El Evaluador Educativo

Número 2, Año II
Octubre 2010

MODELO DE DISEÑO PARA PRUEBAS OBJETIVAS: UNA EVIDENCIA SOBRE LA VALIDEZ DE ESCALA

Agustín Tristán L. **Rafael Vidal U.**

Agustín Tristán

Director del Instituto de Evaluación e Ingeniería Avanzada, S.C. Ingeniero Civil de la Facultad de Ingeniería, UNAM. México, con Doctorado en Ingeniería, especialidad Mecánica de Materiales, de la Escuela Nacional de Puentes y Caminos de Paris, Francia. Ha ocupado diversos puestos relacionados con la Ingeniería, Educación y la Evaluación Educativa en la Facultad de Ingeniería de la UNAM, Universidad La Salle, Institutos Tecnológicos, Instituto de Investigaciones Eléctricas, etc.

Desde 1975 ha realizado un gran número de proyectos y trabajos de investigación relacionados con la evaluación educativa, tanto con modelos clásicos como con el modelo de Rasch. Cuenta con más de 80 publicaciones en temas de Ingeniería y de Evaluación.

Ha asesorado a varias instituciones en la puesta en marcha de proyectos de evaluación del aprendizaje desde nivel preescolar hasta la certificación profesional por competencias. Participa como meta-evaluador de diversas instituciones a través del servicio de diagnóstico de calidad de pruebas estandarizadas, con base en estándares de calidad para pruebas objetivas, del cual ha publicado un libro (2006) y es Coordinador responsable del Examen de Certificación para Profesores de Educación Media Superior (ECPEMS) y de los Exámenes Nacionales de Certificación Profesional de Enfermería, Medicina General, Ingeniería Mecánica, entre otras especialidades.

Cuenta con más de 17 años de experiencia en enseñanza, como instructor y capacitador. Autor de la Familia de Programas KALT utilizada ampliamente en México y en otros países para la calificación y análisis de pruebas objetivas, gestión de banco de reactivos y aplicación de exámenes en computadora. Es Consultor Tecnológico Nivel 5 en el Consejo Nacional de Certificación, CONOCER, Consultor de Análisis de Rasch, University of Chicago e Institute of Objective Measurement de Chicago, Analista Registrado en NAFEMS (Reino Unido). Es autor de modelos logísticos para análisis de pruebas referidas a criterio y fórmulas para el análisis del error y la confiabilidad de pruebas de conocimientos.

Correo electrónico: kaltcomun2@yahoo.com

Rafael Vidal Uribe

Es egresado de la carrera de Filosofía de la Universidad Nacional Autónoma de México (UNAM) y tiene una Maestría en Filosofía de McGill University (Montreal). A partir del año 1982 comenzó a trabajar en asuntos relacionados con la medición y la evaluación. Desde la creación del Centro Nacional de Evaluación para la Educación Superior (CENEVAL, 1994) trabajó en dicho centro, primero como Coordinador del Examen Nacional de Ingreso a la Educación Superior (Exani II) y después como Director Técnico. Desempeñó el cargo de Director General Adjunto del Instituto Nacional para la Evaluación de la Educación de México (INEE), teniendo a su cargo responsabilidades técnicas y de decisión sobre los principales proyectos nacionales e internacionales de evaluación.

Desde mayo de 2006 ocupa el cargo de Director General del CENEVAL, donde se tiene el compromiso de aplicar estándares de calidad para diversas pruebas producidas por dicho centro. Es autor de artículos relacionados con medición y evaluación y traductor del libro sobre la técnica de Rasch: Diseño de mejores pruebas de Wright, B. y Stone, M., entre otras publicaciones del acervo del CENEVAL. Ha sido organizador de numerosas reuniones técnicas y ha dictado conferencias tanto en México como en el extranjero. Desde 1981 es catedrático de Filosofía de la Ciencia en la Facultad de Filosofía y Letras de la UNAM.

Correo electrónico: rafael.vidal@ceneval.edu.mx

Boletín Mensual del Programa Ramal No.10 del MINED

AÑO I: *Referentes teóricos de la línea de investigación*

- *Evaluación Educativa, calidad, equidad y eficacia.*
- *Investigación cuantitativa y cualitativa.*
- *TRI, HLM y SEM.*

AÑO II: *Experiencias evaluativas de la educación en Cuba*

- *Grupos Provinciales*
- *Proyectos Asociados*
- *Educador Evaluador*
- *Tesis Doctorales y de Maestría*
- *Artículos y Ponencias*

Editores: Dr. C. Paul Torres y Fdez. Dra. C. Teresa León Roldán, Jefe y Secretaria del Programa Ramal No.10 del MINED.

Presentación

Cuando se diseña una prueba educativa se debe tener en cuenta que detrás de ella debe haber un modelo. No es razonable construir una prueba como "colección de ítems" sin disponer de un diseño o de especificaciones previas, porque sería equivalente a construir un edificio sin planos de diseño. ¿Puede construirse un edificio sin especificaciones ni diseño? Seguramente es posible pero no deseable.

La palabra "prueba" es muy elocuente, porque permite recordar que el instrumento pretende "probar" la hipótesis (implícita o explícita) que se tiene para un proyecto de evaluación. El evaluador debe plantearse la pregunta: ¿qué se desea probar? Por ejemplo se pueden tener los siguientes casos: (1) se desea comprobar la hipótesis de que los estudiantes van a aprender como resultado de enfrentar un conjunto de experiencias de aprendizaje; (2) se espera probar la hipótesis de que las personas llegan a dominar una competencia al término de un proceso de capacitación; (3) se desea demostrar que la planeación didáctica de un curso fue efectiva para atender los propósitos institucionales; (4) se pretende probar la pertinencia de una intervención pedagógica para atacar las deficiencias detectadas en un grupo de jóvenes. Estos ejemplos, entre otra multiplicidad de posibilidades, corresponden a la misma lógica: probar una hipótesis. Al tratar de medir un rasgo latente se utilizan preguntas o ítems que solicitan una respuesta de la persona, pero esto no quiere decir que una prueba debe verse solamente como una colección de preguntas, sino como el instrumento que sirve para el propósito de "probar" las hipótesis de trabajo.

