PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-17

Pontificia Universidad Católica de Chile
Facultad de Educación

Evaluación de calidad de calificaciones en exámenes escritos a través del modelo multifocal de lente y la teoría de medición de Rasch

Using a Multifocal Lens Model and Rasch Measurement Theory to Evaluate Rating Quality in Writing Assessments

Jue Wang y George Engelhard, Jr.
The University of Georgia, USA

Resumen

Los modelos de lente han sido usados extensivamente para examinar los juicios humanos. Las aplicaciones de estos modelos han ocurrido en una variedad de contextos que incluyen juicios de logros de lectura, diagnósticos clínicos y características personales. Los exámenes escritos sujetos a calificaciones mediadas involucran juicios humanos hacia los ensayos de los estudiantes. Los procesos de juicio determinan la calidad de las calificaciones, y esto afecta directamente la imparcialidad y validez de las calificaciones. El modelo de lente provee un marco teórico con el que se puede evaluar los juicios en exámenes sujetos a calificaciones mediadas. La teoría de medición de Rasch ofrece un acercamiento metodológico alternativo con el modelo multifocal de lente. Para poder ilustrar nuestro acercamiento a esta teoría, fueron examinadas las calificaciones de tres expertos calificadores y 20 calificadores operacionales de un programa estatal de apoyo a la escritura en los Estados Unidos. Hay solo un 35% de estudiantes con competencias comparables en escritura, medidas en base a la comparación de calificaciones de calificadores operacionales y calificadores expertos. La combinación de un modelo de lente multifocal con la teoría de medición de Rasch ofrece nuevas maneras de entendimiento sobre la calidad de calificaciones mediadas por evaluadores.

Palabras clave: Modelo multifocal de lente; Teoría de medición de Rasch; evaluaciones

Correspondencia a:
Jue Wang
126C Aderhold Hall,
110 Carlton St., Athens, GA 30602
Email: cherish@uga.edu
Tel: 1.706.255.5379

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.3

Abstract

Lens models have been used extensively for the examination of human judgments. Applications of lens model have occurred in a variety of contexts including judgments of reading achievement, clinical diagnosis, and personality characteristics. Rater-mediated writing assessments involve human judgments toward student essays. The judgmental process determines the quality of ratings, and this directly affects validity and fairness of ratings. Lens models provide a theoretical framework to evaluate rater judgment in rater-mediated assessments. Rasch measurement theory offers an alternative methodological approach for studying human judgments, and we propose combining Rasch measurement theory with a multifocal lens model. In order to illustrate our approach, the ratings of three expert raters and 20 operational raters from a state writing assessment program in the United States are examined in this study. There are only 35% of the students with comparable writing proficiency measures based on a comparison of ratings from operational raters and expert raters. The combination of a multifocal lens model with Rasch measurement theory offers a new pathway for understanding the quality of ratings for rater-mediated assessments.

Keywords: Multifocal lens model; Rasch measurement theory; rater-mediated assessments; writing assessments

Los juicios humanos juegan un rol fundamental en la evaluación de exámenes escritos a través de correctores, siendo la calidad de los juicios determinante para la fiabilidad, validez e imparcialidad de las correcciones. Los modelos de lente proporcionan un marco teórico para evaluar los juicios de los correctores en exámenes escritos. El modelo de lente originalmente fue presentado por Brunswik (1952, 1955a, 1955b, 1956), siendo adaptado por Hammond (1955a) enfocándose en el análisis de constancia perceptual. El modelo de lente fue adaptado por Hammond para evaluar el juicio clínico, y esta línea de investigación estableció el potencial para usar el modelo de lente para investigar los juicios humanos, incluyendo la teoría del juicio social (Doherty, 1996). Un metaanálisis realizado por Kaufmann, Reips, and Wittmann (2013) documentó el uso exhaustivo del modelo de lente para la examinación del juicio humano y la toma de decisiones en variados contextos.

Una característica que define a los modelos de lente es que incluyen una comparación entre dos sistemas: *un sistema ecológico y un sistema de juicios* (Cooksey, 1996; Hammond, Hursch, & Todd, 1964; Hursch, Hammond, & Hursch, 1964; Tucker, 1964). Por ejemplo, Cooksey (1986) usó el modelo de lente como método para comparar juicios de profesores respecto al rendimiento en lectura de estudiantes, en contraste a las calificaciones obtenidas por el rendimiento en la lectura en un test estandarizado, a través de una serie de indicadores. Cooksey (1986) definió el sistema ecológico en términos de la relación entre los puntajes de tests estandarizados de lectura y 3 indicadores (i.e. nivel socioeconómico, habilidad para la lectura, y habilidad oral en el idioma). De modo similar, definió el sistema de juicios para establecer la relación entre los juicios de los profesores y el mismo set de indicadores. El modelo de regresión es típicamente aplicado en estudios de modelos de lente, y los indicadores basados en la regresión se usan para comparar los sistemas ecológicos y de juicios. Cooksey, Freebody y Wyatt-Smith (2017) también utilizaron un modelo de lente para estudiar el juicio de los profesores en el rendimiento de la escritura.

En este estudio los juicios de expertos y de correctores operacionales han sido comparados, usando un modelo de lente multifocal; en este caso, hay dos sistemas de juicios siendo examinados. Específicamente, se presentan dos modelos de lente, y cada modelo incluye el sistema de juicios de cada grupo de examinadores, los cuales se muestran en la Figura 1. El Modelo 1 representa el modelo de lente para los calificadores expertos, los que definen el sistema de criterios en base a una serie de indicadores (dominio y escala de calificaciones). Los expertos se enfocan en los indicadores para realizar juicios (calificaciones expertas) mostrando como el rendimiento de los estudiantes en escritura debería definirse. El Modelo 2 en el panel de la derecha es una imagen espejo del Modelo 1. El Modelo 2 refleja los juicios de los calificadores operacionales en las evaluaciones escritas, indicando como la evaluación de la competencia de la escritura está efectivamente siendo realizada. Estos modelos son llamados modelos de lente ya que se asemejan a la forma en que la luz pasa a través de un lente, con los indicadores facilitando el enfoque a los juicios sobre la competencia en la escritura.

