Construcción y validación de una rúbrica para medir la expresión escrita en estudiantes universitarios*

Construction and Validation of a Rubric for Measuring Writing in University Students

Recibido: 27 de diciembre de 2013 | Aceptado: 02 de febrero de 2016

Armel Brizuela Rodríguez**
Universidad de Costa Rica, Costa Rica

Resumen
En este artículo se presenta el proceso de construcción y validación de una rúbrica para medir la expresión escrita de estudiantes universitarios de primer año de ingreso. Se recogieron 117 ensayos escritos por estudiantes universitarios (75 mujeres y 42 hombres) de la carrera de Filología Española de la Universidad de Costa Rica. Los ensayos fueron calificados por dos evaluadoras con base en una rúbrica, después de lo cual se calcularon diferentes índices de concordancia entre jueces para evaluar la calidad técnica de ésta. Entre los resultados más importantes se observó un porcentaje de acuerdo absoluto [±3 errores] del 90% en la mayoría de los rubros evaluados. Por otro lado, se observaron correlaciones pequeñas y moderadas entre los aspectos incluidos en la rúbrica. A partir de estos resultados se discute la utilidad de utilizar la lingüística textual como base teórica para construir rúbricas de evaluación.
Palabras clave
Confiabilidad; rúbrica; expresión escrita; estudiantes universitarios

Abstract
This article describes the process of construction and validation of a rubric to measure the writing of first-year university students. 117 essays written by Spanish Philology students (75 women and 42 men) of the University of Costa Rica were collected. Two evaluators, based on a rubric, rated the compositions after which different raters agreement coefficients were calculated to assess the technical quality of the rubric. Among the most important results, a 90% of absolute agreement was observed [± 3 errors] in most of the areas evaluated. On the other hand, small and moderate correlations between the aspects included in the rubric were observed. From these results the utility of using text linguistics as a theoretical basis for constructing written expression rubrics are discussed.
Keywords
Reliability; rubric; writing; university students

doi: 10.11144/Javeriana.upsy15-1.cvrm

Para citar este artículo: Brizuela Rodríguez, A. (2016). Construcción y validación de una rúbrica para medir la expresión escrita en estudiantes universitarios. *Universitas Psychologica, 15*(1), 349-360. http://dx.doi.org/10.11144/Javeriana.upsy15-1.cvrm

* Artículo de investigación. La financiación se obtuvo de la Universidad de Costa Rica.

** Correo electrónico: armel9@gmail.com

Introducción

En este artículo se exponen los resultados obtenidos en el proceso de recabar evidencias de confiabilidad para una nueva rúbrica analítica con la que se busca evaluar la competencia comunicativa a través de la escritura. La perspectiva teórica que subyace a la creación de este instrumento de evaluación se basa en la lingüística textual y su aplicación en el ámbito de la medición educativa (Lomas, 1999). En este sentido, la competencia comunicativa se entiende como lo que un hablante necesita saber para comunicarse de manera eficaz en contextos culturalmente significantes. Para el presente estudio, el interés fundamental fue evaluar esta competencia en la producción de textos adecuados al contexto universitario.

La importancia de esta competencia es fundamental si se toma en cuenta que buena parte de las labores de un estudiante universitario consiste en plasmar por escrito sus conocimientos. Ya sea en exámenes de respuesta construida, informes de investigación, reportes breves de laboratorio, cartas formales, correos electrónicos, entre otros, en los ambientes universitarios es imprescindible poder transmitir información a diferentes tipos de lectores para lograr desempeñarse adecuadamente. No obstante, la expresión escrita en el estudiantado es una de las grandes debilidades de la población universitaria (Rodino & Ross, 1997; Sánchez, 2004a, 2004b, 2005a, 2005b, 2006a, 2006b, 2007).

Inclusive, este problema ha llegado a generar toda una línea de investigación sobre las mejores estrategias para enseñar a producir textos en contextos académicos. Un ejemplo de ello es el trabajo de Parodi y Núñez (2002), quienes, desde una perspectiva cognitivo-textual, proponen evaluar la capacidad para producir textos argumentativos por parte de estudiantes de secundaria. Tras la investigación, se determinó que el principal problema de los textos argumentativos creados por los estudiantes radica en que estos desconocen las categorías esquemáticas del discurso argumentativo, lo que conlleva a inconvenientes como la ausencia de cohesión global en el texto, carencia de relaciones retóricas entre ideas y párrafos y ausencia de tesis, argumentos y conclusiones a la hora de estructurar el texto. Esta investigación permitió a los investigadores determinar que la enseñanza de la escritura enfatiza solamente las habilidades y conocimientos en un nivel microestructural, dejando de lado el macroestructural, es decir, solamente se abordan aspectos como ortografía, sintaxis o léxico, mientras que los aspectos de cohesión o coherencia se dejan de lado.

Ahora bien, el objetivo fundamental de esta investigación no es indagar sobre las posibles causas de esta problemática. Sin embargo, para comprender los contenidos evaluados en la rúbrica mencionada anteriormente, es necesario repasar someramente algunos hallazgos sobre la producción textual y lo que ésta implica en términos de procesamiento cognitivo. Finalmente se expondrán algunas investigaciones en las que se ha intentado generar un baremo o rúbrica para evaluar la expresión escrita con base en su relación con la comprensión y producción del lenguaje.