En el caso de las pruebas se debe contar por lo menos con la tabla de validez de contenido, TVC (o con la tabla de especificaciones) que establece el conjunto de evidencias para garantizar, como su nombre lo indica, que el contenido al que se refiere el instrumento cuenta con la validez necesaria y suficiente desde el diseño. Junto con la TVC, se debe contar con evidencias sobre el constructo que se desea evaluar (p. ej. Cronbach L.J. y Meehl, P.E., 1955) y también sobre la escala utilizada para medir tanto el constructo como el contenido.

Supóngase que se dispone de un conjunto de ítems o ítems calibrados (cuando se habla de calibración se entiende que todos los ítems han pasado por diversos criterios de validación y fueron aprobados en función de criterios estadísticos, psicométricos y edumétricos), el problema que se plantea es: ¿resulta suficiente contar con ítems calibrados para que la prueba que se construya con ellos esté bien diseñada? Resulta evidente que es necesario contar con ítems calibrados, pero ese conjunto no constituye una condición suficiente, de hecho se puede llegar a construir una mala prueba con buenos ítems.

Para construir un instrumento deficiente se pueden seguir varios procedimientos, por ejemplo, al elegir ítems que no corresponden con el perfil de la población focal a evaluar; también basta construir la prueba bajo la óptica de tener la máxima

confiabilidad a expensas de la validez; cuando se eligen todos los ítems al 50% de dificultad (como se sugiere erróneamente en varios libros sobre medición), o cuando el conjunto de ítems no esté balanceado conduciendo a veces a pruebas fáciles y otras veces a pruebas difíciles, lo cual obliga a hacer igualación de formas en cada aplicación; igualmente será deficiente el instrumento con ítems insuficientes o no válidos para medir a cabalidad el conjunto de habilidades que forman el constructo propósito de la prueba.

El análisis individual de los ítems puede realizarse a través de varias herramientas (por ejemplo el modelo clásico dificultad-discriminación, el modelo de Rasch u otros modelos logísticos) y para auxiliar el estudio de validez de contenido (y otras evidencias de validez como la validez de constructo o de criterio) se tienen modelos en diversos programas. Sin embargo, no se ha difundido suficientemente el modelo para la validez de escala, que incide directamente en la construcción de la prueba y en su validez de constructo, lo cual lo hace de utilidad para cualquier tipo de prueba (referida a norma o referida a criterio). El modelo que se presenta en este trabajo ofrece una herramienta que permite identificar qué tan bien está construida una prueba, de acuerdo con un conjunto de parámetros de diseño.

El concepto de validez de escala ha sido tratado bajo diferentes ópticas por varios autores (Suchman, 1950; Dawis, 1987; DeVellis, 1991; Byers y Byers, 1998; McDonald, 2004; O'Connor, 2004; Wright y Stone, 2004) y se puede definir como el atributo de calidad de una prueba centrada en la validez para medir el conjunto formal de objetivos característicos de un dominio teórico que corre desde "poco" hasta "mucho" en un atributo específico para una población focal.

Por ejemplo, no tiene calidad en términos de validez de escala, un reloj despertador común, que solo puede proporcionar el tiempo entre las 8 y las 11:30 de la mañana, o con una buena precisión para los primeros 20 minutos de cada hora, o cuya carátula marque solo los minutos 5,15,17,32,44,45,47 y 51. Un reloj con validez de escala se enfoca entonces a medir el tiempo, en todo el rango de utilidad para el proyecto deseado, mostrando marcas uniformemente distribuidas. ¿Se desea contar con un cronómetro para pruebas olímpicas? Entonces un reloj despertador común no es el instrumento apropiado para ello, en cambio deberá tenerse un aparato que permita medir hasta fracciones de segundo, pero no sería útil si solo pueden leerse entre 8 y 11:30 segundos, o los primeros 20 segundos de la competencia, o cuando la carátula solo marque las fracciones de segundo 5, 15, 17, 32,44, 45, 47 y 51. Un cronómetro con validez de escala debe medir el tiempo en todo el rango de utilidad para la competencia olímpica deseada, mostrando marcas o medidas uniformemente distribuidas en décimas, centésimas o milésimas de segundo. Del mismo modo, se espera que los ítems de una prueba (que se interpretan igual que los minutos de la carátula del reloj) estén distribuidos uniformemente para medir el atributo deseado en todo el rango esperado.

Requisitos de diseño

Para el diseño de una prueba desde el punto de vista de la validez de escala se sugieren los siguientes requisitos que corresponden con cualquier instrumento de medida, como pueden ser una regla, un reloj, un barómetro o una prueba objetiva para medir conocimientos de historia o de matemática:

A) **Ámbito de medida.**

Debe garantizarse que la escala del instrumento cubra el rango más amplio posible del atributo a medir, de tal manera de no dejar fuera a un sustentante, lo cual permitirá estimar la medida de cualquier persona con una mejor precisión. Las diferencias individuales de los sujetos están en proporción directa a la amplitud del ámbito: si el ámbito de medida es reducido, entonces las diferencias individuales estarán menos marcadas, lo cual induce una separación poco clara entre los individuos, dando lugar a problemas de identificación de rasgos distintivos entre las personas. Es posible medir a las personas que caen fuera del ámbito de medida, pero el error estándar será más grande conforme más alejada esté la posición real de la persona en el atributo medido; por ello es importante que el ámbito cubra a todos las personas de la población focal, para evitar grandes errores al estimar sus medidas.

