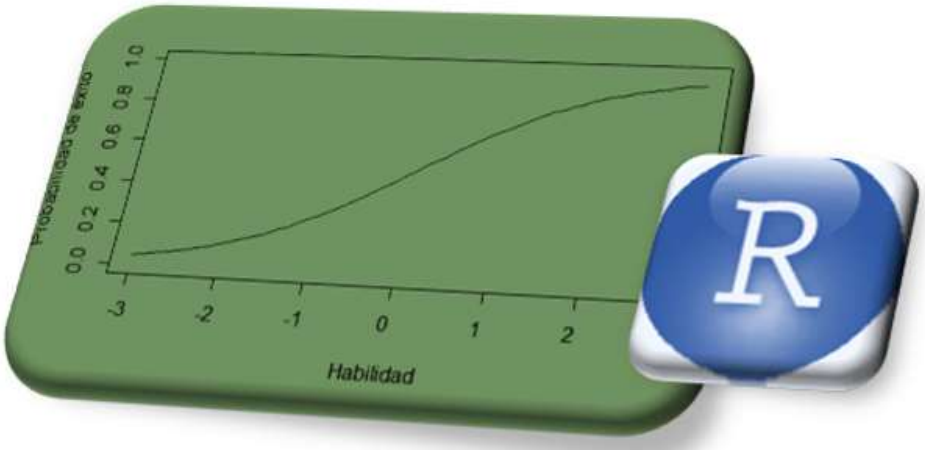
Serie de Posts

Número 5

Los 'Grupos de Calidad' en las Universidades Cubanas: ¿cuánto más se puede hacer por la Autoevaluación? (2° parte)



En la primera parte de este Número 5 de la actual Serie de Posts (de naturaleza multi temática, razón por la cual ha sido desarrollada bajo la denominación genérica de 'Extensionismo universitario') comenzamos a responder la pregunta de qué más se pudiera hacer para que los 'Grupos de Calidad', que se han venido creando en las universidades cubanas, puedan realizar una contribución aún mayor a la calidad de los procesos de autoevaluación que deben preceder las acciones de 'evaluación externa' y de 'acreditación', que dirige en Cuba la 'Junta de Acreditación Nacional' (JAN).

Básicamente, en esa primera parte del Número, nos centramos en la importancia de ampliar las escalas de evaluación (instrumentos de evaluación) con la utilización de reactivos (preguntas, órdenes) que intenten aislar los errores que más frecuentemente se presentan en el cabal cumplimiento de las exigencias

curriculares universitarias (tanto en lo instructivo, como en lo formativo), lo cual es clave para emprender con eficacia las acciones de **mejora escolar**, después. Recordarán que para ello sugerimos la incorporación de **distractores** (en tanto respuestas no evidentemente erradas) en esos reactivos; y sugerimos, además, que tales reactivos deberían ser, por tanto, preferentemente preguntas de 'selección múltiple'.

También analizamos allí que estas sugerencias no debieran ser vistas como un obstáculo per se ante las marcadas limitaciones materiales imperantes, pues pueden adaptarse a un sistema de **evaluación digital**, para lo cual recomendamos especialmente la plataforma **Moodle**. De hecho, vimos que –lejos de generar inconvenientes– con ella se obtienen ventajas adicionales, como la posibilidad de disponer de los parámetros psicométricos del '**índice dificultad**´ y del '**índice de discriminación**' para cada instrumento aplicado.

Sin embargo, a los efectos de la 'autoevaluación' lo más importante no es conocer de la calidad de los instrumentos diseñados, sino el nivel real de preparación de sus respondientes; mas los parámetros psicométricos proporcionados por Moodle están estimados a partir de la llamada '**Teoría Clásica del Test**'.

¿Qué significa esto?... Pues que esa información está subordinada a la muestra (de ocasión) que se utilice. ¡Y esto es inapropiado!... Cierto, porque si la mayoría de los respondientes a los que se le aplica el instrumento (determinado por la escala construida) tuvieran altamente desarrollada la habilidad medida, el instrumento va a resultar 'sencillo'; mientras que en caso contrario, el instrumento resultará 'difícil'. De modo que, de acuerdo con el instrumento, no sabremos con certeza cómo se encuentran los respondientes en relación con la habilidad medida. Y estas opciones extremas pueden suceder con cierta probabilidad; pequeña pero posible.

Es por ello que me gusta hablar mejor de '**muestra de ocasión**' (esto quiere decir que al cambiar la muestra, se suelen obtener medidas estadísticas diferentes, aun cuando la selección muestral haya sido al azar). ¿Qué hacer entonces?...

El necesario cambio de teoría psicométrica

Se necesita realizar los análisis psicométricos (esto es, referidos a los instrumentos y a sus **escalas**) desde otra perspectiva. ¿Cuál?... Pues una que independice el '**índice dificultad**´ y el de '**índice de discriminación**' de la muestra utilizada.

Eso es posible a través de un cambio al paradigma psicométrico conocido como '**Teoría de Respuesta al Ítem**'. Como lo indica su nombre, el foco ya no estará en la puntación global de cada individuo en la prueba (**test**), sino en la obtenida por los respondientes en cada uno de sus reactivos, por separado. La otra modificación de consideración (y esto es muy importante para cumplir el propósito de buscar independencia de la '**muestra de ocasión**') será no trabajar con la respuesta en sí misma, sino con la **probabilidad de ofrecer la respuesta** de acuerdo con el **grado**

de desarrollo de la habilidad latente (como se trata de ilustrar en la imagen que sirve de portada a la presente parte del Número en marcha).