En el centro de la Figura 1 hay un set de ensayos de los estudiantes, que forman los puntos focales para los calificadores operacionales para estimar la competencia de los estudiantes en la escritura. Le llamaremos a esto, el modelo de lente multifocal porque el foco se encuentra en ensayos de estudiantes con dos sistemas de juicios (experto y operacional) que necesitan garantizar la congruencia entre los ensayos de los estudiantes y la competencia escrita. Engelhard (2013) propuso un modelo de lente para examinar la calidad de la evaluación, observando en sus trabajos previos, escalas usadas para examinar facetas de los juicios de los calificadores basándose en la teoría de medición de Rasch (Engelhard, 1992, 1994). El modelo de lente propuesto en este trabajo puede ser visto como una extensión de su trabajo, en el cual se conceptualizan dos sistemas de juicios separados, reflejados en mapas de Wright, que son a su vez estimados por separado para calificadores expertos y operacionales.

Estudios sobre modelos de lente previos usaron correlación y regresión múltiple como análisis para examinar los modelos. Hammond (1996) sugirió que los estudios actuales que utilizaban el enfoque de modelo de lente sobre enfatizaban el rol de la técnica de regresión múltiple, y que "el modelo de lente es indiferente es a priori indiferente, y su principio organizacional se utiliza en ciertas tareas bajo ciertas condiciones; considerando esto como un asunto empírico" (p. 245). En este estudio alentamos el uso de la teoría de medición de Rasch como un *principio organizacional*, debido a las múltiples ventajas de uso de este modelo.

Primero que todo, las calificaciones observadas en las evaluaciones escritas son ordinales, ya que considera más apropiado el uso de técnicas de análisis de datos categóricos, como el modelo de Rasch (Andrich, 1988; Linacre, 1989; Rasch, 1980; Wright & Masters, 1982). Segundo, el modelo de medición de Rasch provee de análisis a nivel del ítem, del calificador y personales, ofreciendo información detallada para cada dominio y categoría en el modelo de lente, así como también estimados lineales de la competencia en escritura. Esto refleja una transición desde la teoría de medición clásica hasta la moderna, utilizando nuevas reglas de medición (Embretson, 1996). Tercero, las propiedades invariables del modelo de Rasch proveen la oportunidad de examinar la comparabilidad entre dos mapas de Wright de los juicios de calificadores expertos y operacionales. Si un nivel apropiado de ajuste de modelo de datos es obtenido, pueden ser obtenidas propiedades invariables de la medición de Rasch, como por ejemplo las mediciones invariables de datos, e invariables de calibración de calificadores.

Objetivo

El objetivo de este estudio es describir el modelo de lente multifocal que puede ser usado para evaluar calidad de las calificaciones en evaluaciones escritas mediadas por un corrector. Investigamos la correspondencia entre calificadores expertos y operaciones, al conceptualizar la competencia escrita basada en los mapas de Wright. Los mapas de Wright son esquemas visuales que representan el modelo de lente usado por los calificadores en dos grupos distintos (experto y operacional) para evaluar la competencia de los estudiantes en escritura. En este estudio, modelos de medición de Rasch fueron usados para calibrar los datos (e.g. dominios y escalas de calificación). Para el modelo de lente multifocal, y para comparar los estimados de competencia escrita obtenidos de las calificaciones de los correctores expertos y operacionales.

Metodología

Participantes

La información fue recolectada de un programa de evaluación estatal de escritura para estudiantes de séptimo grado en el sudeste de los Estados Unidos. Una muestra al azar de 100 ensayos fue utilizada para este estudio, y las calificaciones operacionales fueron obtenidas de 20 evaluadores bien entrenados, en la revisión de estos 100 ensayos. Los expertos suelen ser especialistas en contenido, con muchos años de experiencia y un profundo entendimiento en el área de la escritura. Ellos diseñan y entregan el entrenamiento y práctica a otros calificadores antes de pasar a la etapa de calificación operacional. Los calificadores operacionales son aquéllos que reciben entrenamiento a manos de calificadores expertos, realizando una evaluación operacional.

Procedimientos

Dos modelos de muchas facetas de Rasch (MFRM) por separado fueron utilizados para examinar calificaciones expertas y operacionales en el programa computacional FACETS (Linacre, 2015). Estos MFRM representan un modelo de lente multifocal estimado por separado para ambos tipos de calificadores. La invariancia de las dificultades juzgadas en los dominios y estructuras de la escala de calificaciones fue comparada en ambos modelos. El primer modelo analizó las calificaciones de un panel de tres expertos, y el segundo analizó las calificaciones operacionales. En este estudio, cada MFRM incluyó tres facetas: competencia en la escritura, calificadores y dominios. La ecuación para modelo 1 y 2 puede ser expresada del siguiente modo:

$$\ln\left[\frac{P_{injk}^{(g)}}{P_{inj(k-1)}^{(g)}}\right] = \lambda_i^{(g)} - \theta_n^{(g)} - \delta_j^{(g)} - \tau_{jk}^{(g)} \qquad (1)$$

En donde

$P_{injk}^{(g)}$ = probabilidad de un estudiante *n* de recibir una calificación *k* en el dominio *j* por el calificador *i*.

$P^{(g)}_{inj(k-1)}$ = probabilidad de un estudiante *n* de recibir una calificación *k-1* en el dominio *j* por el calificador *i*.

$\lambda^{(g)}_{i}$ = ubicación lógito-escalar (i.e., severidad) del calificador *i,*

$\theta^{(g)}_{n}$ = ubicación lógito-escalar (i.e., competencia de escritura evaluada) del estudiante *n,*

$\delta^{(g)}_{j}$ = ubicación lógito-escalar (i.e., dificultad evaluada) del dominio *j,* y

$\tau^{(g)}_{jk}$ = t= parámetro de umbral que indica la dificultad de la categoría *k* relativa a la categoría *k*-1
para cada dominio;

(g) = indicador de grupo que posee dos valores 1 y 2. El valor 1 se refiere al **Modelo 1** con calificadores expertos y el valor 2 se refiere al **Modelo 2** con calificadores operacionales.

