UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

INTRODUCCIÓN AL ANÁLISIS DE REACTIVOS A PARTIR DE LA TEORÍA CLÁSICA DE LOS TESTS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I O

P R E S E N T A :

CESAR ANTONIO CHÁVEZ ALVAREZ

DIRECTOR DE TESIS:
M. en C. JOSÉ SALVADOR ZAMORA MUÑOZ
Cd. Universitaria, D.F. 2013

1. Datos del alumno
Chávez
Alvarez
César Antonio
5531363748
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
098502329

2. Datos del tutor
M en C
José Salvador
Zamora
Muñoz

3. Datos del sinodal 1
Doctora
Ruth Selene
Fuentes
García

4. Datos del sinodal 2
Actuario
Jaime
Vázquez
Alamilla

5. Datos del sinodal 3
Doctor
Ricardo
Ramírez
Aldana

6. Datos del sinodal 4
Actuario
Erick
Mier
Moreno

7. Datos del trabajo escrito
Introducción al análisis de reactivos a partir de la Teoría Clásica de los Tests
124 p
2014

Contenido

Índice de tablas

Índice de figuras

Presentación

En este trabajo se presentan las técnicas de análisis de reactivos que se usan en el desarrollo y en la aplicación de instrumentos de evaluación. Se trata de herramientas relativamente comunes en las instituciones que hacen medición educativa. En particular las técnicas básicas que forman parte de la llamada Teoría Clásica de los Tests (TCT) que son aplicadas en pruebas de bajo impacto y muy pequeña escala, acotada incluso a un salón de clase; pero que también son de mucha utilidad para pruebas aplicadas a gran escala como es el caso de las desarrolladas en el Centro Nacional de Evaluación para la Educación Superior A.C. (Ceneval).

Sin omitir los detalles técnicos más relevantes de los modelos para el análisis de reactivos, el objetivo principal del trabajo es hacer una presentación intuitiva de la racionalidad que subyace al análisis cuantitativo de los instrumentos de medición, recalcando su importancia dentro de la metodología utilizada para el desarrollo de las pruebas. En el área de evaluación el análisis de los datos empíricos es de fundamental importancia, ya que permite garantizar la confiabilidad de las inferencias realizadas a partir de los resultados obtenidos en la aplicación de exámenes o cuestionarios.

El documento se divide en cuatro capítulos. El primero es una pequeña introducción sobre el Ceneval y sobre el tipo de evaluación que en él se desarrolla. El segundo describe los principales usos y herramientas del análisis de reactivos y se explican los pasos más importantes para llevarlo a cabo. El tercer capítulo es la exposición formal del modelo de la Teoría Clásica de los Test (TCT), entrando en detalle sólo en algunas de las deducciones más importantes del modelo; también incluye una explicación general de los modelos de la Teoría de la Respuesta al Ítem. El cuarto y último capítulo es un ejemplo que permite interpretar el análisis de reactivos por TCT a partir de los resultados de un programa desarrollado en el software “Mathematica” (Wolfram, 2010).

La exposición y los ejemplos se centran en preguntas (reactivos) de opción múltiple con una respuesta correcta (calificación dicotómica). Este tipo de preguntas, por su versatilidad y por sus ventajas operativas, son de uso común en los exámenes objetivos

y en la evaluación educativa, al grado que se han convertido en la columna vertebral de la mayor parte de las pruebas que se elaboran en el Ceneval.

Capítulo 1. El Ceneval y la evaluación que lleva a cabo

El Centro Nacional de Evaluación para la Educación Superior (Ceneval) es una asociación civil que surgió por mandato de la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES) en 1994. Desde su creación se ha caracterizado por ser una institución eminentemente técnica, cuya misión es promover la calidad de la educación mediante la realización de evaluaciones válidas, confiables y pertinentes. Sus actividades se orientan, principalmente, a evaluar los conocimientos y habilidades adquiridos por los individuos en procesos de enseñanza-aprendizaje formales o no formales de los niveles educativos medio superior y superior. El Centro ofrece sus servicios a instituciones educativas, organizaciones de profesionistas, así como a entidades privadas y gubernamentales. Con el propósito de dar cumplimiento a su misión, el Ceneval persevera permanentemente en el mejoramiento de las evaluaciones que practica, rigiendo sus actividades bajo los siguientes principios básicos:

- **Imparcialidad**. El Centro no favorece a ninguna persona, organismo o institución en particular.

- **Calidad técnica**. Los procesos de diseño, construcción, aplicación, calificación y reporte de resultados de los exámenes están apegados a rigurosos estándares de calidad reconocidos por agencias de evaluación nacionales e internacionales para garantizar su validez y confiabilidad.

- **Actualidad**. Los exámenes se revisan continuamente para ajustar los contenidos que se evalúan a los cambios que sobrevienen en las áreas disciplinares o profesionales.

- **Confidencialidad**. Las personas y las instituciones evaluadas tienen derecho a que sus resultados no sean revelados sin su consentimiento.

- **Transparencia**. Los procesos concernientes a la elaboración, aplicación y calificación de los exámenes están abiertos a escrutinio, siempre y cuando no se contrapongan al principio de confidencialidad.

Apegados a estos principios, los cuerpos colegiados y el personal técnico del Ceneval desarrollan las estrategias de evaluación que permiten generar información sobre los conocimientos, habilidades y competencias que los individuos han logrado adquirir.

Esta información fehaciente permite a las instituciones educativas usuarias contar con elementos sólidos para orientar acciones de mejora educativa. A las personas que responden la prueba (sustentantes) se les ofrece equidad y pertinencia en la evaluación, y a los investigadores, un vasto universo de datos con los que pueden diseñar múltiples modelos explicativos de los aprendizajes logrados.

Los Exámenes del Ceneval

Los exámenes que diseña y elabora el Ceneval tienen las siguientes características:

- **Validez**. Los instrumentos se diseñan y construyen para medir un dominio claramente identificado. Los reactivos de cada instrumento de medición son congruentes con las definiciones del objeto de medida sancionado por el Consejo Técnico del examen. Por ello, las inferencias que se realizan de los resultados son indicadores indudables del rasgo medido.

- **Confiabilidad**. Los resultados de la evaluación son consistentes, de tal forma que los cambios en los niveles de competencia de los sustentantes reflejan cambios en los puntajes obtenidos.

- **Equidad**. Los instrumentos de medición se construyen de forma tal que no benefician a un grupo determinado. Se revisa minuciosamente la redacción y el contenido de los reactivos a fin de evitar sesgos culturales (de género, religiosos, políticos y regionales).

- **Pertinencia**. Los instrumentos de medición tienen una cobertura suficiente de los dominios en los que el sustentante debe mostrar su competencia.

- **Objetividad**. Las respuestas de los sustentantes se califican siguiendo el algoritmo autorizado para el examen correspondiente, por lo que todos los sustentantes de un examen en particular son calificados de la misma forma. Asimismo, todos los instrumentos de medición se califican en un área independiente a la que construye el examen y a la que lo aplica.

Tipología de exámenes

Cada instrumento de medición se diseña y construye de acuerdo con un propósito que delimita los usos de sus resultados de evaluación. El Ceneval desarrolla los siguientes exámenes:

Exámenes de admisión

Su propósito es elegir a los mejores candidatos para ingresar a una institución educativa o seleccionar a los mejores individuos para una posición laboral. Los sustentantes elegidos son aquellos que obtuvieron los mejores puntajes en la prueba (evaluación normativa). Estos exámenes se orientan a la medición de los conocimientos y habilidades que son predictivos de un buen desempeño del nivel por el que concursa. Se aplican antes de que inicien las acciones educativas o laborales y pueden fungir como el corte basal de evaluaciones longitudinales.

En esta categoría se encuentran el EXANI-I, que se utiliza en la selección de los estudiantes que entrarán a cursar el nivel medio superior; el EXANI-II, en la selección de sustentantes que entran a estudios superiores, y el EXANI-III, en la selección para estudios de posgrado.

Exámenes de egreso

Identifican el nivel de conocimientos y habilidades que logran los sustentantes después de haberse expuesto a un proceso formal de instrucción. Generalmente, estas

evaluaciones son un indicador del logro alcanzado, de las metas o de los currículos planeados. Los exámenes de egreso son evaluaciones sumativas o finales.

En el Ceneval estos exámenes se utilizan para evaluar si los sustentantes adquirieron los conocimientos y las habilidades esenciales para el inicio del ejercicio profesional. En esta categoría están los Exámenes Generales de Egreso de la Licenciatura (EGEL).

Exámenes de diagnóstico

Su objetivo es identificar las fortalezas y debilidades de un programa o proceso al momento de la evaluación. Estos exámenes se alinean a currículos o descripciones puntuales de competencia, por lo que necesariamente son instrumentos criteriales[1].

Los exámenes de diagnóstico se pueden aplicar al inicio, durante o al final de un proceso o ciclo escolar. En esta categoría se encuentran la prueba de Evaluación Nacional de Logro Académico en Centros Escolares de Educación Media Superior (ENLACE-EMS), el Examen Intermedio de Licenciatura en Ciencias Básicas de Ingenierías (EXIL), que permite identificar en qué medida los futuros ingenieros, en la fase intermedia de su licenciatura, cuentan con los conocimientos y habilidades intelectuales que se consideran básicas para su formación profesional, entre otros.