No creo que sea necesario ahondar en los procedimientos requeridos para concretar este otro paradigma, pues de eso se ocupan con notable rapidez y facilidad muchos de los entornos estadísticos actuales, especialmente 'mi genio de la lámpara': R-Project (y cuyo ícono aparece también presente en la imagen de la portada, como queriéndonos alertar que: 'No nos preocupemos por las estimaciones de los parámetros, pues de eso se ocupa el software; nosotros, los usurarios, debemos prestar atención solo de los datos que le proporcionaremos y a las interpretaciones de sus devoluciones'). ¡Así de sencillo!

De acuerdo con ello, cabe señalar que el 'índice dificultad´ del reactivo (ítem) estará determinado por el valor del eje horizontal que corresponde al punto de la curva (en forma de "S") donde la probabilidad es 0.5 (la mitad del rango de probabilidades); lo cual es lógico, pues constituye el 'punto de inflexión' a partir del cual es más probable acertar en la respuesta correcta del reactivo, que errar en ella.

Pero esto nos proporciona una ventaja más. ¡Tenemos en la misma escala (la horizontal del gráfico) tanto la magnitud del nivel de desarrollo (o dominio) de la habilidad de los respondientes en ese reactivo (en lo adelante: theta), como el índice de dificultad de ese mismo reactivo (representado comúnmente con la letra b)!

Y ahora, con las estimaciones de todos los 'índices de dificultad', correspondientes a los reactivos que conforman el instrumento, podemos además identificar para cuáles niveles de desarrollo (probables) de la habilidad de los respondientes aquel es (probablemente) más 'fácil' o más 'difícil' para la prueba toda. De momento no diré más sobre esta otra teoría psicométrica, para no complicarle las cosas a los que les perturba la Matemática. Mejor ilustremos el proceso con un par de ejemplos.

Supongamos que, en preparación para un proceso de Acreditación Universitaria, la dirección de una Facultad le encarga a su 'Grupo de Calidad' que realice una autoevaluación centrada en una disciplina curricular de una de sus carreras, en un año académico determinado. Asistido de especialistas en esa disciplina, el 'Grupo de Calidad' confecciona una prueba de doce reactivos (todos de selección múltiple con una única respuesta correcta y tres distractores en cada uno).

La prueba es convertida en una Actividad 'Cuestionario' de la plataforma Moodle y se le aplica a 150 estudiantes de ese año académico, utilizando sus teléfonos celulares; o un laboratorio de computación (cuyas PC están conectadas en red), en el caso de una parte de los estudiantes convocados que, o no tenían celulares personales, o presentaban problemas de conexión a la plataforma.

La siguiente tabla (generada con R-Project) muestra las opciones de respuestas a cada reactivo que eligieron los primeros quince estudiantes que la aplicaron.

Studio Source Editor

extensionismo5b_data_test_letras

Filter

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
1	C	D	A	B	C	B	B	B	D	B	C	B
2	C	D	A	C	C	B	D	B	D	D	C	C
3	C	D	A	C	B	B	B	D	D	B	A	B
4	C	D	A	B	B	B	D	B	C	B	C	B
5	D	D	A	B	B	C	B	D	D	B	C	B
6	C	D	A	A	B	B	D	B	C	B	C	B
7	C	D	A	B	C	B	D	B	C	B	A	A
8	A	D	A	B	C	B	D	B	D	B	A	B
9	C	D	A	B	B	B	D	B	C	B	C	C
10	C	D	A	B	B	B	A	B	C	B	A	C
11	C	D	A	B	B	A	A	B	D	B	B	A
12	C	D	A	B	B	B	A	B	D	B	B	B
13	C	D	A	B	C	B	D	B	D	B	A	A
14	C	D	A	C	B	B	D	B	D	B	A	A
15	C	D	A	B	B	B	D	B	C	B	C	C

También, utilizando el listado de opciones de respuestas correctas otorgado por los elaboradores de la prueba, R-Project calificó automáticamente las respuestas de los 150 respondientes, otorgando 1 a la respuesta correcta y 0 a los distractores.