Uno de los aspectos cruciales de un texto escrito es su coherencia. De acuerdo con Louwerse, McNamara, Graesser, Jeuniaux y Yang (2006), la coherencia es la consistencia de los elementos en la representación mental de quien lee o escucha un texto, esto es, la coherencia remite a los vínculos semánticos y pragmáticos entre las diferentes porciones informativas de un mensaje verbal (Knoch, 2007). Por su parte, McNamara, Kintsh, Butler y Kintsh (1996) afirman que medir el grado de coherencia requerido en un texto para que este propicie una comprensión adecuada representa una tarea compleja, pues esta característica no es únicamente un asunto de la macroestructura del texto (de Beaugrande, 1984; Parodi & Núñez, 2002), sino que también incluye elementos externos al escrito, como el nivel de conocimiento previo con el que cuenta el lector acerca del tema específico del texto.

Ahora bien, la capacidad de un texto para transmitir información no solamente depende de los conocimientos y habilidades de quien lo lee, sino que también se basa en algunas características del texto en sí. McNamara, Crossley y McCarthy (2010) encuentran que aspectos como la complejidad sintáctica, la diversidad léxica y algunas caracterís-

ticas morfológicas de las palabras tienen un gran impacto en la forma en que el receptor procesa un texto escrito. Por tal motivo, en principio, tiene sentido intentar evaluar la competencia comunicativa mediante la expresión escrita, por la existencia de elementos observables en el texto que permiten inferir las capacidades, conocimientos y habilidades de quien lo produce.

En los ambientes educativos, para fines de selección y diagnóstico, muchas veces es importante evaluar la expresión escrita del estudiantado. En este contexto, es fundamental contar con evidencia de que un estudiante podría desempeñarse adecuadamente, abstrayendo toda la complejidad inherente a la producción del lenguaje a nivel cognitivo (Olive, Kellog, & Piolat, 2008; Olive, 2004). De este modo, aunque no se puede concluir que exista una única habilidad para expresarse por escrito (Carroll, 1993), sí es posible operacionalizar una serie de indicadores lingüísticos importantes en la producción textual que permitan la toma de decisiones en contextos aplicados como la selección de personal o el diagnóstico de posibles carencias para escribir textos (Crossley & McNamara, 2010; Beauvais, Olive, & Passerault, 2011).

Con base en lo anterior, es posible concluir que la capacidad para comunicarse mediante el lenguaje escrito es ciertamente compleja, pero no por ello imposible de evaluar en contextos aplicados (laborales, educativos, etc.). No obstante, sí es necesario aclarar que la medición de esta competencia es mucho más compleja que la de otras habilidades, ya que para ello se requiere de calificadores expertos que proporcionen juicios consistentes y concordantes entre sí. En esta línea, Eckes (2005) realiza una investigación donde intenta determinar la influencia que pueden tener la severidad e incluso los prejuicios de género de un juez sobre la calificación que da a un texto producido por un estudiante o candidato a un puesto. Para ello, seleccionó a jueces expertos y los entrenó en el uso de la rúbrica empleada en la calificación del *TestDaf* (examen de alemán para hablantes no nativos). Este investigador identificó la presencia de una gran variabilidad en el nivel de severidad de los jueces, hecho que tiene consecuencias en la calificación de los examinados.

Debido a esta discordancia que suele existir entre jueces expertos que evalúan pruebas de respuesta construida (Bejar, 2012), la recolección de evidencias de validez de las pruebas de desempeño en la producción textual se basa en la necesidad de garantizar un mínimo de concordancia entre los evaluadores antes de utilizarlas para la toma de decisiones. En este sentido, a diferencia de los test de selección única, en el desarrollo de una prueba nueva de expresión escrita los esfuerzos van encaminados principalmente a garantizar la confiabilidad del instrumento: creación de una rúbrica, entrenamiento de los calificadores, creación de un manual de calificación (con definiciones y ejemplos de cada aspecto evaluado), monitoreo constante durante todo el tiempo que tarden los jueces en calificar los textos, reuniones posteriores a la calificación para recibir recomendaciones por parte de los evaluadores y, en general, un control estricto de todos aquellos aspectos que puedan provocar sesgos no deseados en las puntuaciones de quienes realizan la prueba (Johnson, Penny, & Gordon, 2009; Breland, Bridgeman, & Fowles, 1999). Para una revisión más detallada, particularmente de los requisitos psicométricos que debe cumplir una prueba de este tipo, véase el trabajo de Jonsson y Svingby (2007).

Al respecto, Brown (2009) plantea que para lograr resultados válidos en el uso de rúbricas es necesario que los jueces califiquen de manera equitativa y con base en los criterios presentes en la rúbrica. Para ello, los evaluadores (con una formación base que les permita realizar la labor de manera adecuada) han de ser entrenados con rigurosidad. Para este autor, la cantidad de ensayos que se califican, las distracciones presentes en el espacio físico donde se califica, la fatiga y el cambio constante de un tema a otro influyen en gran medida sobre los jueces. Respecto de la cantidad de jueces mínima requerida para obtener una calificación confiable, Brown recomienda la participación de al menos dos jueces para calificar una misma prueba. No obstante, es posible utilizar otros diseños más complejos de juzgamiento, en función de la cantidad de textos por revisar, los recursos económicos para pagarles a los jueces y el tiempo disponible para llevar a cabo esta tarea (Eckes, 2009).