El modelo que se propone tiene como ámbito de diseño el intervalo (20, 80) % en grado de dificultad clásico, o su equivalente en lógitos: (-1.386, +1.386). En este documento se presenta el modelo en términos de grado de dificultad clásico, sin pérdida de generalidad, pudiendo hacer el lector todas las transformaciones a medidas IRT en caso de desearlo, debiendo notarse que para el ámbito en lógitos puede trabajarse más cómodamente entre -1.5 y +1.5 sin ningún problema teórico o práctico.

B) **Uniformidad de distribución de los ítems**

Una buena medida se obtiene en cualquier posición de la escala cuando los ítems están uniformemente distribuidos en el ámbito propuesto. Cuando se tienen ítems con distribuciones azarosas el error no es uniforme produciendo medidas de precisión variable dependiendo de la posición de cada persona. La uniformidad es deseable porque se consiguen estas ventajas:

- Contar con ítems a iguales distancia, lo cual evita dos problemas: el primero es cuando se aglomeran ítems en iguales dificultades (apilamientos), el segundo es cuando no se tienen elementos de medida en algunas dificultades dentro de la variable a medir (saltos).
- Disponer de ítems (elementos de medida) en una escala uniforme que facilita las interpretaciones.

- Contar con un error teórico del instrumento uniforme en todo el ámbito de medidas.
- Mejorar la separación de los ítems, lo cual tiene una incidencia en la confiabilidad del instrumento (separación y confiabilidad son dos conceptos que se estudian desde el punto de vista de análisis de Rasch o IRT).

C) **Dosificación de ítems para cada nivel de desempeño**

Se debe contar con un número comparativamente similar en cualquier nivel de desempeño donde se hagan cortes para definir un constructo, es decir, un constructo se define en función de un conjunto de ítems que conforman un cierto nivel de desempeño. Evidentemente, resulta inconveniente definir un constructo en un nivel formado por un solo reactivo (o muy pocos ítems). Cuando la dosificación no es comparable se puede llegar a tener una descripción muy detallada del dominio de una persona al tener varios ítems y, en contraste, una descripción deficiente o hasta inexistente cuando no se cuenta con ítems suficientes que describan en qué consiste el dominio en una posición específica dentro de la escala. Las ventajas de esta propiedad son:

- Facilita la asignación de los puntos de corte al disponer de una escala lineal de ítems.
- Permite interpretar cada nivel en términos de competencias, habilidades, conocimientos u otro atributo que se esté midiendo.
- Favorece el análisis de validez de constructo por opinión de jueces, como el modelo bookmark muy utilizado en la práctica, pero deficiente si no parte de una dosificación correcta.

D) **Medida de referencia.**

Como resultado de las cualidades anteriores se deriva una más que consiste en que el instrumento debe estar referido a una medida centrada en la media de las habilidades, competencias o dominio esperado. Con este requisito la escala queda centrada con relación al rango total de medidas de dominio de las personas o dentro del rango del constructo medido. Esta condición proporciona los siguientes beneficios:

- Se evita el sesgo de diseño en la medida general del instrumento, balanceada en el constructo, la cual no quedará cargada hacia el lado fácil o hacia el lado difícil. Un diseño balanceado evita que la medida de los sujetos dependa de la dificultad relativa del instrumento, porque obliga a hacer procesos de ajuste de puntuaciones, transformación a medidas estandarizadas Z u otros procedimientos numéricos que balancean el instrumento artificialmente.

- La medida general del instrumento estará balanceada en el dominio de los sustentantes y tendrá una mejor distribución de las respuestas de los sustentantes.
- Facilita contar con la más amplia desviación estándar, tanto para los puntajes de los sustentantes como para la calibración de los ítems.
- Evita tener que hacer igualación (*equating*) entre versiones o módulos de la prueba

Obsérvese que el ámbito de 20 a 80% tiene como medida central 50% (o su equivalente en 0 lógitos al correr de -1.5 a +1.5 lógitos), cumpliendo la necesidad de una escala centrada como estipula el primer parámetro de diseño.

Por tratarse de requisitos de diseño, no se puede aceptar cualquier distribución experimental que se obtenga de una prueba cualquiera. Se trata, por el contrario, de verificar que la distribución experimental corresponda con las especificaciones de diseño: si la distribución de ítems experimentales ajusta al modelo, entonces el diseño de la prueba será aceptable, en caso contrario deberá mejorarse.

Por ello, cuando se pretende aceptar cualquier distribución experimental obtenida en una aplicación, con el argumento de que la "evidencia empírica" manda sobre cualquier diseño teórico, se entraría en el campo de aceptar cualquier edificio sin diseño. Para el modelo que se presenta, no es aceptable una prueba si no cuenta con un diseño. Seguramente puede ocurrir que al aplicar el instrumento los resultados no se parezcan al diseño, la comparación proporciona numerosas opciones de mejorar y corregir la prueba, para que en aplicaciones sucesivas se vaya pareciendo más al diseño deseable con las características indicadas por el modelo.