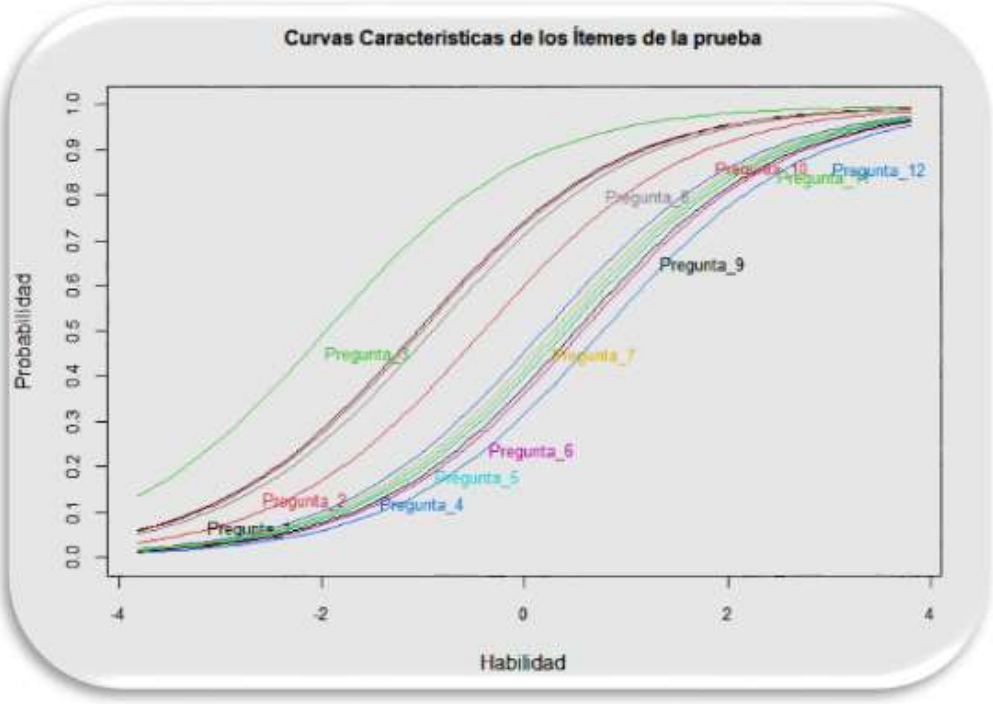
Studio Source Editor

extensionismo5b_data_binario

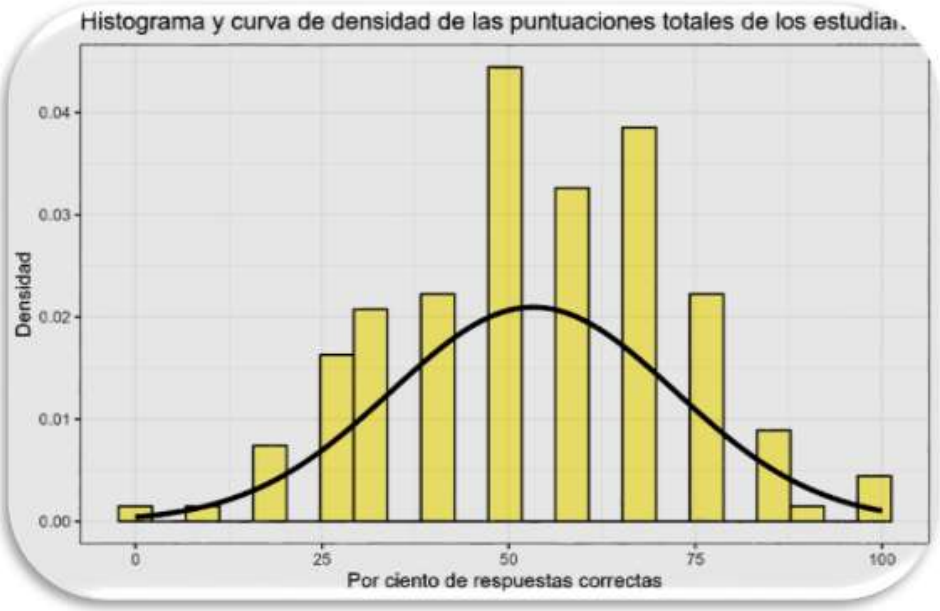
Filter

	Pregunta_1	Pregunta_2	Pregunta_3	Pregunta_4	Pregunta_5	Pregunta_6	Pregunta_7	Pregunta_8	Pregunta_9	Pregunta_10	Pregunta_11	Pregunta_12
1	1	1	1	1	0	1	0	1	0	1	1	0
2	1	1	1	0	0	1	1	1	0	0	1	1
3	1	1	1	0	1	1	0	0	0	1	0	0
4	1	1	1	1	1	1	1	1	1	1	1	0
5	0	1	1	1	1	0	0	0	0	1	1	0

Un primer análisis de interés para el 'Grupo de Calidad', seguramente, es conocer en cuáles de los contenidos de enseñanza evaluados los estudiantes presentan mayores dificultades. Para ello se le pide a R-Project que determine los parámetros de dificultad de cada uno de los doce reactivos evaluados. El siguiente gráfico los ilustra, desde la perspectiva de la 'Teoría de Respuesta al Ítem'.



Es evidente que los contenidos de enseñanza evaluados por los reactivos del 8 al 12 son los que demandan mayores niveles de desarrollo de la habilidad latente (aquí, preparación en la disciplina curricular evaluada), pues para obtener una alta probabilidad de éxito (o sea, de respuesta correcta) en ellos se requiere de valores de 'theta' muy altos (entre 2 y 4 en la escala horizontal). En el extremo izquierdo, en cambio, aparecen los reactivos 'más fáciles' (del 1 al 4), pues tienen una alta probabilidad de respuesta exitosa, aun con bajos niveles de desarrollo de la habilidad latente. Mientras que, el comportamiento de los puntajes totales de los estudiantes en la prueba aplicada se muestra en los siguientes gráficos, generados con R-Project.