Así las cosas, se puede concluir que la tarea de crear una prueba de producción textual está sometida a dificultades particulares que es necesario resolver si se desea contar con un instrumento confiable (Knoch, 2007; Sasaki & Hirose, 1999; East, 2009). De esta manera, cabe señalar que el interés por evaluar la competencia comunicativa mediante un rúbrica fundamentada en las teorías contemporáneas sobre el procesamiento del lenguaje escrito, involucra necesariamente una labor integradora de distintos enfoques, a saber: psicología cognitiva, lingüística y psicometría. Mientras que por un lado las perspectivas cognitiva y lingüística permiten comprender de qué manera los sujetos abordan la tarea de redactar un texto, por el otro las herramientas de la psicometría enfatizan un aspecto fundamental en todo esfuerzo por utilizar pruebas para la toma de decisiones: garantizar la confiabilidad y la validez de las puntuaciones derivadas del uso de la prueba.

Tal y como se expuso al inicio, el objetivo fundamental de esta investigación es el de generar una rúbrica para evaluar la producción textual de estudiantes en contextos académicos. En un estudio anterior (Moreira, 2008) se presentaron los resultados de un intento por desarrollar una rúbrica holística para medir la expresión escrita, sin embargo no fue posible alcanzar el nivel mínimo necesario de concordancia entre los evaluadores. En dicho estudio se recomendó mejorar los procesos de capacitación de los jueces, así como la operacionalización de los aspectos por evaluar. A continuación se expondrán los resultados obtenidos en el desarrollo de una rúbrica analítica para medir la misma habilidad (competencia comunicativa), con base en los aportes teóricos de la lingüística textual.

Características de la escritura académica

En el contexto del Proyecto de Expresión Escrita (encargado de construir pruebas asociadas a la producción textual en el contexto del Programa de Pruebas Específicas), interesa evaluar las habilidades y conocimientos que posee el estudiante para utilizar un registro formal adecuado respecto del contexto universitario. En este sentido, de un alumno de educación superior se espera el dominio de todas las estructuras lingüísticas propias de la norma de uso estándar del español de acuerdo con los criterios de la Real Academia Española. Ciertamente otras normas dialectales son igualmente útiles para comunicarse en el contexto universitario nacional, sin embargo, internamente se espera que el educando (tanto en su expresión oral y escrita, y en los niveles fonético, morfosintáctico, léxico y discursivo) utilice un conjunto específico de recursos idiomáticos asociados al registro formal.

Dado que el objetivo de la rúbrica es el de ser utilizado en contextos universitarios, es importante exponer cuáles son las características de la escritura académica. Cabe señalar que la siguiente exposición corresponde a los rasgos principales del texto escrito académico (artículos científicos, ensayos, informes, cartas formales, etc.), el cual "se constituye como ejemplo de una escritura reflexiva que ha de cumplir los requisitos de imparcialidad, desapasionamiento, neutralidad y distancia" (Calsamiglia & Tusón, 2007, p. 82). Finalmente, en todo momento se mantiene la oralidad como punto de comparación para identificar los rasgos propios del texto escrito.

En el nivel gráfico se utilizan los signos gráficos normalizados e impuestos desde un centro de poder (en el caso del español, estos corresponden a los lineamientos de la Real Academia Española). Al respecto, Rodino y Ross (1997) enfatizan que en el registro formal escrito se espera un estricto cumplimiento de la norma estándar sobre el uso "correcto" de la lengua. Por tal razón, existe la exigencia de un empleo riguroso de las letras para representar los fonemas, convenciones de espaciamiento, mayúsculas y otros signos especiales de notación.

En el nivel sintáctico se observa una preponderancia de oraciones cuyo orden se ajusta al canónico (sujeto - verbo - complementos). Asimismo, es frecuente el uso de construcciones impersonales y pasivas con las que se intenta expresar objetividad borrando la presencia de la primera y segunda personas gramaticales. A diferencia del discurso oral espontáneo, en el registro escrito formal también se suelen emplear las oraciones subordinadas.

Por su parte, en el nivel léxico se espera, por parte del escritor, la explicitud del mensaje. Dadas

sus características, el texto escrito debe presentar un mayor grado de autonomía respecto de quien lo enuncia, por lo que es necesario utilizar secuencias más explícitas (aplicables a menos cosas) y, por ello, que evite en la medida de lo posible la ambivalencia. Esto implica también el uso mayoritario de piezas léxicas semánticamente *llenas* ("hacer el examen" Vs. "contestar el examen"). Así pues, en general, el carácter diferido de la comunicación escrita impone la necesidad de un mayor grado de explicitud con respecto a la comunicación oral. Al respecto, Núñez y del Teso (1996) enfatizan los dos factores definitorios de la incertidumbre involucrada en la producción y recepción de mensajes escritos:

