

**EVALUACIÓN DEL RENDIMIENTO
ACADÉMICO**
**INTRODUCCIÓN A LA TEORÍA DE
RESPUESTA AL ÍTEM**



Autor:

Andrés Burga León

Lima 4 de Octubre del 2005



EVALUACIÓN DEL RENDIMIENTO

INTRODUCCIÓN A LA TEORÍA DE RESPUESTA AL ÍTEM

Mg. Andrés Burga León
UPCH – Facultad de Psicología
UMC – Ministerio de Educación

1. LA EVALUACIÓN DEL RENDIMIENTO

La evaluación es un instrumento sumamente importante dentro del ámbito educativo. A partir de los años 90 se da un importante cambio en la concepción de la evaluación, pasando de estar centrada en los exámenes y calificaciones, para convertirse en un mecanismo de orientación y formación (Cerda, 2003). En la actualidad puede considerarse que la evaluación educativa, cumple cuatro funciones fundamentales (Reátegui, Arakaki y Flores, 2001):

- Toma de decisiones: están referidas a la marcha del proceso pedagógico. Se decide, si un alumno debe pasar o no un curso, y continuar con su proceso de instrucción.
- Retroinformación: se busca conocer las debilidades y fortalezas del alumno en cuanto a sus logros
- Reforzamiento: implica convertir a la evaluación en una actividad satisfactoria, mediante el reconocimiento de su esfuerzo y rendimiento.
- Autoconciencia: se busca que el alumno reflexione respecto a su propio proceso de aprendizaje, cómo entendiéndolo, y que elementos le están causando dificultades.

Dentro de este contexto, se consideran como muy importantes las pruebas de aprovechamiento o rendimiento, que son todas aquellas que buscan evaluar el nivel de habilidad o logro de un alumno luego de un proceso de instrucción (Aiken, 1996). Es decir, el propósito fundamental de estos instrumentos es la evaluación académica, que responde a la pregunta ¿Qué conocimientos o destrezas ha adquirido el alumno tras un periodo de instrucción? (Prieto y García, 1996). Basándonos en la propuesta de Pizarro, Clark y Allen (1987), la medición del rendimiento académico puede ser entendida, como una cantidad que estima lo que una persona ha aprendido como consecuencia de un proceso de instrucción o formación; es la capacidad del alumno para responder al proceso educativo en función a objetivos o competencias.

Además, puede ser entendido en relación con un grupo social que fija los niveles mínimos de aprobación ante un determinado cúmulo de conocimientos, procedimientos o aptitudes (Carrasco, 1985). El rendimiento académico, sin ser el único indicador de la calidad educativa, es uno de los más importantes; y su

estudio ha sido separado, por lo menos desde un punto de vista teórico, en factores cognitivos y afectivo-motivacionales que lo afectan (Marchesi y Martín, 1999). Por ejemplo, en el modelo de aprendizaje autorregulado de Mckeachie, Printich y colaboradores (1992, citado en García, 2002) se demuestra que los factores cognitivos, los motivacionales y la relación entre ambos, ejercen una influencia directa en la implicación del estudiante en el aprendizaje y en su rendimiento académico. Por ello es preciso considerarlo dentro de un marco complejo de variables como los condicionamientos socio-ambientales, factores intelectuales, variables emocionales, aspectos técnicos y didácticos (Capella y colaboradores, 2003).

Como indicamos eanteriormente, la evaluación del rendimiento de los estudiantes es un indicador sobre la calidad del sistema educativo. Podemos esperar que un sistema de calidad logre que los estudiantes alcancen niveles de desempeño suficientes en las diversas áreas evaluadas. En términos de evaluaciones de sistema, se ha puesto énfasis en el logro en matemáticas, lenguaje y ciencias. Para ello se aplican pruebas estandarizadas a muestras representativas de alumnos a fin de conocer el nivel de desempeño que han alcanzado. Thorndike (1989) señala que el método que se centra sobre el nivel de desempeño que se tiene en alguna área del conocimiento o habilidades, corresponde a las llamadas pruebas de aprovechamiento con referencia al criterio.

Un aspecto muy importante de las pruebas de aprovechamiento o rendimiento, es que el contenido de lo evaluado debe estar acorde con el contenido de lo enseñado (Anastasi y Urbina, 1997). Al respecto Good y Brophy (1997) sostienen que de manera típica, las pruebas sólo cubren una muestra pequeña del contenido y los objetivos enseñados y tienen que tomarse decisiones respecto a qué incluir. Los items o preguntas de un test son seleccionados para ser representativos aunque de forma imperfecta, del saber básico que se puede esperar de un alumno (Ingebo, 1997). Es decir, se hace necesario un adecuado muestreo del dominio, el cual debe ser realizado en función a los objetivos de la instrucción, que a su vez se estructuran sobre la base de taxonomías (Prieto y Gracia, 1996). Es importante muestrear la gama completa del contenido enseñado e incluir suficientes ítems para que la medición sea confiable (Cortada, 1999). De esta manera se podrá dar cuenta de cuáles son las áreas que presentan debilidades, además los estudiantes percibirán como injusta la evaluación si se centra sólo en uno pocos contenidos (Good y Brophy, 1997).

Por todo lo anterior, García y Prieto (1996) sostienen que es muy importante que al construir una prueba para evaluar el rendimiento, se defina adecuadamente el dominio o conjunto de indicadores a partir de los cuales se infiere el nivel de logro de las personas en aquella materia que se quiere evaluar. Una prueba de rendimiento queda compuesta por indicadores que se conectan con ítems o tareas significativas, asociadas al dominio a través de definiciones semánticas.

Esto constituye la matriz de contenidos que servirá como base para la construcción de los ítems.

Haciendo una síntesis de diversos autores (Aiken, 1996; Cortada, 1999; Prieto y García, 1996; Thorndike, 1989), se puede considerar que los principales tipos de ítems y sus características, mediante los cuales se evalúa el rendimiento académico, son los siguientes:

1. Ensayo o preguntas abiertas: Se orientan a evaluar la capacidad del alumno para organizar, relacionar y comunicar sus conocimientos. Cuando estas preguntas son usadas de forma exitosas, le piden al alumno mucho más que simplemente reproducir información. Otra ventaja adicional es que no dan lugar a la adivinación. Es recomendable tener una matriz de calificación para este tipo de reactivos. Es decir, especificar de forma clara y demostrable, qué necesita una respuesta para ser considerada como adecuada y recibir el puntaje completo. En cuanto a su confección debemos:
 - Definir los ítems de forma clara, de tal manera que no haya dudas respecto a lo que se pide para su resolución.
 - Poner énfasis en preguntas que pidan solucionar problemas o ejemplificar, más que reproducir información.
 - Utilizar una cantidad reducida de ítems que deben ser respondidos por todos los alumnos.
 - Incluir preguntas que varíen en cuanto a su dificultad
2. Completar oraciones: Se le presenta al alumno un enunciado o párrafo pequeño al cual le faltan algunas palabras. La tarea de la persona consiste en rellenar dichos espacios, de tal manera que le dé sentido al enunciado, además de poseer un contenido correcto según el dominio que se busca evaluar. La principal desventaja de este tipo de ítems es que no miden objetivos complejos. Hay que considerar lo siguiente cuando se construye este tipo de ítems:
 - Procurar que en el caso de tener un solo espacio en blanco, este quede al final
 - Evitar el uso de varios espacios en blanco en el mismo concepto, especialmente si éstos hacen que el ítem pierda significado.
3. Verdadero / Falso: A la persona respondiente se le pide que identifique la verdad o falsedad, de un conjunto de enunciados presentado. El problema de estos ítems es que tiene una alta probabilidad ($p = .50$) de adivinación. Se recomienda considerar lo siguiente en el momento de construirlos:

