denarius
revista de economía y administración

Análisis de ítemes
Teoría clásica y teoría de respuesta al ítem

Miguel Ángel Rosado Chauvet

Resumen

Se presentan tres modelos de análisis de ítemes haciendo énfasis en las diferencias mostradas y las repercusiones que implica la decisión sobre el empleo de cualquiera de ellos en procesos de admisión de alumnos en instituciones educativas y de empleados en empresas.

El modelo Kalt es un porcentaje de dificultades entre 27 y 73 de aciertos, sin permitir proporciones explicadas al azar, pero impidiendo análisis sobre los grupos de rendimientos extremos alto y bajo al no presentar ítemes por debajo del mínimo o por arriba del máximo. El valor crítico de la discriminación está asociado a la dificultad, sin que sea justificable la relación entre ambas, y no depura los casos antes de evaluar los ítemes del instrumento.

El modelo Rasch utiliza la metodología logística, depurando a los sujetos mediante iteraciones que incluyen sujetos e ítemes. No incluye el índice de discriminación, sino que se basa en el ajuste de las respuestas con respecto a la logística de la dificultades de los ítemes y las habilidades de los

sujetos. Sugiere, como límites extremos lógicos entre ± 2.00 con un rango entre 12 y 88 para la dificultad.

El modelo Siseval es clásico e incluye el índice de dificultad con un rango entre 0.12 y 0.88 y con 0.20 de índice de discriminación. Depura los casos previos al análisis de estímulos y es el único que trabaja con diferentes valores en función del número de opciones del ítem, incluyendo las proporciones aleatorias de acierto al asumir que el azar también favorece al sujeto que tiene conocimientos.

Palabras clave: Kalt, Rasch, análisis de ítemes, teoría clásica, teoría de respuesta al ítem.

Presentación

La educación no es la parte conceptual de las escuelas, aunque se incluye en el proceso. La educación es la vida, es el entorno donde transcurre nuestra existencia, son los padres, la familia, los amigos, las figuras de identificación que nos comprometen, los valores que, en nuestro acontecer, nos han impregnado y forman parte profunda de nosotros.

Educarse no es memorizar algunas reglas, sino vivir y hacer nuestro el entorno conocido, imaginado o fantaseado, sentido o actuado; es ver más allá de las cosas aparentes, es más creación que permanencia, es más duda que certeza y es más entrega que posesión.

Asistir a una escuela no es llenar nuestra ignorancia con recuerdos que podamos ofrecer a quien mejor los pague, sino la búsqueda interminable del propio desarrollo, la participación en la solución de las dudas comunes y ancestrales, el planteamiento de escenarios mejores y la búsqueda de los insumos que requiera su logro; es vivir con los pies plantados en la tierra y la mirada puesta en las estrellas.

Justificación

Lo anterior forma parte de mi *concepto de la educación*, pero verificar los logros que se obtienen de ella implica un proceso arduo y consistente donde, si lo cualitativo presenta aspectos importantes en la explicación de nuestro desarrollo, no es menos importante precisar los aspectos cuantitativos del mismo. Medir no es evaluar, pero es una condición necesaria, aunque no suficiente, que permite constatar los avances reales en nuestras metas con independencia tanto del evaluado como del evaluador, en busca de un juicio exento de ambos sesgos. Es importante contar con un concepto de la educación que nos permita *poner nuestra mirada en las estrellas*, pero no es menos importante verificar la realidad de nuestras metas viviendo *con los pies plantados en la tierra*.

A este respecto se han dado dos vertientes: la teoría clásica, representada por F. M. Lord y H. Gulliksen del Educational Testing Service de New Jersey; y la teoría de respuesta al ítem, representada por B. D. Wright de la Universidad de Chicago. En ambos casos no se trata de los creadores de las teorías, pero sí de quienes en la actualidad han expandido los aspectos defendibles de las mismas e impulsado una amplia investigación al respecto.

Un aspecto de importancia para la investigación en general y para la investigación educativa en particular, consiste en la precisión con la cual se obtienen los datos a partir de los cuales se llegará a conclusiones que sirvan de base a juicios y decisiones. Si la precisión no es suficiente, o si no se obtiene en forma consistente, cualquier conclusión a la que se llegue sólo tendrá el peso de una opinión, pero carecerá de la ponderación científica. El discurso estará en función de la credibilidad que se tenga en quien emita la opinión, más que sobre los datos que resistan una prueba de hipótesis en condiciones de replicación en la metodología cuantitativa, o de la triangulación en la metodología cualitativa, por diferentes investigadores y en diversos contextos.