Exámenes de acreditación

Identifican si los sustentantes adquirieron los conocimientos y las habilidades que son equivalentes a los estudios que se proporcionan en un nivel escolar (por ejemplo, los exámenes de acreditación de la secundaria, el bachillerato o licenciaturas).

Estos exámenes se aplican en momentos no ligados a procesos de instrucción. Generalmente, los sustentantes son los que deciden tomar una prueba para mostrar que son aptos, por lo que no se aplican en determinados momentos del proceso educativo. El Ceneval participa con la SEP en los procesos de acreditación de la secundaria, el bachillerato y de algunas licenciaturas (exámenes de los acuerdos 286 y 357).

[1] A diferencia de los exámenes normativos, en los exámenes criteriales se establecen estándares de desempeño a partir de los cuales son dictaminados los sustentantes.

Exámenes de certificación

Su objetivo es identificar si los sustentantes cuentan con los conocimientos y habilidades que establecen los colegios de profesionistas como indispensables para el ejercicio de una profesión. Estos exámenes permiten informar a la sociedad sobre la calidad de la instrucción profesional en disciplinas particulares. Entre los exámenes de certificación que realiza el Ceneval figuran el Examen Uniforme de Certificación de la Contaduría Pública (EUC) y el Examen Único de Certificación de la Calidad del Actuario (EUCCA).

Capítulo 2. ¿Qué es el análisis de reactivos?

Por análisis de reactivos se entienden varias técnicas y procedimientos matemáticos encaminados a verificar la calidad y pertinencia de los reactivos o preguntas de una prueba. El análisis permite inferir las características técnicas de una pregunta, determinar si cumple con los requerimientos que se esperan de ella y, finalmente, decidir si puede ser seleccionada para incluirse en una prueba. También ayuda a elegir los mejores reactivos para una prueba determinada.

En la práctica, la construcción de una prueba requiere de la elaboración de un grupo numeroso de reactivos (ítems), por lo general mucho mayor que los necesarios para la construcción de las versiones previstas por el proyecto de evaluación. Este conjunto de reactivos nuevos se aplica a una muestra de sujetos de la población objetivo (estudio piloto), a fin de analizarlos para obtener índices de su calidad, y poder así producir una prueba definitiva que sea confiable y válida.

Para establecer una relación entre las observaciones obtenidas por medio de la aplicación de una prueba y el constructo o rasgo que se pretende medir, es necesario un modelo de medición.

Conocer y estudiar las características individuales de un reactivo es importante si se quiere mejorar una prueba, regularizar su calidad y estandarizarla. Los reactivos son, en última instancia, los “ladrillos” con que se construye una prueba. De su calidad depende en buena medida que se tenga una buena prueba o no. El análisis estadístico de los reactivos es una herramienta útil pues ayuda a corroborar que en la práctica estos “ladrillos” se comporten como se espera que lo hagan.

Análisis de reactivos en la construcción de una prueba

Una prueba sirve para recopilar información sobre el objetivo que se va a evaluar, de ahí que los reactivos se deben diseñar a partir de una base teórica y metodológica que

garantice que la información que de ellos se obtenga sea útil para la toma de decisiones.

La *metodología Ceneval* para la elaboración de pruebas estandarizadas, objetivas y con reactivos de opción múltiple (Ceneval, 2008), se basa en un modelo secuencial en el que los productos de una fase se convierten en insumos de la siguiente.

Las pruebas son objetivas cuando su calificación no depende del evaluador, toda vez que, sin excepción, hay una única respuesta correcta. Son estandarizadas porque se aplican y se califican manteniendo las mismas condiciones para todos los sustentantes evaluados. Están conformadas por reactivos de opción múltiple cuando incluyen preguntas o problemas que tienen una respuesta correcta y varios distractores, generalmente entre dos y cuatro. Se les llama distractores a las opciones de respuesta que no son correctas pero que representan los errores más comunes al tratar de responder el reactivo, esto es, son respuestas incorrectas plausibles que están relacionadas directamente con lo que se está preguntando.

Procesos de una estrategia de evaluación

En el Centro se utilizan diez procesos para el desarrollo de las pruebas: diseño de la evaluación, delimitación del objeto de medida, construcción del banco de reactivos, verificación cuantitativa, ensamble de cuadernillos y formas, aplicación, procesamiento de la información y calificación, emisión de resultados, mantenimiento del examen y elaboración del material complementario.

Durante el diseño de un instrumento de evaluación se definen las características generales de la estrategia de evaluación que serán las directrices para su desarrollo. Con estas bases se elabora un documento rector denominado *perfil referencial*, el cual concentra las características más importantes del instrumento de medición y sus condiciones de aplicación, calificación e interpretación de los resultados.

En el proceso de delimitación del objeto de medida se elaboran los elementos que constituyen la descripción conceptual y operacional de lo que se pretende medir y se realiza su validación por juicio de expertos. En un marco teórico se plantea la definición

del constructo y la delimitación conceptual de los dominios que lo constituyen. Sobre la base de la racionalidad trazada, se seleccionan los contenidos generales y específicos que se incluirán en el examen. Los contenidos se organizan en una estructura jerárquica que muestra con claridad las áreas, subáreas y temas considerados. Con estos elementos se construyen las especificaciones de los reactivos que constituyen una descripción de cómo deben ser los reactivos. La pertinencia de los contenidos y su congruencia lógica debe ser validada por especialistas.

En el proceso de construcción del banco de reactivos se lleva a cabo la materialización del objeto de medición en los reactivos que conformarán el instrumento de evaluación. Una medida de control en esta fase es que los reactivos elaborados cumplan con estándares de forma y contenido. Para conformar el banco, los reactivos elaborados por los especialistas deben someterse a una revisión efectuada por el personal técnico del Ceneval; posteriormente, otro grupo de expertos realiza la validación del contenido. Para esta distribución de tareas, los especialistas se agrupan en dos comités académicos independientes y participan en talleres de capacitación sobre la elaboración y validación de reactivos.

El objetivo de la verificación cuantitativa es comprobar la calidad psicométrica de los reactivos. El Centro se vale de diferentes procedimientos para poner a prueba los reactivos, con el propósito de calibrarlos y observar su comportamiento estadístico bajo condiciones próximas a la situación real en la que serán resueltos por los sustentantes. Aquí se ubica el piloteo y el análisis de los reactivos que compondrán la prueba. Esto no quiere decir que el análisis de reactivos que aquí se describe pueda aplicarse solamente a evaluaciones educativas y con reactivos de opción múltiple. Exámenes con otros objetivos y características son analizados rutinariamente con esta metodología.

Durante el proceso de ensamble de cuadernillos y formas se integra el examen incorporando las preguntas que formarán parte del instrumento, siguiendo reglas técnicas de diseño para la selección de los reactivos que formarán parte de cada versión, cuidando las características psicométricas del instrumento. Una vez que está conformada, cada versión debe ser revisada editorialmente antes de su impresión en cuadernillos o la colocación de las formas en los servidores, para su aplicación en línea.

En el proceso de aplicación se trata de contar con los procedimientos adecuados para que los sustentantes respondan el instrumento de evaluación en condiciones controladas, estandarizadas, óptimas y equitativas.

En el siguiente paso de la metodología, se procesa la información recolectada durante la fase de aplicación y se realiza la calificación de los sustentantes de manera precisa y libre de errores. Para ello, el Ceneval ha establecido los mecanismos para corroborar los insumos y productos, cuenta con ciclos de autocontrol que alertan sobre la necesidad de auditar la calificación de alguna aplicación.

La entrega de resultados de los exámenes del Ceneval tiene dos vertientes fundamentales: los reportes individuales de los sustentantes y los reportes institucionales. La entrega a las instituciones y la que se hace a la población evaluada deben respetar las políticas de seguridad y confidencialidad para la salvaguarda de la evaluación de cada persona. La información que se genera debe resguardarse para las aclaraciones necesarias, para la elaboración de diversos informes especiales o para su posterior análisis con fines de investigación.

En el proceso de mantenimiento del examen se agrupan diversas actividades que hacen posible el adecuado funcionamiento de cada examen a lo largo de su permanencia operativa.

Finalmente como un proceso paralelo e independiente, pero íntimamente ligado al desarrollo del examen, la elaboración de materiales complementarios es determinante para el éxito de la evaluación, de su validez y operación.

Como puede verse, el proceso de construcción de pruebas consta de varias etapas, las cuales deben estar bien alineadas entre sí para obtener buenos resultados. No hay una etapa más importante que otra. Cada una interactúa con las demás y las complementa.

Todo el proceso de construcción debe ser transparente y se sustenta en la idea de que se requieren verificaciones continuas a lo largo del proceso. Una y otra vez se corrobora que el paso anterior se haya llevado a cabo correctamente. Y una nueva opinión o una forma de análisis adicional son bienvenidas siempre. En este sentido, es importante

mencionar que la estructura y el contenido de los exámenes son determinados con el apoyo de cuerpos colegiados externos. Cada prueba cuenta con un consejo técnico y con comités académicos que proporcionan validez y legitimidad a las pruebas. Hay comités para el diseño de la prueba, otros para la elaboración y la validación de reactivos. Así también comités para el establecimiento de los niveles de desempeño y los puntos de corte.

El análisis de los reactivos es una manera más de comprobar que el trabajo que se ha invertido en la prueba va por buen camino, y otras maneras cobran ese mismo sentido a lo largo del proceso.