- Preguntar sólo cosas importantes.
 - Redactar enunciados cortos y sin ambigüedad.
 - Evitar la doble negación.
 - Evitar el uso de términos como "todos" o "ninguno".
 - En caso de poner "afirmaciones" citar las fuentes de donde fueron tomadas.
 - Redactar los enunciados verdaderos y los falsos con longitudes similares.
4. Opción múltiple: para resolverlos, la persona tiene que elegir entre las diversas opciones de respuesta, cuál de ellas es la adecuada, según el enunciado del problema. En su formulación debe considerarse lo siguiente:
- El enunciado debe ser una sola frase y estar en consonancia formal y de contenido con todas las opciones de respuesta.
 - Ordenar los ítems de forma aleatoria. Aunque algunos autores consideran que es mejor ordenarlos según la temática.
 - El número óptimo de alternativas de respuesta es de tres a cinco.
 - Redactar todas las alternativas de respuesta con longitudes similares.
 - Todas las alternativas de respuesta deben ser gramaticalmente correctas y estar enunciadas de forma similar.
 - Usar sólo alternativas de respuesta posibles. Es decir, no emplear distractores cuyo contenido erróneo resulte obvio.
 - En la medida de lo posible evitar las alternativas "todas las anteriores" y "ninguna de las anteriores", pues estas suelen ser muchas veces la opción correcta.
 - Evitar que la respuesta correcta incluya una palabra clave, que pueda servir como indicio para detectarla.
5. Emparejamiento: en este tipo de ítems se presenta un enunciado y dos columnas, una de ellas representa a los estímulos y la otra a las respuestas. La tarea de la persona respondiente consiste en emparejar, usualmente conectando con una línea, cada uno de los estímulos, con la respuesta correcta, sobre la base de la comparación postulada en el enunciado. Respecto a su construcción es recomendable:
- Especificar claramente cuál es la base del emparejamiento que la persona debe usar.
 - Mantenerse la homogeneidad en el tipo de material presentado.

- Las opciones de los estímulos deben identificarse con números y las de las respuestas con letras.
- Deben tenerse entre 6 y 15 estímulos, con 2 o 3 respuestas extras.
- El ítem debe aparecer completo en una página, no debe cortarse.

El uso de este tipo de ítems, si bien suele ser frecuente en las pruebas estandarizadas, no constituye la única manera de evaluar el rendimiento académico. Helmke y Van Akem (1995) sostienen que se realiza una mejor evaluación si se combinan varios tipos de evaluaciones, para tener una calificación final que refleje el logro de los estudiantes. A nivel de aula, se puede hacer una evaluación más completa del logro del estudiante si se emplean otros métodos (Díaz-Barriga y Hernández, 2002; Good y Brophy, 1997):

- Pruebas de desempeño: ponen énfasis en los componentes procedimentales, pidiendo que la persona ejecute alguna conducta: pronunciar un discurso, pintar, construir, hacer un experimento, etc. Es importante que las personas evaluadas sepan qué tienen que demostrar y los criterios mediante los cuales serán evaluados. Además se puede mejorar la confiabilidad de la calificación si se usan por lo menos dos observadores, que asignan puntuaciones de manera independiente, comprobando luego la confiabilidad de dichas calificaciones (Suen, 1990)
- Portafolios: es una serie organizada de trabajo del alumno que tiene como objetivo mostrar el progreso de este a lo largo del tiempo; fomentando la autoevaluación y autorreflexión sobre dicho progreso, más que una calificación. Este portafolio incluye sólo una parte del trabajo del estudiante, eligiendo el mismo que deberá incluirse dentro del mismo. En ese sentido es muy importante que los profesores les enseñen a los estudiantes cómo usar los portafolios, cómo presentarlo, poniendo énfasis en su función de mostrar la maduración personal, así como la del producto.
- Mapas conceptuales: Sirven para evaluar los conocimientos declarativos del alumno, pudiendo aproximarnos a la forma como el alumno organiza la información, mediante la representación jerárquica de conceptos y proposiciones
- Evaluación informal: se hace generalmente durante la conducción regular del curso, y no ha sido programada de antemano. Implica aprovechar las situaciones disponibles, para evaluar, mediante la observación de las actividades de los alumnos o preguntas en clase. Su función se orienta generalmente hacia la retroinformación, y no suele tener asociada una calificación.

2. PROCESOS DE MEDICIÓN

Nunnally y Bernstein, (1995) nos dicen que la definición clásica de medición pertenece a Stevens, quien en 1957 afirma que medir en un sentido amplio es asignar numerales a objetos o eventos de acuerdo a reglas. Es decir, consiste en reglas para asignar símbolos a objetos de tal manera que:

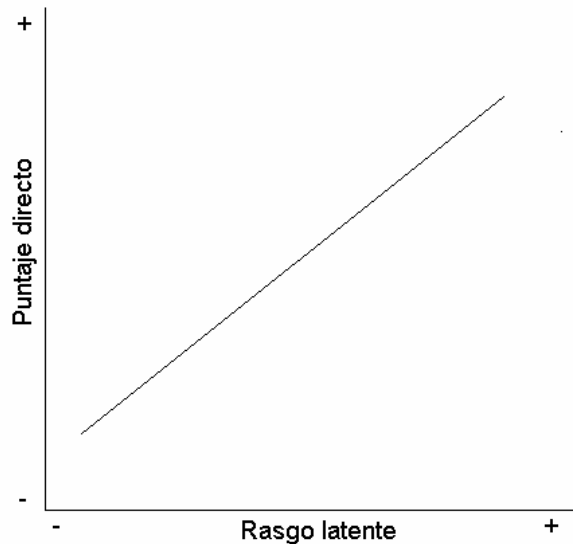
- Representen cantidades o atributos de forma numérica. Indican que tanto del atributo está presente en el objeto
- Definan si los objetos caen en las mismas categorías o en otras diferentes con respecto a cualidades esenciales

Además podemos distinguir dos tipos de procesos de medición: los directos y los indirectos. En los directos, se pone en correspondencia directa un instrumento de medida con la propiedad del objeto medido. Por ejemplo, si queremos saber la longitud de una pieza de madera, ponemos en correspondencia uno de sus lados con una regla y haremos afirmaciones como: "esta pieza mide 27 cm. de largo." Esto gracias a que la escala de medición está contenida en el instrumento de medida.

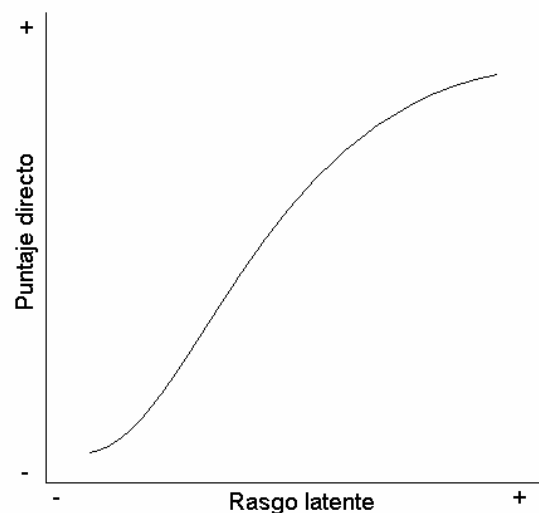
Muchas de las características que evaluamos no son directamente observables, son mas bien construcciones hipotéticas (rasgos latentes) que elaboramos para registrar la uniformidad de la conducta de una persona (Thorndike, 1989). Por lo común se piensa que los rasgos latentes son cuantificables, o sea, que tienen propiedades de cantidad o grado, en el sentido de que una persona puede tener más del rasgo que otra, o que una persona tiene más del mismo en un momento dado que en otro. Por ejemplo, un alumno puede estar más motivado que otro por los contenidos de la clase de matemática, mientras que un mismo alumno puede estar muy motivado por esos temas en primaria, y perder la motivación en secundaria.

