Para citar este artículo, le recomendamos el siguiente formato:

Martín, N., Díaz, C., Córdoba, G. y Picquart, M. (2011). Calibración de una prueba química por el modelo de Rasch. *Revista Electrónica de Investigación Educativa*, *13*(2), 132-148. Consultado el día de mes de año en: http://redie.uabc.mx/vol13no2/contenido-martindiazetal.html

ISSN: 1607-4041

Revista Electrónica de Investigación Educativa

Vol. 13, No. 2, 2011

Calibración de una prueba de química por el modelo de Rasch

Calibration of a Chemistry Test Using the Rasch Model

Nancy Martín Guaregua (1)
mgnc@xanum.uam.mx

Consuelo Díaz Torres (2)
ditc@xanum.uam.mx

Gilberto Córdoba Herrera (1)
gil@xanum.uam.mx

Michel Picquart (3)
mp@xanum.uam.mx

Universidad Autónoma Metropolitana-Iztapalapa
(1) Departamento de Química
(2) Departamento de Matemáticas
(3) Departamento de Física

Av. San Rafael Atlixco 186. C.P. 09340
D. F., México

(Recibido: 9 de diciembre de 2010; aceptado para su publicación: 29 de julio de 2011)

Resumen

Se aplicó el modelo de Rasch para calibrar una prueba de química general con el fin de analizar las ventajas y la información que proporciona el modelo. La muestra fue de 219 alumnos del primer año universitario. Se logró un buen ajuste del modelo en 10 reactivos de un total de 12. Además se evidenció que la prueba cuenta con reactivos de diferentes índices de dificultad, con intervalos vacíos en la escala para los que se tendrán que diseñar nuevos reactivos para que la prueba sea completa.

Palabras clave: Teoría de Respuesta al Reactivo (TRR), modelo de Rasch, Curva Característica del Reactivo (CCR), química.

Abstract

The Rasch model was used to calibrate a general chemistry test for the purpose of analyzing the advantages and information the model provides. The sample was composed of 219 college freshmen. Of the 12 questions used, good fit was achieved in 10. The evaluation shows that although there are items of variable difficulty, there are gaps on the scale; in order to make the test complete, it will be necessary to design new items to fill in these gaps.

Key words: Item Response Theory (IRT); Rasch model; Characteristic Graphical of Item (CGI), chemistry.

I. Introducción

Un instrumento de medición de buena calidad debe tener validez, confiabilidad y objetividad. Tiene validez si mide lo que se pretende medir, es confiable si es independiente del tiempo y es objetivo si es independiente de evaluadores y evaluados. Por lo tanto, la validez, la confiabilidad y la objetividad de una prueba son aspectos importantes a considerar cuando se lleva a cabo su elaboración. En educación se considera al instrumento de medición como el medio a través del cual se recaba información y se registran los datos que permiten una valoración sobre la habilidad o el conocimiento de los alumnos.

Entre las diferentes pruebas de evaluación que existen están: oral, escrito y práctico-procedimental. Todos ellos pueden ser complementarios. La evaluación escrita es la más comúnmente usada como prueba de conocimiento y dentro de esta clasificación, se pueden mencionar, a las pruebas abiertas, las de opción múltiple, estudios de casos, etc.

Las pruebas de opción múltiple están compuestas por reactivos o ítems cerrados con varias opciones de respuesta de las que una sola es correcta y las otras son distractores. Éstas presentan ciertas ventajas sobre las pruebas abiertas en que permiten rapidez al calificar y pueden ser aplicadas a grupos grandes de alumnos. Entre las desventajas están, que pueden ser respondidas al azar y la dificultad

para diseñar reactivos de buena calidad.

De allí, la importancia que tiene la calibración de la pruebas de evaluación que se aplica a fin de asegurar su calidad, de forma que se pueda evaluar con eficiencia el conocimiento.

En la Universidad Autónoma Metropolitana-Iztapalapa (UAMI) no se tiene establecido ningún tipo de norma para el diseño de los exámenes. Tampoco se aplica ningún tipo de valoración o análisis que permita comprobar la validez y la confiabilidad de la evaluación que hacemos a los alumnos.

El propósito del presente estudio es analizar y calibrar los reactivos de una prueba de opción múltiple usada como examen diagnóstico de química general en los alumnos del primer año universitario de la División de Ciencias Básicas e Ingeniería (DCBI) de la UAMI. Después de la revisión de los diferentes métodos que existen para el análisis de instrumentos se decidió aplicar el modelo de Rasch con el fin de analizar el modelo y su ajuste a los reactivos de esta prueba.

II. Antecedentes

Entre los modelos estadísticos más antiguos para análisis de instrumentos está el de la Teoría Clásica de la Medida (TCM) desarrollado por Spearman (Spearman, 1904, Stevens, 1946, Muñiz, 1997), el cual es un modelo de regresión lineal que ha sido aplicado a pruebas sicométricas. (Embretson y Reise, 1986).

La principal limitación de la TCM consiste en que las características de la prueba y las puntuaciones de los evaluados no pueden ser separadas, ya que las características de los reactivos dependen del grupo de personas al que se han aplicado y la puntuación de una persona depende del conjunto particular de reactivos utilizados.