La importancia del análisis estadístico de los reactivos radica en que abre una especie de diálogo entre quienes diseñan la prueba y elaboran los reactivos y quienes estudian la evidencia empírica que resulta de aplicar las pruebas en situaciones reales. Llevar a cabo adecuadamente esta forma de análisis aporta información para poder confirmar en la práctica que el comportamiento teórico de las preguntas (aplicadas a un grupo de sustentantes) y del examen en general es el que se espera, y asimismo para identificar problemas potenciales. En este sentido, el análisis de reactivos establece un vínculo entre la teoría y la práctica.

¿Qué se analiza y qué se obtiene?

El insumo para el análisis de reactivos es precisamente el conjunto de las respuestas de cada uno de los sustentantes a las preguntas de una prueba. Los resultados son indicadores de las características psicométricas de cada uno de los reactivos.

A manera de simplificación, lo más común es que se obtengan al menos dos indicadores básicos de las características del reactivo:

- La *dificultad*, es decir, qué tan fácil o difícil resulta el reactivo para la población que tomará el examen. Tradicionalmente, esto se puede medir considerando la proporción de los sustentantes que aciertan a la respuesta.

- La *discriminación*, o qué tan eficiente es el reactivo para ayudar a diferenciar entre quienes saben y los que no saben.

En algunos casos también se busca información sobre el comportamiento del reactivo con respecto a una población específica (funcionamiento diferencial del ítem o DIF, por sus siglas en inglés). Este tipo de análisis es necesario para identificar comportamientos diferenciados en el funcionamiento de los reactivos, no solamente entre hombres y mujeres sino también entre distintos grupos de sustentantes, por ejemplo: grupos étnicos, sustentantes de diferentes regiones del país, etc.

¿Cuándo se hace el análisis estadístico de los reactivos?

El análisis estadístico de los reactivos de una prueba es un control de la calidad de los reactivos. Una vez que se conoce bien lo que se va a medir, el siguiente paso es identificar las preguntas que se comportan conforme a ese objetivo y aquellas que podrían mostrar comportamientos no esperados, sea por deficiencias en la elaboración del propio reactivo o porque simplemente lo que se pregunta no está permitiendo obtener evidencia de la habilidad o conocimiento evaluado. Las técnicas estadísticas del análisis aportan información útil para este efecto y por eso son de uso rutinario en todos los programas de evaluación educativa en el mundo.

Aun cuando en este documento se hace énfasis en el análisis estadístico de los reactivos, a partir del cual se toman decisiones para incluir o no un reactivo determinado en una prueba, en la práctica el análisis puede hacerse en tres contextos distintos, definidos por la aplicación de la prueba y por la entrega de calificaciones a los sustentantes. Estos contextos son:

1) Análisis en un piloteo o prueba de campo previo a la aplicación. Una práctica común de las agencias de evaluación es probar los reactivos antes de usarlos en las pruebas, procurando hacerlo en grupos de sustentantes que sean lo más parecido que se pueda a aquellos que usarán la prueba real (prueba operativa).

En la práctica, tener acceso a un grupo de sustentantes potenciales que quieran participar en número suficiente y con la misma motivación con la que contestarían la prueba real, no es siempre fácil. Por ello, se ofrecen incentivos como puntos extra en algunas asignaturas escolares.

En programas que se aplican constantemente, como los EXANI o los EGEL que desarrolla el Ceneval, la forma más sencilla de obtener condiciones reales para el análisis de los reactivos es generalmente la de incluir algunos reactivos extra en cada cuadernillo para ser piloteados. Así, los sustentantes que enfrenten esos reactivos especiales los contestarán exactamente igual que lo hacen con el resto del examen, sin que ello afecte sus calificaciones.

Lo importante en cualquier caso es obtener información relevante sobre la calidad del reactivo y tenerla en el mejor momento para decidir, antes de que el reactivo se aplique y tenga un impacto en la calificación de los sustentantes.

2) Análisis después de la aplicación, pero antes de que se entreguen calificaciones. Los reactivos pueden haber sido verificados por expertos o en ejercicios piloto previos, y aún así presentar comportamientos inesperados. Por ejemplo, puede ser que alguien se haya equivocado al asignar la respuesta correcta y haya escrito una “A” en donde debió de haber puesto una “B”. De tal manera que al momento de calificar, una sola de estas razones puede significar que un sustentante pierda puntos por una respuesta que debería ser contada como acierto y fue calificada como error o viceversa. El análisis de reactivos en esta fase se convierte en un valioso auxiliar para detectar problemas y corregirlos antes de que las calificaciones se den a conocer. Después de hacer el estudio, tal vez bastará cambiar una “A” por una “B”. Si el problema es más grave y hace falta hacerlo, podría ser necesario eliminar ese reactivo de la versión. De cualquier manera, el análisis de reactivos ofrece una oportunidad más de verificar y detectar errores que pueden ser corregidos a tiempo.

3) Análisis después de dar las calificaciones. Elaborar un examen no es siempre un acto de una sola vez, que tras armar el cuadernillo se olvida. Con la práctica y el ensayo continuos, las pruebas se van mejorando, dejando lo que ha funcionado bien y

eliminando lo que no funciona. Los bancos de reactivos maduran y se fortalecen; los exámenes se perfeccionan. Revisar los reactivos después de que se aplicó una prueba y se dieron las calificaciones sirve para elegir los mejores reactivos que formarán parte de nuevas versiones.

Principales metodologías de análisis

Son diversas las metodologías para hacer el análisis de los reactivos, según las necesidades de la evaluación, las características del estudio y las capacidades del análisis. Las dos corrientes principales en el análisis de reactivos son la Teoría Clásica de los Tests, de amplia tradición, y la más moderna Teoría de Respuesta al Ítem.

En la *Teoría Clásica de los Tests* (TCT) se conjuntan varias técnicas de lo que es, probablemente, el modelo de análisis más conocido en la psicometría. Aunque sus antecedentes se remontan seguramente más atrás, se reconoce que el modelo clásico fue introducido por el psicólogo británico Charles Spearman quien a principios del siglo XX publicó una serie de argumentos matemáticos en los que expuso que los puntajes de una prueba son medidas inexactas de los rasgos humanos (Spearman 1904, 1907, 1913). Cien años después, la TCT ha avanzado y se ha adaptado a las necesidades actuales. Sin embargo, sus principios básicos se siguen manteniendo como la forma de análisis más popular de la evaluación educativa moderna.

La *Teoría de la Respuesta al Ítem* (TRI) es en realidad un conjunto de modelos matemáticos que buscan describir el comportamiento de los reactivos. A diferencia de la TCT, el eje sobre el que gira la teoría no es la prueba sino el reactivo propiamente dicho. Los modelos de la TRI tratan de descubrir qué relación existe entre la habilidad del sustentante y la probabilidad de contestar correctamente el reactivo que enfrenta, y de describirla a partir de un modelo matemático.

Los modelos básicos de la TRI se desarrollan a partir de cuestionarios con preguntas de respuestas dicotómicas (acierto/error, verdadero/falso, correcto/incorrecto). Hay extensiones de la TRI para otros tipos de variables, ordinales o politómicas, que son utilizadas rutinariamente en los casos en que se utilizan reactivos con respuestas

graduadas. Van der Linden y Hambleton (1997) realizan una exposición detallada de varios de estos modelos.

Los antecedentes del modelo básico pueden trazarse al menos hasta 1927 con el trabajo de Thurstone sobre percepción del estímulo. En las décadas siguientes, y después de otros posibles candidatos para la distinción de haber sido los primeros en penetrar las bases del modelo, parece haber consenso en que los progresos más importantes en términos teóricos y prácticos fueron los de George Rasch, Frederick Lord y Allan Birnbaum, desarrollados entre las décadas de los cincuenta y los sesenta (Van der Linden y Hambleton, 1997).

El llamado modelo de Rasch fue desarrollado por el danés Georg Rasch durante los años cincuenta. Si bien para algunos es un modelo más de los que conforman la TRI, sus adherentes consideran que sus características técnicas y desarrollo histórico lo convierten en un modelo original y que requiere tratamiento por separado. Ciertamente el modelo y la metodología que lo acompaña efectivamente tienen una tradición propia, aunque en términos prácticos comparte muchas de sus características con los otros modelos de la TRI.

El presente trabajo se centrará en los indicadores más populares de la TCT para el análisis del comportamiento estadístico de los reactivos y de la prueba en su conjunto, sin embargo, se realizará una explicación general de los modelos de la TRI y de su interpretación.

Capítulo 3. Pasos para el análisis de reactivos

Los pasos para el análisis cuantitativo de una prueba nueva son:

1. Preparativos previos para el análisis
2. Recolección de la información
3. Selección del modelo de análisis
4. Validación de los insumos para el análisis
5. Análisis de distractores
6. Análisis de sesgo
7. Selección de reactivos para la conformación de la prueba

Antes de comenzar con la descripción de los pasos se hará hincapié en que las mejores prácticas sobre la materia insisten en la necesidad de documentar adecuadamente los procesos. Los estándares de la American Psychological Association (APA), la American Educational Research Association (AERA) y el National Council on Measurement in Education (NCME) –tal vez los más conocidos en el área– plantean el asunto directamente y con claridad:

> *"Se deben documentar los procedimientos utilizados para desarrollar, revisar y probar los ítems y para seleccionar ítems del banco de reactivos. Si los ítems fueron clasificados en diferentes categorías o subconjuntos de acuerdo con las especificaciones de la prueba, los procedimientos utilizados para la clasificación y la pertinencia y precisión de la clasificación también deben ser documentados."* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999)

Paso 1. Preparativos previos para el análisis

Antes de comenzar con el análisis de reactivos es necesario corroborar que se han llevado a cabo adecuadamente los pasos previos al piloteo. Esto incluye el diseño mismo de la prueba, la definición de su estructura y, no menos importante, la elaboración de los reactivos y su revisión. De cada una de las etapas deben existir elementos que tienen que verificarse antes de proceder al análisis.