Por ello debemos inferir su cantidad por medio de indicadores manifiestos. En esos casos, nos encontramos frente a un segundo tipo de procesos de medición: los indirectos. Por ejemplo se pueden utilizar las respuestas a un conjunto de preguntas para determinar la cantidad de conocimientos sobre historia que posee una persona. En este sentido, se hacen indispensables los instrumentos de medición (pruebas), que pueden definirse como aquellas herramientas que permiten la asignación numérica a las magnitudes de la propiedad o atributo, ya sea por la comparación directa con las unidades de medida o provocando y cuantificando las manifestaciones del atributo cuando este es indirecto (Nunnally y Bernstein, 1995). En general dicha cuantificación se realiza por medio de los puntajes directos obtenidos en la prueba. El puntaje directo se determina generalmente como la sumatoria de las puntuaciones obtenidas a cada ítem. Por ejemplo, si un alumno ha resuelto correctamente 15 preguntas, y cada pregunta vale dos puntos, su puntaje directo es 30.

Al momento de establecer las relaciones entre los puntajes directos de las pruebas y la cantidad del rasgo latente, se puede planear que dicha relación es monotónica¹ lineal, como lo hace la Teoría Clásica de los Test (Muñiz, 1996). En el gráfico presentado a continuación se quiere representar que iguales cantidades de aumento en el puntaje directo corresponden a aumentos de la misma magnitud en el rasgo latente:



Además se puede pensar que la relación es monotónica no lineal, como los modelos de Teoría de Respuesta al Ítem (Hambleton, Swaminathan y Rogers, 1999; Muñiz, 1997). Como apreciamos en el siguiente gráfico, aumentos de la misma magnitud en el puntaje directo no corresponden a la misma magnitud de aumento en el rasgo latente:



Del gráfico anterior se puede deducir que la cantidad de rasgo latente que se necesita para pasar de un puntaje directo de 2 a otro de 4 no es la misma que se

¹ Se plantea que una relación es monotónica cuando un aumento de los valores en una de las variables se encuentra asociado al aumento de los valores de la otra variable.

necesita para pasar de 15 a 17 puntos. Por lo tanto, las puntuaciones directas no constituyen una verdadera escala de intervalo. Sin embargo, existe un conjunto de modelos matemáticos que sirven para linealizar dichas relaciones, construyendo medidas a partir de los puntajes directos, de tal manera que tengan las propiedades de una escala de intervalo. Estos modelos matemáticos en general se conocen como Teoría de Respuesta al Ítem

3. TEORÍA DE RESPUESTA AL ÍTEM

El nombre proviene del hecho que estos modelos se centran fundamentalmente en las propiedades de los ítems. Concretamente tratan de modelar matemáticamente² que ocurre cuando una persona con una habilidad determinada se enfrenta a un ítem específico.

Una gran ventaja de estos modelos es que dan la posibilidad de tener mediciones invariantes respecto de los instrumentos utilizados y las personas evaluadas. (Muñiz, 1997). Este es un problema central en Teoría Clásica de los Tests. Por ejemplo, si un alumno responde sólo a ítems difíciles, obtendrá un puntaje directo bajo. En cambio, otro alumno que haya respondido sólo a ítems fáciles, tendrá una puntuación directa alta. Erróneamente, se podría pensar que el alumno con un puntaje directo superior posee más del rasgo latente evaluado. Sin embargo no se estaría considerando los efectos de la dificultad de los ítems (y de la prueba en general) para establecer el puntaje de un alumno en particular. Un alumno puede tener un puntaje directo más bajo en un test difícil que el obtenido por otro alumno en un test fácil, y a pesar de ello tener una medida más alta que la del segundo alumno (Ingebo, 1997).

Además los parámetros obtenidos para describir los ítems en Teoría Clásica de los Tests dependen de la muestra de personas en las cuales fueron calculados. Por ejemplo, el típico índice de dificultad para ítems dicotómicos se obtiene al dividir el número de aciertos al ítem entre la cantidad de personas que rindieron la prueba. Si ese ítem es respondido por personas con alta habilidad, la proporción de aciertos (valor p) será alta, por lo tanto el ítem parecerá fácil. Por otro lado, si ese mismo ítem es respondido por una muestra de personas con baja habilidad, la proporción de aciertos será baja y el ítem parecerá difícil. Como bien señala Ingebo (1997) el valor p únicamente nos dice que tan difícil fue ese ítem para el grupo de personas en el que fue calculado. No se puede generalizar y dar información útil sobre la probabilidad de un estudiante de alto o bajo rendimiento de acertar a ese ítem.

Frente a estos problemas encontrados en la Teoría Clásica de los Tests surgen los modelos de Teoría de Respuesta al Ítem que permiten (Muñiz, 1997):

² Como se verá más adelante, estos modelos matemáticos son probabilísticos.

- Obtener mediciones que no varíen en función del instrumento utilizado, que sean invariantes respecto de los ítems empleados.
- Disponer de ítems cuyos parámetros no dependan de los objetos medidos, que sean invariantes respecto de las personas evaluadas.

Estas dos características son esenciales a fin de poder desarrollar adecuados modelos de medición, y son llamadas por algunos autores "objetividad"³. Cuando se logra el principio de objetividad la comparación del desempeño de dos personas no depende del conjunto particular de ítems usados para compararlas (Stenner, 1990; Wright y Linacre, 1987).

A parte de la invarianza, hay dos requerimientos muy importantes en los ítems que constituyen un test a fin de aplicar los modelos de Teoría de Respuesta al Ítem: la unidimensionalidad y la independencia local.

La unidimensionalidad implica que un solo rasgo latente o constructo se encuentre en la base de un conjunto de ítems (Hattie, 1985). En otras palabras, un instrumento será unidimensional si las respuestas dadas a él son producidas en base a un único atributo. Wright y Linacre (1998) señalan que en la práctica, ningún instrumento puede ser perfectamente unidimensional; lo que buscamos es tener instrumentos que en esencia muestren unidimensionalidad. Por ejemplo, muchos factores como la motivación, ansiedad, velocidad de respuesta tienen un impacto sobre el desempeño de una persona en un conjunto de ítems (Hambleton, Swaminathan y Rogers, 1991). Lo importante es que un instrumento de medida, represente a través de sus puntuaciones un solo factor dominante. Con esto lo que se quiere implicar, es que la mayor cantidad de la varianza observadas en las respuestas a los ítems, sea explicada por un sólo atributo latente (Embretson y Reise, 2000).

Es muy importante tener un instrumento unidimensional, ya que esto será para muchos un requisito indispensable a fin de generar buenas medidas (Wright y Masters, 1982; Wright y Stone, 1998). Las puntuaciones obtenidas de la aplicación de un instrumento psicométrico, dentro de la Teoría Clásica de los Tests, siguen un modelo monotónico lineal, es decir, se asume que existe una relación lineal entre el puntaje directo obtenido y el nivel del rasgo o atributo que se está midiendo. A más puntaje directo, más de ese rasgo o atributo posee la persona evaluada. ¿De donde proviene ese puntaje directo o puntaje global? De la suma de los puntajes obtenidos en cada uno de los ítems. Como señala Cuesta (1996), el obtener los puntajes globales sumando las calificaciones de cada ítem supone que se está midiendo con ellos un solo constructo, de lo contrario no habría ningún fundamento que soporte las operaciones aritméticas realizadas con los ítems. De la misma manera si se pretende medir la cantidad de una variable,

³ Objectivity

esta no debe estar contaminada por las cantidades que posee la persona evaluada en otras variables (Stout, 1987; citado en Cuesta, 1996).