Estas limitaciones han llevado a la propuesta de modelos alternativos. Uno de ellos, surgido en los años 60 para complementar el primero es el atribuido a Cronbach (1971) y llamado de la generalizabilidad, el cual gracias al uso específico del análisis de variancia hace posible analizar las distintas fuentes de error que se presentan en los puntajes mediante el concepto de faceta, término introducido por Cronbach para designar cada una de las características de la medición y puede modificarse de una ocasión a otra; por tanto, hace variar los resultados obtenidos (por ejemplo, los reactivos de una prueba, las formas de codificar las respuestas, los tipos de examen, etc.).

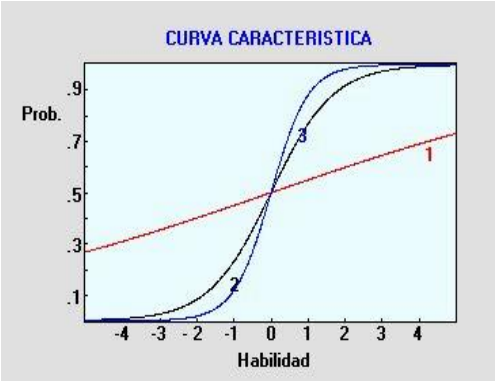
Un método adicional para la validación de un instrumento de opción múltiple es el de Sympson y Haladyna (1988) quienes desarrollaron un método de ponderación múltiple de las respuestas a los reactivos de una prueba (Backhoff, 2000).

Un enfoque más reciente es el de la Teoría de la Respuesta al Reactivo (TRR)

(Hambleton y Swaminathan, 1985, Embretson y Reise, 2000) que se centra más en las propiedades de los reactivos individuales que en las propiedades globales de la prueba. Se trata de llegar a la elaboración de instrumentos de medición cuyas características no sean demasiado influenciadas por un grupo de referencia dado. Se basa en el postulado de que la respuesta de un individuo al reactivo, en particular su probabilidad de dar una respuesta correcta, está determinada por dos tipos de factores: por una parte, algunos atributos del sujeto, su competencia por ejemplo, que no son directamente accesibles a la observación y a la medición y que son generalmente calificados de rasgos latentes y por otra, las propiedades del reactivo, en particular su dificultad, su poder de discriminación o el azar en algunos casos. Por lo tanto, se considera la respuesta al reactivo como una función de las características del individuo y las del reactivo.

La propiedad de invarianza es la característica principal de la TRR. Nos dice que las estimaciones relativas a los reactivos (parámetro de dificultad, de discriminación y de azar) son independientes de la muestra particular de individuos y que las estimaciones relativas a los individuos (nivel de competencia, de habilidad) son independientes de la muestra de reactivos utilizada. Por esto se tiene que asegurar que el ajuste de los datos al modelo sea satisfactorio. Así, la relación matemática se puede representar como la probabilidad de contestar satisfactoriamente el reactivo *i* en función del nivel de habilidad (θ_S) para un sujeto *s*. El índice de discriminación del reactivo se define como una medida para determinar si las competencias o habilidades que mide la prueba también las mide el reactivo. Un buen reactivo debe discriminar entre aquéllos que obtuvieron altas calificaciones en la prueba y aquéllos que obtuvieron bajas calificaciones.

La relación anterior se expresa por una función matemática logística (función característica del reactivo) representada por una gráfica de forma de sigmoides (Rojas y col. 2004). En la figura 1 se representa un ejemplo de este tipo de gráfica llamada Curva Característica del Reactivo (CCR) que muestra la función de probabilidad (P_{is}) en función de la habilidad (θ_S). Las tres curvas presentan diferentes pendientes que corresponden a diferentes índices de discriminación. Las curvas deseables son las que presentan una forma de “S”, como las 2 y 3, lo que implica que el paso de acertar o fallar debe ser gradual. En tanto que, la curva 1, de forma lineal, es la menos discriminatoria y menos deseable. Estas curvas pueden variar según el ajuste del modelo.



Rojas, *et al.* 2004.

Figura 1. Ejemplo de una curva CCR

El Modelo de Rasch

El modelo matemático más sencillo en el marco de la TRR, es el propuesto por Georg Rasch (1960), conocido como modelo de un parámetro. Es un modelo matemático sencillo que permite analizar la medición conjunta en una misma escala, de las personas y de las puntuaciones obtenidas en una prueba dada.

Las ventajas del modelo de Rasch con respecto a otros modelos de la TRR es que es muy simple de aplicar. Otra característica, es que permite analizar las interacciones entre los alumnos y los reactivos. Además, las medidas que se obtienen no dependen de las condiciones específicas de cómo son obtenidas. Otra propiedad, es la unidimensionalidad del instrumento, es decir, que todos los reactivos que lo componen contribuyan a evaluar una sola característica o competencia. Finalmente, la independencia local postula que la respuesta a un reactivo no está influenciada por las respuestas a los otros reactivos del instrumento.