El logaritmo de la probabilidad de que un estudiante recibiese una calificación en categoría *k* en vez de *k* -1 dada su ubicación en la competencia escrita, la severidad del calificador y la dificultad del dominio fueron calculados. El parámetro umbral refleja la estructura de la escala de calificaciones, y no fue considerado como una faceta en el modelo.

Instrumento

El instrumento usado en este estudio es una evaluación escrita de estudiantes de séptimo grado. A los estudiantes se les pidió que escribieran un ensayo en base a cada tema. Una escala de calificación analítica fue utilizada para cada dominio: (a) desarrollo de la idea, organización y coherencia (Dominio IDOC), y (b) uso de lenguaje y convenciones (Dominio LUC). La escala de calificaciones para el Dominio IDOC tiene 4 categorías, mientras que LUC tiene 3.

Resultados

Los dos MFRM se muestran con mapas Wright para reflejar dos modelos de lente separados en la Figura 2. El mapa Wright para el modelo 1 con calificaciones expertas se muestra en el panel izquierdo, y el mapa Wright para el modelo 2, con calificaciones operaciones se muestra en la derecha. Cada mapa Wright muestra la distribución de la competencia en escritura y ubicación de dominio, así como también el uso de las categorías de calificación para cada dominio. En la columna de la competencia escrita (WP) se muestran las frecuencias de los examinadores en cada ubicación, y éstos se encuentran ordenados desde el examinador menos competente (medición logarítmica más alta) hasta el más competente (medición logarítmica más baja). En la columna de dominio, se muestran las ubicaciones del dominio IDOC y LUC, siendo ordenadas desde la más difícil (medición lógita más alta) hasta la menos difícil (medición lógita más baja) para que los examinadores consigan mayores puntuaciones. La estructura de la categoría de uso se realizó en dos columnas – RS.11 para Dominio IDOC y RS.2 para el Dominio LUC en los mapas Wright. Al comparar los dos mapas Wright y las medidas de los MFRM, se puede explorar las correspondencias entre los juicios expertos y de los calificadores, constituyendo la base para la comparación entre los dos sistemas de juicios.

La estadística general para cada faceta se muestra en la Tabla 1. La faceta del calificador y la faceta de dominio están centradas en 0. Las mediciones de competencia escrita estimadas obtenidas desde expertos tienen una media de 0,91 lógito con una desviación estándar de 3,06. La media de las mediciones de la competencia escrita en calificadores operacionales es de 0,55 logits con una desviación estándar de 2,95 logits. El ajuste próximo (Infit) y ajuste lejano (Outfit) de los errores (MSE) se usa para evaluar el ajuste del modelo (Linacre, 2015). Para ambos, Infit y Outfit, mientras más cerca estén las estadísticas del valor esperado de 1.00, más ajustado está el modelo. Los valores promedios para las mediciones de Infit y Outfit para todas las facetas fueron de 0,98 o 0,97 como se muestra en la Tabla 1, indicando buen ajuste del modelo. Un chi^2 significativo para la separación de ensayos en el Modelo 1 de $\chi^2(99)=797{,}2$ con $p< 0{,}05$, que la competencia en escritura tiene una variabilidad estadísticamente significativa (Linacre, 2015). Del mismo modo, los calificadores operacionales tienen un chi^2 de $\chi^2(99)=5483{,}2$ con $p< 0{,}05$, lo cual también indica variabilidad significativa en la competencia escrita entre estudiantes.

Basado en el modelo de lente multifocal, definimos fiabilidad como la correlación entre las competencias en escritura obtenidas de los dos modelos de lente. Estos índices tienen un alto nivel de fiabilidad con un coeficiente de correlación de 0,94. El Panel A de la Figura 3 muestra una relación lineal aproximada entre las competencias estimadas de los dos modelos. A pesar de que la relación es alta, el Panel B muestra importantes diferencias entre las competencias estimadas por los calificadores expertos y los operacionales. Usando una banda de diferencia de ± 0,50 logits para determinar la significancia substancial de las diferencias (Tennant & Pallant, 2007), hay discrepancias no despreciables entre los dos modelos para algunas de las mediciones estimadas de competencias. Hay sólo un 35% de estudiantes que presentan una diferencia relativamente baja en las mediciones de competencia entre -0,5 y 0,50 logits. La gran mayoría de los estudiantes presenta grandes diferencias: un 44% de los estudiantes recibió mediciones más bajas de parte de los calificadores operacionales al ser comparadas con las calificaciones de expertos, mientras que el 21% de los estudiantes recibió calificaciones más altas por parte de los evaluadores operacionales, al ser comparados con los calificadores expertos. Idealmente, el objetivo del entrenamiento para los calificadores operacionales es que se comporten como calificadores expertos con diferencias despreciables al estimar la competencia de los estudiantes. La desviación en la medición basada en calificaciones operacionales en relación a los expertos, indica que las operacionales podrían llevar a mediciones sesgadas de la competencia de los estudiantes, incluso con calificadores profesionales bien entrenados. Esto podría deberse a las diferencias entre los sistemas de juicio y modelo de lente utilizado por ambos tipos de evaluadores, que no se resuelven por completo durante el entrenamiento.

Estos análisis también muestran un funcionamiento distinto en la información entre los dominios y categorías de uso de los dos modelos. Al observar los mapas de Wright (Figura 2) se ve cómo los calificadores operacionales consideran la dificultad de dos dominios, de forma distinta a los expertos. La Tabla 2 muestra las mediciones de ubicación del Dominio IDOC (0,92 logits) y del Dominio LUC (-0,92) en el Modelo 1. Una medición más alta en una faceta de dominio indica que es más difícil para los estudiantes alcanzar calificaciones altas; por lo tanto, el Dominio IDOC es más difícil que LUC. La medición de la ubicación en el Modelo 2 es de 0,50 y -0,50 logits para IDOC y LUC, respectivamente. Aunque el Dominio IDOC parece ser más difícil que LUC en el Modelo 2, la ubicación de los dominios se encuentra más cercana, de modo que los calificadores operacionales evalúan la dificultad de dos

dominios de forma distinta a los expertos. El MSE Infit y Outfit es cercano a 1,00 en ambos, indicando un buen ajuste del modelo.