La independencia local implica que cuando el rasgo latente causante de las respuestas a los ítems de un test se mantiene constante, las respuestas de los examinados a cualquier par de ítems son estadísticamente independientes (Hambleton, Swaminathan y Rogers, 1991). La probabilidad de responder correctamente a un ítem, controlando la habilidad del examinado, no tiene nada que ver con la probabilidad de responder correctamente a otro ítem.

Es decir, la probabilidad de responder correctamente, tanto al ítem uno como al dos, es igual al producto de la probabilidad de responder correctamente a cada uno de los ítems. Esto implica que no hay ningún tipo de dependencia en los ítems a parte de la atribuible al rasgo latente evaluado (Kolen y Brennan, 2004). Por ello Embretson y Reise (2000) señalan que se logra la independencia local cuando las relaciones observadas entre los ítems son explicadas casi en su totalidad por un modelo de Teoría de Respuesta al Ítem.

Cuando el conjunto de ítems utilizados se comporta unidimensionalmente, se obtiene simultáneamente la independencia local (Hambleton, Swaminathan y Rogers, 1991). Como lo señalan Embretson y Reise (1991), la unidimensionalidad implica independencia local, pero la independencia local no necesariamente implica unidimensionalidad. En otras palabras, un conjunto de ítems multidimensionales puede mostrar independencia local.

3.1 NUEVAS REGLAS DE LA MEDICIÓN

Al disponer de nuevos modelos psicométricos, cambian las reglas que se manejaban respecto a la medición de rasgos latentes. Embretson y Reise (2000) proponen diez nuevas reglas de la medición. A continuación presentamos una síntesis de las que consideramos las más importantes:

A) PRIMERA

- Vieja: los errores estándar de medición⁴ se consideran iguales para todos los puntajes en una población en particular. Este error estándar se deriva directamente de la confiabilidad del instrumento y la varianza de las puntuaciones observadas en una muestra en particular.
- Nueva: los errores estándar de medición difieren con los puntajes, En general los puntajes cercanos al centro de la distribución de puntuaciones

⁴ El error estándar de medición se refiere a las fluctuaciones esperadas por azar en el puntaje directo obtenido. Todo proceso de medición indirecta tiene asociado un margen de error, no existe la medida perfecta. Lo importante es tener medidas con el menor error posible, es decir, medidas confiables.

tienen un menor error de medición, en comparación al de los puntajes de los extremos superior e inferior.

B) CUARTA

- Vieja: para obtener parámetros insesgados de los ítems se necesitan muestras representativas. En Teoría Clásica de los Tests la dificultad y discriminación de los ítems dependen de las características de las personas evaluadas.
- Nueva: se pueden obtener parámetros insesgados de los ítems con muestras no representativas usando modelos de Teoría de respuesta al Ítem. Esto se debe a la propiedad de invarianza de los parámetros, es decir, sus valores no dependen del grupo de personas evaluadas.

C) SEXTA

- Vieja: las propiedades de las escalas de intervalo se obtienen cuando se ha logrado una distribución normal de puntajes. En Teoría Clásica de los Tests el supuesto de intervalo no se puede demostrar, se asume.
- Nueva: las propiedades de las escalas de intervalo se obtienen al aplicar modelos de Teoría de Respuesta al Ítem. Esto permite utilizar con mayor certeza todo el conjunto de métodos de análisis estadísticos que presuponen variables medidas a nivel de intervalo.

D) SÉPTIMA

- Vieja: el tener formatos mixtos de ítems tienen un impacto no balanceado sobre el puntaje total. Es decir, los ítems dicotómicos tendrán un menor peso en la calificación total frente a los ítems politómicos.
- Nueva: el tener formatos mixtos de ítems puede producir puntajes óptimos. Se han desarrollado diferentes métodos para construir medidas cuando los ítems del instrumento no tiene la misma cantidad de categorías de respuesta.

Es importante señalar que existe un gran número de modelos de Teoría de respuesta al Ítem, como los modelos logísticos de 1, 2 y 3 parámetros, el modelo de Respuesta graduada de Samejima, el de Respuesta Graduada Modificada de Muraki, el de escala de calificación de Andrich, el de Créditos Parciales de Masters, el de créditos Parciales modificado de Muraki, y el de respuesta Nominal de Bock, entre otros (Embretson y Reise, 2000; Hambleton, Swaminathan y Rogers, 1991; Muñiz, 1997)

3.2 MODELOS RASCH⁵

Estos modelos se centran en el análisis de cada ítem, concretamente de la interacción entre una persona y un ítem. Establecen la probabilidad de respuesta de una persona ante un ítem en términos de la diferencia entre la medida de rasgo o habilidad latente de la persona (B) y la medida del ítem utilizado en términos de su dificultad (D)⁶. Por este motivo se les denomina usualmente modelos de un parámetro (Hambleton, Swaminathan y Rogers, 1991; Muñiz, 1997).

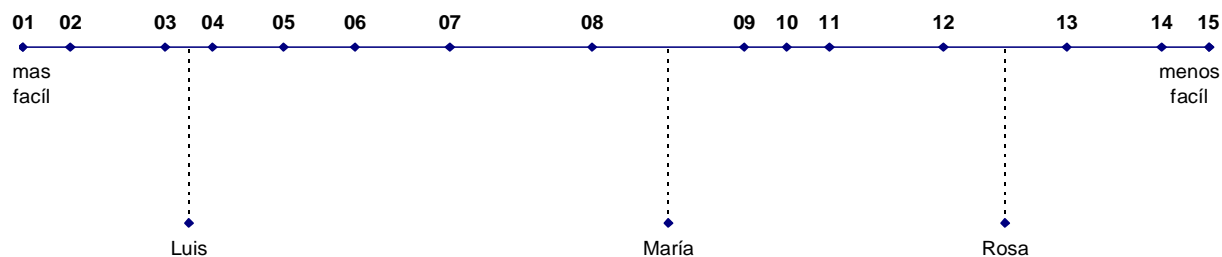
George Rasch, matemático danés, se dio cuenta que los resultados de la interacción entre personas e ítems no pueden estar totalmente predeterminado, sino que implica siempre un elemento de impredecibilidad (Wright y Linacre, 1989). Esto conlleva al requerimiento que en términos probabilísticas, mientras mayor habilidad, mayor probabilidad de acertar a un ítem; mientras más difícil un ítem, menos probable para cualquier persona acertarlo. Para ello se establece un modelo matemático de tipo probabilístico que vincula la habilidad o rasgo latente de una persona, con la probabilidad de respuesta correcta a un ítem. Pensar con probabilidades implica un salto de lo observable y fijo, a lo relativo y probable. (Ingebo, 1997).

En los modelos Rasch la habilidad de las personas y las dificultades de los ítems se ubican en la misma métrica. Al respecto Smith y Kramer (1989) nos recuerdan que la existencia de una métrica común permite combinar la habilidad de la persona y la dificultad del ítem para predecir el desempeño de una persona en un ítem cualquiera e identificar respuestas inesperadas. La idea central del análisis Rasch es poder construir una escala conformada por los ítems ordenados según su dificultad. Ello implica que a mayor habilidad, la persona tendrá una mayor probabilidad de acertar a los ítems y, por lo tanto, un mayor número de respuestas correctas. Es muy importante tener en cuenta que la medida estimada de la persona no es igual al puntaje directo (número de ítems correctos) que posee, este será solo un insumo a partir del cual se construirá la medida Rasch.