El modelo propuesto por Rasch proporciona una solución para calibrar pruebas de evaluación y se basa en las siguientes suposiciones: a) el instrumento a medir se representa en una dimensión, en la que se ubican de manera conjunta a los alumnos y a los reactivos de la prueba; b) el nivel del alumno en el instrumento y la dificultad del reactivo determinan la probabilidad de que la respuesta sea correcta.

Haciendo un ajuste adecuado de los datos es posible obtener pruebas más eficientes. La relación matemática se puede representar como la probabilidad de que un reactivo *i* tenga una respuesta satisfactoria, para un nivel de habilidad θ_s del sujeto *s*, según:

$$P_{is} = \frac{1}{1+e^{-(\theta_s-\beta_i)}} \qquad (1)$$

La ecuación (1) se puede representar como el cociente entre la probabilidad de una respuesta correcta a un reactivo P_{is} y la probabilidad de una respuesta incorrecta ($1-P_{is}$) y la diferencia entre el nivel de habilidad de una persona (θ_s) y el nivel de dificultad de un reactivo (β_i):

$$\frac{P_{is}}{1-P_{is}} = e^{(\theta_s - \beta_i)} \qquad (2)$$

Entonces, cuando una persona responde a un reactivo equivalente a su umbral de competencia (o habilidad), tendrá la misma probabilidad de una respuesta correcta y de una respuesta incorrecta. Esto es, $[P_{is}/(1-P_{is})] = 0.50/0.50 = 1.0$. En este caso, se tiene que la dificultad del reactivo es equivalente al nivel de la habilidad de la persona (θ_s-β_i) = 0. Si la habilidad de la persona es mayor que la requerida por el reactivo la probabilidad de una respuesta correcta será mayor que la de una respuesta incorrecta [(θ_s-β_i)>0]. Por el contrario, si la habilidad de la persona es menor que la requerida por el reactivo la probabilidad de respuesta correcta será menor que la de una respuesta incorrecta [(θ_s-β_i)<0)]. Este modelo ha sido aplicado acertadamente en muchas áreas como, psicología, educación, medicina y socio-economía (Golia, 2011). Un buen ajuste de las medidas obtenidas depende de las suposiciones planteadas y la calidad de la prueba.

Existen distintas escalas métricas de los valores de las personas y los reactivos. La más utilizada es la escala lógitos, que es el $Ln[P_{is}/(1-P_{is})]$, es decir (θ_s-β_i). La localización del punto cero de la escala es arbitraria, pero por tradición el punto cero indica la habilidad media de los sujetos (Embretson y Reise, 2000). El intervalo de probabilidad, en la gran mayoría de los casos, se ubica entre -5 y 5.

Por tanto, el interés es estimar a los parámetros, (θ_s) habilidad de las personas, (β_i) índice de dificultad e índice de discriminación de los reactivos. Como el ajuste es a un solo parámetro, el índice de discriminación toma un valor de uno y el índice de dificultad (β_i) es el que se estima en el modelo para explicar las características de cada reactivo. La interpretación de la prueba se fundamenta en la probabilidad, alta o baja, que tiene un alumno de contestar correctamente un reactivo. Los procedimientos de cálculo de éstos son largos por lo que es necesario usar programas de computadora como, Winsteps, Bigsteps o Rascal, entre otros. Los procedimientos de análisis permiten detectar a los reactivos y a los alumnos que no se ajustan al modelo.

Por tanto, además de estimar los parámetros antes mencionados es necesario determinar el grado en que los datos obtenidos se ajustan al modelo. Existen dos medidas de bondad de ajuste en el modelo de Rasch: el INFIT que se interpreta como ajuste interno, es un valor sensible al comportamiento inesperado que afecta a los reactivos cuya dificultad está cerca del nivel de habilidad de una persona y el OUTFIT que se interpreta como ajuste externo, es un valor sensible al comportamiento inesperado que afecta a los reactivos cuya dificultad está lejos del

nivel de habilidad de una persona. Estos estadísticos se reportan como medias cuadráticas de residuales (MNSQ) y como residuales estandarizados (ZSTD). En la práctica el criterio que se aplica es que los valores de la media cuadrática deben estar entre 0.8 y 1.3 y los valores estandarizados deben estar entre -2 y 2, lo cual indica un ajuste razonable (González, 2008). Los valores INFIT o OUTFIT fuera de este intervalo indican una falta de ajuste al modelo, valores de la media cuadrática menores a 0.8 o valores estandarizados menores a –2 indican datos con demasiado determinismo o poco estocásticos, mientras que valores de la media cuadrática mayores a 1.3 o valores estandarizados mayores a 2 indican alta posibilidad de azar (Tristán, 1998).

Otro estadístico útil en la calibración de los reactivos es el coeficiente de correlación punto-media que mide el grado de asociación entre el puntaje particular observado para el reactivo individual y el puntaje total observado en la prueba. Valores altos de esta correlación indican que el reactivo "trabaja en la misma dirección que el conjunto de reactivos" al que pertenece la prueba (González, 2008).

