LA ELABORACIÓN Y APLICACIÓN DE CRITERIOS DE EVALUACIÓN UNIFORMES EN GRAMÁTICA Y FONÉTICA INGLESAS

Negrelli, Fabián
Facultad de Lenguas, UNC
Córdoba, Argentina
fabiannegrelli09@gmail.com

Capell, Martín Salvador
Facultad de Lenguas, UNC
Córdoba, Argentina
mscapell@unc.edu.ar

Introducción

Numerosos autores (González, 2005; Anijovich, 2010; Mottier López, 2010; Matute Vázquez y Muriel Gómez, 2014; entre otros) consideran que evaluar es uno de los temas más difíciles de la pedagogía, y 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" (36). Sin lugar a dudas, 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 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? 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, adquirir aquellas competencias y/o conocimientos que aún no han sido asimilados para superar la presente etapa y ser promovido 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 es lo que nos impulsó a realizar un análisis profundo de la situación que permitiese garantizar objetividad y consistencia por parte de los evaluadores al momento de calificar lo que denominamos genéricamente la "prueba" y que, al mismo tiempo, garantizara imparcialidad en la representación de los intereses de los alumnos. En este contexto, durante el periodo 2016-2017, llevamos a cabo en la Facultad de Lenguas, Universidad Nacional de Córdoba, un estudio que nos condujera a lograr prácticas evaluativas justas en las áreas de gramática y fonética inglesa, dotando a la evaluación de confiabilidad y transparencia. En este trabajo, nos focalizaremos en los objetivos, metodología empleada y principales resultados de dicha investigación.

Acerca de la fiabilidad y la objetividad en las correcciones de las evaluaciones escritas

Lograr prácticas evaluativas justas implica trabajar en pos de dotar a la "prueba" de confiabilidad y transparencia. En este sentido, diversos autores han argumentado extensamente acerca de la fiabilidad de las calificaciones y de la influencia negativa que ejerce la subjetividad en la percepción y actitud de los que corrigen (Bachman, 1990; Gass, 1994; Alderson, Clapham y Wall, 1995; Bachman y Palmer, 1996; Bachman y Cohen, 1998; Bachman, 2002; Hughes, 2003; Brown, 2004; Alderson, 2005; Fulcher y Davidson, 2007; Bachman y Palmer, 2010; Arnal-Bailera, Muñoz Escolano y Oller-Marcén, 2016). Sabemos que la objetividad es intrínseca a la tarea de medir. Decimos, por tanto, que una medición es objetiva si dos o más agentes, aplicando las mismas técnicas y procedimientos, llegan a resultados similares con un grado mínimo de desvío. 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, 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" que fue 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 nuestro 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 calificación que impacta definitivamente en la decisión del docente de promover a los estudiantes a niveles superiores, lo cual, se torna, al menos, cuestionable.