Cuando se nos presentan datos cuantificados como resultado de una investigación, la tendencia general es a la credibilidad de las hipótesis de apoyo. En las ciencias exactas existe un mayor isomorfismo entre el dato observado y su representación numérica en un instrumento de medición, pero en las ciencias sociales el isomorfismo es relativo porque presenta una mayor variabilidad.

Aun cuando la educación escolarizada no constituye todo el universo de la educación, sí ha sido la base del escalamiento social; dadas las condiciones económicas por las que transitamos, se ha hecho necesario presentar los resultados de esta educación de manera incuestionable, como un valor social prioritario y no como dispendios, cambiando el enfoque de costo-utilidad o costo-beneficio a objetivos de costo-efectividad y costo-eficiencia. Esto es necesario para responder a los requerimientos de una sociedad en la solución de sus problemas, con independencia de la índole de los mismos, dado que el costo educativo gravita sobre los ingresos de quien la propicia mediante la vía impositiva.

Desde el punto de vista humano, es cuestionable negar la oportunidad a quien propugna por su beneficio personal, pero dadas las posibilidades de oferta educativa y su contingencia en nichos laborales específicos, ofrecer una oportunidad a una persona con escasa oportunidad de culminar sus estudios interfiere en las posibilidades de ofrecerla a dos o tres personas que podrían ser útiles a la sociedad que las subsidia, excluyéndolas de los beneficios de la decisión. Por ello, es necesario que los instrumentos que reflejen las potencialidades de solución social de los estudiantes tengan la suficiente precisión, de modo que sirvan para tomar las decisiones necesarias tanto para quien ofrece como para quien demanda los beneficios del bien parcial de la educación escolarizada.

Si nuestras formas de evaluar los contextos donde nos movemos, los insumos con los que contamos, los procesos que ocurren, los productos que se obtienen y la estabilidad de los beneficios logrados no cuentan con un referente objetivo, válido y confiable que permita juicios claros y decisiones precisas, los cambios estarán destinados al fracaso. De aquí la importancia de evaluar los métodos de evaluación y la trascendencia de contar con un reflejo de la realidad que permita juicios y decisiones transparentes.

No obstante, existe la conciencia de que los métodos, así como los instrumentos que de ellos emanen, tienen la misma connotación que se le adjudica a las armas: el resultado final no depende de ellos, sino principalmente de quién, cómo y para qué se haga uso de los mismos.

La costumbre en nuestro medio de elaborar pruebas educativas o de importar pruebas de selección en empresas sin realizar un análisis exhaustivo de cada uno de los ítemes en cuanto a su dificultad, discriminación y ajuste, así como utilizar instrumentos que no cuentan con la validez y confiabilidad adecuadas, dificultará tomar decisiones racionales respecto a la idoneidad de los sujetos en relación con su potencialidad educativa o su desarrollo en la empresa, en perjuicio tanto de las personas que cuentan con las características deseables como de aquellas que ingresan al estar sometidas a presiones insalvables, y de la institución que los admite al disminuir su eficiencia educativa o laboral.

Errores de medición[1]

El resultado de una medida y el valor real de la cantidad medida no suele ser precisamente igual. La diferencia entre estos dos valores puede tener su origen en errores del azar o en errores sistemáticos. Los errores del azar son los que ocurren en el acto de medición en sí mismo; los errores sistemáticos ocurren como resultado de las fallas del instrumento y las deficiencias de calibración.

Para obtener una medida significativa se debe especificar siempre la precisión con la que se hace, es decir, los límites de rechazo. El intervalo dentro del cual las fallas de valor de medida reales determinan el error absoluto de la medida, normalmente se trabaja con un 5% de error, que es el punto que permite mayor control tanto del error tipo Alfa como del error tipo Beta, es decir, controla mejor las posibilidades de rechazar nuestra hipótesis cuando de hecho es verdadera, así como de aceptar nuestra hipótesis cuando de hecho es falsa. El error relativo es igual al error absoluto dividido entre el valor medido, y normalmente se expresa como un porcentaje. La media aritmética y la dispersión caracterizan la distribución

[1]Miguel Ángel Rosado Chauvet, traducción y edición del artículo de Steven J. Dick basado en M. D. Anthony, "Engineering Methodology", 1999; A. O. Dilke, *Mathematics and Measurement,* 1987; O. E. Doeblin, *Measurement Systems,* 3ª ed., 1982; V. J. Drazil, *Quantities and Units of Measurement,* 1983; Steven Geczy, *Basic Electrical Measurements,* 1984; P. L. Hewitt, *Modern Techniques in Methodology,* 1984; J. F. Liebman y A. Greenberg (eds.), *Physical Measurements,* vol. 2, 1986; M. U. Reissland, *Electrical Measurement,* 1989; R. S. Sirohi y R. Krishna, *Mechanical Measurements,* 1983, para Enciclopedia Multimedia Grolier, 1997.

de los posibles resultados de la medida. La media aritmética de la distribución de probabilidad coincide con el valor de la cantidad medida si no hay error sistemático.