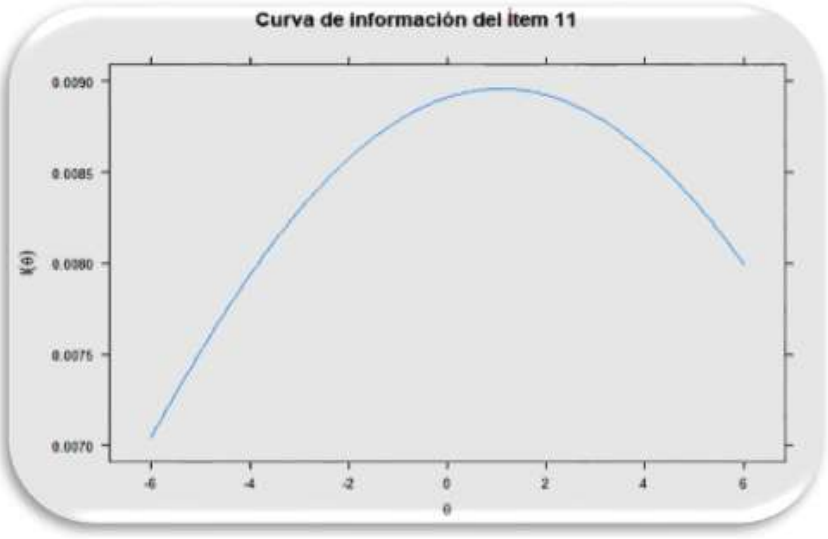


Ahora bien, ya hemos señalado (en la parte anterior de este Número) que pudiera suceder que, al construir el instrumento de evaluación, interese también incluir

uno o dos reactivos en forma de **pregunta de desarrollo** (es decir, que no todos sean del tipo de selección múltiple, para evitar en alguna medida la incertidumbre de obtener aciertos por adivinación). El inconveniente aquí es que esos ítems no podrán ser calificados automáticamente por un software; de modo que se requerirá de algún tipo de tribunal de calificación conformado por especialistas en la materia.

Si es así, se supone que no se desaproveche potencialidades reduciendo la escala de calificación a una puntuación binaria, como hasta ahora, sino que **se emplee una rúbrica de más opciones**. Con frecuencia, en estos casos se suelen emplear tres categorías de calificación: 'respuesta sin crédito', 'respuesta con crédito parcial' y 'respuesta con crédito total'.

En consecuencia, el **modelo matemático-estadístico** que se ha estado utilizando en el ejemplo no funciona, pero afortunadamente se tienen modelos específicos para estas otras situaciones también. Veámoslo... Supongamos que en la prueba del ejemplo anterior el reactivo onceno y el duodécimo (los dos últimos) fueron creados como 'preguntas de desarrollo' por los especialistas que asisten al 'Grupo de Calidad' de la Facultad; así como que ellos mismos los calificaron, siguiendo la escala antes descrita (de tres categorías ordenadas ascendentemente). Entonces, **R-Project** sí puede aplicar el nuevo modelo requerido y ofrecernos las 'salidas' correspondientes.

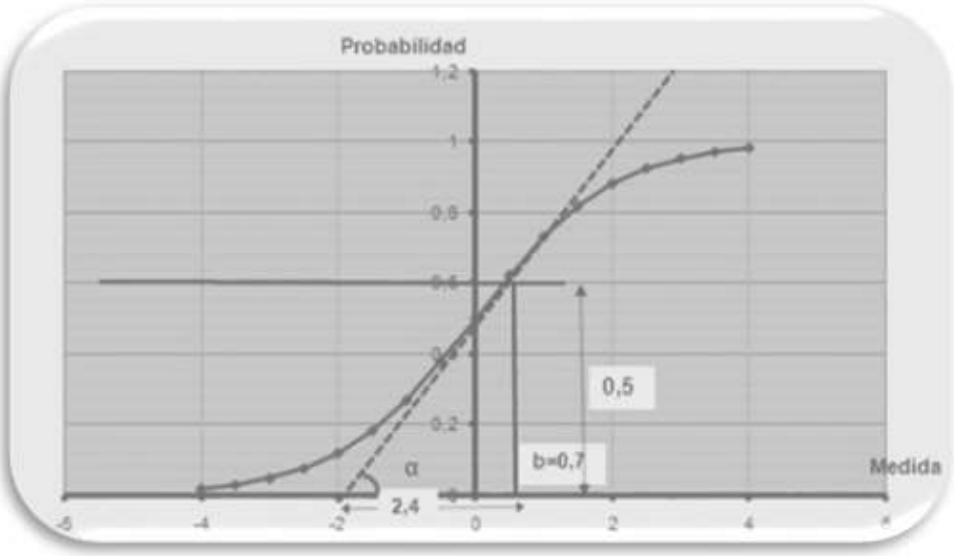


Por ejemplo, aquí aparece graficada la **'Curva de información'** del ítem 11 (ahora convertido en una 'pregunta de desarrollo'). Nótese que no asume la forma de 'S' de las curvas características de los ítems de respuesta binaria, pero nos sigue hablando de cierta dificultad del reactivo para estudiantes con un bajo nivel de desarrollo de la **habilidad latente** (o sea, con 'theta' entre -6 y -3); mientras que lo más frecuente es que demande un nivel de competencia de alrededor de un 'theta'. Algo parecido se obtiene para el reactivo duodécimo.

Bueno, seguramente un lector avispado se habrá percatado que, con el paso de la **'Teoría Clásica del Test'** al nuevo paradigma psicométrico de 'Teoría de Respuesta al Ítem', he dejado de referirme al **'índice de discriminación'**, concentrándome solo

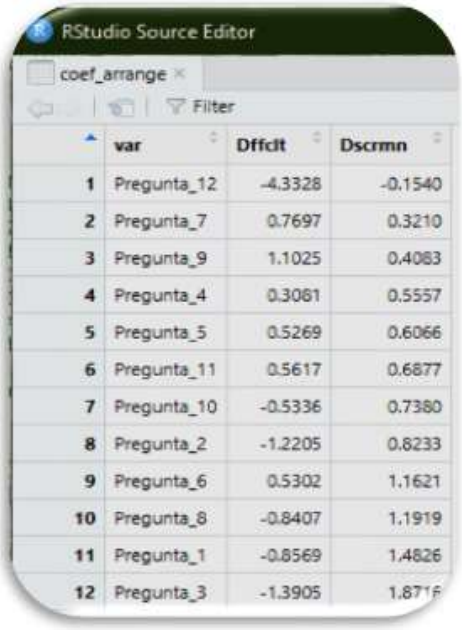
en el par dialéctico: 'índice de dificultad' – 'habilidad latente'. Pues, es momento de aclarar que, aunque con un significado diferente, también podemos hablar de aquel bajo la perspectiva del nuevo paradigma.