Acerca del marco teórico

La evaluación es fundamental en los procesos de enseñanza y de aprendizaje ya que determina la acreditación necesaria para que los estudiantes puedan participar en la comunidad educativa. El impacto de los modos de evaluar es una problemática susceptible de ser analizada y explorada en profundidad debido a que, en muchos casos, son las características de los procesos de evaluación las que determinan las prácticas pedagógicas concretas. Sin lugar a dudas, el campo de la evaluación es parte de una realidad sobre la cual docentes e investigadores debemos reflexionar de manera sistemática en los procesos mencionados. Cantón Mayo y Pino Juste (2011) arguyen que la evaluación juega un papel de gran relevancia ya que "[…] además de conocer el grado de dominio alcanzado por el alumno en relación con los objetivos propuestos, la evaluación también sirve para determinar si el proceso de enseñanza ha sido adecuado para alcanzar dichos objetivos" (18). La fiabilidad de las calificaciones en la evaluación de lenguas está condicionada por distintas fuentes de variación. El resultado de un examen debería reflejar lo que sabe el alumno; sin embargo, según Bachman (1990), las mediciones lingüísticas son una interacción compleja entre los aspectos controlables y menos controlables del contexto evaluador, y las características del estudiante que es examinado. Para dicho autor, los aspectos controlables son el marco y los atributos de la prueba, tales como el tipo de examen, las consignas, la respuesta esperada, y los parámetros y criterios de evaluación. Aspectos menos controlables son la suficiencia de la muestra, la precisión de la escala con la que hay que juzgarla y la mediación de los jueces entre la escala y la muestra. Las características del estudiante son su experiencia en el idioma, su cultura y madurez, entre otros aspectos. 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. Por otra parte, 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. En este orden de ideas, el concepto *medición* se refiere a un amplio rango de tareas de aprendizaje, entre ellas las 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, también 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 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 y Hatch, 1999; Stemler, 2004; Johnson, Penny y Gordon, 2009; Gwet, 2014). 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. Diferentes estudios que se han llevado a cabo han demostrado que las calificaciones de jueces competentes e independientes pueden variar (Hill y Parry, 1994; Wiley y Haertel, 1996; Black y William, 1998; Moscal y Leydens, 2000; Silvestri y Oescher, 2006; Sadler, 2009; Rezaei y Lovorn, 2010). Lo que es importante para un juez, puede ser menos para otro. En este contexto, muchos errores en la calificación son atribuibles a la variable "juez". Según Wolf (1990) y Watts y García Carbonell (1999), se han identificado tendencias en los errores, tales como la gravitación hacia el punto medio de una escala, evitando los extremos; a otorgar puntuaciones similares a características que el evaluador considera similares, pero que no lo son; y al efecto "halo" o tendencia a permitir que la impresión general del ejercicio influya en la evaluación de lo específico. A través de lo expuesto, no quedan dudas de que la tarea de corregir y calificar una prueba dista mucho de ser sencilla y requiere de un entrenamiento continuo y profundo. En este sentido, Huitrado y Climent (2013) sostienen que la formación académica de los profesores-correctores es determinante a la hora de calificar las tareas. Dichos autores agregan que el colectivo de profesores en formación es generalmente el que presenta una mayor dispersión de calificaciones y, por ende, señalan la necesidad de una formación específica. Respecto de los aspectos del entorno que influyen en la calificación, podemos mencionar la hora del día, las distracciones en la sala donde se califica y el orden de presentación de los ejercicios. Los aspectos influyentes que surgen de los propios ejercicios son la organización, el desarrollo, la calidad argumental, la longitud, la ortografía y la caligrafía. La velocidad con la que los jueces se ven obligados a evaluar y su actitud personal ante la presión a ceñirse a criterios prefijados también pueden influir. Por otra parte, de exámenes que derivan consecuencias significativas para los alumnos, igualmente puede provenir una presión adicional sobre los jueces. Estos

problemas de tipo operativo, cuya magnitud e importancia no pueden ignorarse de modo alguno, se complican cuando se unen a las preocupaciones teóricas por el "constructo" o concepto objeto de evaluación. En la evaluación de una lengua, se evalúa la competencia en esa lengua junto con la habilidad en la comunicación oral y escrita y la adecuación del contenido. En este orden de ideas, Bachman y Palmer (1996) proponen el modelo de *habilidad comunicativa en la lengua* (*Commuicative Language Ability*), en el que se definen una serie de competencias que componen dicha habilidad. Es decir, la habilidad comunicativa se puede desglosar en dos: por un lado, los mecanismos 'psicofisiológicos' y, por otro, las 'estrategias metacognitivas'. Estas incluyen una competencia estratégica y una competencia lingüística. Esta última, a su vez, se desglosa en una competencia para la organización -gramatical y textual- y otra pragmática – funcional y sociolingüística. Los conocimientos temáticos y el contexto de la situación completan, en una descripción muy breve, lo que para estos autores es la habilidad comunicativa en la lengua y lo que debe ser objeto de evaluación en una lengua. Otro aspecto a considerar en pos de 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 estudiantes en los diferentes niveles de concreción curricular - constituyen el referente fundamental para determinar el grado de consecución de los objetivos generales de la asignatura. 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 y García Carbonell, 2005; Robb Singer y 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.