Como ejemplo, supongamos que se ha construido una prueba con quince ítems que se ajustan a un modelo Rasch y se les ha ordenado según su dificultad, del ítem más fácil (01) al más difícil (15):

⁵ Existe una controversia entre los seguidores del análisis Rasch y los modelos de Teoría de Respuesta al Ítem. En general, el análisis Rasch plantea que sus modelos son una definición de medición (Wright, 1989). Lo importante es ver hasta que punto los datos se ajusten a dicha definición. Si se da este ajuste, se habrá construido una buena medida. En cambio, la Teoría de Respuesta al Ítem se orienta a encontrar que modelo matemático se ajusta mejor a los datos. Mas información sobre esta controversia puede encontrarse en Shaw (1991).

⁶ En algunos textos se utiliza la letra theta (θ) para referirse a la habilidad de las personas y la letra b para referirse a la dificultad de los ítems



Si sabemos que Luis tiene una habilidad mayor que la dificultad del ítem 03, pero menor que las del ítem 04, lo más probable es que haya acertado al ítem 03 y todos los más fáciles (01 y 02), y haya fallado al ítem 04 y todos los más difíciles (05 al 15). Por su parte María tiene una habilidad mayor que la dificultad del ítem 08, pero menor que la del ítem 09. Por lo tanto lo más probable es que haya acertado al ítem 08 y todos los más fáciles (01 al 07) y haya fallado al ítem 09 y todos los más difíciles (10 al 15). Finalmente, Rosa probablemente habrá acertado al ítem 12 y todos los más fáciles, y habrá fallado el 13 y los más difíciles. Nótese que hemos dicho que es probable que haya acertado todos los más fáciles y fallado los más difíciles. No estamos afirmando que en la realidad se encontrará este tipo de patrones de fallos y aciertos. Lo usual es, por ejemplo, que María haya acertado los ítemes 01 a 05, haya fallado el 06, acertado los 07 y 08, fallado el 09, acertado el 10 y fallado todos los demás.

3.2.1 MODELO PARA ÍTEMS DICOTÓMICOS

Un ítem dicotómico tiene una sola respuesta correcta, por lo tanto se puede acertarlos y recibir un punto ($X=1$) o fallarlos y no recibir ningún puntaje ($X=0$).

La relación entre la habilidad y dificultad puede graficarse por medio de las curvas características del ítem (CCI) que nos dan información concreta sobre la probabilidad de respuesta de una persona ante un ítem. Al trazar dichas curvas se dan las siguientes relaciones en el caso de tener ítems dicotómicos:

1. $B > D$; $p(X=1 | B, D) \in]0,5 ; 1,0]$
2. $B < D$; $p(X=1 | B, D) \in [0,0 ; 0,5[$
3. $B = D$; $p(X=1 | B, D) = 0,5$

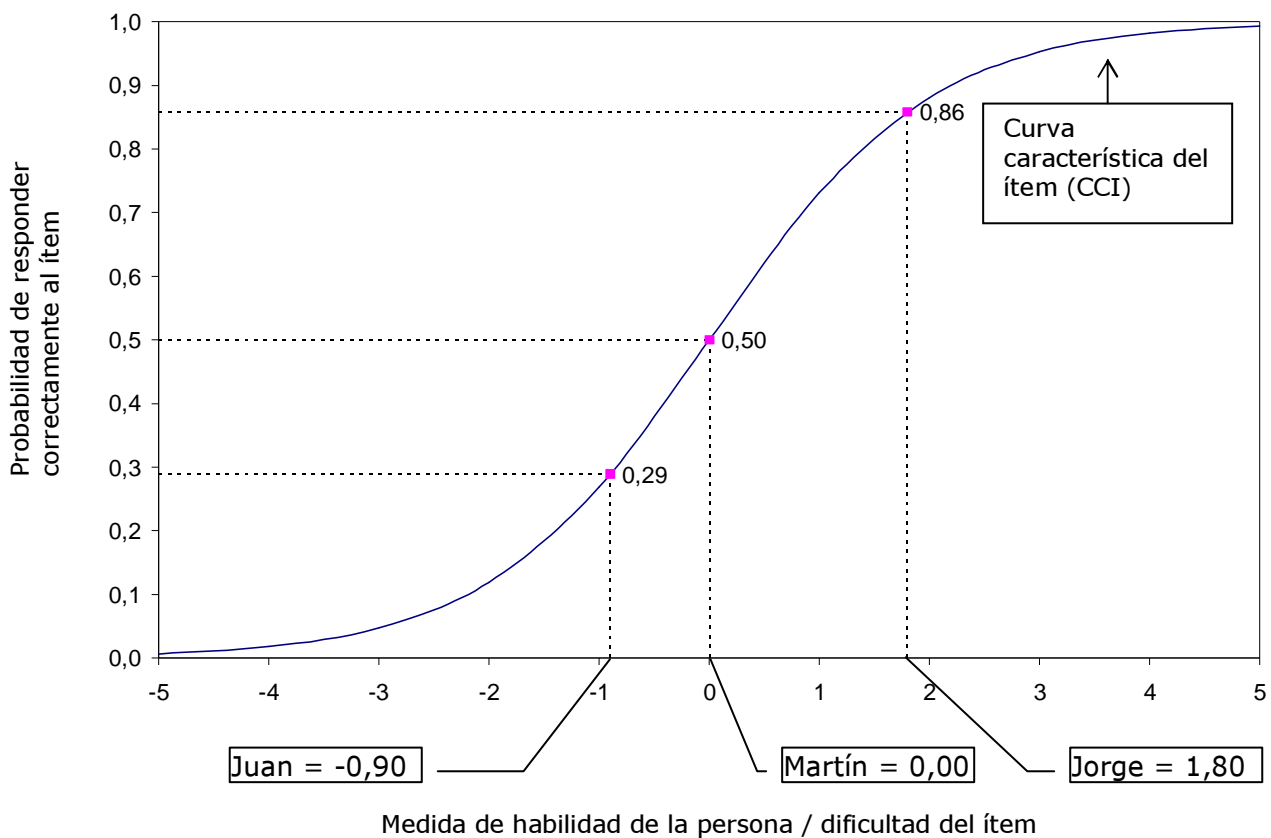
El primer caso nos dice que si la habilidad de la persona es mayor que la dificultad del ítem, la probabilidad de responder correctamente a dicho ítem es mayor que 0,5 (50%). La segunda situación indica que si la habilidad de la persona es menor que la dificultad del ítem, la probabilidad de responder correctamente a dicho ítem es menor que 0,5 (50%). Finalmente, si la habilidad de la persona es igual que la dificultad del ítem, la probabilidad de responder correctamente a dicho ítem es igual a 0,5 (50%). Como señala Ingebo (1990),

de esta manera se puede comprobar empíricamente la teoría que los estudiantes con mayores conocimientos tienen una mayor probabilidad de responder correctamente a una pregunta, frente a los estudiantes con menor conocimiento.