Paso 2. Recolección de la información

El siguiente paso para el análisis cuantitativo de los reactivos consiste en recabar la información necesaria (básicamente de las respuestas de los sustentantes). En el Ceneval generalmente se trabaja con un archivo de texto plano (*.DAT) que contiene estos datos. En cada renglón aparecen el folio del sustentante, la versión del examen[2] que respondió y las respuestas que eligió, cuando se trata de exámenes con reactivos de opción múltiple. Para recabar este tipo de información se deben tomar en cuenta los siguientes elementos:

a) **Definición del objetivo y alcances.** Se reconocen varios tipos de ejercicios que varían en tamaño y alcance. En ocasiones, la verificación estadística del comportamiento de los reactivos se ve precedida por uno o más ejercicios de menores dimensiones que permiten ir depurando el conjunto de reactivos a partir de criterios cualitativos, para luego avanzar hacia estudios de un marcado carácter estadístico.

   En las etapas iniciales de la comprobación de las cualidades de los reactivos es común que se sometan a ejercicios de dimensiones reducidas, quizá entre 10 y 30 personas, con los que se busca hacer un diagnóstico general, suficiente para detectar los errores más evidentes, como pueden ser la falta de respuesta correcta, problemas en la redacción de la pregunta o valores extremos en los

[2] Cuando una prueba es aplicada en repetidas ocasiones, es necesario elaborar pruebas diferentes que midan el mismo constructo, para evitar la copia o el plagio de las preguntas de una aplicación a otra. A cada una de las colecciones distintas de reactivos se le denomina “versión de examen”.

estadísticos del reactivo. Aunque el enfoque del ejercicio puede variar, casi desde un análisis de tipo “grupo de enfoque”, con esta información se pueden identificar los reactivos notoriamente inadecuados y corregir las faltas evidentes.

En realidad, cuando el impacto del examen se estima bajo –como en el caso del diagnóstico que hace el profesor a sus alumnos en la clase–, un estudio de tan baja escala puede ser suficiente al menos para darse una idea de cómo están los alumnos.

Sin embargo, el “piloteo” propiamente dicho, es un ejercicio más exigente en términos del tamaño y representatividad de las muestras: a partir de él se determinan las características métricas definitivas con las que los reactivos ingresarán al banco operativo, es decir, al repositorio del cual pueden ser tomados para formar parte de la prueba definitiva. Esto es especialmente cierto en instituciones como el Ceneval, en donde los exámenes pueden tener un alto impacto y se espera que las diferentes versiones se comporten con estabilidad.

b) **Diseño de la muestra.** Debido a que el principal objetivo del estudio piloto es identificar las características psicométricas de los reactivos que pueden servir para nutrir las pruebas, resulta indispensable asegurar que la muestra elegida de sustentantes sea representativa de la población objetivo, es decir, de la población a la que se dirige el examen o la que normalmente lo contesta. La representatividad de la muestra está determinada por ciertas características que identifican a esos sustentantes como parte de la población objetivo, aunque otros factores también intervienen, por ejemplo, el número de sustentantes que participan en un estudio piloto y su motivación para resolver los reactivos que se les presenten.

Siempre debe procurarse que las pruebas piloto y los análisis se realicen en ejercicios bien controlados, que respeten las condiciones normales de aplicación y que se hagan con sujetos que representen lo más fielmente posible a la población objetivo. Idealmente debe utilizarse el muestreo probabilístico para garantizar la representatividad, por ejemplo, si se tiene un conocimiento preciso

de la población objetivo a la que está dirigida la prueba, podría obtenerse una muestra estratificada de los sustentantes de la prueba piloto, que tenga una distribución similar a esta población objetivo. Sin embargo, en los casos en los que las condiciones no lo permitan, será suficiente emplear una estrategia de muestreo intencional o de conveniencia que refleje la heterogeneidad de la población objetivo del examen. Es importante que se distingan las aplicaciones de alto impacto para la institución y el sustentante de aquellas con otros fines. Lo mismo deben evitarse las aplicaciones que en ocasiones se hacen, por ejemplo, a los maestros de quienes serán sustentantes y a todo tipo de poblaciones que pudieran ser distintas a aquellas a las que se dirige el examen.

El mejor escenario para el estudio piloto consiste en incluir los reactivos que se quieren probar en versiones finales aplicables en condiciones normales de operación. De esta forma se garantiza, en buena medida, que se trata de la misma población que utiliza el examen y que los individuos están bien motivados para contestarlo.

c) ***Determinación del tamaño de la muestra.*** El número mínimo de sustentantes en una muestra es difícil de determinar para todos los casos, ya que depende de varios factores, entre los que destacan las características de la población involucrada, la homogeneidad del constructo que se mide, el número de reactivos incluidos en el estudio y su calidad. Y el objetivo que se busca con el análisis también es un factor importante.

En la tabla 1 aparecen los tamaños de muestra mínimos sugeridos para el cálculo de estadísticos estables de la TCT y de los parámetros de los reactivos en los tres modelos de la TRI descritos más adelante (Downing y Haladyna, 2006). Los tamaños de muestra presentados en la tabla 1 son propuestas que, de acuerdo con la experiencia de diversos autores, han sido adecuados para los diferentes tipos de análisis. Es recomendable que para cada programa de evaluación se realicen los estudios necesarios para obtener las conclusiones particulares sobre los errores de estimación de los estadísticos de los reactivos y de la prueba.

Tabla 1. Propuesta de tamaños de muestra mínimos para distintos modelos de análisis de reactivos

Modelo		Núm. de sustentantes requerido
Teoría Clásica de los Tests		Alrededor de 100
Teoría de la Respuesta al Ítem	1 parámetro (Rasch)	100
	2 parámetros	250 – 500
	3 parámetros	Más de 1000

Siempre será deseable contar con el mayor número de sustentantes posible, invariablemente por arriba de los valores indicados en la tabla 1.

d) **Diseño de las versiones para piloteo.** En programas que se aplican constantemente, como los EXANI o los EGEL del Ceneval, la forma más sencilla de obtener condiciones reales para el análisis de los reactivos es generalmente la de incluir algunos reactivos extra en cada cuadernillo para ser piloteados. De este modo, el sustentante los resolverá sin que pueda diferenciarlos de los que sí se calificarán y los resolverá con toda honestidad, como haría con los otros.

En ocasiones sucede que los reactivos que van a ser probados se entregan al sustentante en un cuadernillo por separado. Esto –que muchas veces se justifica por restricciones operativas– tiene la desventaja de que podría llegar a cambiar la actitud del sustentante al enfrentarse a un cuadernillo diferente, separado de la prueba principal, sobre todo si por alguna razón puede sospechar que esas preguntas son de piloteo y no influirán en su calificación.

Cuando se trate de aplicaciones exclusivas para piloteo deben buscarse condiciones e incentivos que estimulen a los sustentantes a hacer su mejor esfuerzo. Además, deben construirse versiones cortas, con el número de reactivos que puedan contestarse en una sesión breve, incluyendo el tiempo de preparación y la lectura de las instrucciones.

Las versiones que contienen exclusivamente reactivos piloto pueden seguir la estructura original del examen o adoptar una estructura distinta. Seguir de cerca la estructura de la prueba es recomendable cuando lo que se busca es analizar las características del conjunto, como la confiabilidad de la prueba. Elaborar versiones con estructuras distintas a las establecidas es especialmente recomendable para facilitar la administración cuando se trata de exámenes largos y para aquellos que cuenten con secciones o áreas de pocos reactivos. Como desventaja, debe tenerse en cuenta que este método puede requerir de más cuadernillos y, por lo tanto, complicar la aplicación, además de que implica contar con más sustentantes y una mayor participación de instituciones para poder probar todos los reactivos. A manera de ejemplo, en el *anexo 1* se expone un esquema en el que se distribuyen reactivos de todos los temas contemplados en la estructura del examen, a lo largo de la muestra de piloteo (*Diseño matricial*).

e) **Número de reactivos a pilotear**. Se debe tomar en cuenta que en ocasiones se pierde hasta la mitad de los reactivos, al determinarse que no cumplen con las características psicométricas requeridas. Por lo tanto es recomendable que los reactivos que se sometan a piloteo hayan pasado por alguna forma de validación previa, de modo que no se desperdicien recursos en el análisis y calibración de reactivos de mala calidad y que al mezclarse con el resto de los reactivos en el análisis puedan incluso llegar a distorsionar las medidas.

   Para obtener una calibración adecuada es de esperar que las versiones cuenten con un número suficiente de reactivos que tengan un comportamiento estadístico adecuado como para mantener una buena referencia del constructo por medir; se recomienda realizar el análisis con al menos 20 reactivos con parámetros estadísticos “aceptables”. Es verdad que pueden usarse menos reactivos si hace falta. Sin embargo, mientras más sean estos reactivos, mejores y más estables serán los resultados del análisis.

f) **Número de sesiones de aplicación.** Las pruebas piloto pueden extenderse en varias sesiones, especialmente si se cuenta con pocos sustentantes o si lo que se busca es mantener las sesiones de aplicación en tiempos razonables.

g) **Condiciones de aplicación, capacitación a aplicadores e instrucciones para los sustentantes.** En un estudio piloto deben conocerse con anticipación las condiciones en que habrá de llevarse a cabo la aplicación y ajustar las acciones operativas en concordancia.