Descripción del modelo

El modelo propuesto que satisface estos requisitos de diseño es muy sencillo y de fácil aplicación y se denomina "***Recta de diseño de la prueba***". El modelo se basa en la propuesta original de Wright y Stone (1979 y 1988), que posteriormente ha sido retomada por otros autores (por ejemplo Bond y Fox, 2001), con la diferencia que en la propuesta de los autores citados no se cuenta con un modelo de referencia. El modelo de "*Recta de diseño de la prueba*" que se propone a continuación permite disponer de un criterio objetivo que sirve referencia y de aplicación general. Se trata de un modelo que garantiza que los ítems se distribuyen en todo el ámbito deseado para la población focal, con una distribución uniforme y centrada en el atributo objetivo del proyecto de evaluación.

Para facilitar el trabajo de construcción del modelo se plantea una herramienta geométrica que facilita además la visualización tanto del diseño teórico como del experimental. Para diseñar la "*Recta de Diseño de la Prueba*" constrúyase el "*Plano de Diseño*", cuyo eje horizontal tiene los grados de dificultad y el eje vertical

tiene a los ítems ordenados en forma secuencial del más difícil al más fácil (en medidas logísticas se ordenan del más fácil al más difícil). El modelo establece que los ítems deben alojarse sobre la recta que une el punto de dificultad 20% para el primer reactivo con el punto de dificultad 80% en el último reactivo de la prueba.

El modelo se muestra en la Figura No.1 para una prueba de 50 ítems.

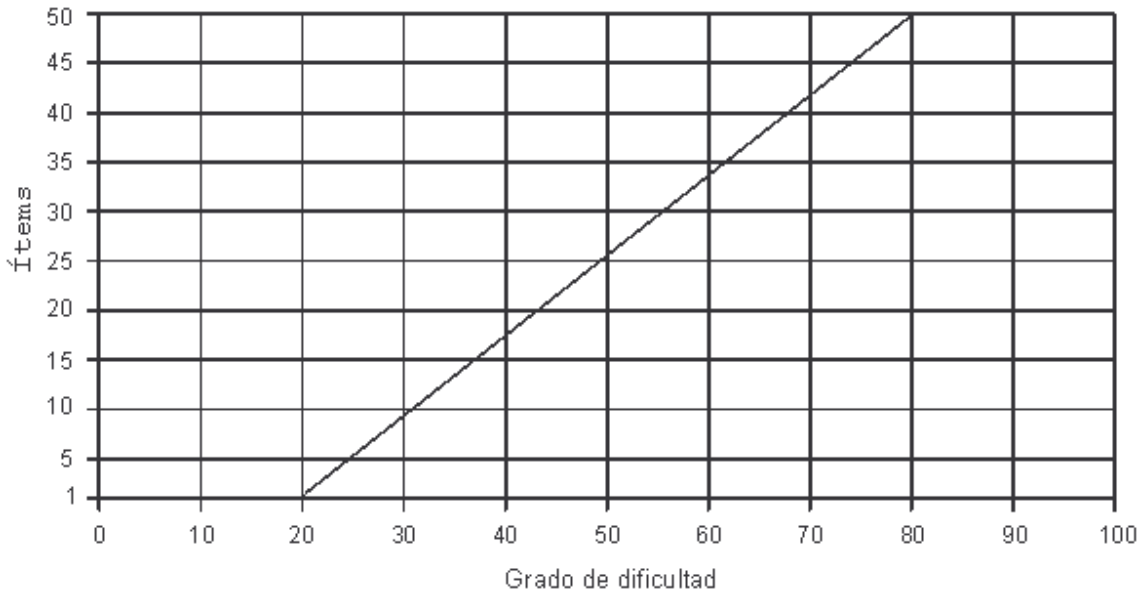


Figura No.1: Plano de diseño y Recta de diseño de la prueba[1].

La recta mostrada (denominada en lo sucesivo **RECTA DE DISEÑO 20-80**) cumple con los requisitos de diseño (A) a (D), de tal modo que si se construye una prueba cuyos ítems se alojen en ella, se tendrán todas las ventajas previstas.

La distancia entre la dificultad teórica y la dificultad experimental puede ser medida fácilmente por medio de su diferencia, con lo que se tiene un valor objetivo de cotejo de aceptación o de rechazo para una prueba dada.

Complementariamente a lo anterior, se dispone de un indicador objetivo que proporciona un valor de ajuste o desajuste de cada uno de los ítems con respecto al diseño, se trata de una "distancia" entre la dificultad del reactivo y el modelo de diseño, de tal modo que mientras menor sea el desajuste la prueba estará mejor diseñada.

[1] Nota: El modelo de la figura 1 deberá escalarse en el eje vertical al número de ítems que tenga la prueba o variable medida.

Expresión matemática del modelo

La ecuación de la **Recta de Diseño 20-80** depende del número de ítems que se estén utilizando en una prueba y tiene esta forma:

$$GD_{esperado} = 20 + 60\,(y-1) / (N-1) \qquad (1)$$

Donde $GD_{esperado}$ es el grado de dificultad esperado para el reactivo "y", N es el numero de ítems totales de la prueba o de la sección de prueba que se esté analizando. Para este modelo los ítems deben ordenarse por dificultad, siendo "y" la posición que ocupa el reactivo dentro de la variable de diseño y no su número de identificación dentro de la prueba.

Al igual que la recta del modelo, se puede determinar también la *Recta de Diseño* observada y=m x +b (donde "x" es el grado de dificultad del reactivo "y"), por medio de las ecuaciones generales:

$$m = (n \Sigma xy - \Sigma x \Sigma y) / (n \Sigma x^2 - \Sigma x \Sigma x) \qquad (2)$$

$$b = (\Sigma x^2 \Sigma y - \Sigma x \Sigma xy) / (n \Sigma x^2 - \Sigma x \Sigma x) \qquad (3)$$

En cuyo caso se pueden hallar los límites de dificultad inferior y superior observados:

$$Gd_{min} = (1-b) / m \qquad (4)$$

$$Gd_{max} = (n-b) / m \qquad (5)$$

Estos valores se pueden comparar contra los valores del ámbito de diseño 20% y 80% respectivamente, con ello el evaluador puede estimar que tan amplio está el ámbito.