> Uno es la no coincidencia en el espacio ni en el tiempo de los procesos de emisión y recepción. Otro es el carácter no evanescente de las señales escritas, que hace que lo expresado por escrito pueda ser leído un número indefinido de veces por un número indefinido de receptores en un número indefinido de situaciones. (p. 28)

En el nivel textual, un texto escrito se caracteriza por explicitar las relaciones semánticas y conceptuales entre los enunciados a través de marcadores discursivos, uno de los mecanismos más importantes para elaborar textos cohesivos. Además, hay poca o nula presencia de redundancias y palabras repetidas innecesariamente. Se espera que cada pieza informativa sea pertinente (relevante), ya que la permanencia propia del medio escrito hace innecesaria la repetición constante de los referentes. En este sentido, Rodino y Ross (1997) plantean que, aun cuando sea necesario cierto grado de repetición para permitir que el lector no pierda de vista los referentes del texto, "este grado aceptable y necesario de redundancia de significados *no implica repetición idéntica de los significantes* (léxico y estructuras sintácticas)" (p. 31). En este sentido, quien escribe debe procurar un avance informativo continuo, lo cual provoca que la densidad informativa sea sumamente alta en comparación con los textos orales.

Finalmente, en cuanto al nivel textual, cabe señalar que en todo discurso escrito se segmenta la información en párrafos, capítulos y apartados, lo cual obliga al uso de signos de puntuación para favorecer la comprensión. Desde una perspectiva normativista tradicional, estos han sido abordados como parte del nivel ortográfico de la lengua, sin embargo, desde un enfoque funcionalista (Halliday & Matthiessen, 2004), cumplen la función de marcar la estructura jerárquica de los textos: grupos oracionales o frases (comas), cláusulas (punto y coma o punto y seguido), complejos clausales (puntos finales de párrafo) y apartados (punto final).

En síntesis, la presente investigación plantea el desarrollo y validación de una rúbrica que permita operacionalizar el enfoque teórico expuesto. Por tal razón, es necesario acudir a las herramientas y metodologías que la psicometría moderna ha desarrollado en aras de construir instrumentos de medición de alta calidad que no solo se fundamenten en una propuesta teórica sólida, sino que también permitan un proceso de evaluación riguroso y objetivo.

Método

Participantes

Los datos fueron recogidos en marzo de 2013. La muestra está conformada por 117 estudiantes universitarios (75 mujeres y 42 hombres) matriculados en el curso Introducción a la filología de la carrera de Filología Española de la Universidad de Costa Rica. Esta materia pertenece al bloque curricular correspondiente al primer semestre del primer año de estudio, por lo cual una gran cantidad de estudiantes eran de primer año de ingreso (41%) o llevaban a lo sumo un año (19%) o dos (11%) en la institución. De este modo, aproximadamente el 70% de los participantes habían estado poco tiempo en la universidad.

Instrumentos

El instrumento utilizado para recolectar los datos constó de las instrucciones, dos hojas para que los estudiantes hicieran una versión preliminar de un ensayo y dos hojas más para la versión final. En la hoja de intrucciones se le indica al estudiante que dispone de una hora y 30 minutos para escribir un

ensayo de al menos 100 palabras con alguno de los siguientes temas: "la lectura como ventana al mundo", "un personaje histórico importante", "¿qué se puede esperar de la formación en la Universidad de Costa Rica?", "Las tecnologías de comunicación actuales afectan la ortografía del español" y "la cultura ambiental y el desarrollo representan un desafío para la humanidad".

Además, se construyó una rúbrica para evaluar los ensayos. Con base en el enfoque expuesto en la introducción, se decidió evaluar los siguientes aspectos: ortografía, morfología, sintaxis, léxico, construcción de párrafos, cohesión, registro y coherencia. Para cada uno de estos rubros se creó un espacio donde se pudiera anotar el número de errores cometidos. Asimismo, en la rúbrica se presenta una serie de casillas para consignar el nivel alcanzado por el estudiante en una escala del 1 (deficiente) al 5 (excelente), de modo que, a menor cantidad de errores, mayor es el nivel asignado en cada rubro.

Finalmente se elaboró un manual de calificación en el que se explica en cómo debe interpretarse cada aspecto de la rúbrica y se brindan ejemplos de errores representativos para cada rubro. El objetivo de este documento fue el de servir de guía en todo momento a las calificadoras, quienes lo tuvieron a su disposición (junto con la rúbrica) durante todo el proceso que se explicará a continuación.

Procedimiento

Una vez recogidos los ensayos durante la segunda semana de marzo de 2013, se procedió a contratar a dos calificadoras para que los evaluaran. Las personas encargadas de esta labor fueron dos estudiantes avanzadas (graduadas del bachillerato en Filología Española).

Antes de iniciar con la calificación de los ensayos, ambas se reunieron con el investigador principal de este proyecto, quien les explicó cómo debía utilizarse la rúbrica, así como también la forma de utilizar el manual. Se realizaron dos sesiones para estas labores previas: en la primera, el investigador les presentó a las calificadoras los instrumentos y les proporcionó tres ensayos para que los evaluaran; en la segunda, se llevó a cabo un grupo de discusión para determinar si estos habían sido calificados correctamente, de acuerdo con la rúbrica.