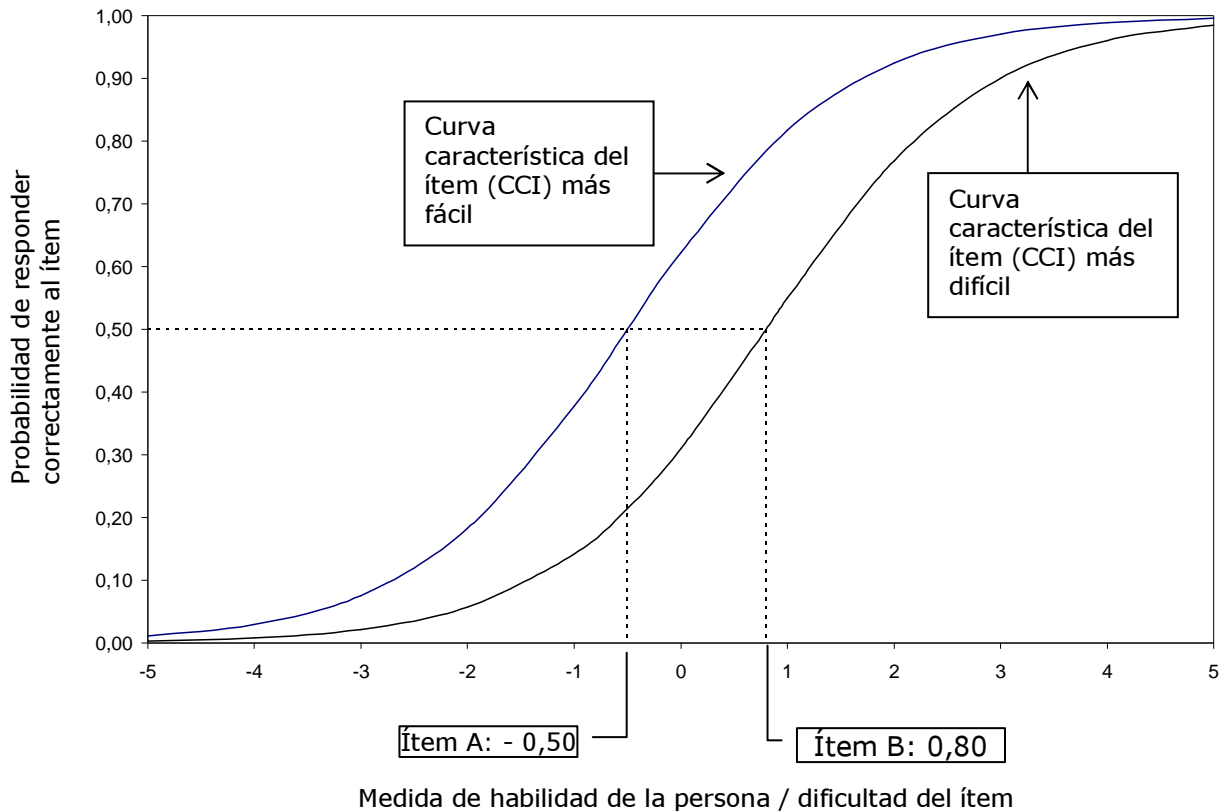
Matemáticamente la CCI se grafica con la siguiente función:

$$P(X_{is} = 1 | B_s, D_i) = \frac{e^{B_s - D_i}}{1 + e^{B_s - D_i}}$$

Esta relación entre la habilidad de una persona y la dificultad de un ítem se presenta en el siguiente gráfico:



En el gráfico se ve que, al enfrentarse a este ítem en particular, Juan, cuya habilidad es de -0,90 tiene una probabilidad de 0,29 de acertar a este ítem, es decir, lo más probable es que lo falle y obtenga 0 puntos. En cambio, Jorge, cuya habilidad se ha estimado en 1,80, tiene una probabilidad de 0,89 de acertar a este ítem, por lo tanto es más probable que lo acierte y reciba un punto. Finalmente, Martín tiene una habilidad igual a la dificultad del ítem; por eso se afirma que tiene iguales posibilidades de acertar o de fallar el ítem.

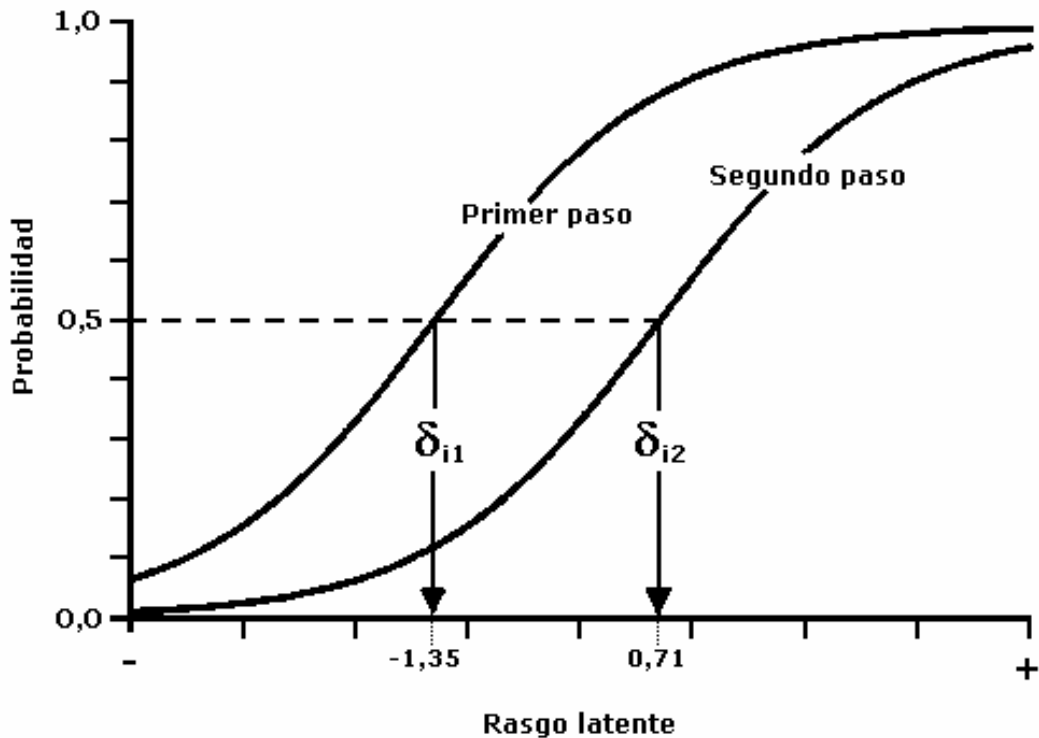


Al comparar dos o más curvas características de ítems, se puede decir que, mientras más a la derecha se encuentra una curva, más difícil es el ítem. Además, se expresa numéricamente la dificultad de un ítem, como aquel valor de la habilidad que posee una probabilidad de 50% de acertar a dicho ítem. En el caso que se presenta a continuación, el ítem más fácil tiene una dificultad de -0,50, y el más difícil de 0,80:

3.2.2 MODELO DE CRÉDITOS PARCIALES

Masters introduce en 1982 el Modelo de Créditos Parciales para trabajar con ítems politómicos de categorías ordenadas (Verhelst y Verstralen, 1997). El modelo especifica que cada ítem tiene su propia estructura de calificación. Se deriva de los tests de opción múltiple en los que hay respuestas incorrectas, pero que indican algún conocimiento, y se les da un crédito parcial (Wright, 1999). Fox (1999) señala que el Modelo de Créditos Parciales es una generalización del modelo Rasch que se puede aplicar a situaciones en las cuales los ítems pueden variar en el número de alternativas correctas y cantidad de opciones de respuesta en un mismo test.