La información de categoría entregada por los dos modelos también varía. La Tabla 3 compara las categorías estadística y estructura en dos modelos para el dominio IDOC. Los expertos usaron una calificación 2 con mayor frecuencia que los calificadores operacionales (49% versus 39%). Los calificadores operacionales otorgaron una calificación de 4 de forma más frecuente que los expertos (10% versus 4%). La distancia entre los estimados umbrales adyacentes debería estar entre 1,40 y 5,00 logits, para que puedan ser considerados como distintivos (Linacre, 1998; Engelhard, 2013). En este análisis, las distancias de umbrales para el Dominio IDOC en los dos modelos se encuentran en este rango; por lo tanto, se puede decir que los umbrales son distintivos, y que cada categoría contiene una información única con respecto a la competencia en escritura. Como fue sugerido por Engelhard y Wind (2013), la estructura de la escala de calificaciones para el dominio IDOC usando los coeficientes de las categorías estimadas fue graficada (Tabla 3, Panel B). Aquí se muestran diferentes usos de distintas categorías entre calificadores expertos y operacionales. Para nuestro análisis, los umbrales estimados en el Modelo 1 se encuentran más dispersos que aquéllos del Modelo 2. Las curvas de probabilidad de categoría (Figura 4) y las curvas de información de categoría (Figura 5) del Dominio IDOC entre ambos modelos, confirma que el uso de la categoría por expertos es más disperso que el de los calificadores operacionales.

La Tabla 4 compara la información de categoría para el Dominio LUC en los dos modelos. Las proporciones de uso de cada categoría son comparables entre expertos y calificadores operacionales. Los umbrales estimados para la misma categoría en ambos modelos son también muy cercanos. La distancia entre los estimados umbrales adyacentes para el Dominio LUC en ambos modelos varía entre 1,40 y 5,00 logits, de modo que las tres categorías son importantes para la escala completa. Las estructuras de la escala de calificaciones del Dominio LUC (Tabla 4, Panel B) muestran usos bastante similares de las categorías de calificación entre ambos tipos de calificadores. Las curvas de probabilidad de categoría (Figura 4) y las curvas de información de categoría (Figura 5) del Dominio LUC entre los dos modelos también es comparable, indicando usos similares de la escala para el Dominio LUC entre los calificadores operacionales y expertos. En general, la categoría de uso de los calificadores operacionales es relativamente precisa al ser comparada con los calificadores expertos en el Dominio LUC, sin embargo, más discrepancias entre expertos y calificadores operacionales fueron encontradas en el Dominio IDOC.

Discusión

El modelo de lente multifocal representa una perspectiva prometedora y un enfoque para la evaluación de la calidad de las calificaciones en el contexto de tareas corregidas por un calificador. La teoría de medición de Rasch proporciona una nueva metodología para la examinación del modelo de lente multifocal, así como para la comparación de sistemas de juicios de calificadores expertos y operacionales. Como Karelaia (2008) dijo “la belleza simple del modelo de lente de Brunswik recae en el reconocimiento de que el juicio y criterio de una persona pueden considerarse como dos funciones separadas de datos disponibles en el ambiente de esa decisión” (p. 404). En este estudio, comparamos mediciones de competencia en escritura realizada por calificadores operacionales y expertos mediante el mismo set de datos (i.e medición de dominio y categoría de uso) en dos modelos de lente. Los

análisis entregan información para explorar la calidad de la calificación y los juicios de los calificadores con gran detalle.

Los resultados de este estudio indican que los calificadores operacionales otorgaron distintas calificaciones a las entregadas por los expertos, y que las escalas pueden ser vistas como calificaciones poco precisas que resultaron en distintos estimados de la competencia escrita de algunos estudiantes. Sólo un 35% de las mediciones de la competencia en escritura fue comparada bajo dos sistemas de juicios mediante evaluadores expertos y operacionales. Por lo tanto, calificaciones poco precisas podrían sesgar los estimados de la competencia escrita para el resto de los estudiantes. Los análisis de los datos revelaron distintas apreciaciones para la dificultad de los dominios y diferentes categorías de uso entre los calificadores operacionales y los expertos. Primero, los operacionales consideraban los dos dominios como cercanos en términos de dificultad en comparación a los calificadores expertos. Segundo, los calificadores operacionales tenían distintas categorías de uso de las escalas de calificaciones para el Dominio IDOC al ser comparados con los expertos. La estructura de la escala de calificaciones para el Dominio IDOC presentó categorías más amplias en el Modelo 1 (expertos) que en el Modelo 2 (operacionales). Para el Dominio LUC, los evaluadores operacionales fueron más consistentes en el uso de la escala de calificaciones que los expertos.

Investigaciones previas, como la de Sulsky y Balzer (1988) sugieren diversas formas de obtener "calificaciones verdaderas" en evaluaciones mediadas por un calificador, incluyendo el uso de escalas desarrolladas por un panel de expertos y las calificaciones promedio de los calificadores operacionales. En nuestros análisis, estimamos la competencia en escritura en el Modelo 1 basado en las calificaciones de los expertos, y la de los operacionales en el Modelo 2. En otras palabras, comparamos estas dos formas de crear "calificaciones verdaderas" para la competencia en escritura, concluyendo que no entregaban resultados comparables. Los sistemas de juicio usados por los expertos y calificadores operaciones no son iguales luego de haber pasado por un exhaustivo entrenamiento.

Una de las principales ventajas de la teoría de medición de Rasch, en comparación a las técnicas de regresión múltiple, es la obtención de mediciones estimadas de competencia en la escritura en una escala de intervalos. La propiedad de independencia local de la teoría de medición de Rasch asegura que las comparaciones sean válidas y razonables en base a mediciones invariantes. Por supuesto, esta propiedad de independencia local es contingente a la obtención de un buen ajuste de modelo, el cual fue garantizado en nuestro estudio. Además, el modelo de Rasch, como técnica para analizar datos categóricos, es más apropiado al momento de lidiar con respuestas con mediciones ordinales. Aún más importante, las clasificaciones de la información en estudios sobre juicio humano son categóricas en su estructura.