Finalizada la capacitación, en la segunda semana de abril de 2013 las calificadoras iniciaron con la evaluación de los ensayos. Esta etapa concluyó en la segunda semana de agosto de 2013. Es importante señalar que cada calificadora realizó su labor de manera independiente, sin que en ningún momento el investigador principal interviniera en el proceso ni hubiera ningún tipo de intercambios entre las calificadoras sobre la forma en que estaban evaluando los ensayos. Para garantizar la independencia de las evaluaciones, se les solicitó a estas personas que la calificación solamente podría hacerse dentro de las instalaciones del proyecto de investigación.

Análisis

Para evaluar la calidad técnica de la rúbrica, se calcularon algunos estadíticos descriptivos, los porcentajes de acuerdo absoluto entre las calificadoras para cada rubro así como los coeficientes correlación entre los diferentes aspectos evaluados. El objetivo de los tres tipos de análisis utilizados es el de brindar evidencias de la confiabilidad de la rúbrica a la hora de ser utilizada para evaluar textos escritos.

Resultados

Estadísticos descriptivos

En la Tabla 1 se presentan algunos estadísticos descriptivos para los diferentes aspectos incluidos en la rúbrica. Para interpretar adecuadamente la Tabla 1 se debe tomar en cuenta que cada rubro se evaluó de acuerdo con el número de errores identificado por cada calificadora (señaladas como C1 y C2 en la Tabla 1). Así, por ejemplo, en la primera columna de la Tabla 1 se presenta la media de errores identificados por cada evaluadora, mientras que en la segunda columna se puede observar la mediana para cada rubro.

Es importante tomar en cuenta los altos valores de asimetría y curtosis en la gran mayoría de rubros. En general, la mayoría de sujetos cometió pocos errores (entre 0 y 10); sin embargo, en algunos en-

sayos fue posible identificar una gran cantidad de errores. Esta situación provocó que la distribuición de las variables (los rubros) no fuera normal, así como también la variabilidad observada se redujo en gran medida debido a esta poca cantidad de errores de redacción en la gran mayoría de estudiantes. Otro aspecto que se debe destacar en la tabla es la consistencia entre los valores para cada calificadora, lo cual indica que la concordancia entre ellas fue aceptable en la mayoría de los aspectos evaluados.

Aún cuando hay una cierta consistencia entre los estadísticos descriptivos calculados para cada evaluadora, es necesario utilizar una técnica más informativa sobre el nivel de concordancia entre estas. Este aspecto es de suma importancia, ya que para garantizar un nivel mínimo de confiabilidad en el proceso de calificación se requiere que los encargados de calificar difieran lo menos posible en la calificación asignada a cada ensayo. Como se verá a continuación, algunos rubros generaron un alto nivel de concordancia entre las calificadoras mientras que para otros, dicho nivel fue relativamente bajo.

Porcentajes de acuerdo absoluto

En la Tabla 2 se presenta un análisis descriptivo inicial de la concordancia observada entre las calificadoras. La primera columna contiene el código asignado a cada subrubro, a saber: O (ortografía), S (sintaxis), M (morfología), P (construcción de párrafos), CR (coherencia), CS (cohesión), L (léxico) y R (registro formal o informal). Estos se organizaron en función de la concordancia observada: al principio los que permitieron un mayor acuerdo entre las calificadoras y al final los que generaron menos corcondancia. En el encabezado, se presentan las diferencias en cuanto al número de errores identificados por ambas calificadoras, de manera que cuanto más a la derecha, mayor es la cantidad de errores de diferencia para cada subrubro. Finalmente, los valores de esta tabla corresponden al porcentaje de estudiantes (para cada subrubro) en los que se presenta una discrepancia determinada. Por ejemplo, en el 65.5% de los ensayos no se observó ninguna discrepancia en el rubro S3 respecto del número de errores cometidos por los participantes. Como se puede observar, la máxima discrepancia en este aspecto de la sintaxis fue de dos errores.

Un indicador de que los ensayos fueron evaluados de manera confiable es que en todos los rubros (excepto el CS2, correspondiente a los errores de signos de puntuación) aproximadamente el 90% de los ensayos muestran una discrepancia de apenas tres errores. En otras palabras, las calificadoras lograron identificar los errores con un alto grado de concordancia entre ellas. Por otro lado, son espe-

Tabla 1.
Estadísticos descriptivos (N = 117)

Rubro	Media		Mediana		Desviación estándar		Asimetría		Curtosis		Mínimo		Máximo	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
Ortografía	5.6	6.2	3	4	7.3	7.6	2.8	2.7	10.8	10.2	0	0	48	50
Morfología	0.9	1.6	1	1	1.3	1.4	2.6	2	9.5	1.5	0	0	8	7
Sintaxis	3.6	5	3	4	2.7	3.5	0.9	0.8	0.8	0	0	0	12	14
Vocabulario	2.2	2.3	1	1	2.3	2.5	1.8	1.7	4.7	3.1	0	0	14	13
Párrafos	3.2	4.2	3	4	2.2	3.1	0.8	0.8	0.3	0.6	0	0	10	16
Cohesión	10.5	8.1	9	7	7	5.7	2.5	1.6	13.7	4.4	0	0	56	35
Registro	1.3	0.9	0	0	2.4	1.7	3.9	2.9	21	10.1	0	0	18	10
Coherencia	2.2	4.5	2	4	2.3	3.9	1.8	1.2	5.9	1.8	0	0	14	18
Total	29.5	32.8	26	29	16	17.9	1.5	1.1	3.1	1.3	7	3	96	92

