

LOS ANÁLISIS SICOMÉTRICOS DE LOS RESULTADOS EDUCATIVOS Y SU INTERPRETACIÓN

MÁRIA LALINSKÁ, TOMÁŠ JABLONSKÝ
Y KATARÍNA JABLONSKÁ ¹

Fecha de recepción: septiembre de 2012
Fecha de aceptación y versión definitiva: octubre de 2012

RESUMEN: El artículo pone sobre la mesa las necesidades actuales de la investigación pedagógica relacionadas con la medición y la interpretación de los resultados de la educación. Está orientado a un método moderno de evaluación mediante tests basado en la Teoría de la respuesta al ítem (IRT-Item Response Theory) y al uso del software que desarrolla sus modelos básicos.

PALABRAS CLAVE: Educación, Medida, Evaluación, Teoría de la respuesta al ítem.

Psychometric analysis of educational results and their interpretation

ABSTRACT: The article deals with the actual needs of pedagogical research related to measurement and interpreting of the educational results. It is focused on a modern approach of testing through the Item Response Theory (IRT) and its basic models using the appropriate software.

KEY WORDS: Education, Measurement, Testing, Item Response Theory.

¹ Doc. PaedDr. Tomáš Jablonský, PhD. Catholic University in Ružomberok. Faculty of Education. Hrabovská cesta 1. 034 01 Ružomberok. Slovakia. E-mail: tomas.jablonsky@ku.sk

PaedDr. Mária Lalinská. Catholic University in Ružomberok. Faculty of Education. Slovakia. E-mail: maria.lalinska@ku.sk

Mgr. Katarína Jablonská. Catholic University in Ružomberok. Faculty of Education. E-mail: katarina.jablonska@ku.sk

En la teoría pedagógica tradicional se ha dedicado una mayor atención a los procesos educativos que a sus productos, es decir, que a sus resultados. La investigación en este ámbito de la pedagogía ha sido casi inexistente. Sin embargo, actualmente la situación está cambiando y se está desarrollando cada vez más el área de la evaluación pedagógica. A pesar de que, en la pedagogía moderna, tanto en la teoría como en la práctica de la evaluación son consideradas clave las ciencias integradas, todavía queda en ellas un considerable espacio para la investigación. Entre los componentes finales de los procesos educativos están los resultados educativos. Estos resultados se podrían definir como lo que los alumnos aprenden en las escuelas y, si deseamos conocer la calidad de la educación, es imprescindible medirlos e interpretarlos de manera adecuada. Para poder establecer mediciones son necesarias herramientas de medición y de interpretación que se consideren más o menos objetivas. En nuestros días aumenta la necesidad de enseñar cuestiones relativas a la medición y a la evaluación de los resultados educativos también a los estudiantes de todas las especialidades de la carrera pedagógica.

La evaluación del esfuerzo de los alumnos se puede considerar un instrumento estimulante de gran importancia educativa solamente si la evaluación es correcta, objetiva y justa, con un valor declarativo concreto respecto a cada alumno. Una de las posibilidades que tiene el profesor para expresar la evaluación de los resultados del aprendizaje es mediante *las notas*, las calificaciones que se añaden a los documentos escolares oficiales y que reflejan el nivel general del conocimiento de alumno, así como el de sus capacidades y habilidades. Desde el punto de vista taxonómico, la calificación escolar se puede considerar como uno de los factores básicos de motivación para aprender nuevos conocimientos, pese a que en nuestra época es venerada en muchos casos de manera excesiva, ya que con frecuencia la calificación influye en la estructura social y en la posición del alumno dentro de su escuela, su clase, su familia, etc. Para muchos alumnos, las notas constituyen precisamente la finalidad de su esfuerzo, lo que de ninguna manera debería ser entendido como el sentido de la calificación escolar.