Adicionalmente, al disponerse de la media de dificultades observadas, se puede comparar contra el valor de diseño centrado al 50%.

La comparación reactivo a reactivo contra la **Recta de Diseño 20-80** proporciona una "distancia" o "desajuste" dado por:

$$Desajuste = GD_{observado} - GD_{esperado} \qquad (6)$$

Donde $GD_{observado}$ es el grado de dificultad observado u obtenido experimentalmente en el reactivo "y", en tanto que $GD_{esperado}$ se obtiene de la ecuación (1). Es claro que si $GD_{observado} = GD_{esperado}$ se tiene un desajuste nulo y el reactivo estará alojado en la recta de diseño 20-80.

Es muy importante y útil tener en cuenta que el desajuste tiene signo, de tal modo que:

Si Desajuste es mayor que 0 entonces el reactivo es más fácil que el valor esperado.

Si Desajuste es menor que 0 entonces el reactivo es más difícil que el valor esperado.

La combinación de los desajustes individuales de los ítems brinda dos posibles medidas globales de desajuste de la prueba en su conjunto:

Desajuste absoluto medio: $DAM = (\Sigma \left| \text{desajuste} \right|) / N$ (7)

Desajuste cuadrático medio: $DCM = (\Sigma \text{desajuste}^2) / N$ (8)

Ambas medidas son positivas y brindan un estimado de la diferencia de los valores observados respecto de los esperados, se trata por lo tanto de valores objetivos contra los cuales puede compararse una prueba dada, respecto a la *Recta de Diseño 20-80*.

No hay valores teóricos para la aceptación del desajuste, pero la experiencia de múltiples pruebas verificadas con el modelo desde 1999 señala que es aceptable una DAM de 5% por diseño y hasta 10% por revisión. Cuando se habla de diseño es cuando se toman ítems calibrados de un banco para construir una prueba siguiendo los criterios indicados; una vez administrado el instrumento se calibran de nuevo los ítems para determinar la recta experimental y su valor de revisión. El intervalo teórico para DCM es (50, 225), recordando que son valores obtenidos al elevar el desajuste al cuadrado, por lo que es más sencillo trabajar con DAM que es adimensional.

Aspectos cualitativos del modelo

Respecto a los aspectos cualitativos, el modelo permite identificar estos elementos:

a) División de las dificultades

La **Recta del modelo** define a los ítems "bien distribuidos" por lo cual divide al Plano de Diseño en tres zonas:

- Arriba de la recta (o a su izquierda) se trata de ítems más difíciles que lo que sugiere el modelo.
- Debajo de la recta (o a su derecha) los ítems son más fáciles que lo que indica el modelo.
- Se establece una "**zona de aceptación**" dentro de la distribución esperada para los ítems de la prueba, dentro de una banda de 5% de cada lado de la Recta de Diseño 20-80.

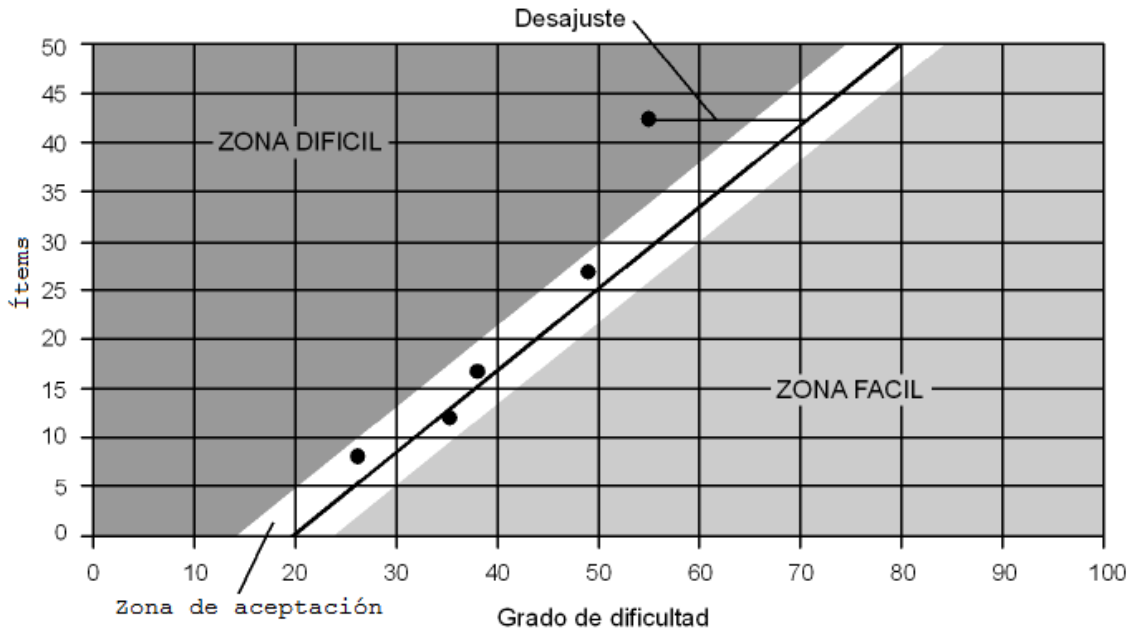


Figura No. 2: Zonas en el plano de diseño.