III. Metodología

Con el fin de realizar un diagnóstico de ideas básicas en química general se diseñó un examen diagnóstico (Anexo 1) con algunas preguntas tomadas del Journal of Chemical Education (JCE) y otras tomadas de libros de texto correspondientes a los temas de los primeros cursos de química de la UAMI. Los temas que se incluyeron fueron: concepto de mol, relaciones molares, nomenclatura de sales, teoría cinética de los gases, conservación de la masa y transformación de fases.

La prueba fue de opción múltiple con 12 reactivos y 5 opciones de respuestas cada una. Sólo se tenía una respuesta correcta y cuatro distractores. Se dio un tiempo de 40 minutos para la resolución de la prueba, el cual se consideró suficiente dado que el 95% de los participantes terminó en ese tiempo.

Participó una muestra de 219 alumnos del primer año universitario de la DCBI de la UAMI. La aplicación de la prueba tuvo lugar durante primavera de 2009. Los resultados fueron recopilados en una base de datos.

IV. Resultados y Discusión

En primer lugar, se calcularon estadísticos descriptivos con los datos obtenidos de la aplicación de la prueba de evaluación (Anexo 1) con el fin de conocer los resultados del grupo (N=219). Se considera este tamaño de muestra (N>100) apropiado para su análisis estadístico

La prueba fue de 12 reactivos, los cuales fueron clasificados en aciertos y

desaciertos, a través de un conjunto de "distractores" o respuestas incorrectas. La prueba no fue planificada con pesos relativos a los aciertos en los diferentes reactivos. Al respecto, se ha reportado (Golia, 2011) que la precisión del modelo de Rasch va en relación lineal con el tamaño de la prueba. Esto es, al aumentar el número de reactivos, se logra un mejor ajuste. Sin embargo, el autor demostró con una prueba con 10 reactivos que el modelo produce medidas de ajuste precisas y estables.

En la Figura 2 se presenta la distribución de frecuencia porcentual de aciertos de los alumnos del grupo en cada reactivo. El grupo obtuvo un promedio de aciertos de 45% con una desviación estándar de 19.7%. Cabe mencionar que hubo un menor número de aciertos que en grupos previamente evaluados (50%) (Martín *et al.*, 2009).

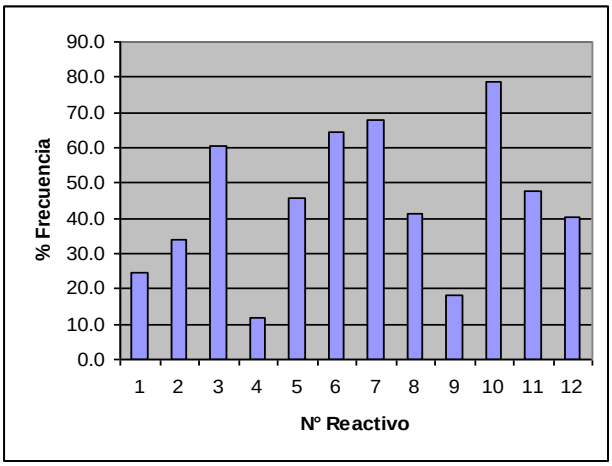


Figura 2. Histograma de los aciertos del grupo

Como se expuso anteriormente lo que se busca es la objetividad de la prueba, en consecuencia, se procedió a analizar el ajuste de estos datos con el modelo de Rasch mediante el programa Winsteps versión 3.69.1.4.

Se estimaron los índices de dificultad (β_i) de los 12 reactivos del examen diagnóstico (Anexo), así como los valores INFIT y OUTFIT y la correlación punto-media. En la Tabla I se muestran los resultados obtenidos al ajustar el modelo a los datos, los reactivos están ordenados, de acuerdo al índice de dificultad (β_i), de mayor a menor valor.

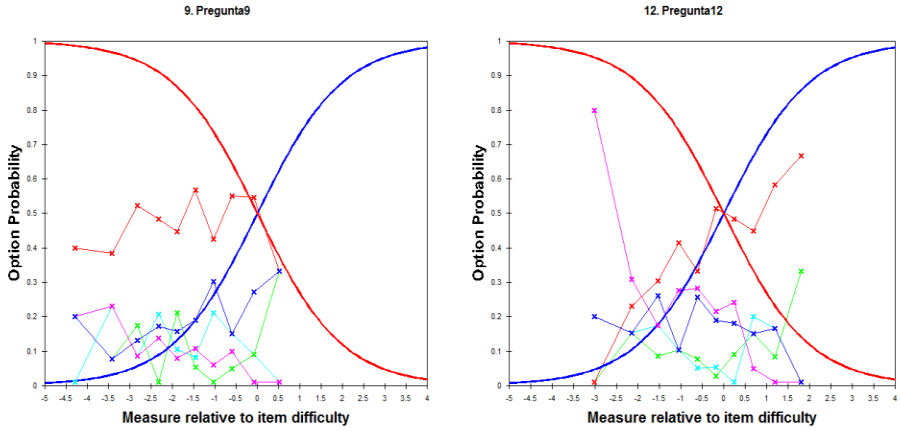
Se observa que 10 de los 12 reactivos de la prueba presentan valores INFIT dentro de los intervalos descritos anteriormente, de 0.8 a 1.3 para la media cuadrática y de -2 y 2 para los valores estandarizados. Sin embargo, el reactivo 12 presenta un valor estandarizado INFIT mayor a 2, lo que indica una falta de ajuste debido a respuestas al azar ó a una falta de precisión en el enunciado, mientras que el reactivo 9 presenta un valor OUTFIT mayor a 1.3 para la media cuadrática y en el límite para el valor estandarizado. También se puede observar en la Tabla I que estos dos reactivos son los que tienen una correlación punto-media más baja, 0.21

y 0.10 respectivamente, mientras que los demás reactivos muestran correlaciones mayores a 0.30. Es decir, que es conveniente revisarlos ya que no se ajustan adecuadamente al modelo.