Dicho muy intuitivamente, si en la teoría anterior el 'índice de discriminación' estaba determinado por la diferencia de proporciones de aciertos en el grupo superior de respondientes (o sea, con mayores porcentajes de aciertos en el test), y la proporción análoga en el grupo inferior (es decir, de respondientes con menores porcentajes de aciertos en la prueba en su conjunto), acá –en la 'Teoría de Respuesta al Ítem'– el análisis se traslada a cada reactivo por separado, y expresa la medida en que este logra separar a los respondientes con mayor desarrollo de la habilidad latente de los de menor desarrollo en ella.



Por tanto, y como se muestra en el gráfico anterior, este otro parámetro está asociado aquí con el grado de inclinación (o pendiente) de la 'Curva Característica del Ítem' en su punto de inflexión (recuérdese, el que se ubica en la intercepción de la recta paralela al eje horizontal a la altura de una probabilidad de éxito de 0.5; recta esta con la cual se podía ubicar el valor del 'índice de dificultad' del ítem, b=0.7). Es decir, se trata de un valor vinculado con el ángulo de inclinación α que se muestra en la parte inferior de la figura.

Y es razonable, pues en la medida que ese ángulo de inclinación sea mayor, mayor será también la separación de las 'colas' extremas de la 'curva S', que representan los niveles de desarrollo de la habilidad medida más bajos (a la izquierda) y más altos (a la derecha), respectivamente. R calcula ese indicador también ('Dscrmn').


RStudio Source Editor

coef_arrange

Filter

	var	Dffclt	Dscrmn
1	Pregunta_12	-4.3328	-0.1540
2	Pregunta_7	0.7697	0.3210
3	Pregunta_9	1.1025	0.4083
4	Pregunta_4	0.3061	0.5557
5	Pregunta_5	0.5269	0.6066
6	Pregunta_11	0.5617	0.6877
7	Pregunta_10	-0.5336	0.7380
8	Pregunta_2	-1.2205	0.8233
9	Pregunta_6	0.5302	1.1621
10	Pregunta_8	-0.8407	1.1919
11	Pregunta_1	-0.8569	1.4826
12	Pregunta_3	-1.3905	1.8716

Así, en este otro gráfico (de arriba) se han ordenado de manera ascendente los reactivos, a partir del 'índice de discriminación' de cada uno de ellos; aparece a su lado también el 'índice de dificultad', recalculado con el nuevo modelo matemático-estadístico que trabaja con los dos parámetros a la vez.

Llegado a este punto, es bueno aclarar que todos los 'valores de salida' que hemos estado utilizando ('b', 'theta', y ahora 'a', el parámetro de discriminación) están expresados en una escala con media cero y desviación estándar uno; es decir, son valores estandarizados. De modo que un puntaje cero en esa escala no significa un valor bajo, sino medio; los valores negativos sí lo son (especialmente, los menores que -3); mientras que los valores positivos representan puntuaciones altas (en particular, los mayores que 3).

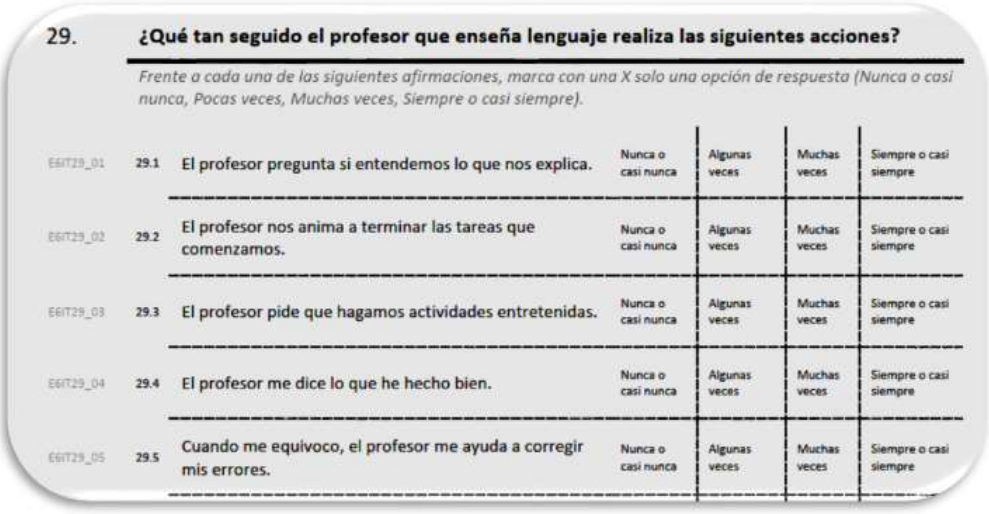
¿Y qué hay con las respuestas a los cuestionarios?