La distancia horizontal de un reactivo a la Recta de Diseño 20-80 es la medida de desajuste. Se considera desajustado a un reactivo fuera de la banda establecida en la zona de aceptación.

b) Apilamientos

Cuando los ítems no están bien distribuidos se pueden tener dos problemas. El primero de ellos es el apilamiento (Figura No.3). Este problema ocurre cuando hay varios ítems en una misma dificultad, en cuyo caso no aportan un valor substancial a la escala de medida o al constructo, al estar explorando una misma medida o posición en la escala. Se sugiere cambiar estos ítems "repetidos" o "apilados" por ítems de otras dificultades, para equilibrar el instrumento. Estas sugerencias han sido planteadas por Wright y Stone (1979) bajo otras consideraciones convergentes con este modelo.

c) Saltos

El segundo problema de la mala distribución es la presencia de saltos o huecos en la escala. Se trata de intervalos de dificultad donde no hay ítems y se manifiestan en "saltos horizontales" entre dos ítems sucesivos. Un salto está asociado con un desajuste importante de los ítems, incrementando a su vez el desajuste medio de la prueba.

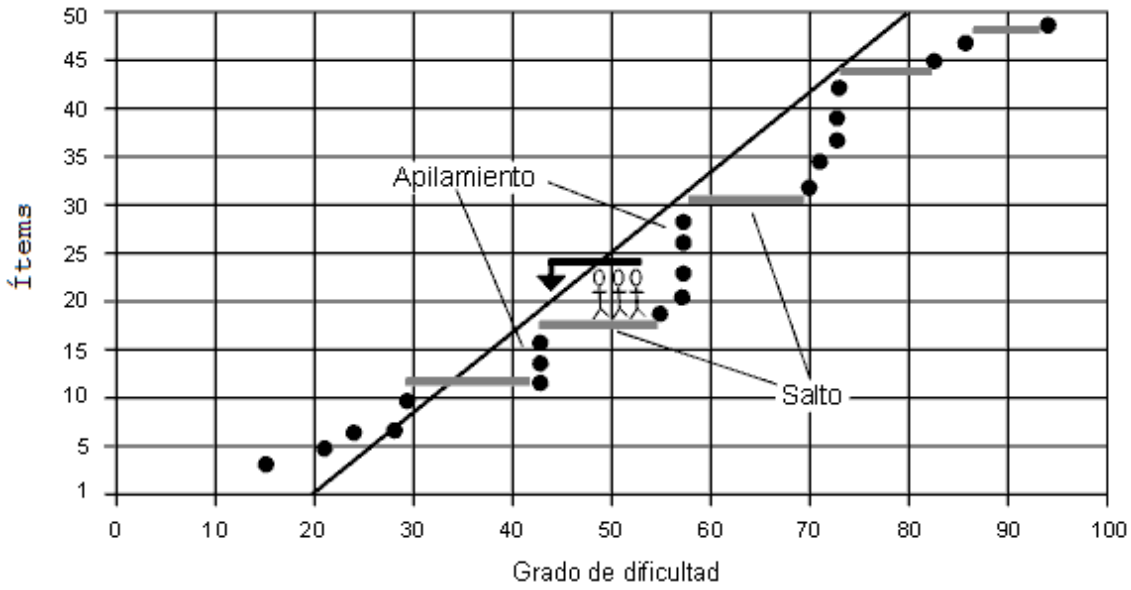


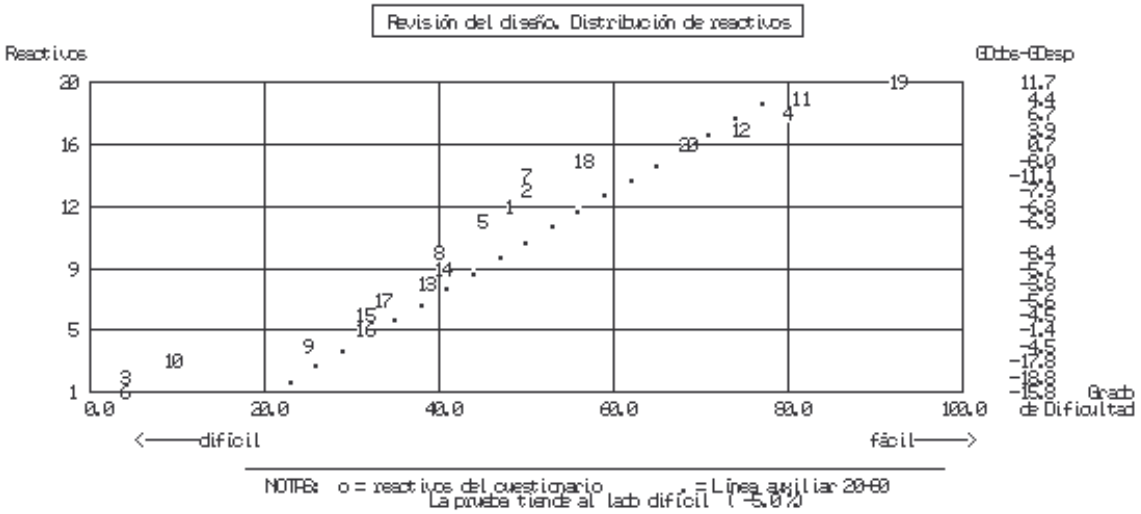
Figura No. 3: Ejemplo de apilamientos y saltos en la distribución de ítems.

Cuando hay saltos se puede decir que "faltan" ítems que miden una cierta parte de la escala. Al faltar elementos de medida, entonces los sustentantes que tienen un nivel de dominio en dicha zona serán medidos por defecto en el intervalo inmediato anterior y se mezclarán o confundirán con otros sustentantes de ese intervalo, incrementando el error de medida en estos casos (Figura No.3).