Pueden establecerse dos pautas para los errores del azar: 1) repetir una medición que proporcione información sobre la magnitud de los errores de azar y 2) repetir una medición que reduzca el error de medida donde el resultado final es la raíz cuadrada de *n*, siendo *n* el número de medidas tomado. El error del azar disminuye rápidamente al principio, pero después es más lento, y no puede reducirse más cuando los errores sistemáticos empiezan a predominar.

La magnitud de los errores sistemáticos es más difícil de estimar y reducir. El proceso de medición debe analizarse con todo cuidado en cada caso. Cada tipo de medida tiene sus propios errores sistemáticos característicos, pero pueden enumerarse los más importantes:

1. Errores del punto nulo, causados por un error de medición en la condición nula o por una parte nula defectuosa del instrumento, lo cual producirá a menudo un cambio moderado constante de todos los valores.

2. Los errores de calibración se presentan cuando las condiciones bajo la medida de referencia (calibración) no se toman lo más próximas posible a las condiciones de medida real.

3. El propio instrumento de medición casi siempre influye en la magnitud del signo que se va a medir.

4. La histéresis y el movimiento perdido son fenómenos en los que la indicación de un instrumento de medición depende de la lectura previa.

5. Los errores de paralaje son el resultado de que en la mayoría de las carátulas de los instrumentos el indicador se localiza a una distancia ligera de la escala, y la lectura depende del ángulo del que se toma.

La mejor forma de corregir los errores sistemáticos consiste en convertirlos en errores al azar. Esto puede hacerse introduciendo tantas variaciones como sea posible en el método de medición y en el instrumento.

Generalidades de la medición[2]

Una medición es la asignación de numerales a objetos o acontecimientos según ciertas reglas, pero en las ciencias sociales las medidas como aprendizaje, inteligencia, actitud, honestidad, cohesión, etc. no pueden tratarse como en las ciencias naturales con medidas como peso, longitud o volumen. No obstante, si se establecen reglas sobre una base racional o empírica, la medición es teóricamente posible teniendo en cuenta que ninguna medición es mejor que sus reglas.

La asignación de numerales no tiene un significado cuantitativo si no le damos tal significado. Un numeral adquiere la característica de número sólo cuando se le ha asignado un significado cuantitativo, es decir, cuando los objetos de un conjunto se proyectan en los objetos de otro conjunto, donde la función es una regla de correspondencia. Una *regla* es una guía, método u orden que nos dice qué hacer, con tal que en los objetos por medir se considere el dominio, y los numerales que se asignan o proyectan sobre ellos consideren la extensión. Por esto, la medición es una relación o conjunto de pares ordenados que implican una *función*. La *función*, o regla de correspondencia, es igual al conjunto de pares ordenados (x, y) tal que x es un objeto y cada y correspondiente es un numeral. En medición debemos cuestionar si los procedimientos empleados tienen cierta correspondencia racional y empírica con la *realidad*. Si observamos un objeto en ciencias sociales y le asignamos un numeral, la relación debe ser *isomórfica*, es decir, que haya una buena correspondencia entre los dos órdenes de valores.

Cuando medimos objetos, ponderamos sus propiedades o características; no medimos el objeto en sí, sino los *indicadores* de las propiedades de los objetos.

[2] Miguel Ángel Rosado Chauvet, (1999), edición y síntesis de F. N. Kerlinger en *Investigación del comportamiento. Técnicas y metodología*, 1975; y de S. Stevens en *Mathematics measurement and psychophysics* 1951, citado por Kerlinger en su obra.

En ciencias naturales las propiedades están estrechamente vinculadas con la observación directa, pero en ciencias sociales las propiedades son más complejas y elusivas, por lo que debemos *inferir* las propiedades o características por la observación de presuntos indicadores de las propiedades, que son especificados por definiciones operacionales que indican las actividades u "operaciones" necesarias para medir la propiedad de las variables o construcciones.

Cualquier procedimiento de medición se inicia con la definición de los objetos del universo de discurso. A continuación, deben definirse las propiedades de los objetos del universo, divididas en dos subconjuntos cuando menos, mutuamente excluyentes y exhaustivos; cada objeto se asigna a un subconjunto, y sólo a uno. Para esto contamos con tres reglas o postulados básicos:

1. Para clasificar debemos ser capaces de afirmar que un objeto *es igual* (a ≡ b) en una característica a otro, o que *no es igual* (a ≠ b), de acuerdo con uno o más criterios que no sean ambiguos.