Hammond (1996) enfatizó que diversas metodologías pueden ser utilizadas en el marco del modelo de lente. Creemos que la teoría de medición de Rasch debería formar parte de la familia de métodos que actualmente se están usando con los modelos de lente. En este estudio, la teoría de Rasch fue usada para información separada en distintas categorías para cada dominio. Una familia de modelos de medición de Rasch models (Andrich, 1988; Linacre, 2015; Rasch 1980; Wright & Masters, 1982) puede ser usada para realizar los análisis, usando el marco propuesto para los diversos propósitos de investigación. Futuras investigaciones deberían explorar otras metodologías basadas en la teoría

de Rasch. Nestler y Back (2015) propusieron un modelo de ecuación estructural de clasificaciones cruzadas para evaluar un modelo de lente para juicios de personalidad. El modelamiento de ecuación estructural (SEM) puede incorporar algunos modelos de mediciones basados en la teoría de respuesta al ítem (IRT; Baker & Kim, 2004; Muthén & Asparouhov, 2013). A futuro, las investigaciones podrían enfocarse en la estimación de modelos de lente multifocales basadas en IRT con el marco de SEM.

En este estudio, no nos enfocamos en calificadores individuales. Cooksey (1986) en un estudio previo evaluó la predicción en general de los profesores sobre la competencia lectora en niños. El modelo de lente multifocal como marco teórico para juicios de calificadores puede ser usado para comparar calidad de las calificaciones entre evaluadores expertos y operacionales. Las discrepancias entre ambos modelos de lente pueden ser identificadas mediante el modelo de Rasch como una metodología empírica. La combinación de la teoría de medición de Rasch y el enfoque del modelo de lente puede ayudarnos a entender los juicios de los evaluadores, entregando soporte para mejorar el entrenamiento de los calificadores; por lo tanto, ayudando también a mejorar la confiabilidad, validez e imparcialidad de las calificaciones en el contexto de las evaluaciones realizadas por medio de correctores.

En resumen, este estudio describió la teoría de medición de Rasch para evaluar el modelo de juicio a través del uso del modelo de lente. El modelo de lente multifocal puede ser utilizado para juicios de calificadores para comparar la calidad de las calificaciones y juicios entre calificadores expertos y operacionales. Las discrepancias entre los dos modelos de lente pueden ser identificadas por el modelo de Rasch, como una metodología empírica. La combinación de la teoría de Rasch y el modelo de lentes puede ayudarnos a entender los juicios de los correctores, entregando orientación para mejorar el entrenamiento, mejorando por lo tanto la calidad de la corrección, así como el nivel de confianza, validez e imparcialidad, en el contexto de tareas escritas corregidas por un calificador.

El artículo original fue recibido el 4 de octubre de 2016
El artículo revisado fue recibido el 16 de marzo de 2017
El artículo fue aceptado el 18 de octubre de 2017

Referencias

Andrich, D. (1988). *Rasch Models for Measurement.* Newbury Park, CA: Sage.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Brunswick, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Brunswik, E. (1955a). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193-217.

Brunswik, E. (1955b). In defense of probabilistic functionalism: A reply. *Psychological Review, 62*(3), 236-242.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal*, *23*(1), 41-64.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. Academic Press.

Cooksey, R.W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analyzing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401-434.

Doherty, M. E., & Kurz, E. M. (1996). Social judgement theory. *Thinking & Reasoning*, *2*(2-3), 109-140.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56-70.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.

Engelhard, G., & Wind, S.A. (2013). *Rating Quality Studies using Rasch Measurement Theory.* College Board Research Report 2013-3.

Hammond, K.R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, 62, 255–262.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological review*, *71*(6), 438.

Hammond, K. R. (1996). Upon reflection. *Thinking & Reasoning*, *2*(2-3), 239-248.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological review*, *71*(1), 42.

Karelaia, N., & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens studies. *Psychological Bulletin, 134*(3), 404-426.

Kaufmann, E., Reips, U. D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PloS one*, *8*(12), e83528.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA press.

Apéndice

Tabla 1.
Resumen de estadísticas para los modelos Facets

	Modelo 1 (Expertos; n=3)			Modelo 2 (Calificadores; n=20)		
	Ensayos	Expertos	Dominios	Ensayos	Calificador	Dominios
Medición						
Media	.91	.00	.00	.55	.00	.00
SD	3.06	.86	1.31	2.95	.51	.70
N	100	3	2	100	20	2
Ajuste próximo MSE						
Media	.98	.98	.98	.98	.98	.97
SD	.64	.11	.04	.26	.11	.04
Ajuste externo MSE						
Media	.97	.97	.97	.97	.97	.97
SD	.69	.18	.07	.28	.12	.02
Estadísticas de separación	.89	.96	.99	.98	.91	>.99
Confiabilidad de separación	797.2*	56.8*	98.1*	5483.2*	217.3*	213.0*
Chi cuadrado (χ^2)	99	2	1	99	19	1
df						

Note: *p<.05; MSE representa el error cuadrático medio.

Tabla 2.
Domain statistics for Facets models

	Modelo 1 (Expertos; n=3)			Modelo 2 (Calificadores; n=20)		
Dominios	Medición (SE)	Ajuste próximo MSE	Ajuste externo MSE	Medición (SE)	Ajuste próximo MSE	Ajuste externo MSE
Desarrollo de ideas, organización y cohesión (IDOC)	.92 (.13)	1.01	1.02	.50 (.04)	1.00	.99
Uso de lenguaje y convención (LUC)	-.92 (.14)	.95	.93	-.50 (.05)	.95	.96

Note: SE representa errores estándar; MSE representa el error cuadrático medio.

Tabla 3.