Muchos pedagogos consideran la calificación escolar como poco conveniente, puesto que se basa en un criterio comparativo de los resultados de los alumnos, es decir, se trata de una oposición *alumno / alumno* o *alumno / estándar previamente determinado*. Además, este método comparativo de evaluación puede causar dificultades, sobre todo podría inspirar en los alumnos falta de generosidad, envidia, rivalidad no deseada o competición destructiva, lo que podría reflejarse en una baja autoestima y una baja apreciación de sí mismo por parte del alumno, así como en su respeto hacia los demás. La calificación sirve para medir la cantidad de información memorizable, así como la validez de las soluciones. Sin embargo, no sirve para

ofrecer detalles sobre otros «valores añadidos»: no informa sobre lo que es realmente valioso, sobre cómo enfrentarse a las dificultades, sobre cómo actuar y cooperar, sobre el entusiasmo, el esfuerzo ni la motivación. Mediante una escala de cinco notas, de sobresaliente a suspenso, es imposible evaluar de modo gradual y sistemático cómo el alumno supera los problemas y las dificultades. [4]

La evaluación en el campo de las humanidades no se apoya en la existencia de unas escalas fijas a las que el alumno sea obligado a ajustarse. Al contrario, su finalidad es el niño como tal, es decir, el avance en el desarrollo de sus actitudes, capacidades y conocimientos respecto a un estado pasado. [8]

Existe una única respuesta a la pregunta de por qué y para qué los pedagogos califican diariamente a los alumnos: para que aprendan a evaluarse a sí mismos y al mundo que les rodea. Dicho con otras palabras, el sentido de la evaluación no es la evaluación por la evaluación, sino el aprendizaje de la capacidad de evaluar las propias fuerzas y habilidades, para que el alumno pueda de esta manera establecerse metas alcanzables. Lo más importante para la vida cotidiana que puede aprender un alumno o un adolescente durante el proceso de la evaluación es, según Kosová (2000), a realizar una autoevaluación objetiva. [5]

El alumno aprende a realizar esta autoevaluación, junto con el sentimiento de satisfacción proveniente de su empeño, durante todo el proceso de su ontogénesis, desde su temprana niñez. Es evidente que, en su vida cotidiana, no obtiene una buena nota por aprender a montar en bicicleta o por atarse los cordones de los zapatos. Su premio siempre ha sido la satisfacción al completar una tarea o al resolver un problema, lo que significa que estos han sido los momentos que han originado sus factores motivadores internos que conforman la base del concepto de la evaluación sin calificación. Ésta es la diferencia más importante cuando se compara este tipo de motivación con la motivación exterior utilizada en el sistema de calificación escolar tradicional, del que podría abusarse para obligar, castigar o premiar.

LA TEORÍA MODERNA DE LOS TESTS COMO ALTERNATIVA PARA MEDIR, EVALUAR E INTERPRETAR LOS RESULTADOS EDUCATIVOS

Durante las últimas dos décadas del siglo xx, ha tenido lugar una considerable revolución investigadora en el campo de la medición de los resultados educativos. Se trata de conocimientos, habilidades, capacidades y competencias, y también de algunas actitudes y cualidades afectivas que tienen

su expresión en todos los niveles: el individual, el de la clase o de la escuela, el del sistema escolar; así como en el nivel de la comparación internacional. Se trata de un fenómeno complejo, influido por una gran variedad de componentes y factores que reflejan y afectan al pensamiento del alumno. Los métodos de medición y evaluación modernos dan importancia a la medición basada en criterios frente a la medición tradicional basada en normas de comparación entre los individuos.

Recientemente en este campo han sido desarrollados varios procedimientos y modelos estadísticos, como por ejemplo la Teoría G (Generalizability Theory) y la Teoría de la respuesta al ítem, TRI (Item Response Theory, IRT). La Teoría G se originó a partir de la Teoría clásica de los tests, TCT (Classical Test Theory, CTT), y se convirtió en una herramienta útil en la preparación y creación de mediciones fiables. La Teoría de la respuesta al ítem se basa en un procedimiento nuevo: en la medición basada en criterios donde la capacidad del alumno no se deriva de su comparación con otros alumnos, sino de la probabilidad de acierto o de error al responder a los ítems de una prueba estandarizada.