Fuente: elaboración propia

Tabla 2.
Porcentajes de acuerdo absoluto entre calificadoras

Rubro	Discrepancias										
	0	1	2	3	4	5	6	7	8	9	≥ 10
S3	65.5	89.7	100								
S5	46.6	92.2	99.1	100							
S1	52.6	81	96.6	99.1	100						
S2	52.6	85.3	95.7	99.1	100						
S4	56.9	88.8	95.7	98.3	100						
P2	51.7	83.6	94.8	98.3	100						
M1	35.3	71.6	88.8	94.8	99.1	100					
P3	47.4	78.4	89.7	94.8	98.3	99.1	100				
O3	75.9	90.5	94	97.4	98.3	99.1	100				
CR1	33.6	56	73.3	89.7	95.7	99.1	100				
P1	52.3	80.2	90.5	94.8	98.3	99.1	99.1	100			
O1	56	83.6	92.2	97.4	97.4	98.3	99.1	100			
O2	37.9	69	87.9	96.6	97.4	98.3	99.1	100			
CS3	39.7	65.5	91.4	98.3	98.3	98.3	98.3	100			
L2	44	77.6	89.7	94.8	96.6	98.3	99.1	99.1	100		
S6	34.5	74.1	86.2	93.1	96.6	98.3	98.3	99.1	100		
CS1	36.3	69.8	84.3	92.2	96.6	98.3	99.1	99.1	99.1	100	
L1	39.7	73.3	83.6	92.2	93.1	93.9	94.8	98.3	98.3	100	
CR2	32.8	62.9	75.8	89.7	94	96.6	97.4	98.3	98.3	98.3	100
CS2	12.9	34.5	51.7	63.8	79.3	85.3	89.7	94	95.7	97.4	98.3*
R1	46.6	78.4	87.9	94	96.6	96.6	97.4	99.1	99.1	99.1	99.1**

*13 errores de discrepancia para abarcar al 100% de los sujetos.
**16 errores de discrepancia para abarcar al 100% de los sujetos.
Fuente: elaboración propia

rables los bajos porcentajes correspondientes a una discrepancia de 0 errores, ya que la escala utilizada por las calificadoras es continua (cantidad de errores), sin que se restrinja *a priori* el rango posible de errores que podían ser identificados.

En este punto es relevante señalar que los coeficiente kappa (Cohen, 1960, 1968), tanto el que se utiliza en escalas nominales, como el ponderado, no son apropiados en este caso para determinar el acuerdo entre las calificadoras debido, precisamente, a la corrección del azar que implica su uso. Dado que a las evaluadoras no se les dio una escala predeterminada para calificar, no es plausible que existan acuerdos por azar entre ellas. Para que tal situación hubiera sido posible, sería necesario haberles indicado que los errores oscilarían en un rango de 0 a *n* errores.

Coeficientes de correlación de Spearman.

En este punto es importante señalar que se utilizó el coeficiente de correlación de Spearman porque los rubros presentan distribuciones sumamente asimétricas, con valores altos de curtosis y con un rango reducido. Dado que esta situación afecta considerablemente el cálculo del coeficiente de correlación de Pearson, se prefirió utilizar la técnica de Spearman para brindar una idea más clara sobre la asociación entre los diferentes contenidos de la rúbrica.

En la Tabla 3 se presentan los coeficientes de correlación de Spearman entre los aspectos evaluados en la rúbrica. Para brindar una idea más clara de los patrones de correlación entre los contenidos evaluados, se sumaron los 21 subrubros presentados en la Tabla 2 de acuerdo con su relación con las diferentes

Tabla 3.
Coeficientes de correlación de Spearman

	Ortografía	Morfología	Sintaxis	Vocabulario	Párrafos	Cohesión	Registro
Morfología	**0.33** **0.37**						
Sintaxis	**0.34** **0.36**	**0.25** **0.38**					
Vocabulario	0.18 0.14	-0.05 0.25	**0.27** **0.50**				
Párrafos	0.04 0.04	0.07 0.06	0.18 **0.24**	0.03 0.15			
Cohesión	**0.44** **0.41**	**0.20** **0.31**	**0.35** **0.41**	**0.35** **0.25**	0.05 0.15		
Registro	0.01 0.01	-0.16 0.03	0.07 **0.27**	-0.05 0.04	**0.21** **0.21**	0.10 **0.26**	
Coherencia	0.02 0.03	0.03 0.06	**0.21** **0.33**	**0.28** 0.10	0.04 **0.49**	**0.21** **0.20**	0.13 **0.25**

*En cada rubro, la correlación superior corresponde al Juez 1 y la inferior, al Juez 2.
****Los coeficientes en negrita son estadísticamente significativos ($p < 0.05$)**
Fuente: elaboración propia

áreas incluidas en la rúbrica. Por ejemplo, los rubros S1, S2, S3, S4 y S5 en su conjunto se relacionan con los errores sintácticos cometidos por los estudiantes, lo cual justifica crear una única variable denominada *Sintaxis* en la Tabla 3. Por otro lado, esto mejora los cálculos de los coeficientes de correlación dado que una variable calculada como la sumatoria de dos o más variables siempre mostrará una mayor variabilidad que la de las variables que la componen.