Categorías estadísticas para el desarrollo de ideas, organización y dominio de cohesión

Panel A. Categorías estadísticas

	Modelo 1 (Expertos; n=3)			Modelo 2 (Calificadores; n=20)		
Categoría de puntuación	Proporción de uso	Umbral de medición Rasch-Andrich	Distancia	Proporción de uso	Umbral de medición Rasch-Andrich	Distancia
1	21%			23%		
2	49%	-4.27		39%	-3.47	
3	27%	.17	4.44	28%	.25	3.72
4	4%	4.10	3.97	10%	3.22	2.97

Panel B. Estructura de estimados de categorías de puntuación

Puntuaciones	1 (Más bajo)	2	3	4 (Más alto)
Umbrales estimados con calificaciones de expertos		τ_1 (-4.27 logits)	τ_2 (.17 logits)	τ_3 (4.10 logits)
Umbrales estimados con calificaciones de calificadores operacionales		τ_1 (-3.47 logits)	τ_2 (.25 logits)	τ_3 (3.22 logits)

Nota: La distancia se refiere a la diferencia entre la medición de umbrales de dos categorías adyacentes.

Tabla 4.

Categorías estadísticas para el uso de lenguaje y el dominio de convención

Panel A. Categorías estadísticas

	Modelo 1 (Expertos; n=3)			Modelo 2 (Calificadores; n=20)		
Categoría de puntuación	Proporción de uso	Umbral de medición Rasch-Andrich	Distancia	Proporción de uso	Umbral de medición Rasch-Andrich	Distancia
1	22%			25%		
2	50%	-2.26		48%	-2.29	
3	28%	2.26	4.52	27%	2.29	4.58

Panel B. Estructura de estimados de categorías de puntuación

Puntuaciones	1 (Más bajo)	2	3 (Más alto)
Umbrales estimados con calificaciones de expertos		τ_1 (-2.26 logits)	τ_2 (2.26 logits)
Umbrales estimados con calificaciones de calificadores operacionales		τ_1 (-2.29 logits)	τ_2 (2.29 logits)

Nota: La distancia se refiere a la diferencia entre la medición de umbrales de dos categorías adyacentes.

Modelo 1 Modelo 2

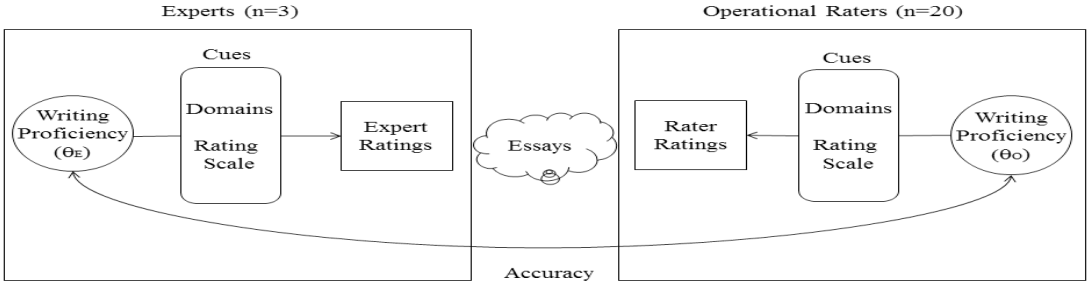


Figura 1. Modelo de lente multifocal para la evaluación de escritura.

Modelo 1
Expertos (N=3)

Modelo 2
Calificadores operacionales (n=20)

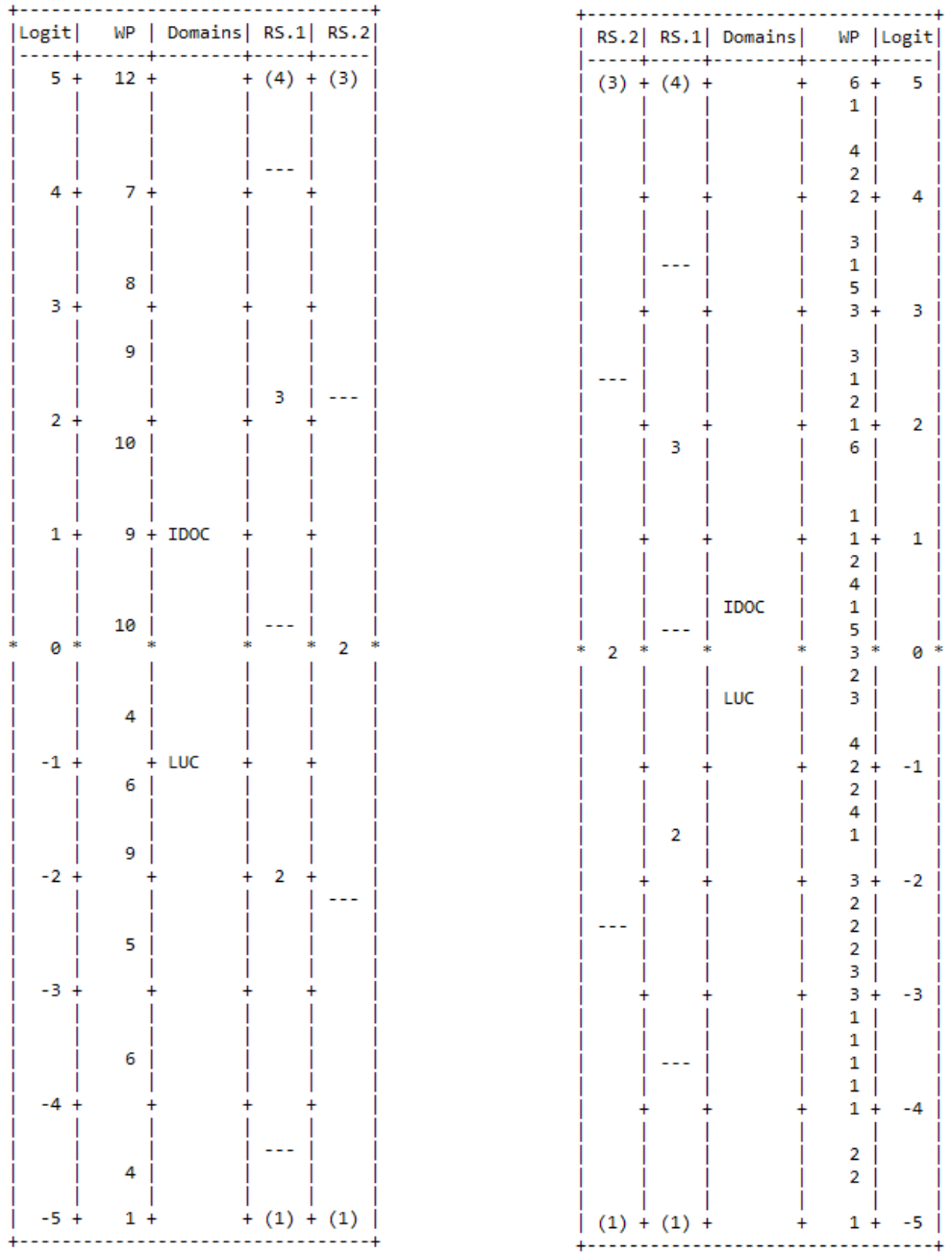
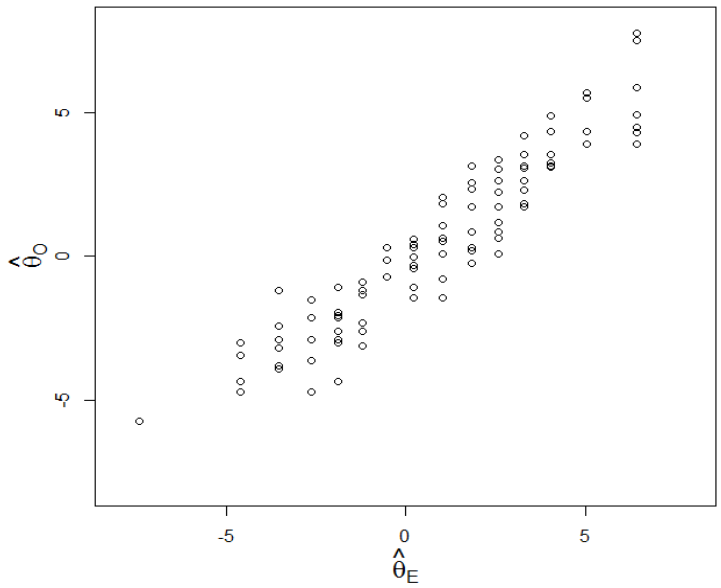