En cuanto a la praxis de la TRI, está creciendo la importancia de los factores que influyen en los resultados de los alumnos. La investigación se centra en tres campos: las preguntas (los ítems) de los tests, los alumnos evaluados y los evaluadores; estos últimos sobre todo en relación con la evaluación de preguntas abiertas o pruebas orales.

La Teoría de la respuesta al ítem se apoya en dos postulados básicos:

1. El resultado, es decir, la reacción de una persona frente a un ítem, puede ser previsible o explicable por un conjunto de factores llamados rasgos latentes (ocultos).
2. La relación entre las reacciones de los respondentes a un ítem del test y el conjunto de signos relacionados con ese ítem pueden ser descritos por una función monótona creciente.

Esta función se denomina *Función característica del ítem*, FCI (Item Characteristic Function, ICF), y viene representada por el gráfico de la llamada *Curva característica de un ítem*, la CCI (Item Characteristic Curve, ICC). La Función característica del ítem muestra cómo, a medida que aumenta la capacidad de una persona, aumenta también la probabilidad de que dé una respuesta correcta a un ítem.

Todos los modelos de la TRI comparten una característica clave que expresa la relación entre los rasgos (capacidades) latentes y las probabilidades de validez de las respuestas a los ítems; es decir, la probabilidad de dar una respuesta válida a un ítem es representada por la función de un rasgo (capacidad) latente. En un contexto donde se miden y se interpretan los re-

sultados educativos, este rasgo latente es *el rasgo* cognitivo de las personas evaluadas.

Existen varios modelos de TRI diferenciados por las respectivas prescripciones matemáticas de la función característica del ítem que vienen dadas por el número de los parámetros específicos. Dentro de la TRI se distinguen dos tipos básicos de parámetros: los parámetros relacionados con la persona y los parámetros relacionados con el ítem. Como se ha indicado más arriba, dentro del contexto educativo, el parámetro de una persona, siendo ésta un alumno o un estudiante, es *el rasgo*. Dentro de los parámetros del ítem, los más frecuentes son: *la dificultad*, *la discriminación* (sensibilidad, selectividad) y *la pseudoadivinación* (pseudo-azar). Según el número de parámetros, se distinguen: el modelo logístico de 1 parámetro (dificultad) (Modelo 1 PL de Rasch), el modelo logístico de 2 parámetros (dificultad, discriminación) (Modelo 2 PL), y el modelo logístico de 3 parámetros (dificultad, discriminación, pseudoadivinación) (Modelo 3 PL de Lord).

Una de las etapas más importantes en la aplicación de cualquier modelo TRI es la estimación de los parámetros. Esta puede ser realizada mediante varios procedimientos (por ejemplo la calibración (estimación a partir de las respuestas) o la relación entre las respuestas y el rasgo latente (estimación de una persona según la Curva característica del test). Para facilitar los cálculos en este proceso ha sido desarrollado software especial (por ejemplo el Winsteps, el Facets, el OPLM, el SAS, el R, etc.). Sin estos programas el análisis y la interpretación de los datos provenientes de la TRI serían complicados y demasiado difíciles. La TRI ha encontrado su aplicación práctica principalmente en procesos relacionados con la creación y la evaluación de tests o de sus variantes respectivas, y en la creación y mantenimiento de bancos de ítems, así como también en los análisis e interpretaciones sicométricas de resultados educativos.

EL MODELO DE RASCH (MODELO LOGÍSTICO DE UN PARÁMETRO, MODELO 1PL)

El Modelo de Rasch se considera el modelo de TRI más básico, más simple y más utilizado en los análisis de los ítems de tests lingüísticos. Su nombre se debe a uno de los iniciadores de la teoría moderna de los tests, al matemático de origen danés George Rasch (1901-1980). Su modelo representa un caso específico de función logística de 1 parámetro, siendo este la dificultad de los ítems. La Función característica del ítem viene definida por la siguiente fórmula:

$$P(X_{ji} = 1) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \qquad P(X_{ji} = 0) = \frac{1}{1 + \exp(\theta_j - b_i)}$$

ó , $i = 1, 2, \dots, n$.