En general, la Tabla 3 muestra que para ambas calificadoras los contenidos evaluados muestran correlaciones moderadas o bajas y en muchos casos cercanas a cero. Tal situación es esperable por cuanto los contenidos (todos ellos indicadores de la competencia comunicativa de los estudiantes) evaluados en la rúbrica difieren en cuanto a los conocimientos y destrezas requeridos para su uso.

Conclusiones

La expresión escrita continúa siendo un aspecto que se valora como importante para el logro académico, sin embargo, es una de las grandes debilidades de la población estudiantil universitaria. De hecho, las mismas calificadoras fueron enfáticas en reportar que la tarea de evaluar los ensayos fue sumamente difícil debido a la falta de claridad que mostraron los estudiantes para exponer sus ideas. Otro aspecto que dificulta la lectura de textos escritos por estudiantes es la caligrafía, la cual en ocasiones se vuelve ininteligible. Por ello, es necesario implementar procesos de enseñanza-aprendizaje más efectivos que se enfoquen en la expresión escrita.

La capacidad para comunicarse mediante el lenguaje escrito es ciertamente compleja, lo que no significa que sea imposible de evaluar en contextos aplicados. No obstante, la medición de esta competencia demanda más recursos que la de otras habilidades, puesto que se requiere de calificadores expertos que proporcionen juicios consistentes y concordantes entre sí. Ciertamente, el interés por desarrollar un instrumento de evaluación que focalice habilidades complejas demanda una mayor cantidad de recursos, por lo cual los formatos de pruebas de selección única prevalecen en el ámbito de pruebas estandarizadas. Sin embargo, en la presente investigación se demuestra que es posible generar evidencias de validez y confiabilidad en pruebas de respuesta construida, sabiendo que la subjetividad de la evaluación de los jueces siempre será un factor que en mayor o menor medida afecta los puntajes de los sujetos.

Evaluar la competencia comunicativa mediante una rúbrica fundamentada en las teorías contemporáneas sobre el procesamiento del lenguaje escrito involucra necesariamente una labor integradora de los enfoques de la psicología cognitiva, la lingüística y la psicometría. El interés fundamental de esta investigación fue el de desarrollar una rúbrica que no solo permitiera llegar a un cierto nivel de concordancia entre los jueces, sino que también tuviera un sustento teórico en las propuestas contemporáneas sobre los aspectos importantes en la producción escrita de lenguaje. Aún cuando existe una gran variedad de constructos y enfoques teóricos en el ámbito de la medición de la expresión escrita (Jeffery, 2009; Behizadeh & Engelhard, 2011), la lingüística textual mostró ser una propuesta lo suficientemente sólida como para que se pueda operacionalizar en rubros de evaluación, lo cual redunda en una mejor integración con los estándares psicométricos que debe cumplir todo instrumento de medición educativa.

Para evaluar la producción escrita mediante jueces, no solamente se requiere de una rúbrica, sino que también se necesita un manual con definiciones y ejemplos de cada contenido evaluado, así como un también un proceso de entrenamiento adecuado. Es importante señalar que el proceso de validación expuesto cumplió en términos generales con las recomendaciones usuales para construir este tipo de instrumentos. Sin embargo, al ser consultadas sobre cómo mejorarían el proceso de calificación de los ensayos, las evaluadores recomendaron lo siguiente: 1) aclarar aún más cómo deben interpretarse los errores de morfología y sintaxis, 2) aumentar el tiempo de entrenamiento previo a la calificación, 3) tomar en cuenta el número total de palabras empleado a la hora de calcular la nota de cada estudiante, 4) incorporar un rubro que evalúe si el estudiante siguió las instrucciones a la hora de redactar el ensayo y 5) sugerir posibles temas para que el estudiante escoja uno y que se le solicite escoger un título relacionado con ese tema.

Finalmente, en lo que respecta a la concordancia entre las calificadoras, es importante mencionar que los valores observados apuntan a que las evaluadoras calificaron básicamente lo mismo. Esto último no solo se relaciona con la confiabilidad del instrumento, sino que también se vincula con el tema de la validez de las inferencias que se pueden realizar a partir de los puntajes obtenidos en esta prueba. A diferencia de los test de selección única, en los cuales la confiabilidad y la validez son temas que deben abordarse por separado, en esta prueba los coeficientes de confiabilidad indican no solamente que el error de medición se mantuvo en un nivel aceptable, sino que también fue posible medir el constructo de interés. Si cada calificadora se hubiera concentrado en evaluar diferentes constructos, es poco probable que hubieran llegado al nivel de acuerdo alcanzado.

En futuras investigaciones se incluirán las observaciones de las calificadoras, así como también otras mejoras para disminuir los desacuerdos observados en los referidos a la sintaxis y a los signos de puntuación. Además, se incorporarán tareas más específicas como las de redactar textos argumentativos, narrativos y descriptivos en aras de indagar sobre posibles carencias para construir diferentes tipos de textos.