Tabla I. Índices de dificultad y medidas de ajuste al modelo de Rasch

Número de Reactivo	Número de respuestas correctas	Índice de Dificultad (β_i),	Medidas de ajuste				
			INFIT		OUTFIT		Correlación punto-media
			MNSQ	ZSTD	MNSQ	ZSTD	
4	26	2.01	0.95	-0.2	0.90	-0.2	0.30
9	40	1.45	1.17	1.1	1.66	2.0	0.10
1	54	1.02	1.04	0.4	1.01	0.1	0.30
2	74	0.50	.85	-1.7	0.81	-1.3	0.52
12	88	0.17	1.19	2.2	1.23	1.7	0.21
8	90	0.13	.95	-0.6	0.92	-0.6	0.44
5	100	-0.10	1.02	0.3	1.00	0.1	0.38
11	105	-0.21	1.03	0.4	1.03	0.3	0.37
3	133	-0.84	.94	-0.7	0.86	-1.1	0.48
6	141	-1.03	1.02	0.2	0.99	-0.1	0.38
7	149	-1.23	.86	-1.4	0.83	-1.1	0.52
10	172	-1.89	.98	-0.1	0.87	-0.5	0.39
Media	97.7	0.00	1.00	0.0	1.01	-0.1	
Desv. Est.	43.2	1.09	0.10	1.0	0.22	1.0	

Con el fin de analizar, del punto de vista de las respuestas, a los reactivos 9 y 12, en la Figura 3 se muestran las gráficas que corresponden al análisis de distractores (o desaciertos) de estos dos reactivos. Se observa la falta de ajuste de los datos de los dos reactivos al modelo. En ambas curvas (respuesta correcta y distractores) los datos están muy dispersos y se alejan de la curva de ajuste.



(a) (b)

(a) **Reactivo 9**. Aciertos: opción A, Azul. Desaciertos, opciones: B Roja, C Rosa, D Verde, E Azul Turquesa.
(b) **Reactivo 12**. Aciertos: opción B, Roja. Desaciertos, opciones: A Azul, C Rosa, D Verde, E Azul Turquesa.

Figura 3. Curvas de análisis de distractores

Cabe mencionar, que estos reactivos corresponden a los temas de nomenclatura y relación de moles, respectivamente, los cuales se suponen son conocidos por los alumnos. En el reactivo 9 la opción (B) que es incorrecta, es la que responden con mayor frecuencia, esto es, que no tienen conocimientos de las cargas de los elementos químicos. En tanto que el reactivo 12, sobre relación molar, parece que si discrimina pues los alumnos con mayor habilidad eligen la respuesta correcta, sin embargo, no se ajusta al modelo de Rasch. Por tanto, es importante hacer un análisis más profundo de estos dos reactivos tomando en consideración además de los aciertos, a cada uno de los distractores de los reactivos y al enunciado de los mismos.

Se repitió el análisis sin considerar a los reactivos 9 y 12, los cuales como comentamos antes, deben ser revisados. Los resultados se muestran en la tabla II. Se observa que con los 10 reactivos los valores INFIT y OUTFIT están dentro de los límites establecidos, lo que indica un ajuste razonable al modelo, y las correlaciones son mayores a 0.30. Por otro lado, los índices de dificultad de los reactivos toman valores de -1.84 (más fácil) a 2.35 (más difícil) con un promedio de 0 y una desviación estándar de 1.17.

Tabla II. Índices de dificultad y medidas de ajuste al modelo de Rasch sin considerar los reactivos 9 y 12 que deben ser revisados

Número de Reactivo	Número de respuestas correctas	Índice de Dificultad (β_i)	Medidas de ajuste				
			INFIT		OUTFIT		Correlación punto-media
			MNSQ	ZSTD	MNSQ	ZSTD	
4	26	2.35	1.04	0.3	1.22	0.7	0.31
1	54	1.26	1.14	1.3	1.11	0.6	0.32
2	74	0.70	0.88	-1.4	0.83	-1.2	0.53
8	90	0.30	0.95	-0.6	0.89	-0.9	0.50
5	100	0.06	1.08	1.0	1.14	1.3	0.39
11	105	-0.05	1.10	1.4	1.13	1.2	0.38
3	133	-0.72	0.96	-0.5	0.87	-1.0	0.49
6	141	-0.93	1.04	0.5	0.98	-0.1	0.42
7	149	-1.14	0.86	-1.6	0.78	-1.5	0.54
10	172	-1.84	1.00	0.1	0.86	-0.6	0.41
Media	102.6	0.00	1.01	0.0	0.98	-0.2	
Desv. Est.	44.2	1.17	0.09	1.0	0.15	1.0	

La Figura 4 es una representación gráfica del escalamiento de la dificultad de los reactivos y de los alumnos en una sola escala (-3 a +3 lógitos) y es llamada mapa de alineación de los reactivos y alumnos del grupo evaluado. Del lado izquierdo están representados los alumnos y del lado derecho los reactivos.