2. Para establecer igualdad entre los miembros de un conjunto, debemos satisfacer que pueden asignarse objetos ordinariamente no susceptibles de observación a subconjuntos de un universo. Si (a = b) y (b = c), entonces (a = c).

3. La transitividad permite formar expresiones ordinales o de orden de clasificación. Si [(a.>.b) y (b.>.c)], entonces (a.>.c).

Características de un instrumento de medición

Tras asignar numerales a objetos o acontecimientos según reglas, el investigador debe enfrentarse a problemas de *confiabilidad* y *validez*. Los datos de todos los instrumentos de medición en ciencias sociales contienen errores de medición y, en la medida en que los contengan, en esa misma medida los datos que proporcionen no serán fidedignos.

Es posible enfocar la definición de confiabilidad en tres formas: a) si medimos el mismo conjunto de objetos una y otra vez con el mismo instrumento de

medición y obtenemos el mismo valor, será *estable*; b) si medimos el mismo conjunto de objetos con dos instrumentos similares y obtenemos el mismo valor, será *equivalente,* y c) si cada una de las partes del instrumento mide la misma característica que el instrumento completo, éste será internamente *consistente*.

La validez se refiere a que un instrumento mida lo que se pretende medir. La mayoría de las estructuras (*constructos*) en ciencias sociales son invenciones que no pueden definirse en forma objetiva mediante su observación directa (metro, kilo, litro), sino que deben medirse a través de sus indicadores definidos de manera operacional (aprendizaje, aptitud, actitud, inteligencia).

Si el indicador se define de manera exhaustiva, contamos con un *constructo;* y si se obtiene una muestra de éste en una forma suficiente y representativa, contamos con un buen *contenido* del mismo. Si se cumplen en forma satisfactoria ambos requisitos, un análisis de consistencia del instrumento permitirá obtener su validez *teórica* (validez de *construcción* y *contenido*).

Si los resultados del instrumento mantienen un isomorfismo directo con el objeto o acontecimiento medido, o con otro instrumento que sea isomórfico con el acontecimiento u objeto medido, contamos con una validez *concurrente*, y si el isomorfismo se refiere a una medición actual relacionada con un objeto o acontecimiento futuro, se contará con una validez *predictiva*. Ambas formas de validez corresponden a la validez *empírica*.

La medición en instituciones educativas y empresas

Toda institución educativa escolarizada, incluyendo en ésta tanto la educación *in situ* como la educación abierta y la educación a distancia, así como la capacitación, el entrenamiento y el desarrollo en empresas por lo que de institucional poseen, debe tener planes y programas de estudio aceptados con respecto a los objetivos educativos que se espera cumplir. El contenido del programa de una asignatura o un módulo constituye el *constructo* del que partimos para la evaluación del logro de objetivos, y el muestreo suficiente y representativo del contenido del programa nos permite inferir si el universo del *constructo* se ha logrado, dentro de un continuo taxonómico que implique desde aspectos básicos hasta complejos asociados con el aprendizaje.

Si lo anterior se cubre a cabalidad, debemos centrarnos en verificar si cada una de las partes que integran el instrumento mantienen congruencia con la población a la que se pretende medir (evaluación con enfoque a normas), así como con la obtención de los objetivos del programa (evaluación con enfoque a criterios), es decir, debe cumplir con las expectativas del programa tanto como con los aprendizajes esperados en el grupo, respetando las diferencias individuales.

Cada una de las partes que integran el instrumento debe representar objetivos del programa, pudiendo tratarse de preguntas que impliquen una respuesta específica, declaraciones que soliciten la manifestación de actitudes o conductas observables y exhaustivas. Cualquiera de estas manifestaciones de la obtención de un aprendizaje cognoscitivo, afectivo o psicomotor queda incluida en el término inclusivo de ítemes.

Se llama *test* a la situación experimental normalizada que sirve de estímulo a un comportamiento que se evalúa por una comparación estadística con el de otros individuos colocados en la misma situación, permitiendo clasificar al individuo examinado en términos ya sea cuantitativos o bien tipológicos. La calidad y la adecuación del test dependen directamente de los ítemes que lo conforman, quedando el todo en función de las partes.