Paso 3. Selección del modelo de análisis

Con el objetivo de establecer criterios que ayuden a seleccionar el modelo para el análisis de los reactivos piloteados, a continuación se enlistan algunas de las ventajas y desventajas de la Teoría Clásica de los Tests, con relación a los modelos de la TRI.

Ventajas de la Teoría Clásica

La Teoría Clásica tiene las siguientes ventajas sobre los modelos de la Teoría de la Respuesta al Ítem (TRI):

1) Es más fácil de utilizar. El análisis puede hacerse incluso sin usar software especializado. Una calculadora o una hoja de cálculo son muchas veces todo lo que se necesita. Y para calificar, por lo general basta con contar las respuestas correctas.

2) Se explica con mayor sencillez a un público no especializado pues sus principios básicos son fácilmente comprensibles para la gran mayoría. Cualquiera entiende cuando se habla de proporción de sustentantes que respondieron correctamente y, a la hora de calificar, el número de aciertos puede interpretarse sin recurrir a fórmulas sofisticadas.

3) Para una estimación precisa de los parámetros se requieren muestras modestas, incluso de menos de 100 sustentantes. Y aún con muestras menores se puede

utilizar la Teoría Clásica como indicador aproximado de la calidad y características generales de los reactivos.

4) Cuando se trabaja con poblaciones relativamente homogéneas y con un número reducido de versiones, los resultados son suficientemente precisos para tomar decisiones.

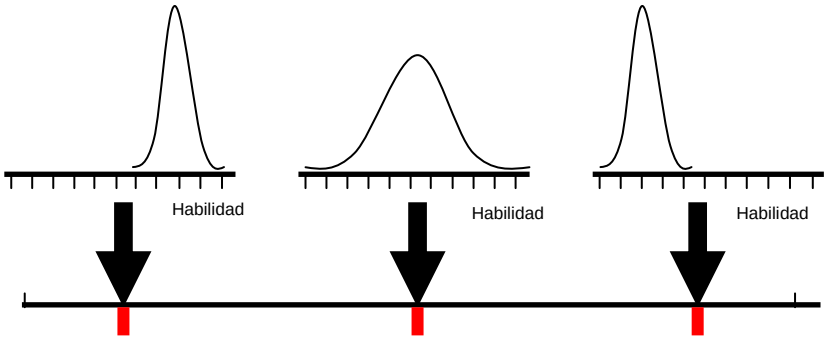
5) Es una metodología flexible, con supuestos menos rígidos. Esto quiere decir que es razonablemente apropiada incluso cuando el contenido de las áreas de la prueba es moderadamente heterogéneo y cuando los supuestos de la TRI, siempre más exigentes, no se cumplen.

Desventajas de la TCT

1) Probablemente la mayor desventaja de la TCT puede resumirse como una "dependencia circular": los estadísticos del sustentante (por ejemplo su puntuación en la prueba) son dependientes de los reactivos que le tocan en suerte en esa prueba en particular, mientras que los estadísticos de los reactivos dependen de los sustentantes que contestaron en esa ocasión.

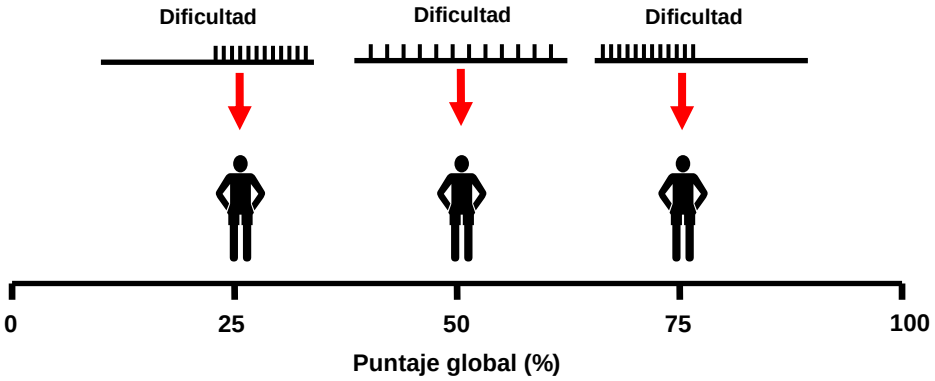
   Como se muestra en la figura 1, la dificultad estimada para los reactivos varía de acuerdo con la habilidad del grupo de sustentantes que se incluya en la muestra. Si los sustentantes fueron hábiles en general, los reactivos resultarán fáciles. Sin embargo, si sucede que los sustentantes no fueron tan hábiles, se tendrá el resultado contrario, y los mismos reactivos parecerán más difíciles.

Figura 1. La dificultad del reactivo depende de la distribución de la habilidad de los sustentantes con los que se calcule



La habilidad de los sustentantes –estimada a partir del puntaje observado– depende de la dificultad de los reactivos que componen la prueba. Así, un mismo sustentante obtendrá puntuaciones diferentes en tres versiones distintas de la prueba si éstas difieren en la distribución de dificultad de los reactivos que las componen. En la figura 2, el mismo sustentante puede ubicarse en tres calificaciones distintas (25, 50 y 75 por ciento de aciertos) dependiendo de si la dificultad de los reactivos es alta, media o baja.

Figura 2. Los puntajes observados de los sustentantes dependen de la distribución de dificultad de los reactivos de la prueba



2) Otro problema de la TCT es que supone que el error estándar de medida es igual a lo largo de toda la escala de habilidad. Esto no es necesariamente válido. Una prueba puede ser más precisa en algunos rangos de puntación que en otros.

Esto depende de la distribución de la dificultad de los reactivos de la prueba de que se trate. Metodologías más modernas (como la TRI) pueden aprovechar el conocimiento sobre el error para elaborar pruebas en las que se controle mejor en las regiones de puntuación que más interesen.

Ventajas de la Teoría de Respuesta al Ítem (TRI)

Los supuestos sobre los que se funda la TRI son más estrictos y, si se cumplen adecuadamente, le dan a la metodología una flexibilidad y un poder que difícilmente se logra con la TCT. En condiciones apropiadas los modelos de la TRI superan algunos de los problemas de la teoría clásica. En particular, hay tres importantes ventajas de la TRI sobre la TCT:

1) Invarianza de grupo: Los parámetros de los reactivos (por ejemplo: dificultad y discriminación) son independientes del grupo particular de sustentantes utilizados para su estimación. Las muestras no requieren ser tan representativas u homogéneas como en la TCT para obtener una representación matemática del reactivo que permita modelar su comportamiento estadístico.

2) Invarianza de ítem: La estimación de la habilidad de los sustentantes es independiente del conjunto de reactivos con los que se estime. Esto está relacionado con el punto anterior y en cierta forma resuelve el problema de la dependencia circular de la TCT.

3) Es posible estimar el error asociado a cada nivel de habilidad, en lugar de estimar un solo error estándar de medida para todo el rango, como es el caso en la TCT.

Desventajas de la TRI

1) Es relativamente más compleja, lo que la hace más difícil de aplicar en la práctica y de explicar a los sustentantes. Requiere generalmente de software especial para hacer el análisis y la calificación.

2) El número de sustentantes deseable para una calibración estable de los reactivos es mayor. Si bien en el modelo de un parámetro las muestras necesarias son similares a las de la TCT, cuando se usa el modelo de dos parámetros las condiciones comienzan a ser significativamente más exigentes. La literatura habla de entre 250 y 500 sujetos para obtener estimaciones aceptables, dependiendo de las condiciones de aplicación y las características de la prueba (Downing y Haladyna, 2006). Y cuando se usa el modelo de tres parámetros, las necesidades son mucho mayores, con mínimos de entre 1,000 y 1,500. Y aun con muestras sustancialmente mayores es posible encontrar problemas de indeterminación en el cálculo del segundo y el tercer parámetro, especialmente cuando se trata de contenidos incluso moderadamente heterogéneos, como sucede corrientemente en algunos instrumentos de evaluación de conocimientos.

3) Funciona bien si los supuestos se cumplen. La TRI no es siempre apropiada para todos los casos y para todos los exámenes. Es una técnica bastante flexible y el cumplimiento de los supuestos que la sustentan tiene suficiente tolerancia como para adaptarse a muchos casos. Sin embargo, no es para todos los exámenes, no lo es especialmente para aquellos con contenidos complejos.

TRI o Clásica. ¿Cuál elegir?

La elección del modelo depende de varias circunstancias, teóricas y prácticas, aunque es común que se usen ambas metodologías al mismo tiempo, ya que no son del todo excluyentes.

Por simplicidad, el modelo de la TCT es la ruta a seguir en primera instancia. Es más sencillo y fácil de explicar, y los resultados son muchas veces similares a los que se logran con modelos más sofisticados. La prudencia aconseja mantenerse con la TCT hasta que las circunstancias justifiquen acudir a otra metodología.