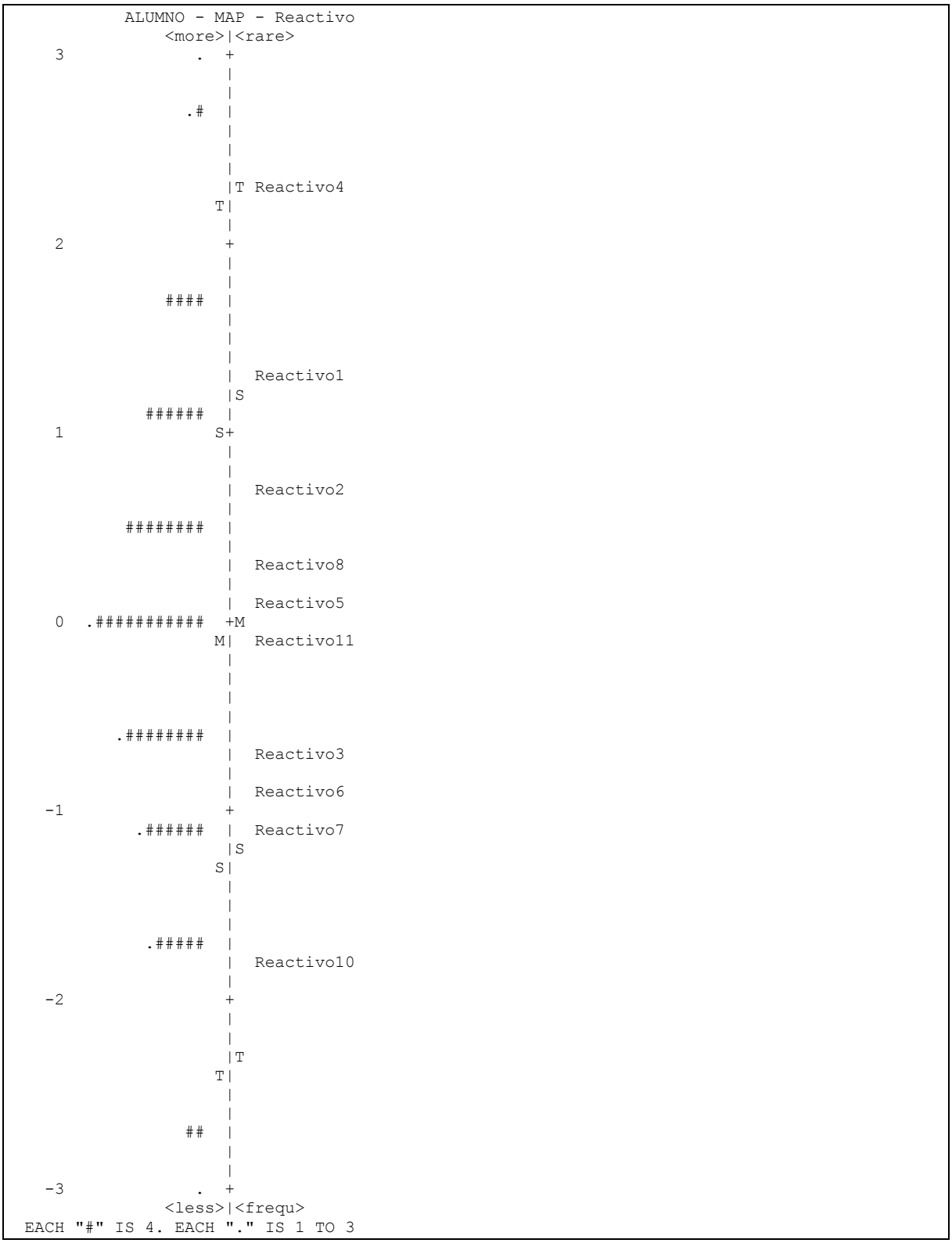


Figura 4. Mapa que muestra la alineación entre reactivos (derecha) y alumnos (izquierda). Cada signo "#" corresponde a 4 alumnos y cada punto "." a 3 alumnos

La estimación de la habilidad de los alumnos está entre -2.66 lógitos (menor habilidad) a 2.73 lógitos (mayor habilidad), con un promedio de -0.12 y una desviación estándar de 1.21. Se observa que la distribución de la habilidad de los alumnos es aproximadamente normal, que la diferencia entre la media de habilidad de los alumnos y la media de dificultad de los reactivos es pequeña y que hay alumnos con habilidad por debajo de la dificultad de los reactivos y también por arriba. Además, la dificultad de los reactivos no es uniforme, ya que se observan huecos entre los reactivos 10 y 7, 3 y 11, así como entre los reactivos 1 y 4. Esto es, que es necesario ampliar la prueba diseñando más reactivos.