Tanto la teoría clásica (TC) como la teoría de respuesta al ítem (TRI) parten de un cuidadoso análisis de los contenidos por parte de varios expertos en el dominio por medir y en aspectos relativos a la medición, iniciándose un examen crítico del material presentado para decidir sobre los contenidos y aspectos formales de los ítemes, asegurando la congruencia del dominio. El test puede ser unidimensional o multidimensional, según pertenezca a un solo dominio o a una diversidad de ellos. Si no existe claridad sobre los límites y dimensiones de este dominio (*constructo*), se estará violando un aspecto básico de la validez teórica del instrumento.

En ambas teorías, una vez clarificados los alcances y limitaciones del dominio, deberán elaborarse ítemes suficientes que incluyan los aspectos relativos con diversos niveles de dificultad para cada punto, formando de esta manera un *banco de ítemes* lo más amplio y completo posible. Para delimitar el dominio se han desarrollado taxonomías diversas, así como formatos que nos permiten

mejorar la calidad de los ítemes. A partir de este banco podremos seleccionar los que formen el test, incluyendo una muestra suficiente y representativa del dominio por medir para responder al segundo aspecto de la validez teórica.

· Aquí se da la primera diferencia entre las teorías, pues la TC toma en cuenta tanto la dificultad como la discriminación del ítem, mientras que la TRI sólo se basa en la dificultad.

Posteriormente debemos conformar instrumentos que se aplicarán sin fines evaluativos para quien responde, y cuyo único propósito es el de obtener una validez empírica relativa, mediante el ulterior análisis estadístico de los ítemes. Este proceso recibe el nombre de *piloteo* del instrumento, el cual lleva a una primera confrontación entre el *deber ser* y la *realidad educativa* que permite detectar un sesgo no observado en el análisis previo, el cual se genera por diferente cultura, entorno social, nivel socioeconómico, etc., de los sujetos que responden, así como la falta de compromiso al no encontrarse en juego ninguna acción contingente al resultado, o bien por alguna falla imputable directamente al instrumento, como errores al asignar la respuesta correcta (*clave*), respuestas que no sean mutuamente excluyentes, lenguaje que no corresponda a los sujetos que responden, etc. Cuando una variable es contaminada por otra, puede sesgarse la medida en función de la variable contaminadora, por ejemplo un ítem de aptitud numérica que exija un alto nivel de comprensión verbal por la forma como se presenta el problema.

Parámetros del ítem

a *Índice de discriminación*, que en la curva logística de la TRI, si es invariante, también se ajustará a una recta que atraviese el origen, con una pendiente recíproca del parámetro *b*. Desde el enfoque de la TC es la diferencia de aciertos entre un subgrupo de "alto rendimiento" y otro de "bajo rendimiento", en proporción al total de aciertos de ambos subgrupos.

b *Índice de dificultad*, donde la suma de los valores refleja el número de sujetos que acertó al ítem en proporción al número total de respuestas del total del grupo, yendo desde los mínimos que indican gran dificultad hasta

los máximos que indican una gran facilidad. Si existe invarianza perfecta, los ítemes se ubicarán sobre una recta y, a medida que se alejen de dicha recta, calculada mediante la correlación producto-momento de Pearson, la varianza se resentirá.

c La *probabilidad de acertar al ítem por azar* cuando se desconoce la respuesta correcta, y está dada por la probabilidad de 1/k, donde k es el número de opciones. Este parámetro no se ve afectado por la elección del origen de la escala y sus unidades en una curva logística, por lo que su estimación será idéntica para las muestras.

Desde el enfoque estadístico se sugiere utilizar $z.=.\pm 1.96$ con un mínimo de error del 5%, considerando que los límites críticos admisibles suponen que *el mejor ítem es aquel que es respondido en forma correcta por la mitad del grupo y en forma incorrecta por la mitad complementaria, siempre y cuando la mitad que acierta corresponda al grupo de máximo rendimiento y la mitad que falla corresponda al grupo de mínimo rendimiento*.

· La diferencia entre la TC y la TRI, con respecto a la discriminación, consiste en que la TRI plantea que no hay bases para determinar a qué corresponde un grupo de máximo y mínimo rendimiento, mientras que la TC supone que, si el instrumento ha cumplido con los supuestos de la validez teórica, el grupo de máximo rendimiento será el que acierte con mayor frecuencia a los ítemes del instrumento y viceversa.

El modelo de un parámetro de Rasch[3] de la TRI sugiere que los valores máximos admisibles en los extremos de la logística se sitúan entre –2.000 y +2.000, lo cual corresponde a dificultades entre 0.1192 y 0.8808 calculadas como el logaritmo natural de p/q y q/p. No obstante, el modelo sugiere que la mayor precisión se encuentra entre –1.000 y +1.000, con valores de dificultad entre 0.2689 y 0.7311 respectivamente.