La probabilidad de validez de una respuesta se representa como $P(X_{ji} = 1)$, mientras que $P(X_{ji} = 0)$ es la probabilidad de una respuesta incorrecta de una persona j en un ítem i . El símbolo θ_j representa el rasgo de una persona j (variable independiente); la letra b_i viene definida como el parámetro de la dificultad de un ítem i , y el \exp es una función exponencial donde la n es el número de los ítems del test.

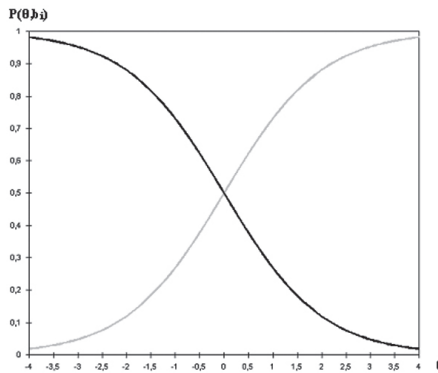


GRÁFICO 1: Modelo logístico de Rasch con un parámetro (curva azul: probabilidad de una respuesta válida de las personas con el rasgo θ , siendo el b_i la dificultad del ítem; curva roja: probabilidad de una respuesta falsa de las personas con el rasgo θ , siendo el b_i la dificultad del ítem).

La probabilidad de acierto de un ítem se define como la función de la diferencia entre el rasgo θ_j y la dificultad b_i . Su gráfico es representado por una curva creciente en forma “S” (dibujo n.º 1) cuyo rango es el intervalo (0, 1). La función decreciente representa la probabilidad de error de una persona j respecto a un ítem i . $P(X_{ji} = 0)$.

El parámetro de la dificultad, el b_i , es un punto situado en el eje del rasgo latente (θ), donde la probabilidad de acertar tiene el valor numérico de 0,5. De ahí que el parámetro de la dificultad sea a la vez el indicador de la posición de la Curva característica de un ítem (la CCI) respecto a la escala del rasgo θ . Por lo tanto, según aumenta el valor de la dificultad de un ítem, debe aumentar también el rasgo latente de la persona para alcanzar al menos el 50% de probabilidad de acertar. Respecto a cualquier ítem, los ítems

de mayor dificultad aparecen situados en la parte más a la derecha del eje θ , mientras que los más fáciles se sitúan a su izquierda.

De lo expuesto más arriba, se derivan estas conclusiones:

- si el grado de dificultad correspondiente a una persona es más alto que el parámetro de dificultad del ítem, existe una probabilidad alta de acierto en la respuesta a ese ítem;
- si el grado de dificultad correspondiente a una persona es más bajo que el parámetro de dificultad del ítem, existe una probabilidad baja de acierto en la respuesta a ese ítem, y
- si el grado de dificultad de una persona es el mismo que el parámetro de dificultad del ítem, la probabilidad de acierto es de 0,5.

Es importante señalar que, en el Modelo de Rasch con un parámetro, la dificultad del ítem es el único factor que influye en la respuesta de la persona testada. La asíntota en la parte baja del gráfico de la Función característica del ítem es nula ($y = 0$), es decir, la persona con el nivel del rasgo latente más bajo dispone de una probabilidad de valor cero de acertar en un ítem concreto. Sin embargo, este modelo no considera la posibilidad de acertar al azar (un caso frecuente en los tests con diferentes respuestas posibles). Este problema lo trata de resolver el modelo logístico de tres parámetros, siendo estos la dificultad, la discriminación y la pseudoadivinación (Modelo 3PL de Lord).