En caso de que la prueba tenga muchos saltos, apilamientos, desajustes, etc., el evaluador deberá realizar modificaciones a su diseño de manera de mejorarlo. Una prueba con muchos ítems difíciles o muchos ítems fáciles deberá incluir nuevos ítems que balanceen el diseño, centren y afinen el ámbito.

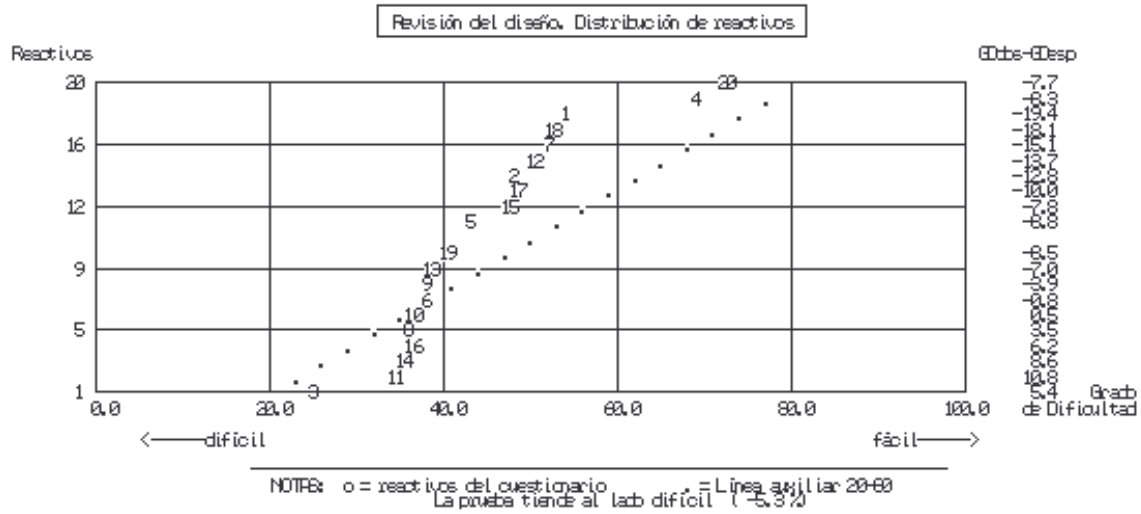
Ejemplos de resultado con el modelo

Los programas que proporcionan directamente el plano de diseño son **Winsteps** (Linacre, 2008) y **Kalt Criterial** (Tristán, 1999-2006). Se presentan dos ejemplos proporcionados por KALT-CRITERIAL en el Reporte Técnico. Se sugiere al lector utilizarlos para verificar los aspectos descritos en esta Nota.



NOTA: o = reactivos del cuestionario . = Línea auxiliar 20-80
La prueba tiende al lado difícil (-5.0%)

Ajuste de recta a datos observados				Diferencia (dDbs-dDesp)	
Límite inferior	3.8	Pendiente	0.28	Absoluta media	7.7
Límite superior	66.3	Ordenada	0.1		
Dificultad media	46.0	R2	0.957	Cuadrática media	88.6



NOTA: o = reactivos del cuestionario . = Línea auxiliar 20-80
La prueba tiende al lado difícil (-5.3%)

Ajuste de recta a datos observados				Diferencia (dDbs-dDesp)	
Límite inferior	24.7	Pendiente	0.48	Absoluta media	8.9
Límite superior	64.6	Ordenada	-10.8		
Dificultad media	44.7	R2	0.873	Cuadrática media	108.1

Figura No.4: Aplicación a pruebas reales.

Aplicación al análisis de constructo

Una aplicación importantísima de este modelo se enfoca a la revisión de la validez de constructo de una prueba. Partiendo de la ventaja de tener los ítems de la prueba ordenados por dificultad, es factible definir "cortes" de niveles de dominio (Figura No. 5) y asociarlos con los ítems que se encuentran dentro de dicho "corte". Una vez identificados, el evaluador podrá trabajar en identificar el constructo involucrado en los ítems dentro de dicho corte.