[3] En los estudios con la técnica de Rasch para un parámetro se utilizó el programa Bigsteps©, versión 2.56 del 13 de abril de 1955, de John Michael Linacre, Chicago, Mesa Press.

Entre los modelos de la TC estudiados, encontramos que el modelo Kalt® sugiere valores de dificultad entre 0.27 y 0.73 (±0.995) y Siseval[4] establece, con base en la prueba binomial con aproximación a *z*, valores entre 0.1166 y 0.8834 (±2.025), ambos representando valores sugeridos en el modelo de la TRI. No obstante, el modelo Siseval indica los valores externos mencionados con la posibilidad de incluir la probabilidad aleatoria para todos los sujetos, así como valores internos de 0.2384 y 0.7166 (±0.928) que excluyen las probabilidades de azar en todo el grupo.

Aquí surge un punto de discusión sobre los límites restringidos o los límites amplios de los modelos. Si optamos por el límite restringido ganamos en exactitud, pero perdemos las respuestas extremas aceptables que nos proporcionan información sobre los individuos que ante una pregunta muy fácil no son capaces de responder con acierto, así como sobre los individuos que frente a una pregunta muy difícil la responden de manera acertada. Si optamos por el límite amplio permitimos la inclusión de respuestas aleatorias en los sujetos que la ignoran, pero nos brinda la posibilidad de medir respuestas de sujetos excelentes; aquí las mediciones son de menor precisión. En todo caso, la oscilación de la medida en el modelo de la TRI es de 0.1497 y la oscilación en el modelo de la TC es de 0.1518 en cada extremo, es decir, cerca de un 30% de oscilación entre el límite laxo y el restrictivo como criterios de inclusión o exclusión.

Por otra parte, ni el modelo de la TRI para un parámetro ni el modelo Kalt® de la TC toman en cuenta la diferencia en la cantidad de sujetos, por tratar con proporciones. Tampoco varían las probabilidades de acierto en función del número de opciones de la pregunta. El modelo Siseval de la TC sí los incluye, y en él los extremos pueden variar de acuerdo con los siguientes ejemplos:

1. Si tenemos 100 casos, o se trabaja con porcentajes,

   · con 5 opciones los límites exteriores estarían entre 0.1166 y 0.8834.

[4] El programa Siseval se encuentra en desarrollo, principalmente para evaluación en aula, con el enfoque de la teoría clásica, mediante técnicas desarrolladas por Miguel Ángel Rosado Chauvet y programado en Turbo Pascal por Eduardo Victor Rosado Colmenares.

· con 4 opciones los límites exteriores estarían entre 0.1601 y 0.8399.

· con 3 opciones los límites exteriores estarían entre 0.2359 y 0.7641.

· con 2 opciones los límites exteriores estarían entre 0.3970 y 0.6030.

2. Si tenemos 40 casos,

· con 5 opciones los límites exteriores estarían entre 0.0635 y 0.9365.

· con 4 opciones los límites exteriores estarían entre 0.1033 y 0.8967.

· con 3 opciones los límites exteriores estarían entre 0.1747 y 0.8253.

· con 2 opciones los límites exteriores estarían entre 0.3325 y 0.6675.

Si siempre tomamos los valores correspondientes a 100 sujetos que responden a ítemes de 5 opciones, tenemos la ventaja de una norma fija pero la precisión de la medición se vicia por falta de flexibilidad en las diferencias específicas. La posibilidad de mantener un solo valor consiste en producir siempre ítemes con la misma cantidad de opciones, estabilizado el valor de 1/k y tratando los resultados de los individuos como proporciones de respuesta en vez de cantidad de sujetos; se supone que en cualquier caso se incluirá un elemento extraño como constante de medición.

En cuanto a la discriminación, hemos observado en estudios de regresión que los valores medios cuadráticos o estandarizados del modelo de la TRI mantienen una relación más estrecha y más alta con los valores de discriminación que con los valores de dificultad de los modelos de la TC. Por otra parte, la discriminación en el modelo Kalt® de la TC tiene como parámetro del índice una proporción del 30% del índice de dificultad. Aquí los resultados obtenidos no son realmente independientes.

Otro aspecto que debe ser común a ambas teorías consiste en la unidimensionalidad, ya que dimensiones dispares dentro de un mismo análisis

pueden llevar a conclusiones erróneas. La comprobación de la unidimensionalidad sigue siendo un campo poco investigado, donde el análisis factorial permanece como técnica indiscutible.