Acerca de los resultados en las asignaturas *Práctica Gramatical del Inglés* y *Gramática Inglesa I*

Tal como fue estipulado en el proyecto presentado oportunamente ante SECyT UNC, a los fines de recolectar la muestra, los miembros del equipo que dictan ambas asignaturas procedieron a diseñar y administrar el examen ("la prueba") en un todo de acuerdo con los contenidos disciplinares y formato de evaluación impartidos durante el ciclo lectivo 2016. En cada una de las asignaturas, la muestra estuvo conformada por 60 exámenes -elegidos de forma aleatoria- correspondientes al turno de Exámenes Finales Regulares de noviembre de 2016. Los exámenes fueron luego divididos en tres grupos de 20. En una segunda fase, se abocaron a realizar la descripción y análisis de los objetivos y perfil del egresado de las carreras de Profesorado, Traductorado y Licenciatura en Inglés (Plan de Estudios vigente), para continuar, luego, con la descripción y el análisis de los objetivos, contenidos disciplinares, metodología de trabajo, formas y criterios de evaluación en cada una de las asignaturas foco de esta investigación. En su informe, llegaron a la conclusión de que si bien tanto los objetivos como la metodología de enseñanza y de evaluación planteados en los respectivos programas respondían a las pautas y descriptores establecidos en el Plan de Estudios vigente, los criterios de evaluación no estaban lo suficientemente descriptos o establecidos para asegurar fiabilidad en la corrección de los exámenes independientemente del docente que corrigiera esa prueba. En consecuencia, se procedió a revisar este punto. Asimismo, se suministró una encuesta a los alumnos, en la cual se hizo hincapié en su experiencia personal en el cursado de *Práctica Gramatical del Inglés* y *Gramática Inglesa I* durante el ciclo lectivo 2016. En términos generales, se

observó un amplio consenso entre los alumnos respecto de que a lo largo de su trayectoria como estudiantes en la Facultad de Lenguas habían experimentado a menudo en las distintas cátedras la falta de uniformidad en cuanto a los criterios de corrección de sus evaluaciones (valoración de errores, respeto por las ideas, distintas exigencias en cuanto a la profundidad en el desarrollo teórico de un tema, entre otros aspectos) con respecto a las de otros compañeros, dependiendo de quiénes habían sido los docentes que habían corregido los exámenes. En cuanto a las asignaturas objeto de estudio en esta oportunidad, las respuestas de los alumnos corroboraron nuestra hipótesis inicial en cuanto a la sensación de los alumnos con respecto a la existencia de diferentes criterios de corrección y calificación utilizados por los distintos docentes de la cátedra. Entre otros aspectos, los alumnos resaltaron el hecho de que la aplicación de diferentes criterios de corrección reflejaba cierta imparcialidad por parte de los docentes, lo cual generaba situaciones de manifiesta imparcialidad. Así, el equipo docente de ambas cátedras procedió a corregir el primer grupo de exámenes que fueran fotocopiados y preservados oportunamente para este fin. Los distintos evaluadores corrigieron en forma individual, y sin establecer ni acordar previamente criterios de corrección (excepto los establecidos en el Programa de cada asignatura), los 20 exámenes que formaban parte del primer grupo de la muestra. Posteriormente, se realizó una reunión con los miembros del equipo de investigación (se aplicó la técnica del *grupo de discusión*) a los fines de verificar la presencia o ausencia de uniformidad de criterios de evaluación entre los distintos sujetos evaluadores. Esta sesión fue coordinada por el director y la codirectora del proyecto. A modo de ilustración, nos referiremos específicamente a los hallazgos respecto de la asignatura *Práctica Gramatical del Inglés*. Entre las conclusiones a las que arribamos, creemos que vale la pena mencionar las siguientes:

(1) en solo dos casos (10%) hubo coincidencia con respecto a la calificación final obtenida; por lo tanto, eran fundadas las quejas y/o percepciones de los alumnos en cuanto a la ausencia de criterios unificados de corrección.

(2) Existía una necesidad imperiosa de mejorar y compartir las prácticas evaluativas de la cátedra, sobre todo con en el caso del evaluador N.º 4 (donde se nota mayor disparidad en cuanto a las notas o porcentajes otorgados en cada caso con respecto a los otros evaluadores), quien se había incorporado a la cátedra recientemente y, por ende, carecía de la experiencia que poseían los otros tres evaluadores, quienes se venían desempeñando en el área de gramática inglesa durante varios años.