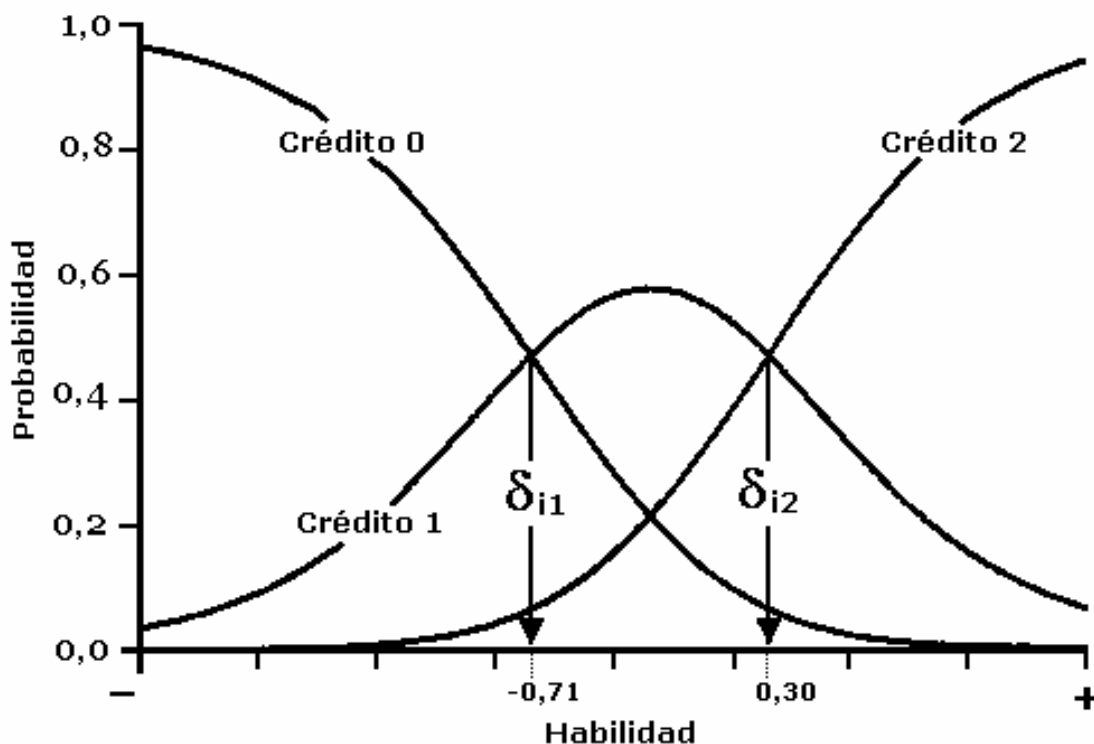
Al plantear este modelo se supone que en el proceso de resolución de un ítem, una persona responde de manera secuencial a un conjunto de subproblemas en el ítem. Los créditos parciales dados equivalen al número de pasos que deben



La curva a la izquierda se refiere a la probabilidad de recibir 1 punto en lugar de 0. La curva de la derecha, por su parte nos indica la probabilidad de recibir 2 puntos frente a 1 ó 0 puntos. Es decir, una persona con una medida de rasgo latente de -1,35 tiene una probabilidad igual al 50% de superar el primer paso, mientras que una persona con una medida de 0,71 tiene una probabilidad del 50% de superar el segundo paso.

Otra manera de graficar el Modelo de Créditos Parciales es mediante las curvas de categorías de respuesta. Estas nos muestran la relación existente entre la cantidad del rasgo latente evaluado y la probabilidad de obtener cada uno de los puntajes. Indican en que lugar del continuo de rasgo latente las respuestas a una categoría son más probables que a otra categoría, dando lugar a las curvas de categorías de respuesta.

Embretson y Reise (2000) señalan que el término δ_{ij} es considerado como la dificultad del paso asociado con el puntaje de una categoría j . Mientras más alto el valor de δ_{ij} , más difícil es ese paso en relación a otros pasos dentro del ítem. Es decir, los términos δ_{ij} , representan la dificultad relativa de un paso:



Los parámetros δ_{ij} se encuentran en la intersección de las curvas de las categorías de respuesta. Si una persona tiene medida del rasgo latente menor a -0,71, lo más probable es que haya recibido 0 puntos (crédito 0) en este ítem. Si su medida se encuentra entre -0,71 y 0,30, lo más probable es que reciba 1 punto (crédito 1). Finalmente si su medida de habilidad es mayor que 0,30 lo más probable es que reciba 2 puntos (crédito 2).

En este modelo puede haber muy pocas o ninguna observación en alguna de las categorías de respuesta de algunos ítems. Por lo tanto las estimaciones de la dificultad de esos pasos no será muy segura (Linacre, 2000). Al trabajar con estos modelos es importante asegurar un buen número de respuestas dentro de cada uno de los créditos probables.

3.2.3 AJUSTE AL MODELO Y CONFIABILIDAD

Ningún conjunto de datos se ajusta perfectamente a un modelo matemático, incluyendo a la curva normal. Ninguna variable se distribuye exactamente según esta distribución. Lo importante es hasta que punto es significativo dicho desajuste de los datos con respecto al modelo matemático (Schutz, 1990).

Wright y Masters (1989) señalan que una ventaja importante de los modelos Rasch es que proporcionan medidas de ajuste de los ítems y las personas. Por ejemplo, una persona con desajuste implicaría un patrón inesperado de respuesta, que puede tener diversas explicaciones (responde al azar, falla los ítems fáciles, pero acierta los difíciles, etc.). Si un ítem muestra desajuste con el

modelo, podría explicarse por su falta de discriminación, o porque este ítem está midiendo algo muy distinto al resto de ítems, es decir, carece de unidimensionalidad. Las dos medidas de ajuste empleadas en los modelos Rasch son:

- Outfit: Outlier sensitive mean square residual goodness of fit statistic. Es una medida sensible al comportamiento inesperado alejado de la medida.
- Infit: Information weighted mean square residual goodness of fit statistic. Es una medida sensible al comportamiento inesperado cercano a la medida.

Una ventaja del uso de estas medidas de ajuste es que no varían significativamente en función al tamaño de la muestra (Schutz, 1990). Tienen un valor esperado de 1,00 y varían entre cero e infinito (Linacre y Wright, 1994). Cualquier valor inferior a 1,00 implica que los datos no muestran mucha aleatoriedad, mientras valores superiores indican que los datos presentan demasiada aleatoriedad.

Hambleton, Swaminathan y Rogers (1991) sostienen que dos fuentes más del descontento con la Teoría Clásica de los Test descansan en la definición de la confiabilidad y lo que se puede pensar como su inverso conceptual: el error estándar de la medida.

Dentro del modelo de la Teoría Clásica de los Tests, Muñiz (1996) señala que las mediciones deben verse libres de errores de medición. Si las evaluaciones efectuadas con un instrumento son consistentes, si carecen de errores de medida, se les considera confiables. Así, el objetivo principal de la confiabilidad es tratar de estimar el error existente en las medidas mediante un indicador denominado coeficiente de confiabilidad (Muñiz, 1996). En esta misma línea, podemos citar a Suen (1990): "La confiabilidad es la fuerza de la relación entre el puntaje observado y el puntaje verdadero⁷. Esto puede ser expresado como la correlación obtenida mediante el coeficiente de Pearson entre el puntaje observado y el puntaje verdadero; eso es ρ_{XV} . Esta correlación es denominada índice de confiabilidad." (p. 28)

La confiabilidad, en este marco, se define también como la correlación entre los puntajes del test en formas paralelas de una prueba⁸. Dos pruebas serán paralelas cuando miden el mismo rasgo latente, con la misma cantidad de ítems,

⁷ El modelo de la Teoría Clásica de los test establece que el puntaje observado (X) es igual al puntaje verdadero (V) más el error de medición ϵ . En la práctica es imposible conocer el puntaje verdadero de una persona, pues ninguna medida se encuentra libre de error de medición.

⁸ Al correlacionar dos pruebas paralelas, teóricamente se debería obtener un coeficiente igual a 1,00, pues al ser paralelas es como si correlacionáramos una variable consigo misma. Sin embargo, la presencia del error de medición hace que las correlaciones con formas paralelas sean diferentes a 1,00.

tienen la misma media aritmética y la misma varianza. Si dos test miden lo mismo, cualquier diferencia de puntuaciones en ambos test, debe ser producto del error de medición. En la práctica, satisfacer los requerimientos de la definición de test paralelos es difícil, si no imposible. Por otro lado, la confiabilidad es reportada como si fuese una característica invariante, cuando no lo es. Depende no sólo del test, sino de la distribución de habilidad de la muestra estudiada y del número de ítems utilizado para evaluarlos (Muñiz, 1996).