ANÁLISIS DE LOS ÍTEMS MEDIANTE LA TRI UTILIZANDO EL SOFTWARE WINSTEPS (RASCH MODEL, LINACRE, 2009)

Con frecuencia, para realizar un análisis sicométrico de los ítems basado en el Modelo 1PL de Rasch, se suele utilizar el software Winsteps. Este software fue desarrollado por Benjamin Wright y John M. Linacre, de la Universidad de Chicago, en los años 80 del siglo pasado. Este sistema, con el paso de años, ha sido mejorado y actualizado y, en nuestra época, representa una de las soluciones más cómodas para obtener los análisis de los ítems, así como la estimación de los parámetros.

Antes de empezar con el análisis de datos en el Winsteps, el usuario está obligado a crear un archivo básico, es decir, un script, para especificar los parámetros del modelo, las estructuras de datos, el tipo de los archivos de salida y su formato utilizando un lenguaje de programación especial creado para el Winsteps. Este archivo básico es guardado como un documento de texto y debe ser abierto por el Winsteps. Los datos para el análisis de Rasch

(los resultados del test) son sucesivamente guardados bien en el archivo básico o, en caso de una base de datos más extensa, se archivan en un documento especial en formato de texto o de tabla. No obstante, este archivo debe ser inscrito en el archivo básico usando la sintaxis correcta. Algunos ejemplos de archivos de base forman parte del paquete Winsteps, de manera que este software puede ser usado también por personas que no hayan sido entrenadas en su programación.

Aparte de los análisis estándares según Rasch, el usuario puede elegir entre varios tipos de archivos y formatos de salida, dependiendo de sus preferencias (por ejemplo: Excel, WordPad, SAS, SPSS, etc.). Entre las salidas más utilizadas para los tests lingüísticos están: ITEMS (análisis completo de los ítems), DISTRACTORS (análisis completo de los distractores), PERSONS (análisis completo de personas-candidatos), RESIDUALS (análisis complementarios, como son las estimaciones de probabilidad de cada persona en cada ítem del test) y varios tipos de escalas (por ejemplo la Escala de Guttman), mapas de datos y tablas. La unidad de medida para todos los datos se denomina *lógito* (*logit*) y es la derivada del logaritmo neperiano del índice de probabilidad de una respuesta correcta o incorrecta a un ítem dado.

En este programa, cada ítem puede ser representado en un gráfico. El Winsteps es capaz de dibujar los modelos de las curvas teóricas de cualquier ítem perteneciente a la misma categoría de respuestas, así como los modelos de las curvas empíricas (estas curvas pueden ser representadas en un mismo gráfico, véase el gráfico 2).

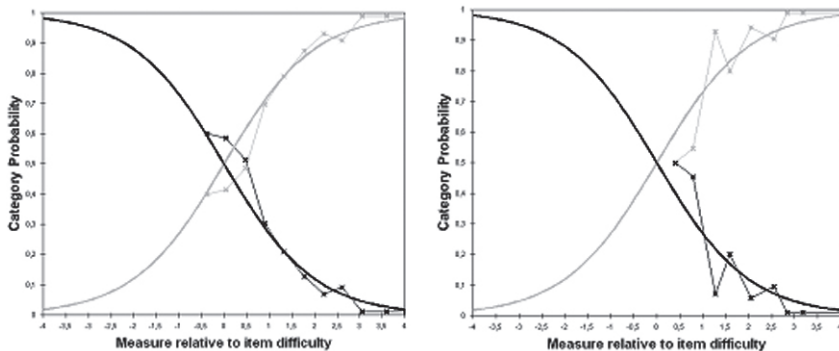


GRÁFICO 2: Representaciones de la curva empírica y su modelo teórico respecto a un ítem respondido por 1.298 personas (gráfico izquierdo) y por 100 personas (gráfico derecho).

Entre las funciones más avanzadas del software Winstep está también el análisis del Funcionamiento diferencial del ítem, el FDI (Differential Item Functioning, DIF). Se trata de analizar el nivel de dificultad del ítem dentro de un grupo representativo. Así es posible, de una manera muy simple, observar las diferencias entre los niños y las niñas; es decir, si un ítem concreto ha causado más dificultades a los niños o a las niñas, o viceversa (un asunto relevante en la pedagogía de género).