Referencias

Beauvais, C., Olive, T. & Passerault, J. M. (2011). Why Are Some Texts Good and Others Not? Relationship Between Text Quality and Management of the Writing Processes. *Journal of Educational Psychology*, *103*(2), 415–428.

Behizadeh, N., & Engelhard Jr, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. Assessing writing, 16(3), 189-211.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9.

Breland, H. M., Bridgeman, B & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework*. New York: College Entrance Examination Board.

Brown, G. T. (2009). The reliability of essay score: the necessity of rubrics and moderation. In L. Meyer et al. (Eds.), *Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice, and*

Research (pp. 43-50). Wellington, New Zealand: Ako Aotearoa.

Calsamiglia, H., & Tusón, A. (2007). *Las cosas del decir*. España: Ariel.

Carroll, J. (1993). *Human Cognitive Abilities*. New York: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Crossley, S. A, & McNamara, D. S (2010). Cohesion, Coherence, and Expert Evaluations of Writing Proficiency. *Proceedings of the 32nd annual conference of the Cognitive Science Society*, 984-989.

De Beaugrande, R. (1984). *Text Production*. Recuperado de http://www.beaugrande.com/text_production.htm

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88-115.

Eckes, T. (2005). Examining raters effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Language assessment quarterly 2*(3), 197-221.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy Division.

Halliday, M. & Matthiessen, C. (2004). *An Introduction to Functional Grammar*. London: Hodder Arnold.

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing, 14*(1), 3-24.

Johnson, R. L. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review 2*(2), 130-144.

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two ratings scales for the assessment of coherence. *Assessing writing, 12*(2), 108-128.

Lomas, C. (1999). *Cómo enseñar a hacer cosas con las palabras*. España: Editorial Paidós.

Louwerse, M., Graesser, A., McNamara, D., Jeuniaux, P., & Yang, F. (2006). Coherence is also in the eye of the beholder. In M. Silva, & A. Cox (Eds.), *Proceedings of the Cognitive Science Workshop "What have eye movements told us so far, and what is next?"*. Recuperado de 129.219.222.66:8080/SoletlabWeb/pdf/CoherenceEyeBeholder.pdf

McNamara, D., Kintsh, E., Butler, N., & Kintsh, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction, 14*(1), 1-43.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Writing Communication 27*(1), 57-88.

Montero, E. (2007). Teoría G: un futuro paradigma para el análisis de pruebas psicométricas. *Actualidades en Psicología, 21*, 117-144.

Moreira, T. (2008). Construcción y validación de pruebas de expresión escrita en la Universidad de Costa Rica. *Avances en medición*, 6(*1*), 85-100.

Núñez, R., & del Teso, E. (1996). *Semántica y pragmática del texto común*. España: Ediciones Cátedra.

Olive, T. (2004). Working Memory in Writing: Empirical Evidence From the Dual-Task Technique. *European Psycologis, 9*(1), 32-42.

Olive, T., Kellogg, R. T., & Piolat, A. (2008). Verbal, visual, and spatial working memory demands during text composition. *Applied Psycholinguistics*, *29*(4), 669-687.

Parodi, G., & Núñez, P. (2002). En búsqueda de un modelo cognitivo/textual para la evaluación del texto escrito. En M. Martínez (Ed.), *Propuesta de intervención pedagógica para la comprensión y producción de textos académicos* (pp. 65-98). Cátedra UNESCO MECEAL: Lectura y Escritura. Recuperado de http://www.unesco-lectura.univalle.edu.co/articulos.html

Revelle, W. (s. f.). *An introduction to psychometric theory with applications in R*. Recuperado de http://personality-project.org/r/book/.

Rodino, A., & Ross, R. (1997). *Problemas de Expresión Escrita del Estudiante Universitario Costarricense*. Costa Rica: EUNED.

Sánchez, C. (2004a). Historiografía de la enseñanza de la redacción en Costa Rica: los libros de texto. *Revista de Filología y Lingüística*, *20*(*1*), 219-246.

Sánchez, C. (2004b). La puntuación y las unidades textuales: una perspectiva discursiva para el estudio de los problemas de su uso y para su enseñanza. *Revista Educación*, *28*(*2*), 233-254.

Sánchez, C. (2005a). Los problemas de redacción de los estudiantes costarricenses: una propuesta de revisión desde la lingüística del texto. *Filología y Lingüística*, *21*(*1*), 267-295.

Sánchez, C. (2005b). Los conectores discursivos: su empleo en redacciones de estudiantes universitarios costarricenses. *Filología y Lingüística*, *21*(*2*), 169-199.

Sánchez, C. (2006a). Historia de un desencuentro: investigación y enseñanza de la redacción en Costa Rica. *Revista de Filología y Lingüística*, *22*(1), 223-245.

Sánchez, C. (2006b). ¿Cuestión de método? Sobre los cursos remediales universitarios de expresión escrita. *Revista Educación*, *30*(1), 65-81.

Sánchez, C. (2007). Los objetivos de la instrucción gramatical en la enseñanza del español como lengua materna. *Filología y Lingüística*, *33*(1), 167-190.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16*(4), 457-478.