Mientras más confiable sea un test, menor será el error estándar de medición que este posee. El problema con la medida del error de estándar, es que se supone que todos los examinados son medidos con la misma precisión, en independencia del nivel de rasgo latente que poseen (Hambleton, Swaminathan y Rogers, 1991). Linacre y Wrigth (1989) sostienen que al ajustar los datos al modelo Rasch para utilizarlos en el establecimiento de medidas, nuestro objetivo es construir un sistema invariante de medidas de intervalo, estimar su precisión (error estándar) y evaluar hasta que punto estas medidas y sus errores son confirmadas por los datos (medidas de ajuste). Una ventaja de los modelos Rasch, es que permiten calcular un error estándar para cada una de las medidas. En general se estiman con mayor precisión las medidas cercanas al promedio, mientras que las medidas de los extremos superior e inferior del rasgo latente son estimadas con menor precisión.

En el marco de los modelos Rasch, se habla también del índice de confiabilidad de separación de personas. Este sirve para indicarnos que tan bien sirven las medidas de un test para diferenciar las cantidades de rasgo latente que poseen los evaluados (Wrigth y Masters, 1982). Un índice menor a ,50 indica que las diferencias entre las medidas son producidas principalmente por el error de medición. (Fisher, 1992)

4. REFERENCIAS

- AIKEN, L.
1996 *Tests psicológicos y evaluación*, 8ª ed. México: Prentice Hall.
- ALDWIN, C.
1994 *Stress, coping and development. An integrative perspective*. Nueva York: The Guilford Press
- ANASTASI A. y S. URBINA
1998 *Tests psicológicos*. México. Prentice – Hall.
- CAPELLA J. y COLABORADORES
2003 *Estilos de aprendizaje*. Lima: Pontificia Universidad católica del Perú.
- CARRASCO, J.
1985 *La recuperación educativa*. España: Anaya.
- CERDA, H.
2003 *La nueva evaluación educativa*. Bogotá: Magisterio
- COLOM R.
1998 *Psicología de las diferencias individuales*. Madrid. Pirámide.
- CORTADA, N.
1999 *Teorías psicométricas y construcción de tests*. Buenos Aires: Lugar Editorial
- CUESTA M.
1996 "Unidimensionalidad". En J. Muñiz (ed.) *Psicometría*. Madrid: Pirámide, pp.239 – 291
- DÍAZ-BARRIGA F. y G. HERNÁNDEZ
2002 *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista*. 2.a ed. México: McGraw-Hill.
- EMBRETSON S. y P. REISE
2000 *Item Response Theory for psychologists*. New Jersey: Lawrence Earlbaum Associates.
- FISHER, W.
1992 Reliability statistics En J. Linacre (ed.) *Rasch Measurement Transactions Part 2*, 1996. Chicago: MESA Press, p.238.
- FOX, C.
1999 "An introduction to the partial credit model for developing nursing assessments". *Journal of Nursing Education*, vol.38, nº8, pp. 340-346.
- GARCÍA, L.
2002 "Factores asociados al rendimiento académico en estudiantes de psicología de la UNMSM". *Revista de Investigación en Psicología*, vol.5 n.º1.

- GOOD T. y J. BROPHY
1997 *Psicología educativa contemporánea*. 5.a ed. México: McGraw-Hill.
- HAMBLETON R., H. SWAMINATHAN y J. ROGERS
1991 *Fundamentals of Item Response Theory*. California: SAGE
- HATTIE J.
1985 "Methodology review: Assessing unidimensionality of tests and items". *Applied Psychological Measurement*, vol.9 n.º2, pp.139-164.
- HELMKE A. y M. VAN AKEN
1995 "The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study". *Journal of Educational Psychology*, vol. 87, n.º4, pp. 624-637.
- INGEBO, G.
1989 "Educational Research and Rasch Measurement". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, pp 43-46.
1997 *Probability in the measure of achievement*. Chicago: MESA
- KOLEN, M. y BRENNAN, R.
2004 *Test Equating, Scaling and Linking. Methods and Practices*. 2a ed. Nueva York: Springer
- LINACRE J.
1994 "DIMTEST disminuyendo". *Rasch Measurement Transactions*, vol.8 n.º.3, p.384. Consulta hecha en 27/01/2005. <<http://www.rasch.org/rmt/rmt83n.htm>>.
2000 "Comparing Partial Credit and Rating Scale Models". *Rasch Measurement Transactions*, vol.14 n.º 3, p.768. Consulta hecha en 03/07/2005. <<http://www.rasch.org/rmt/rmt143k.htm>>.
- LINACRE, J. y WRIGHT, B.
1989 "Length of a Logit". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, pp.54-55
1994 "Chi-Square Fit Statistics". En J. Linacre (ed.) *Rasch Measurement Transactions Part 2*, 1996. Chicago: MESA Press, pp.360-361.
- MARCHESI A. y E. MARTÍN
1999 *Calidad de la enseñanza en tiempos de cambio*. Madrid: Alianza Editorial
- MUÑIZ, J.
1996 *Teoría Clásica de los Tests*, 2.a ed. Madrid: Ediciones Pirámide
- NUNNALLY J. y I. BERNSTEIN
1995 *Teoría Psicométrica*. 3.a ed. México: McGraw - Hill.
- PIZARRO R., L. CLARK y M. ALLEN
1987 "El ambiente educativo del hogar". *Diálogos Educativos*, n.º 9-10, pp. 66-83.

- PRIETO G. y A. GARCIA
1996 "Construcción de Ítems". En J. Muñiz (ed.) *Psicometría*. Madrid: Pirámide
- REÁTEGUI N., M. ARAKAKI y C. FLORES
2001 *El reto de la evaluación*. Lima: PLANCAD-GTZ-Ministerio de Educación.
- SCHULZ, E.
1990 "Functional assessment of fit". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, pp.82-84.
- SHAW, F.
1991 "Descriptive IRT vs. Prescriptive Rasch". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, p.131.
- SMITH R. y G. KRAMER
1989 "Response Pattern Analysis with Supplemental Store Reports". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, pp.33-35.
- STENNER, J.
1990 "Objectivity: specific and general". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, p.111.
- SUEN, H.
1990 *Principles of tests theories*. Nueva Jersey: Lawrence Earlbaum
- THORNDIKE, R.
1989 *Psicometría aplicada*. México: Limusa
- VÉLEZ E., E. SCHIEFENBEIN y J. VALENZUELA
1998 "Factores que afectan el rendimiento académico en la educación primaria: Revisión de la literatura de América Latina y el Caribe". *Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI)*. Consulta hecha en 21/02/2004. <<http://www.campus-oei.org/calidad/Velezd.pdf>>.
- VERHELST N. y H. VERSTRALEN
1997 "Modeling Sums of Binary Responses by the Partial Credit Model". CITO. Consulta hecha en 21/02/2004. <<http://download.citogroep.nl/pub/pok/reports/Report97-7.pdf>>
- WRIGHT, B.
1988 "Useful Measurement through One-Step Item Banking." En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, p.24.
1999 "Model selection: Rating Scale or Partial Credit?". *Rasch Measurement Transactions*, vol.12, n.º 3, p. 641-642. Consulta hecha en 03/07/2005. <<http://www.rasch.org/rmt/rmt1231.htm>>.

WRIGHT B. y J. LINACRE

- 1987 "Rasch model derived from Objectivity". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, pp.5-6
- 1989 "The Differences between scores and measures". En J. Linacre (ed.) *Rasch Measurement Transactions Part 1*, 1995. Chicago: MESA Press, pp.63-65.
- 1998 "MESA research memorandum 44". *Archives of Physical Medicine and Rehabilitation*, vol.70 n.º12, pp. 857-860.

WRIGHT B. y G. MASTERS

- 1982 *Rating Scale Analysis*. Chicago : MESA

WRIGHT B. y M. STONE

- 1998 *Diseño de Mejores Pruebas*. México: CENEVAL