La TCT no ha permanecido estática. Los estudiosos y practicantes de la evaluación han desarrollado soluciones prácticas a algunas de las limitaciones teóricas del modelo. Un

ejemplo de ello es la evolución de métodos para equiparar[3] versiones distintas de una prueba en diferentes situaciones de aplicación, como la equiparación lineal o la equipercentil.

A pesar de la popularidad y la eficacia de la TCT, la TRI tiene un papel cada vez más importante en el mundo de la evaluación. Una razón para acudir a modelos de respuesta al ítem puede ser simplemente para complementar el análisis y obtener más información sobre los reactivos, aunque no se usen todas las ventajas del modelo. Pero esas ventajas son las que verdaderamente pueden llegar a justificar el esfuerzo adicional. La TRI es recomendable cuando se esperan diferencias importantes de habilidad entre las poblaciones en que se usará la prueba. Se pueden hacer versiones fáciles para unos y más difíciles para otros, por ejemplo.

En ciertas circunstancias, la TRI es indispensable. Un ejemplo: los grandes exámenes internacionales, en que se aplican muchas versiones distintas entre sí o se utilizan fragmentos de los contenidos de una prueba mayor y aun así se requiere que los resultados sean comparables. Y, por supuesto, es la metodología que hay que seguir en los exámenes adaptativos por computadora, en que los reactivos por utilizar se van seleccionando conforme el sustentante va resolviendo su prueba, esto es, la prueba de cada sustentante se construye secuencialmente, y cada reactivo se selecciona según el desempeño en el conjunto de reactivos de la prueba que se ha administrado hasta ese momento; el algoritmo de aplicación se detiene cuando se ha estimado la habilidad del sustentante con la precisión deseada.

La TRI se convierte por lo tanto en una herramienta invaluable para mantener distintas versiones de la prueba, combinando los reactivos de maneras distintas –versiones más cortas para ahorrar reactivos o tal vez versiones más fáciles o más difíciles para adaptarse a la población que habrá de contestarlas– y aun así obtener una estimación comparable de las habilidades.

[3] Con la finalidad de mantener la equidad de la evaluación, la igualación o equiaración de versiones se refiere al conjunto de técnicas estadísticas enfocadas a corregir diferencias en la dificultad de las diferentes versiones del examen.

Paso 4. Validación de los insumos para el análisis

El análisis de reactivos de opción múltiple en la mayor parte de los exámenes del Ceneval requiere de los siguientes insumos:

- Estructura del examen. Documento que describa la prueba, con el número de reactivos que corresponde a cada área del examen, si hay varias áreas, especificando a qué niveles se ofrece calificación individual.
- Formato de versiones que describa:
    1. Clave de identificación de los reactivos en el banco
    2. El área a la que pertenece el reactivo
    3. La posición que ocupan los reactivos en cada versión
    4. El estatus del reactivo (si es activo y se calificará o si es piloto)
    5. La clave de respuesta correcta para cada uno de los reactivos.
- Archivo de texto con las respuestas de los sustentantes (cadenas de respuestas).
- Cualquier otra fuente de información que pueda ayudar a conocer las características de la aplicación o de la población que se presentó. Entre la información más importante para recabar están las fechas de aplicación de las versiones que se analizan, las instituciones participantes en la aplicación, las sedes de aplicación y el número de sustentantes. Se recomienda registrar información sobre las características de la aplicación y de la población examinada; por ejemplo, si se trató de un ejercicio diagnóstico, de auto-aplicación, si fue una aplicación especial, etcétera.

Antes de proceder al análisis, se validarán todos los insumos para verificar que se cumpla con los siguientes elementos:

- El archivo de la estructura del examen debe corresponder con las versiones aplicadas.

- El formato de versiones debe contener todos los datos requeridos. En particular, se deben verificar las respuestas correctas indicadas en el archivo.
- En cuanto a la validación de respuestas de los sustentantes, se debe verificar que en el archivo no haya sustentantes duplicados y que las cadenas de respuestas tengan sólo caracteres válidos (por ejemplo, en las pruebas de opción múltiple del Ceneval, solamente se permiten las letras A, B, C, D, el espacio para indicar que no fue contestada la pregunta y el asterisco que indica respuesta múltiple).
- Para efectos del análisis, se recomienda descartar del archivo las cadenas de los sustentantes con pocas respuestas. El porcentaje de tolerancia puede variar, pero es de presumir que los sustentantes con más del 30% de las preguntas sin respuesta no terminaron la prueba o que no estaban suficientemente motivados.
- Se verificará y se registrará la proporción de respuestas obtenida por cada reactivo. Si la proporción de sustentantes que no respondieron algún reactivo en particular es alta en relación con el resto de los reactivos, vale la pena explorar las posibles razones. En la mayor parte de las pruebas del Ceneval las tasas de respuesta por reactivo siempre son muy superiores al 90%, así que, sin excepción, sería sospechoso un reactivo con una tasa de respuesta menor. Por supuesto, el porcentaje varía en cada examen y circunstancia de aplicación.
- Si se muestra una disminución generalizada en la tasa de respuesta de los sustentantes hacia el final de su examen –en comparación con el resto de las preguntas de la prueba–, esto puede indicar que la longitud del examen podría ser excesiva. Así, por ejemplo, si el 95% de los sustentantes responde las primeras secciones, pero es el 80% el que responde las últimas, debe sospecharse que el examen es muy largo o que el tiempo es demasiado corto.
- Los insumos verificados y los productos del análisis se deben conservar como evidencias para verificaciones o análisis posteriores.

Paso 5. Análisis de distractores

El análisis de los distractores ayuda a conocer mejor las preguntas y a explicarnos por qué un reactivo se comporta como se comporta. También es un insumo de gran interés para quien quiere conocer el estado del conocimiento en los sustentantes: ¿dónde se equivocan más?, ¿por qué?, ¿qué tipo de conocimientos les causan confusión?, ¿cuáles necesitan profundizarse más?

Los distractores son opciones de respuesta incorrectas plausibles, ya que incluyen los errores más comunes de los sustentantes al enfrentarse a la pregunta planteada. Entre los lineamientos técnicos que permiten la elaboración de un buen grupo de distractores destacan que: pertenecen al mismo tema y campo semántico, tienen el mismo nivel de generalidad o especificidad, ninguno destaca por su extensión respecto al resto ni respecto a la respuesta correcta y son distintos entre sí.

El estudio de los distractores es particularmente útil cuando los análisis previos dan evidencia de que un reactivo no se comporta como se espera. Cuando se han detectado posibles problemas en el funcionamiento estadístico de algún reactivo, el siguiente paso consiste en tratar de averiguar los motivos de estas anomalías, para corregir el reactivo o en su caso desecharlo definitivamente.

Muchas veces presentado de manera gráfica, este análisis permite determinar el patrón de comportamiento tanto de la respuesta correcta como de los distractores del reactivo a lo largo de la escala de puntajes globales.

Paso 6. Análisis de funcionamiento diferencial de los reactivos

La igualdad de condiciones para todos los sustentantes de una prueba es uno de los principios necesarios para garantizar la validez de la evaluación. Entre las tareas que asume el Ceneval para lograr la equidad entre los sustentantes, está la creación de versiones equivalentes de examen y la estandarización de las condiciones de aplicación de las pruebas. Sin embargo, en ocasiones, un reactivo es particularmente más fácil (o

más difícil) para algún grupo de sustentantes (por ejemplo, para las mujeres en comparación con los hombres).

Se dice que un reactivo presenta *funcionamiento diferencial* cuando resulta más fácil (o difícil) para un grupo de sustentantes que para otro, una vez que se han tomado en cuenta las diferencias en el conocimiento o habilidad de los grupos en el rasgo que se está evaluando. Cuando un grupo de sustentantes presenta una mayor habilidad promedio que otro, tenderá a mostrar un mejor rendimiento en todos los reactivos de la prueba. Sin embargo, un reactivo con funcionamiento diferencial mostrará una brecha significativamente más grande que la atribuible a la diferencia en habilidad de los grupos.

Generalmente el estudio del funcionamiento diferencial se enfoca a detectar sesgos de los reactivos para grupos con características distintas, como pueden ser, entre muchas otras, el género o el nivel socioeconómico.

Paso 7. Selección de reactivos para la conformación de la prueba

El primer requisito de selección para los reactivos es que cumplan con los contenidos y las especificaciones establecidos para la prueba por los diseñadores. El respeto por la estructura temática y las especificaciones de los reactivos es fundamental. Ante ellas, los datos estadísticos del instrumento pasan siempre a un segundo plano. De acuerdo con lo anterior, en muchos programas es común que se construyan y prueben dos o tres veces el número de reactivos por especificación, para poder elegir siempre el mejor.

A continuación se describen algunos de los índices más utilizados para el análisis de reactivos mediante la TCT y la TRI y los criterios de selección de reactivos comúnmente utilizados. Se refieren básicamente a la dificultad del reactivo y a lo que se le denomina discriminación. En el siguiente capítulo se explicará con mayor detalle la racionalidad y los algoritmos de cálculo de cada uno de estos indicadores.

Índices de la Teoría Clásica de los Tests

Grado Dificultad

El grado de dificultad se define como el porcentaje de sustentantes que contestaron correctamente el reactivo. En la Teoría Clásica se espera que los reactivos muestren un grado de dificultad medio. Técnicamente, justo en el 50% de grado de dificultad, la varianza de las respuestas al reactivo (ceros y unos) alcanza su máximo[4] y precisamente ahí es donde el reactivo puede diferenciar más entre los sustentantes que saben de los que no saben.