Aparte del programa Winsteps, existe también su versión reducida y gratuita llamada *Ministep*. Esta contiene todas las funciones del Winsteps, la única limitación estriba en un reducido número de ítems analizables, el análisis está limitado a un máximo de 25 ítems y de 70 personas.

Es importante señalar que el número de respondentes en los análisis TRI es crucial. Para realizar un análisis TRI fiable los científicos recomiendan analizar las respuestas de, como mínimo, 500 personas. Hemos elegido como ejemplo de este procedimiento la comparación entre un ítem respondido por 1.298 personas y el mismo ítem sometido a 100 respondentes (seleccionados al azar). Se puede observar que, en el primer caso, la curva empírica característica del ítem se corresponde mucho más con el modelo de esa curva. De ahí se deriva una muy buena posibilidad de aplicación de la TRI en pruebas de certificación nacional, en las que participan grandes cantidades de respondentes.

Dentro del programa Winsteps aparecen datos de salida de dos tipos: el primero viene representado por *el infit* y *el outfit* (ambos forman parte de las tablas ITEMS y PERSONS), y el segundo se denomina *escala de evaluación (rating scale)*. Los dos indican el nivel de diferencia entre los datos teóricos (la curva teórica) y los de la realidad (la curva empírica). En relación con los tests, el *infit* (o el *outfit*) puede marcar, por ejemplo, cómo los alumnos sobresalientes (o los suspensos / casi suspensos) se han equivocado (o han acertado) respondiendo un ítem fácil (o difícil). Resulta complicado diagnosticar las causas de un *infit* o un *outfit* inadecuado (frecuentemente fuera del intervalo 0,7-1,3). Por lo que respecta a los ítems, estos niveles indican que un ítem es excesivamente difícil, está fuera del rango de la prueba o se puede tratar de un ítem ambiguo o redundante, etc. Un *infit* o un *outfit* problemático en relación con los respondentes sugiere el hecho de que han adivinado la respuesta correcta, han copiado, engañado, se han distraído, etc.

INTERPRETACIÓN DE LOS RESULTADOS DE LOS ANÁLISIS SICOMÉTRICOS

En la interpretación de los resultados sicométricos es imprescindible tomar en consideración la influencia de varios factores internos y externos que

pueden reflejarse, de manera deseable o no, en los resultados. En cuanto a la *distorsión de la medición* (*measurement bias*), los factores siguientes son los que aparecen con mayor frecuencia [6]: el género del alumno; el género del encuestador (o del creador de los ítems); la pertenencia del alumno a cierto grupo cultural o socioeconómico; la manera de responder a los ítems del test; factores relacionados con las condiciones situacionales y con las condiciones de los procedimientos del test (la limitación temporal, la manera de formular las instrucciones, la postura del encuestador); factores relacionados con la motivación, la postura y personalidad de alumno (el miedo a ser testado, la motivación para alcanzar un éxito, la autoestima, la manera de pensar, la impulsividad).

Si un análisis sicométrico muestra indicadores de un comportamiento extraño de un ítem, este puede ser eliminado del test o del banco de ítems, corregido o desplazado a otra posición dentro del test (por ejemplo, un ítem difícil puede ser desplazado al final del test). Las causas de estos cambios son las siguientes:

- ninguna respuesta correcta o varias respuestas correctas (mal funcionamiento de la clave y de los distractores, o la existencia de respuestas creativas alternativas en los ítems abiertos);
- los ítems miden algo que está fuera del ámbito del test (ítems que exigen ciertos conocimientos matemáticos o generales, etc.);
- los ítems son inadecuados respecto al nivel del test o del grupo en cuestión;
- los alumnos casi suspensos o suspensos (los que se sitúan debajo del nivel de dificultad en la escala TRI, en el eje θ) responden correctamente con más frecuencia de la que viene dada por el modelo (por causas de adivinación o de poseer un conocimiento especial: el *outfit* es alto), y
- los alumnos sobresalientes (los que se sitúan por encima del nivel de dificultad en la escala TRI, en el eje θ) responden correctamente con menos frecuencia de la que viene dada por el modelo (por causas de distracción, cansancio, prisa: el *outfit* es alto; o por miedo o gran dificultad del ítem: el *infit* es bajo).