Es evidente que, un examen que sea 100% confiable no es siempre posible de alcanzar, de allí que sea necesario complementar la evaluación de los conocimientos de los alumnos con otros tipos de evaluación y que sean continuamente analizadas. De igual forma, aun cuando se demostró la utilidad del modelo de Rash como una herramienta que permite conocer a la habilidad que tiene el alumno en responder a los reactivos que diseñamos, como una función de su dificultad, también permite visualizar factores que con frecuencia ignoramos como son, la elaboración de un buen enunciado y respuestas y distractores adecuados.

En la UAMI, no se tiene la costumbre de analizar los exámenes que se aplican. Siempre se consideran como buenos. Sin embargo, no siempre se tienen exámenes homogéneos y confiables que sean aplicables a cualquier grupo, con cualquier profesor y en cualquier momento; de allí, que sea necesaria su valoración continua.

V. Conclusiones

El grupo obtuvo un porcentaje promedio de 45% de aciertos con una desviación estándar de 19.7%.

A través del modelo de Rasch fue posible conocer la calidad de los reactivos de nuestra prueba mediante los datos estadísticos que reporta el paquete. Se comprobó que sólo dos de los reactivos (9 y 12), de un total de 12 del instrumento de medición, no se ajustaron al modelo de Rasch, los cuales deben revisarse y analizarlos con mayor profundidad a través de los distractores de cada uno para lograr mejorarlos y reutilizarlos.

Además, el análisis permitió comprobar que los reactivos usados presentaron diferentes grados de dificultad, los cuales estuvieron en función de la habilidad de los alumnos del grupo evaluado. Asimismo, se detectó la ausencia de reactivos en algunos intervalos de la escala de índices de dificultad, por lo que deben agregarse nuevos reactivos a la prueba de evaluación para lograr que éste sea más completo.

Es importante señalar que la aplicación del método de Rasch para evaluar la calidad de los exámenes, además de servir para poder mejorar el instrumento, que a veces se considera muy bueno, permite realizar una evaluación más confiable de los estudiantes. En resumen, el modelo de Rasch es una herramienta valiosa para la calibración de pruebas de evaluación.

Referencias

Backhoff E., Larrazolo N. y Rosas N. (2000). Nivel de dificultad y poder de discriminación del examen de habilidades y conocimientos básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa, 2*(1).

Backhoff, E., Tirado F. y Larrazolo N. (2001). Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa*, *3*(1).

Budescu, D. V. (1979). *Differential weighting of multiple choice items*. Princeton: Educational Testing Service.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational measurement* (2a. ed.). Washington: Consejo Americano en Educación.

Embretson S. E. y Reise S. P. (2000). Item response theory for psychologists. Mahwah, NJ: LEA.

Golia, S. (2011). The impact of questionnaire size on the accurace of the Rasch measure. *Journal of Applied Science, 11*(4), 707-712.

González M., M. J. (2008). *El análisis de reactivos en el modelo de Rasch.* Manual Técnico A. Serie Medición y Metodología. México: Universidad de Sonora, Instituto Nacional para la Evaluación de la Educación.

Hambleton, R. K. y Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Martin-Guaregua, N., Córdoba-Herrrera, G., Lomas-Romero, L., Rojas H. A. y Picquart, M. (2009). *Errores conceptuales de química básica en alumnos del primer año universitario*. Enseñanza de la ciencias, VIII CIDEC, pp. 950-951. http://ensciencias.uab.es/congreso09/numeroextra/art-950-951.pdf

Muñiz, J. (1997). *Introducción a la Teoría de Respuestas a los Ítems*. Madrid: Pirámide.

Planinić, M., Ivanjek, L. y Sušac, A. (2009). The Rasch model based analysis of the Force Concept Iventory. *Physical Review Special Topics-Physics Education Research, 6*(1), 1-9.

Prieto, G. y Delgado A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15*(1), 94-100.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenague: Danish Institute for Educational Research.

Rojas, R. M., Manriquez LL., G., Gatica A., Y. y Salcedo A, L. P. (2004). Curso de UML Multiplataforma Adaptativo basado en la Teoría de Respuesta al Item. *Revista Ingeniería Informática*, *10*.

Spearman, C. E. (1904). The proof and measurement of association between two things, *American Journal of Psychology, 15*, 75-101.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 667-680.

Sympson, J. B. y Haladyna, T. M. (1988). An evaluation of "polyweighting" in domain referenced testing. Trabajo presentado en la Reunión Anual de la American Educational Research Association, Nueva Orleans, EE.UU.

Tristán L. A. (2001). *Análisis de Rasch para todos: una guía simplificada para evaluadores educativos*. México: Centro Nacional de Evaluación.