(3) Los docentes de la cátedra debían trabajar en pos de desarrollar métodos o procedimientos de corrección fiables que garantizaran a los alumnos resultados justos.

(4) Era necesario describir los distintos tipos de errores cometidos por los alumnos y acordar la importancia de dichos errores en la calificación final con el objeto de elaborar una normativa de aplicación que lograra la máxima estandarización con respecto a los criterios de evaluación que debían seguir todos los miembros de la cátedra.

Sobre la base de estos resultados, nos abocamos al diseño de una taxonomía descriptiva y explicativa de errores para fijar criterios uniformes de corrección. Luego de cumplir con esta fase, se implementó una segunda etapa de corrección para verificar en qué medida se habían podido unificar dichos criterios. En términos generales, en esta segunda etapa de corrección, no se observaron diferencias significativas, por lo cual podemos concluir que: (1) era necesario unificar criterios de corrección; (ii) los criterios consensuados fueron aplicados de manera uniforme por los distintos evaluadores; y (iii) los resultados del presente proyecto en esta área fueron altamente positivos.

Acerca de los resultados en la asignatura *Práctica de la Pronunciación del Inglés*

En una primera etapa, los miembros de este subgrupo de trabajo se abocaron a la lectura y análisis de los programas de la asignatura bajo escrutinio. Al respecto, pudieron observar que, si bien en el programa de *Práctica de la Pronunciación del Inglés* explicitaban los criterios de

evaluación que se aplicarían al momento de puntuar los exámenes parciales y finales, estos criterios estaban redactados de manera muy general, lo que permitía una multiplicidad de consideraciones posibles al momento de calificar la producción de los estudiantes. En una segunda etapa, recolectaron la muestra que se utilizaría en las distintas instancias de corrección a los fines de verificar la posible ausencia de criterios unificados y las consecuentes soluciones. Esta de tarea de corrección fue llevada a cabo por parte de los miembros del subgrupo como así también de otros docentes de la Cátedra, quienes lo hicieron en calidad de colaboradores externos. En una tercera etapa, se organizaron sendas reuniones en las que se aplicó la técnica de *grupos de discusión*. En estas reuniones se compartieron las particularidades y dificultades experimentadas en el proceso de evaluación de las muestras. Asimismo, se verificó el grado de convergencia o divergencia que hubo entre las calificaciones y/o puntajes dados a las distintas "pruebas" por cada uno de los docentes-calificadores. En algunos casos, se observó cierta disparidad en cuanto a cómo calificar las transcripciones en las que se utilizaron símbolos fonéticos pertenecientes a otra variedad del inglés que no fuera la que se enseña en esta asignatura (por ejemplo, si el estudiante transcribía una palabra representándola en inglés norteamericano, en vez de representarla en inglés birtánico, que es el que se utiliza, al menos en las transcripciones). En otros casos, se advirtió que no existía uniformidad con respecto a cómo tratar un error repetido; en este sentido, algunos de los docentes lo consideraban tantas veces como se repitiera, mientras que otros lo consideraban solo una vez, ya que era un solo error que, casualmente, aparecía un determinado número de veces. En base a estos dos principales problemas que se advirtieron, se creyó necesario definir con mayor precisión los criterios con el fin de proceder a una nueva evaluación de las mismas pruebas. Luego de esta instancia, se procedió a la segunda evaluación. En este caso, la paridad en las calificaciones fue notoria. Estas acciones nos permitieron arribar a ciertas conclusiones: (i) es necesario establecer, como una sección obligatoria de los programas de estudio, una enumeración detallada de todos los criterios que se deberán tenerse en cuenta al momento de corregir los exámenes, a los fines de lograr una calificación fiable, independientemente de quién la lleve a cabo; (ii) es necesario distribuir entre los miembros de la cátedra un instructivo detallado y ejemplos de producciones escritas de alumnos con las correcciones esperadas según el tipo y valoración del error; (iii) de aquí en adelante, será necesario entrenar a los profesores que se sumen al equipo docente de la cátedra en la aplicación de estos criterios unificados de corrección.