Figura 2. Mapa de Wright para expertos y calificadores operacionales

Note: El número mostrado en la columna WP representa la cuenta de ensayos en cada ubicación. WP se refiere a las medidas estimadas de competencias en escritura. RS.1 indica el uso de categorías por calificadores del dominio IDOC, y RS.2 para los del Dominio LUC.

Panel A. Estimado de competencias de escritura para los modelos 1 y 2



Panel B. Diferencias entre los estimados de competencias de escritura entre los modelos 1 y 2

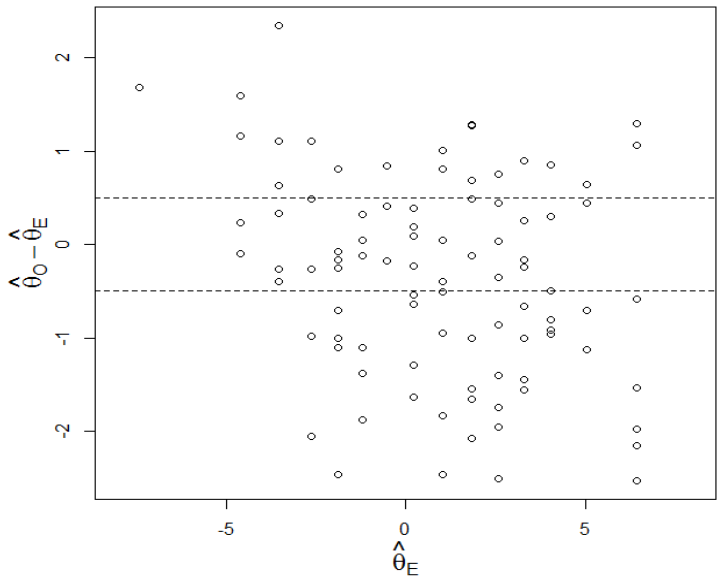


Figura 3. Comparación de competencia en escritura basado en los modelos 1 y 2

Nota: las líneas punteadas en el panel B indicant una banda de ±.5 lógitos.