Los modelos de la TC toman como criterio de calidad de los sujetos la suma de sus aciertos, y al viciar las dimensiones podemos tener como iguales a los sujetos que, por ejemplo, a) son excelentes en matemáticas y pésimos en español, b) son excelentes en español y pésimos en matemáticas y, c) son mediocres en ambas materias. En los tres casos este valor viciado nos dará como resultado sujetos que responden a la mitad de ítemes, manteniéndolos en el centro del parámetro.

Los modelos de la TRI no son ajenos a este influjo porque parten de los aciertos independientes a las preguntas, mediante el índice de dificultad. Sin embargo, al realizar el análisis de ítemes mediante las iteraciones tomamos las respuestas de los sujetos para incluirlos o excluirlos de análisis, asumimos las respuestas como anormales y al eliminar a los sujetos que no corresponder a la forma de respuesta de los sujetos que se someten a la medición, arrastramos el error hacia los resultados de los ítemes. De aquí que deban realizarse análisis por cada uno de los temas incluidos en las pruebas y no por la prueba en su totalidad.

El modelo Siseval controla la influencia de respuestas atípicas mediante un análisis de congruencia de respuestas de los sujetos, a través de tres aspectos: a) suficiencia de respuesta para ser incluidos en el estudio, b) valores de dificultad del instrumento, excluyendo a los sujetos que no se encuentren dentro de límites aceptables, y c) valores de discriminación en función del grupo, excluyendo los casos en los que no respondan en forma consistente.

Conclusiones

Si bien los modelos cuantitativos mantienen en general una mayor precisión que los modelos cualitativos, hemos podido observar, mediante el análisis de los modelos presentados, que representan a la teoría de respuesta al ítem (Rasch

para un parámetro) y a la teoría clásica (Kalt®[5] de dos parámetros relacionados y Siseval de dos parámetros independientes), las dificultades derivadas de suponer que en ciencias sociales contamos con valores reales y absolutos, aunque aparentemente se tengan como tales.

En todo caso, nada sustituye el rigor metodológico ni las replicaciones en los métodos cuantitativos o las triangulaciones en los métodos cualitativos; aquí debe tenerse en cuenta que cualquier variación en la definición de los *constructos* o en la población meta implica un nuevo estudio exhaustivo de los parámetros y criterios que debemos utilizar para mantener una justicia evaluativa.

Una observación cualitativa que no permita conclusiones similares ante una triangulación sólo tendrá el peso de una opinión, sin que sea generalizable más allá del instrumento de observación, del criterio del observador y del sujeto u objeto observado. Un valor cuantitativo que no supere la prueba de replicación sólo tendrá el peso del resultado obtenido para el sujeto o el grupo medido, con el instrumento utilizado y calificado por un evaluador en las condiciones de una situación específica. Y si bien la ciencia se nutre inicialmente de hechos y datos particulares, su objetivo final consiste en llegar a generalizaciones a través de leyes y principios.

Mientras no tengamos la verdad absoluta en ninguno de los extremos, debemos concebir las diversas metodologías como complementarias (no contrarias) y enriquecer las conclusiones con las aportaciones de ambas, sin darnos tregua en la búsqueda de una realidad que, aun siendo relativa, signifique la mejor opción con la que contamos. No obstante, la metodología que elijamos para el análisis de los instrumentos y sus ítemes siempre implicará un sesgo para la decisión que tomemos en función de sus resultados.

[5] Por la aplicación y el efecto que tiene actualmente en nuestro país la utilización del paquete en diversas instituciones, incluyendo la Universidad Autónoma Metropolitana y el Ceneval, nos hemos dado a la tarea de estudiar sus bases y funcionamiento, y hemos encontrado algunos problemas que se contraponen a la teoría psicométrica y a los principios estadísticos y probabilísticos.

Al elegir los modelos Kalt o Rasch podemos incluir una menor variabilidad en las mediciones si eliminamos el posible acierto por azar, y, unido a ello, eliminar también las posibilidades de medición de los sujetos extremos que fallan a pesar de estar frente a ítemes fáciles o que aciertan aun cuando los ítemes sean dificiles. Estos modelos no toman en cuenta el número de opciones y trabajan sólo con aciertos y errores; sin embargo, la probabilidad a la que se enfrenta un sujeto cuando tiene cinco opciones (1/5 = 0.20) no es la misma que cuando tiene sólo dos (1/2 = 0.50).

Por otro lado, el modelo Siseval admite las respuestas al azar suponiendo que en cualquier caso éste favorecerá a los que tienen mayores conocimientos pues ellos dudarán, por ejemplo, entre dos opciones de cinco posibles, en contraposición a los que poseen menores conocimientos, quienes quizá duden entre cinco opciones posibles. Esto permite un estudio más amplio tanto entre estos sujetos como entre los contenidos más extremos del instrumento; sin embargo, el modelo toma en cuenta el número de opciones, cerrando los límites a medida que las opciones son menores.