En la primera parte de este Número 5 de la Serie señalamos también la importancia de utilizar, en los procesos de 'Autoevaluación' ejecutados por los 'Grupos de Calidad' de nuestras universidades, no solo escalas referidas a conocimientos de las esferas instructiva y formativa, sino también 'cuestionarios de contexto' (llamados usualmente 'de factores asociados al logro').

Esto es recomendable porque se necesita explorar las causas de los resultados alcanzados con los primeros con el mismo rigor con que se trata de evaluar con ellos; de lo contrario, se 'abre una puerta' a la especulación y al empirismo. O sea, en materia de evaluación (aquí, educativa) es muy importante apegarse a la utilización del 'método científico' en todo momento; desde la organización del evento y la elaboración de los instrumentos, hasta la interpretación final de los resultados (como veremos con detenimiento en la tercera y última parte de este Número).

En consecuencia, si en el primer grupo de instrumentos (las pruebas) debieran predominar las 'preguntas de selección múltiple', en los cuestionarios debiera

emplearse preferentemente la 'escala Likert'; es decir, el planteamiento de un juicio y varias opciones de acercamiento y no a él, como se muestra en el siguiente ejemplo, tomado de los reportes del ERCE-2019, de la OREALC-UNESCO.



29. **¿Qué tan seguido el profesor que enseña lenguaje realiza las siguientes acciones?**

Frente a cada una de las siguientes afirmaciones, marca con una X solo una opción de respuesta (Nunca o casi nunca, Pocas veces, Muchas veces, Siempre o casi siempre).

E6IT29_01	29.1	El profesor pregunta si entendemos lo que nos explica.	Nunca o casi nunca	Algunas veces	Muchas veces	Siempre o casi siempre
E6IT29_02	29.2	El profesor nos anima a terminar las tareas que comenzamos.	Nunca o casi nunca	Algunas veces	Muchas veces	Siempre o casi siempre
E6IT29_03	29.3	El profesor pide que hagamos actividades entretenidas.	Nunca o casi nunca	Algunas veces	Muchas veces	Siempre o casi siempre
E6IT29_04	29.4	El profesor me dice lo que he hecho bien.	Nunca o casi nunca	Algunas veces	Muchas veces	Siempre o casi siempre
E6IT29_05	29.5	Cuando me equivoco, el profesor me ayuda a corregir mis errores.	Nunca o casi nunca	Algunas veces	Muchas veces	Siempre o casi siempre

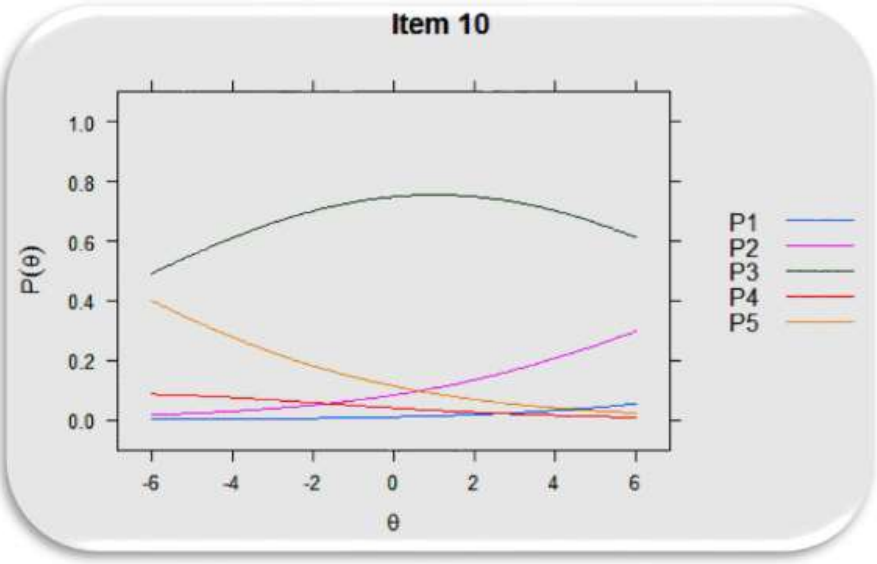
Ahora bien, puesto que en estos casos no necesariamente tiene que existir una única respuesta correcta, entonces debemos utilizar un tercer tipo de modelo matemático-estadístico, diferente a los dos anteriores arriba empleados (y siempre a través de un software estadístico especializado en estos temas, como lo es R-Project).

Ampliemos el ejemplo anterior, del encargo que le realizó la dirección de la Facultad a su 'Grupo de Calidad', de realizar una autoevaluación centrada en una disciplina curricular de una de sus carreras, en un año académico determinado.