Partiendo de las características de la TRI citadas más arriba, podemos sostener que esta teoría, en sus procedimientos respecto a la interpretación de datos, al ser comparada con la Teoría clásica de los tests, presenta numerosas diferencias e incluso ventajas básicas. A diferencia de la TCT, una de las ventajas de la TRI es la invariabilidad de los datos respecto a los parámetros del respondente y del ítem. Las estimaciones de la capacidad de una persona no dependen de un test concreto, así como los parámetros del ítem no se someten a un grupo de encuestadores concreto. De esta manera se

soluciona el problema de la escasa comparatividad entre diferentes grupos testados respecto a diferentes tests o sus variantes.

BIBLIOGRAFÍA

- [1] BUTAŠ, J., y LALINSKÁ, M. (2010), «Application of Exponential Functions in Testing», en *Usta Ad Albim Bohemica* n.º 1/2010. Revista reseñada, editada con ocasión de la XVII Czech, Polish and Slovak Mathematical School en Litoměřice, República Checa, ISSN 1802-825: pp. 13-24.
- [2] BUTAŠ, J., y LALINSKÁ, M. (2011), «La Teoría de la respuesta al ítem y el análisis de ítems en los tests de idiomas extranjeros», en *La medición del nivel de conocimientos con el fin de incrementar la calidad de la educación*. Volumen anual de los discursos presentados durante la conferencia. Trnava: Typi Universitatis Tyrnaviensis, 2011. ISBN 978-80-8082-509-6: pp. 32-38.
- [3] BUTAŠ, J., y LALINSKÁ, M. (2011), «La medición de los resultados educativos como parte de la educación pedagógica», en *La ciencia y la educación como fomento de la sociedad letrada*, volumen anual de la III Conferencia internacional de científicos en la Universidad J. Selye. Komárno: Univerzita J. Selyeho, 2011. ISBN 978-80-8122-017-3: pp. 245-265.
- [4] JABLONSKÝ, T. (2009), «La calificación verbal: ¿sí o no?», en *La Problemática de la calificación verbal en el sistema educativo actual*, volumen de publicaciones de la Cátedra de Pedagogía presentadas en diversas conferencias internacionales. Ružomberok: Centro editorial de la Facultad de Pedagogía de la Universidad Católica de Ružomberok. ISBN: 978-80-8084-415-8: pp. 104-105.
- [5] JAREŠOVÁ, A. (2009), «La calificación en la educación primaria y antes de primaria», en *La Problemática de la calificación verbal en las escuelas de hoy*, volumen de publicaciones de la Cátedra de Pedagogía presentadas en diversas conferencias internacionales. Ružomberok: Centro editorial de la Facultad de Pedagogía de la Universidad Católica de Ružomberok. ISBN: 978-80-8084-415-8: pp.106-108.
- [6] JENSEN, A. R. (1980), *Bias in mental testing*, London: Methuen & Co., Ltd. 786 pp. ISBN 0-416-83230-X.
- [7] PARTCHEV, I. (2004), *A visual guide to item response theory*, Jena: Friedrich-Schiller-Universität: p. 61. sin ISBN.
- [8] ZENTKO, J., y UHRINOVÁ, M. (2009), «La calificación de asignaturas educativas», en *La Problemática de la calificación verbal en las escuelas de hoy*, volumen de publicaciones de la Cátedra de Pedagogía presentadas en diversas conferencias internacionales. Ružomberok: Centro editorial de la Facultad de Pedagogía de la Universidad Católica de Ružomberok. ISBN: 978-80-8084-415-8: pp. 211-215.