ANEXO 1

Examen Diagnóstico

1. Indica ¿cuál de las siguientes afirmaciones permanece igual, antes y después, de una reacción química?

a. La suma de las masas molares de las sustancias involucradas
b. El número de moléculas de todas las sustancias involucradas.
c. El número de átomos de cada tipo involucrados.
d. Ambos, (a) y (c) pueden ser iguales.
e. Cada una de las respuestas de a), b) y c) pueden ser las mismas.

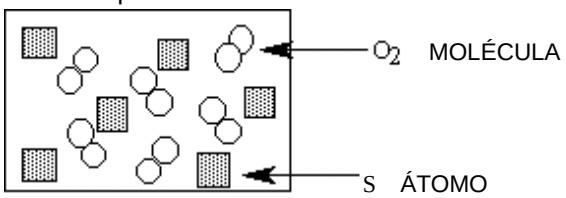
2. En un vaso de vidrio que contiene leche fría se forman gotas de agua en las paredes fuera del vaso (frecuentemente referido como sudor). ¿Cómo puede explicarse este fenómeno?

a. El agua de la leche se evapora y se condensa fuera de las paredes del vaso.
b. El vidrio actúa como una membrana semipermeable que permite que el agua pase a través de ella, pero no a la leche.
c. El vapor de agua del aire se condensa.
d. El frío causa que el oxígeno y el hidrógeno del aire se combinen sobre las paredes del vidrio y formen agua.
e. Todas las anteriores

3. ¿Cuál es la masa en gramos de una solución cuando se mezclan 1 g. de sal con 20 g. de agua?

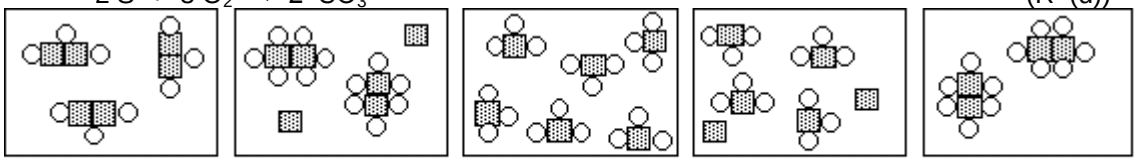
a. 19
b. 20
c. Entre 20 y 21
d. 21
e. Más de 21

4. En el diagrama se representa una mezcla de átomos de azufre (S) y moléculas de oxígeno (O_2) en un recipiente cerrado.



Indica cuál de los diagramas siguientes representa el posible resultado después que la mezcla anterior ha reaccionado completamente según la reacción:

$2\ S + 3\ O_2 \rightarrow 2\ SO_3$ (R=(d))



5. Se tiene un vaso con agua y dos hielos. Después que se derriten los hielos el volumen del agua permanece igual. Este fenómeno se debe a que:

a. La masa de agua desplazada es igual a la desplazada por la masa de hielo.
b. El agua es más densa en su forma sólida (hielo).
c. Las moléculas de agua desplazan más volumen que las moléculas de hielo.
d. El agua del hielo fundido cambia el volumen del agua.
e. Cuando el hielo se funde, sus moléculas se expanden.

6. Una muestra de 1 g. de yodo sólido es colocada en un tubo. El tubo es sellado y todo el aire es evacuado. El tubo y el yodo sólido pesan 27.0 g. El tubo es calentado hasta que el yodo se evapora y se llena con yodo gaseoso. La masa en gramos después del calentamiento será:

a. Menor de 26.0
b. 26.0
c. 27.0
d. 28.0
e. Más de 28.0

7. Si a 3 g. de sal común se agrega agua hasta completar 30 g. la concentración de sal en esta solución es:

a. 1 %
b. 3 %
c. 10%
d. 30%
e. 100%

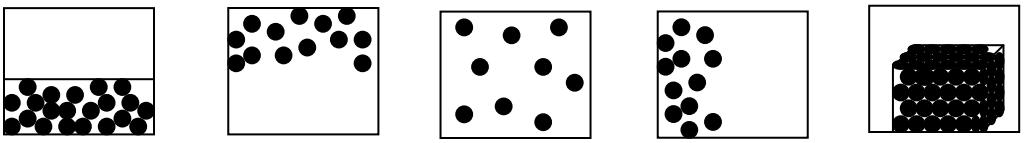
8. Si la densidad de un gas es 0.5 g/m^3, entonces 5.0 g. de ese gas ocuparán un volumen en m^3 de:

a. 0.1
b. 0.4
c. 2.0
d. 2.5
e. 10.0

9. La fórmula del sulfato de sodio es:

a. Na_2SO_4
b. $NaSO_4$
c. $Na(SO_4)_2$
d. Na_2SO_3
e. $NaSO_3$

10. La figura que representa microscópicamente a un gas en equilibrio, de acuerdo con la Teoría Cinético-Molecular, es:



a) b) **c)** d) e)

11. Dos moles de ácido nítrico (HNO_3) contienen _____ moles de átomos de oxígeno:

a. 1
b. 2
c. 3
d. 4
e. 6

12. La relación molar de $HCl/ZnCl_2$ en la reacción: $Zn+2HCl=ZnCl_2+H_2$ es:

a. 1:2
b. 2:1
c. 1:1
d. 2:2
e. 3:2