Los valores alejados de una dificultad media no son recomendables. Los reactivos con un grado de dificultad extremadamente bajo o extremadamente alto no aportan información a la medición de los sustentantes; en otras palabras, de poco sirven los reactivos que todos contestan y aún menos los que no contesta nadie. Por lo tanto, una regla generalmente utilizada para la selección de reactivos para una prueba consiste en aceptar reactivos que estén cerca de la mitad de la escala.

Estas "dificultades óptimas" ayudan a diferenciar a un grupo mayor de sustentantes; sin embargo, el problema con este arreglo es que no permite conocer las diferencias entre sustentantes que se encuentran en las colas de la distribución de la habilidad que se está evaluando (sustentantes de habilidad alta o sustentantes de habilidad baja). Muchas veces es recomendable distribuir las dificultades en un rango más amplio para conocer las diferencias a todo lo largo del espectro. Algunos programas de evaluación procuran que las dificultades de los reactivos de sus pruebas se distribuyan de manera más o menos uniforme a lo largo del rango más amplio posible. Este es el caso de varios de los exámenes del Ceneval, que distribuyen su grado de dificultad entre 20% y 80%. Otros programas tratan de que los reactivos de sus exámenes se distribuyan en el rango, pero que se acumulen más hacia el centro o alguna zona predeterminada, tratando de que haya más reactivos precisamente en la región en la que se acumulan más sustentantes. Los exámenes ACT (American College Testing) y SAT (Scholastic Aptitude Test), desarrollados en Estados Unidos, por ejemplo, distribuyen sus reactivos

[4] Al tratarse de ensayos de una distribución Bernoulli (p), cuya varianza (p*q) se maximiza en p = 0.5.

a manera de campana, asignando un número determinado de reactivos a cada fragmento del rango y acumulando más preguntas hacia el centro de la distribución y menos en los extremos. Como quiera que sea, la selección de los reactivos dependerá necesariamente de ese diseño requerido.

Discriminación

Cuando se aplica una prueba, lo que se espera es que podamos diferenciar adecuadamente entre los sustentantes que saben sobre el tema que se les pregunta y los que no. Por lo tanto, un indicador importante en la selección de reactivos debe ser algún índice que permita determinar si un reactivo *discrimina* de manera efectiva entre ambos tipos de sustentantes.

Hay varios indicadores de discriminación. Algunos se desarrollan con sus fórmulas en el siguiente capítulo. Los más comunes son algunos tipos de correlación entre el ítem y la calificación total. Esta correlación puede medirse con la fórmula de Pearson (correlación punto biserial en el caso de reactivos dicotómicos) o con la correlación biserial. En ambos casos se favorece la noción de que el constructo es unidimensional y que se mide aproximadamente de manera homogénea a todo lo largo de la escala. Cuando esto no es así –porque se midan varias áreas al mismo tiempo, o porque el constructo no es homogéneo o no interesa la continuidad a todo lo largo del rango sino en algunos puntos concretos–, pueden usarse preferentemente otras medidas, como la correlación entre el ítem y algún criterio externo.

Casi todos los índices pueden tomar valores entre -1 y 1. Los valores positivos indican que los sustentantes de alto desempeño tienden a elegir la opción correcta más frecuentemente que los sustentantes de bajo desempeño, mientras que se observan valores negativos cuando ocurre lo contrario. Índices más altos de discriminación se refieren a un reactivo que se ajusta mejor al objetivo que se busca.

Basado en estudios empíricos, Ebel (1965, citado en Crocker y Algina, 1986, p. 315) estableció algunos criterios para la interpretación del índice de discriminación (ver tabla 2).

Tabla 2. Criterios para la selección de reactivos por el índice de discriminación

1.	Si $ID \geq 0.40$	El reactivo discrimina muy bien
2.	Si $0.30 \leq ID < 0.40$	La discriminación del reactivo podría ser considerada como satisfactoria
3.	Si $0.20 \leq ID < 0.30$	El reactivo es marginal y requiere revisión
4.	Si $ID < 0.20$	El reactivo debiera ser eliminado o modificado completamente

Como todo criterio estadístico, estos deben ser tratados con cuidado. El límite inferior de 0.20 es bajo, aunque aceptable, para todos los indicadores que se enlistan. En caso de considerarse necesario –para completar los contenidos de la estructura, por ejemplo–, pueden usarse algunos reactivos con discriminaciones más bajas, siempre que no sean negativas. La discriminación negativa debe evitarse siempre que sea posible.

Aunque más discriminación siempre es mejor, debe también sospecharse de correlaciones altas, por arriba de 0.80 por ejemplo. En estos casos es probable que los reactivos sean redundantes en lo que se pregunta, o poco diversos. Posiblemente podrían obviarse algunos de estos reactivos en la prueba sin que se pierda información importante.

Hacer el cálculo de la discriminación y descartar automáticamente los reactivos que no cumplan con el requisito puede tener consecuencias indeseables: llevarnos a descartar más reactivos de los necesarios o afectar la estructura factorial del constructo. Con esta estrategia se ignora la posibilidad de que unos pocos reactivos con muy mal ajuste puedan afectar los coeficientes de los demás.

Las técnicas que utilizan la calificación total como referencia de la habilidad del sustentante son sensibles a que existan reactivos malos, varios de los cuales pueden desvirtuar la medida y mermar su calidad, haciéndola menos apropiada para utilizarse como un referente óptimo. Si se tienen muchos reactivos aceptables y muy pocos o casi ningún reactivo defectuoso, podría decirse que el problema es insignificante. Sin embargo, al aumentar las proporciones de reactivos malos siempre será más recomendable seguir un procedimiento más parsimonioso, quitando un reactivo a la vez

o sólo unos pocos, los más notoriamente defectuosos, y repetir el procedimiento cuantas veces sea necesario.

El trabajo para descartar reactivos de una prueba no es exclusivamente estadístico. Sin tratar de demeritar las técnicas y los procedimientos, a veces puede decirse que es un proceso artesanal. Por ello, siempre se puede aprovechar el conocimiento del contenido de los reactivos y combinarlos con éstas y otras funciones estadísticas, que pueden incluir, por ejemplo, medidas de consistencia interna como el Alfa de Cronbach (Cronbach, 1951).

Teoría de Respuesta al Ítem

El análisis de reactivos a partir de los modelos de la TRI representa una oportunidad de allegarse información que permita hacer inferencias sobre la verdadera habilidad del sustentante. Sin embargo, en la práctica, el reactivo se convierte en una abstracción, una fórmula matemática con la que "modela" su comportamiento estadístico. Por ello, el reactivo debe cumplir con al menos los siguientes requisitos:

- Que el modelo utilizado muestre un ajuste correcto a los datos, de modo que pueda decirse que la fórmula elegida para modelar el comportamiento de los reactivos refleja adecuadamente la realidad de la prueba. Se trata de un requisito básico. Si la abstracción no se parece a la realidad, de nada servirá entonces construir sobre ella.

  Existen múltiples criterios para evaluar el ajuste. Karabatsos (2002) discute 36 indicadores de ajuste de las respuestas de los individuos distintos y, ciertamente, no son todos. Cualquier criterio de ajuste que se use tiene ventajas y desventajas.

  Para aplicar los modelos de la TRI, uno de los programas computacionales de uso más común es BILOG-MG (Zimowski *et al*., 2003). Este programa utiliza una prueba ji-cuadrada para medir la bondad del ajuste entre las respuestas al reactivo y el modelo elegido para el análisis. Por su parte, en el caso del modelo de Rasch, los programas más comunes utilizan dos indicadores de ajuste denominados "Infit" y "Outfit". Para efectos prácticos, generalmente se esperan las tolerancias mostradas en la tabla 3 para considerar que el ajuste es adecuado.

Tabla 3. Criterios para la selección de reactivos de acuerdo al ajuste del modelo de la TRI a los datos

TRI - Bilog	p-valor de la prueba de bondad de ajuste (χ^2) mayor a 0.05
Rasch	0.7 ≤ Infit ≤ 1.3 0.7 ≤ Outfit ≤ 1.3

- Que el reactivo aporte información suficiente para los objetivos deseados y al diseño estadístico planteado para el instrumento. Cada prueba se diseña para un objetivo particular y –como parte integrante de la prueba– cada reactivo debe aportar algo. Por esta razón, generalmente no son adecuados los reactivos con dificultades muy extremas o discriminaciones muy bajas. En la tabla 4 se proponen límites para los parámetros de los reactivos en la TRI. Cabe mencionar que en caso de la TRI la escala de dificultad de los reactivos depende de cierto marco de referencia (Baker, 2001), por lo que la propuesta de la tabla 4 tiene sentido cuando la escala se presenta estandarizada y no sufre alguna transformación lineal adicional. En el capítulo siguiente se explica de dónde salen los parámetros de dificultad (a), discriminación (b) y pseudo-adivinación (c), así como la racionalidad general de la escala.

Tabla 4. Criterios para la selección de reactivos de acuerdo al valor estimado de sus parámetros en TRI

1) Dificultad	-2.5 ≤ b ≤ 2.5
2) Discriminación	0.5 ≤ a ≤ 2
3) Pseudo-adivinación	c ≤ 0.30

Como se mencionó anteriormente, no hay criterios estrictos para ninguno de los parámetros, especialmente en el caso de la dificultad del reactivo. Para algunas

poblaciones de habilidades extremas, por ejemplo, pueden ser necesarios reactivos con dificultades igual de extremas. Esto es particularmente cierto si se considera que, a diferencia de la TCT, la escala en la que se calibran los reactivos en TRI puede ser muy variable. Cuando se usa la TRI, no importan en realidad los números concretos de la escala que se establezcan (pueden variar mucho, según la población de referencia); lo que sí es fundamental es que, sean cuales sean estos números, se mantengan constantes.