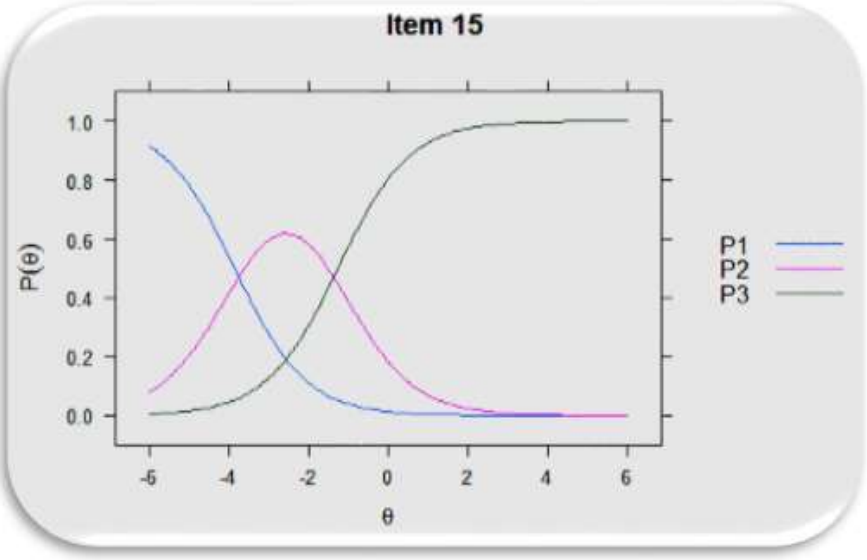
Esta vez, los propios miembros del 'Grupo de Calidad', apoyados en su vasta experiencia en gestión de procesos universitarios, confeccionan un cuestionario de factores asociados a los logros de la disciplina avaluada. Este está compuesto por quince reactivos del tipo 'escala Likert', con 5 opciones de respuesta cada ítem: 'Muy en desacuerdo', 'En desacuerdo', 'Ni de acuerdo, ni en desacuerdo', 'De acuerdo' y 'Muy de acuerdo', en ese orden.

También aquí, el instrumento es incorporado a la Actividad 'Cuestionario' de la plataforma Moodle y se le aplica a 200 informantes relacionados con ese año académico; entre estudiantes, docentes y directivos. Tras procesar la base de datos generada con la aplicación del instrumento, se detectan veintinueve respondientes que no contestaron los quince reactivos, de modo que tuvieron que ser excluidos. La base de datos se procesará, por tanto, con 171 registros.

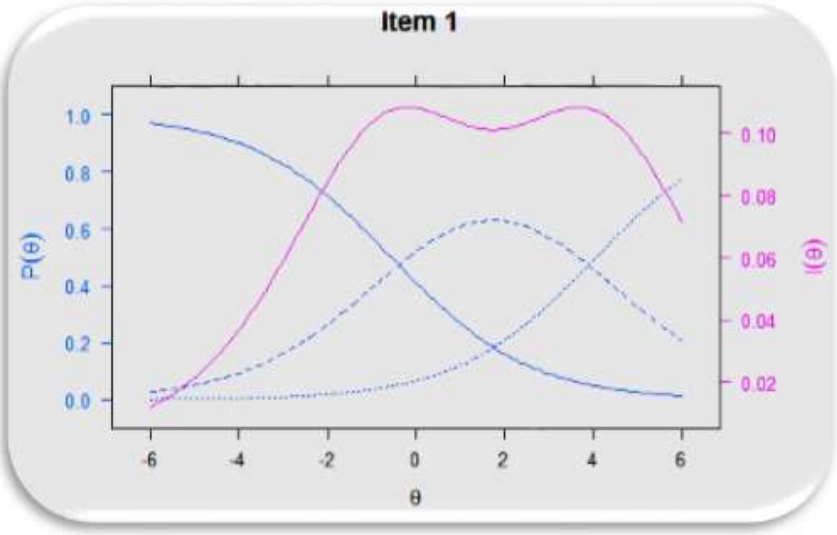
Al trabajar, con R-Project, el modelo-estadístico apropiado para instrumentos con respuestas múltiples, se generan gráficos parecidos a los de 'Curva Característica del Ítem' del primer modelo; estos otros se conocen como 'Función de respuesta de opción' y aparecen representadas en él las curvas de todas opciones elegidas en cada reactivo; veamos el caso del reactivo No.10.



En este caso, la 'habilidad latente' medida no representa dominio de un contenido curricular de la disciplina, sino una inclinación positiva hacia el estado del contexto que describe el juicio que encabeza el reactivo. De acuerdo con los datos acopiados, en el Ítem No.10 del cuestionario lo más probable es que los consultados se inclinen por la opción tercera ("Ni de acuerdo, ni en desacuerdo"), con independencia del nivel de inclinación general que tengan hacia el tema tratado; mientras que la opción segunda ("En desacuerdo") es esencialmente elegida por los sujetos que tienen una predisposición muy positiva hacia el tema tratado en la indagación (o sea, con theta [θ] entre 4 y 6).

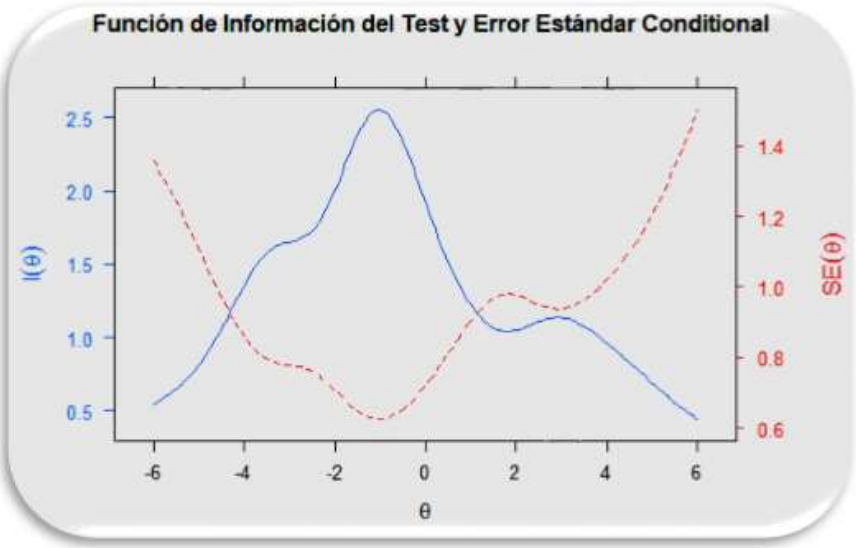


Por su parte, en el reactivo No.15 se da una situación diferente; para empezar, los respondientes solo eligieron las opciones primera, segunda y tercera. Cuando se trata de los que tienen una muy baja inclinación hacia la situación general explorada (theta [θ] entre -6 y -4) se aprecia una alta probabilidad de que elijan la primera opción del reactivo ("Muy en desacuerdo"), mientras que la tercera para los que poseen una inclinación media-alta hacia aquella.

