language Assessment: Principles and Classroom Practices*. New York: Pearson Education.

Celce-Murcia, M. (Ed.) (2001). *Teaching English as a Second or Foreign Language.* New York: Heinle & Heinle.

Díaz Barriga, Á. (1990). *Currículum y Evaluación Escolar*. Buenos Aires: Cuadernos Rei Argentina: IEAS-Aique.

Fulcher, G. & Davidson F. (2007). *Language Testing and Assessment. An Advanced Resource Book*. New York: Routledge.

Gass, S. (1994). The reliability of grammaticality judgements. In E. Tarone, S. Gass and A. Cohen (eds.). *Research Methodology in Second Language Acquisition.* 303-322. Hillsdale, N. J: Lawrence Earlbaum Associates.

González B., (2005). Calificar no es evaluar: Bogotá: Nuevo Horizonte Bogotá DC.

Gwet, K. (2014). *Handbook of Inter-Rater Reliability*. 4th edition. Gaithersburg: StatAxis Publishing.

Hayes, J. & Hatch J. (1999). Issues in measuring reliability. *Written Communication 16* (3), 354-367.

Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of Second Language Writing*, 12, 17-29.

Hughes, A. (2003). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Johnson, R., Penny & Gordon B. (2009). *Assessing Performance: Developing Scoring and Validating Performance Tasks*. New York: Guilford Publications.

Matute Vásquez, A. & Muriel Gómez, L. (2014). *La evaluación formativa en los procesos de aprendizaje de matemáticas*.

Mcnamara, T. (2000). *Language Testing.* Oxford: Oxford University Press.

Mottier López, L. (2010). *La evaluación significativa*. Argentina: Paidós

Purpura, J. (2004). *Assessing Grammar*. Cambridge: Cambridge University Press.

Robb Singer, N., & Lemahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. *Journal of Writing Assessment, 4*.

Stembler, S. (2004). A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation, 9* (4).

Uebersax, J. (1998). Validity inferences from interobserver agreement. *Psychological Bulletin 104*, (3), 405-416.

Watts, F. & García Carbonell, A. (2006). *La evaluación compartida: investigación multidisciplinar*. Valencia: Servicio de Publicaciones de la Universidad Politécnica de Valencia.