Conclusión

A modo de conclusión, deseamos resaltar que en pos de llevar a cabo una evaluación justa en el aula, lo cual implica que sea equitativa, válida y transparente, el objetivo central de nuestro estudio fue asegurar la fiabilidad de las calificaciones de las pruebas escritas en las asignaturas bajo escrutinio. Es fundamental que tengamos en cuenta que desde el momento que somos los docentes quienes planteamos qué, cómo y cuándo evaluar, estamos involucrando nuestra propia subjetividad. 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. Luego de haber concluido nuestra tarea, confiamos en que los resultados de esta investigación aumentarán significativamente la fiabilidad en la corrección de las pruebas escritas en las cátedras objeto de investigación, ya que los alumnos obtendrán calificaciones sustancialmente más discriminatorias al haber una mayor concordancia entre jueces. Por otro lado, si bien nuestro proyecto fue concebido teniendo en cuenta algunas asignaturas de los planes de estudio de las carreras de Profesorado, Traductorado y Licenciatura en Inglés, los resultados muy probablemente sean de interés para

mejorar el proceso de evaluación de otras asignaturas en las que se observe la misma problemática.

Referencias Bibliográficas

Alderson, J. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment*. Londres: Continuum.

Alderson, J. C., Clapham, C. y Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

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

Arnal-Bailera, A., Muñoz Escolano, J. M., y Oller Marcén, A. (2016). Caracterización de las actuaciones de correctores al calificar pruebas escritas de matemáticas. *Revista de Educación,* 371, 35-60.

Bachman, L. F., y Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research.* Cambridge: Cambridge University Press.

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

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

Bachman, L. F. (1990). *Language testing construction and evaluation*. Oxford: Oxford University Press.

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

Black, P. J., y William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice,* 5 (1), 7-73.

Brown, H. D. (2004). *Language assessment: principles and classroom practices*. New York: Pearson Education.

Cantón Mayo, L., y Pino Juste, M. (2011) (Coord.). *Diseño y desarrollo del currículum.* Madrid: Alianza Editorial..

Fulcher, G., y Davidson, F. (2007). *Language testing and assessment. An advanced resource book*. Nueva York: Routledge.

Gass, S. (1994). The reliability of grammaticality judgements. En E. Tarone, S. Gass y A. Cohen (Eds.). *Research methodology in second language acquisition* (pp. 303-322). Hillsdale, N. J: Lawrence Earlbaum Associates.

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

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4ta ed.) Gaithersburg: StatAxis Publishing.

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

Hill, C., y Parry, K. (1994). Models of literacy: the nature of reading tests. En C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 7-34). Harlow, Reino Unido: Longman.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Huitrado, J. L., y Climent, N. (2013). Conocimiento del profesor en la interpretación de errores de los alumnos en álgebra. *PNA*, 8 (2), 75-86.

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

Johnson, R., Penny, J. y Gordon, B. (2009). *Assessing performance: developing scoring and validating performance tasks*. Nueva York: Guilford Publications.

Jorba, J. y Sanmartí, N. (1993). La función pedagógica de la evaluación. *Aula, 20*, 20-23.

Matute Vázquez, A., y Muriel Gómez, L. J. (2014). *La evaluación formativa en los procesos de aprendizaje de matemáticas* (Tesis de licenciatura). Universidad de Antioquía, Facultad de Educación. Recuperado de http://ayura.udea.edu.co:8080/jspui/handle/123456789/1322.

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

Rezaei, A. R., y Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. Assessing Writing, 15 (1), 18–39.

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

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. Assessment and Evaluation in Higher Education, 34, 159-179.

Silvestri, L., y Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. International Electronic Journal of Health Education, 9, 25–30.

Stemler, S. E. (2004). A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation, 9* (4). Recuperado de http://PAREonline.net/getvn.asp?v=9&n=4.

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

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

Wiley, D. E., y Haertel, E. H. (1996). Extended assessment tasks: purposes, definitions, scoring and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: promises, problems and challenges* (pp. 61-89). Mahwah, NJ: Lawrence Erlbaum Associates.

Wolf, R. M. (1990). Rating scales. En Keaves, J. P. (ed.), *Educational Reserach, Methodology and Meassurement: An Internacional Handbook.* Oxford: Pergamon Press.