El modelo Rasch depura a los sujetos por iteración y el modelo Siseval por proporción de respuestas e idoneidad con el grupo, mientras que el modelo Kalt no depura a los sujetos.

La decisión que se tome en función del modelo o de los indicadores utilizados tendrá necesariamente una repercusión en la aceptación de los ítemes y en la admisión de los alumnos o de los empleados.

Bibliografía

Adkins Wood, Dorothy. *Elaboración de tests,* Trillas, México, 1983.

College Entrance Examination Board. *El análisis de ítems. Materiales sobre los fundamentos y prácticas de las admisiones universitarias,* Hato Rey, The College Board, Puerto Rico, 1979.

Educational Testing Service. *Making the classroom test: A guide for the teachers,* Educational Testing Service, New Jersey, 1961.

Gronlund, Norman E. *Elaboración de tests de aprovechamiento,* Trillas, México, 1973.

Kerlinger, Fred N. *Investigación del comportamiento. Técnicas y metodología,* Interamericana, México, 1975.

Muñiz F., José. *Teoría de respuesta a los ítems,* Pirámide, Madrid, 1990.

Rosado Chauvet, Miguel Ángel. (1973), "Prueba de aptitudes múltiples para obreros de máquinas-herramienta y ayudantes de las mismas". Tesis para obtener el grado de Licenciado en Psicología. Facultad de Psicología, UNAM, México, 1973.

Rosado Chauvet, Miguel Ángel (1999), edición y síntesis de F. N. Kerlinger en *Investigación del comportamiento. Técnicas y metodología*, 1975; y de S. Stevens en *Mathematics measurement and psychophysics,* 1951, citado por Kerlinger en su obra.

Rosado Chauvet, Miguel Ángel. "Cinco estudios de estadística para la evaluación", *Cuadernos del CENEVAL*, núm. 96, vol. 2, México, pp. 3-6.

Rosado Chauvet, Miguel Ángel. "Hacia un modelo de evaluación de la docencia en la educación superior". Tesis para obtener el grado de Maestría en Psicología Social. Facultad de Psicología, UNAM, México, 1998.

Rosado Chauvet, Miguel Ángel. "Análisis de ítemes. Teoría Clásica y Teoría de Respuesta al Ítem. Comparación y aplicación de tres modelos". Tesis para obtener el grado de Doctor en Educación. Universidad Autónoma de Tlaxcala, División de Estudios de Postgrado, Departamento de Ciencias de la Educación, 1999.

Siegel, Sidney. *Estadística no paramétrica. Para las ciencias de la conducta,* Trillas, México, 1975.

Stake, Robert E. *La imagen de la evaluación educacional*. Universidad de Illinois. Materiales del Proyecto Multinacional de Evaluación de la OEA: Investigación Evaluativa, Universidad del Valle de Guatemala, Guatemala, 1980.

Stufflebeam, Daniel L. *Enfoques alternativos para la evaluación educativa: una guía de auto-estudio para educadores (Módulos I-II),* Materiales del Proyecto Multinacional de Evaluación de la OEA: Investigación Evaluativa, Universidad del Valle de Guatemala, Guatemala, 1980.

Tavella, Nicolás. M. *Análisis de los ítemes en la construcción de instrumentos psicométricos,* Trillas, México, 1978.

Tristán, Agustín. "Relación entre grado de dificultad y discriminación (1). Primera parte: Estudio del grado de dificultad", *Noticias ICI*, núm. E-10, San Luis Potosí, 8 de marzo de 1995.

Tristán, Agustín. "Relación entre grado de dificultad y discriminación (2). Segunda parte: Estudio de la discriminación", *Noticias ICI*, núm. E-11, San Luis Potosí, 8 de marzo de 1995.

Tristán, Agustín. "Relación entre grado de dificultad y discriminación (3). Tercera parte: El dominio de discriminación", *Noticias ICI*, núm. E-12, San Luis Potosí, 11 de marzo de 1995.

Tristán, Agustín. "Relaciones entre grado de dificultad y discriminación (4). Cuarta parte: Norma discriminativa", *Noticias ICI*, núm. E-13, San Luis Potosí, 12 de marzo de 1995.

Wright, Benjamín D. y Mark H. Stone. *Diseño de mejores pruebas. Utilizando la técnica de Rasch.* Traducción al castellano. México, CENEVAL.