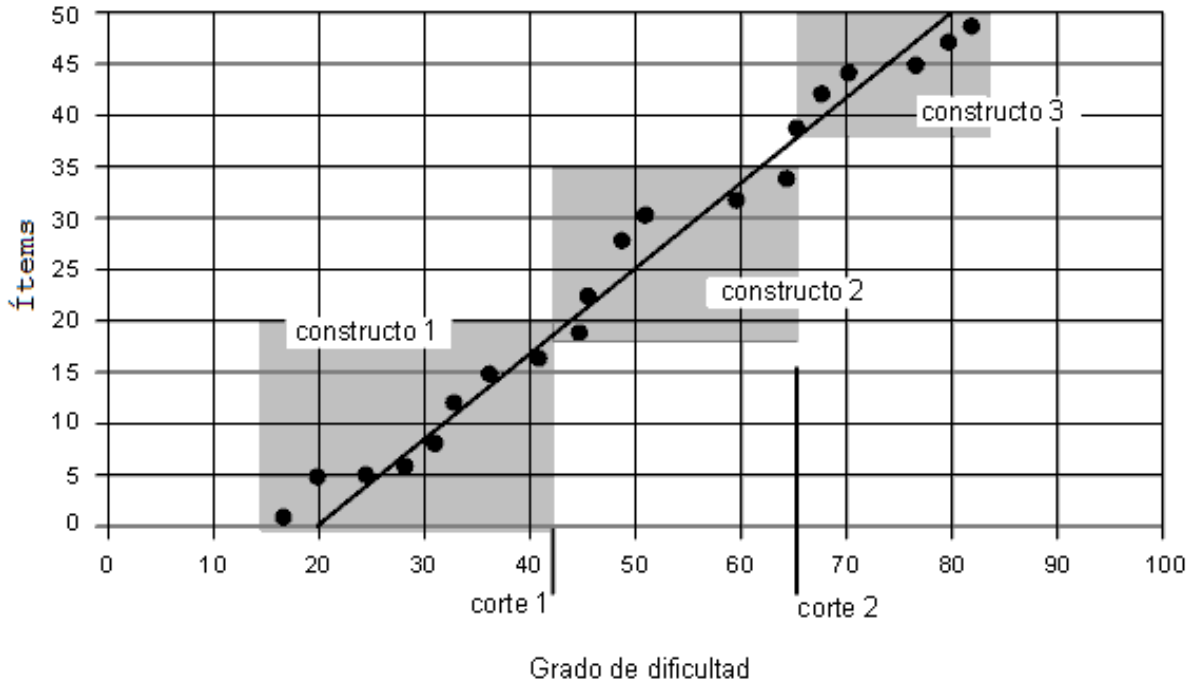


Figura No.5: Organización de ítems por niveles de constructo.

Al hacer la interpretación y análisis de constructo por intervalos de corte, se puede llegar a definir una taxonomía experimental, donde los niveles de complejidad corresponden con niveles de dificultad, lo cual puede ser muy conveniente si se compara contra los niveles de complejidad que se tienen en las taxonomías teóricas (desde la taxonomía de Bloom hasta las más modernas), donde la "complejidad" corresponde con procesos mentales que van de simple a complejo, definidos de manera teórica, *a priori*, por un grupo de expertos, pero que no pueden ser demostrados ni medibles de manera objetiva; en cambio la dificultad corresponde con la frecuencia de respuestas correctas de las personas de la población focal, que se obtiene de manera objetiva por medio de una medición (Tristán y Molgado, 2006). Este enfoque se refiere al modelo de anclaje para definir constructos propuesto por Beaton y Allen (1992) o a la taxonomía **SOLO** (Biggs y Collis, 1982), así como al esquema *bookmark* para interpretar los niveles de corte.

La revisión de lo que mide el constructo en cada nivel de dominio o de desempeño es tarea que podrá dejarse a jueces o expertos, quienes deberán decidir todos los aspectos generales de validez de constructo, aprovechando la información proporcionada por el modelo en función de los grupos de ítems que entran en cada nivel.

Conclusiones

Se presenta el modelo de "***Recta de diseño de una prueba***" que proporciona una herramienta sencilla para diseñar y revisar una prueba y contar con una evidencia de su validez de escala. El modelo se presentó aquí utilizando el Grado de Dificultad del modelo clásico, pero puede utilizarse igualmente con las medidas de modelos logísticos, en particular las medidas procedentes del análisis de Rasch o de la teoría de la respuesta al Ítem (Tristán y Vidal, 2007).

El modelo de "**Recta de diseño de una prueba**" ha sido utilizado desde 1999 con éxito, permitiendo el análisis de calidad de la escala de muy diversas pruebas tanto nacionales como internacionales.

Referencias

1. Beaton, A.E. y Allen, L. (1992) *Interpreting scales through scale anchoring.* Journal of Educational Statistics, 17, 191-204
2. Biggs, J.B., y Collis, R.E. (1982) *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.
3. Bond T.G. y Fox C.M. (2001) *Applying the Rasch model.* Erlbaum, NJ Pp. 4-8.
4. Byers C. y Byers W.A. (1998) *Sliding scale: a technique to optimize the assessment of knowledge level.* Int. Pers. Manag. Ass. Assessment Council. Chicago. June. 5 pp.
5. Cronbach L.J. y Meehl, P.E. (1955) *Construct validity in psychological tests.* Psychological Bulletin, 52, 281-302
6. Dawis R.V. (1987) *Scale construction.* Journal of Counseling Psychology. 34(4), 481-489
7. DeVellis R.F. (1991) *Scale development.* Applied Social Research Methods Series. Vol. 26. Sage, NewBury Park. Pp. 8-11.
8. Linacre J.M. (2006) A user's *guide to Winsteps.* Winsteps.com
9. O'Connor (2004) *Measuring quality of life in health.* Elsevier Science Health Science.
10. McDonald J.A.L. (2004) *The optimal number of categories for numerical rating scales*. PhD Dissert. Coll. Education. Univ.of Denver. 170 pp.
11. Suchman E.A. (1950) *The logic of scale construction.* Educational and Psychological Measurement. Vol. X. 79-93
12. Tristán L.A. (2001) *Contribución al estudio del error de medida.* Notas sobre evaluación criterial, N.13. Instituto de Evaluación e Ingeniería Avanzada, México.
13. Tristán, L.A. y Molgado R.D. (2006) Compendio de taxonomías. Instituto de Evaluación e Ingeniería Avanzada, México.
14. Tristán, L.A. y Molgado R.D. (2007) Limits of measures in tests, a meta-analysis. Internal Report, IEIA. Instituto de Evaluación e Ingeniería Avanzada, México.
15. Tristán L.A. y Vidal R. (2007) *Linear model to assess the scale's validity of a test*. Reunión de AERA. Chicago. Disponible en ERIC: ED501232
16. Tristán, L.A. (1999-2006) Kalt Criterial, Manual de usuario. Instituto de Evaluación e Ingeniería Avanzada, México.
17. Wright B.D. y Stone M.H. (1988) *Validity in Rasch measurement.* Research memorandum 54. MESA. University of Chicago. 12 pp.
18. Wright B.D. y Stone M.H. (1979) *Best test design.* MESA Press. Chicago. pp 133-140
19. Wright B.D. y Stone M.H. (2004) *Making measures*. The Phaneron Press. Chicago, USA. pp. 35-39