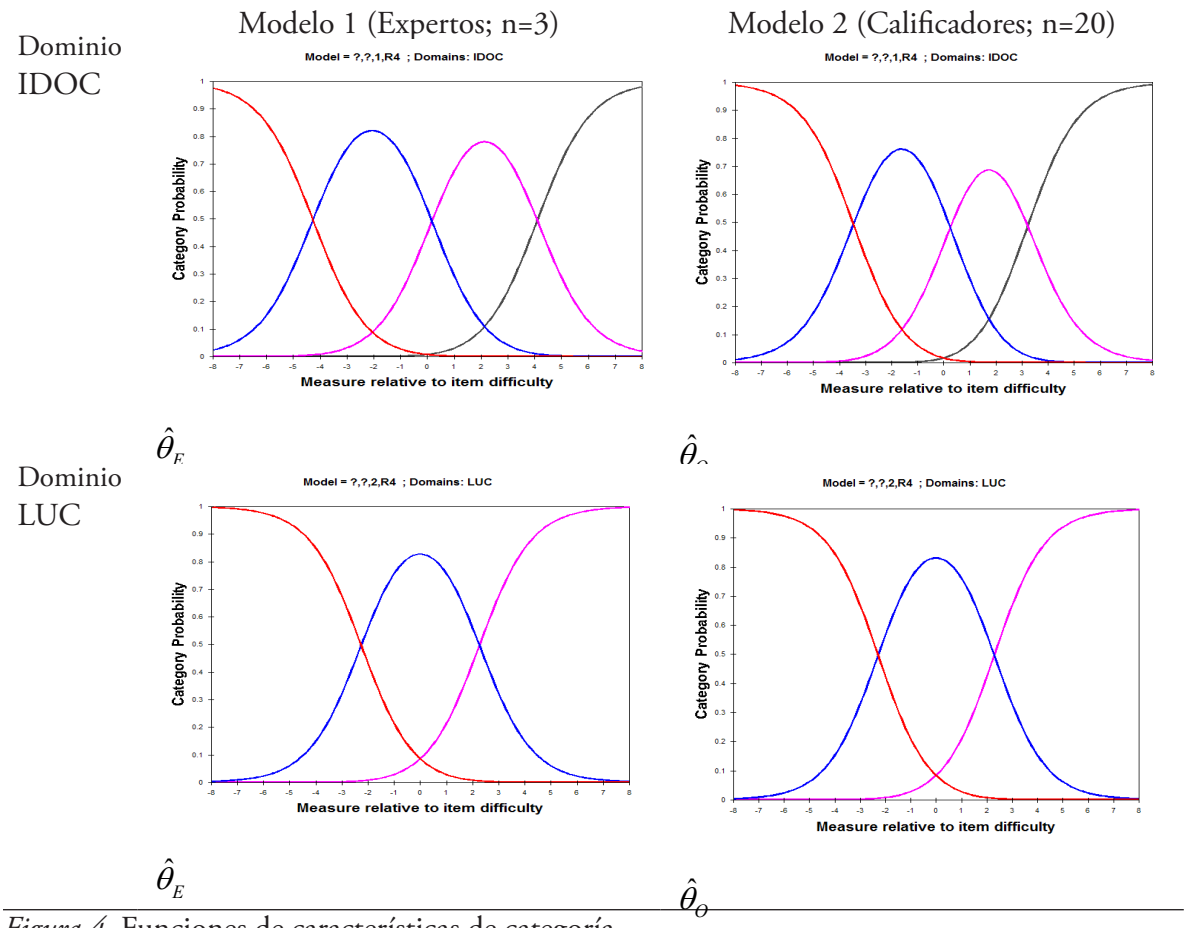


Figura 4. Funciones de características de categoría

Note: Cada curva representa una categoría. El color rojo indica una categoría de uso de 1, el color azul es la categoría de uso de 2, el morado se refiere a la categoría de uso de 3 y la línea negra para el Dominio IDOC es para la categoría de uso de 4.

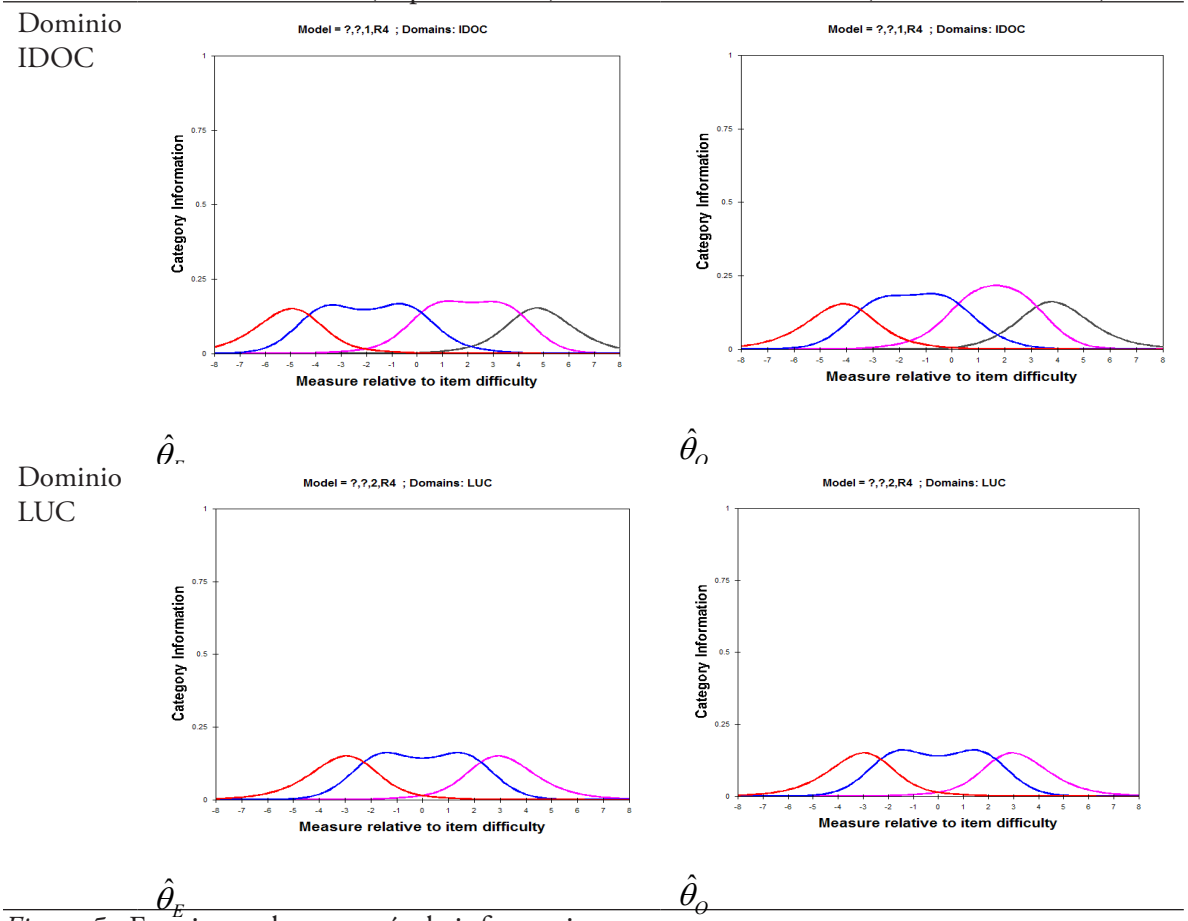


Figura 5. Funciones de categoría de informaciones

Nota: Cada curva representa una categoría. El color rojo indica una categoría de uso de 1, el color azul es la categoría de uso de 2, el morado se refiere a la categoría de uso de 3 y la línea negra para el Dominio IDOC es para la categoría de uso de 4.