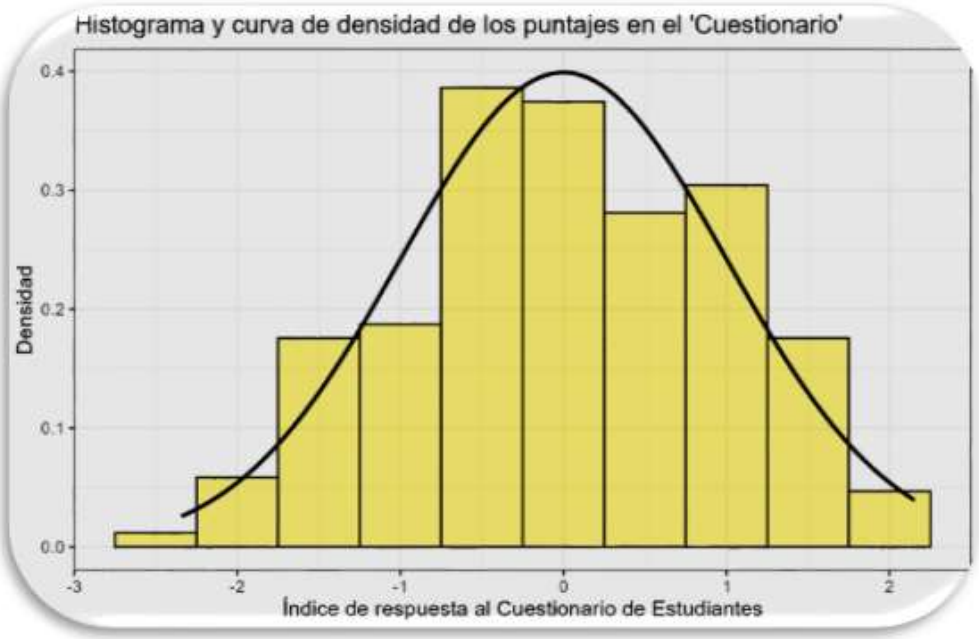


El modelo matemático-estadístico, trabajado con R-Project, permite combinar la **'Función de respuesta de opción'** con la **'Función de Información del reactivo'** en un mismo **'gráfico de salida'**, de modo que se pueda conocer el comportamiento de las opciones de respuesta elegidas, pero además el comportamiento general del reactivo. En la figura de arriba se aprecia que en el Ítem No.1 del Cuestionario es más probable que la mayor cantidad de información se obtenga en el rango de theta de -1 a 5 (es decir, entre los respondientes con una inclinación media a media-alta hacia el tema explorado).

Eso es lo estimado por el modelo para un reactivo en particular (el Ítem No.1), pero él también es capaz de estimar, y se puede representar asistido del software, el comportamiento general de todas las respuestas dadas al **Cuestionario**, además de informar con qué margen de error se producen esas estimaciones, como se aprecia en el **'gráfico de salida'** siguiente, donde se observa que –con mucha certeza– la mayor cantidad de información la proporcionan los individuos con una inclinación (casi) media hacia la situación que se quiere diagnosticar con el **'Cuestionario'**.



Por último, es posible generar un índice estadístico (puntaje) de cada individuo con respuestas al 'Cuestionario' registradas; los de puntuaciones altas indican una mayor inclinación hacia el tema consultado y, obviamente, los de valores bajos lo contrario. Este 'índice estadístico' es clave para relacionar los resultados en la prueba con las respuestas al cuestionario, explicar los resultados alcanzados en la prueba y poder entonces, desde la 'Autoevaluación', tratar de combatir (o fortalecer) las causas de los resultados de aquella, de acuerdo con lo explorado con el cuestionario. Pero eso será tema de análisis en la tercera y última parte de este Número 5.



Bueno, esto es todo por hoy. Seguiremos en unos días. ¡Los espero!...

Dr. Cs. Paul Antonio Torres Fernández

Profesor e Investigador Titular; Bioestadístico

Facultad de Ciencias Médicas 'Salvador Allende'

Universidad de Ciencias Médicas de La Habana

orcid.org/0000-0002-7862-2737

Para profundizar:

Torres, P. A. (2010). Teoría de Respuesta al Ítem. *Boletín Mensual 'El Evaluador Educativo' No.8 / Año I*. Instituto Central de Ciencias Pedagógicas.

https://drive.google.com/file/d/1PbzxrbWFusROa7Q7KneuBgF8XH8FpfrB/view?usp=sharing

_____ (2016). Acerca de las pruebas objetivas y la enseñanza desarrolladora. *Temas de Educación / Vol.22, No.1: 115-129*. https://revistas.userena.cl/index.php/teduacion/article/view/740