Normalmente los valores de la escala de habilidad/dificultad se estandarizan, tomando como referencia la distribución de habilidad de los sustentantes o la distribución de dificultad de los reactivos. Por tal motivo, por lo general se propone para la dificultad un intervalo de -2.5 a 2.5 y así impedir la presencia de reactivos exageradamente fáciles o difíciles, considerando condiciones generales (normales, no extremas) de aplicación. Por razones técnicas, también conviene que el reactivo se encuentre en un rango de dificultad en que haya sustentantes suficientes para poder hacer una estimación precisa de sus características métricas. En los extremos, hay generalmente pocos sustentantes y la estimación de los parámetros del reactivo puede tener un error importante. La existencia de un número suficiente de sustentantes en el punto que marca la dificultad del reactivo puede determinar una calibración adecuada, o defectuosa.

Además del análisis individual de los reactivos, la TRI permite anticipar las características métricas y el nivel de precisión de la prueba según los reactivos que la componen. Por ello, la selección de los reactivos trasciende el hecho de que cumplan o no con ciertos límites; en aplicaciones avanzadas, es necesario también que cada reactivo tenga las cualidades que se requieren para que funcione el conjunto, es decir, que cada “ladrillo” embone en su lugar. Cuando se aborde el tema de la función de información de la prueba, quedará más claro cómo puede adaptarse la prueba –y, por supuesto, la selección de sus reactivos– a los objetivos de la evaluación.

Limitaciones del análisis de reactivos

Los resultados del análisis de reactivos son siempre una propuesta, una fuente de información que debe sopesarse en el contexto de la prueba y de lo que se busca con ella. De partida, como sucede en todo procedimiento estadístico, existen elementos que pueden afectar seriamente los resultados: la representatividad de la muestra, el número de sustentantes involucrados en el estudio, las condiciones particulares de la aplicación, etcétera. Pero más allá de los problemas propios de la estadística, que deberán, como siempre, tratar de controlarse, vale la pena subrayar otras limitaciones.

- **Un mal resultado en el análisis no significa que deba eliminarse el reactivo de la prueba.** Los resultados del análisis son indicadores importantes de la calidad del reactivo, sobre todo cuando lo que se busca es un instrumento que permita diferenciar mejor entre los que saben y los que no; y es claro que casi siempre de esto es de lo que se trata al hacer una evaluación: diferenciar con el mayor detalle posible entre los sustentantes. Sin embargo, éste no es el propósito cuando se trata simplemente de corroborar que los sustentantes saben lo que deben saber. En estos casos, el hecho de que un reactivo no sea particularmente discriminador no resulta fundamental.

  Un ejemplo puede aclarar las cosas. Supóngase que un maestro quiere ver si sus alumnos aprendieron bien lo que les enseñó en clase y hará una serie de preguntas para investigarlo. Sin duda, es perfectamente válido que las preguntas sean sobre todos o la mayor parte de los temas que se vieron en clase, hayan sido fáciles o difíciles. Y supongamos también que todos los alumnos demuestran ser estudiosos y obtienen un buen resultado en la prueba. Como consecuencia, los reactivos tienen indicadores extremos en el análisis: son fáciles y, por lo mismo, no discriminan bien (todos los responden correctamente). ¿Hay algo malo en ello? En principio, nada. Ciertamente, el maestro estará muy complacido si todos salen bien (y muy desdichado si nadie contesta nada), independientemente de que algunos de los muchachos pudieran saber un poco más y otros un poco menos de lo que se vio en clase. Lo importante para nuestro maestro no es necesariamente identificar las

diferencias, sino corroborar que sus alumnos hayan aprendido (o no) lo que les enseñó. Si todos lo aprendieron, el examen será “demasiado fácil” y no discriminará mucho, pero resultará perfecto para nuestro profesor.

La discriminación y la dificultad, a pesar de ser tan valiosas en otros contextos, pasan a un segundo término en estos casos. De hecho, insistir en seleccionar ciertos reactivos porque discriminen mejor que otros puede llegar a tener efectos perniciosos en una prueba. Por razones técnicas, las medidas más comunes para la discriminación favorecen pruebas con contenidos más homogéneos. Esto, que es lo que se busca para hacer una buena medición, puede también, si no se utiliza juiciosamente, promover pruebas “más planas” y favorecer que se eliminen contenidos que se vieron en clase y que podría ser importante evaluar.

- **Un buen resultado en el análisis de reactivos no es suficiente para decidir que un reactivo se debe incluir.** El análisis de reactivos no es garantía de validez. El hecho de que un reactivo, o un grupo de reactivos, presente un comportamiento estadístico adecuado no significa necesariamente que sea válido utilizar ese reactivo en una prueba.

  El análisis del conjunto de los reactivos ayuda a verificar la consistencia interna del criterio que se evalúa; sin embargo, la información que aporta no es suficiente para determinar si lo que se mide, por consistente que sea, es lo que efectivamente se quiere medir. Otros elementos, generalmente criterios externos, deben aportarse para complementar el análisis en términos de validez.

  Por lo anterior, regularmente se añaden al análisis de reactivos otros estudios para verificar la validez del criterio y la de los resultados, de manera predictiva o concurrente. Estos métodos pueden ser estadísticos o a partir de opiniones expertas. En otras palabras, no basta con que el reactivo mida bien lo que mide; además debe verificarse que esto que mide, sea lo que sea, se aproxime a lo que verdaderamente queremos medir.

Que existan limitaciones en el uso del análisis de reactivos no quiere decir que deba obviarse. Por el contrario, estas situaciones deben ser consideradas como excepciones, casos extremos. Y aún en estos casos no es tan fácil decidir si se descartan los indicadores. El análisis de reactivos aporta información importante sobre la adecuación del reactivo a la prueba. Al obviar esta información y esta validación se corre el riesgo de cometer errores importantes. El juicio de los expertos es trascendental, pero muchas veces no es suficiente. También los expertos se equivocan. El análisis de reactivos ofrece una buena dosis de realidad que difícilmente se obtiene de otro modo.

Por lo demás, el hecho de que un sustentante sepa o no lo que se le pregunta en una prueba es un asunto de grado, no de todo o nada. Por ello, siempre es preferible contar con un continuo más amplio en el que se puedan ubicar los sustentantes con distintos grados de habilidad. Las preguntas que ayuden mejor a diferenciar en esta escala resultan preferibles. De poco sirve en un examen hacer preguntas que todos saben o que nadie puede contestar. Por lo general, estas preguntas no valen el espacio que ocupan en la prueba y el tiempo que el sustentante podría utilizar para contestar otras preguntas que aporten mayor información.

Capítulo 4. Fundamentos técnicos del análisis de reactivos

En los capítulos anteriores se definieron los principales objetivos del análisis de reactivos y se trazaron procedimientos básicos para llevarlo a cabo; fueron omitidos, en la medida de lo posible, los detalles más técnicos y las fórmulas matemáticas que definen a la TCT y a los modelos de la TRI. Sin pretender desarrollar de manera exhaustiva todos los elementos de estas teorías, en este capítulo se describirán algunas de sus cuestiones técnicas esenciales, que permitirán una mejor comprensión del análisis. Entre otras cosas, se incluyen formulas matemáticas para algunos de los estadísticos más utilizados y se describen procedimientos específicos para el análisis de distractores y para el análisis de funcionamiento diferencial (DIF).

Teoría Clásica de los Tests

La idea fundamental de este modelo se basa en la siguiente ecuación:

$$X = V + E$$

donde:

$X =$ Puntuación observada
$V =$ Puntuación verdadera
$E =$ Error de medida

Esta ecuación implica que el puntaje que observamos –el que se obtiene a partir de la calificación en la prueba– está compuesto por dos elementos básicos: una cierta medida que representa el puntaje verdadero (o lo que en realidad debería obtener el sustentante de acuerdo con su nivel) y un error de medida, que puede depender de muchos factores.

Aunque en la práctica la aplicación repetida de una prueba a una misma persona resulta inviable, esta idea constituye el soporte conceptual en que descansa la TCT. Si se pudiera administrar una prueba a un mismo sustentante en repetidas ocasiones, suponiendo que en cada aplicación es posible lograr que el examinado olvide la

experiencia ganada de la evaluación inmediata anterior, las puntuaciones observadas de dicho sujeto podrían variar debido a los errores de medición. En un examen de opción múltiple, por ejemplo, parte del error de medición podría atribuirse a los descuidos del sustentante al codificar la respuesta correcta en la hoja de respuestas, este error podría llevarlo a perder puntos; sin embargo, también habría que considerar que el sustentante tiene cierta probabilidad de elegir la respuesta correcta por azar, lo que lo llevaría a ganar puntos en la prueba aun cuando no cuente con los conocimientos específicos que explora esa pregunta. En este sentido, para cada sustentante de una prueba, X y E son variables aleatorias que pueden tomar diferentes valores de una aplicación a otra, mientras que V es un valor constante a lo largo de las aplicaciones repetidas.

Definición del puntaje verdadero

Tomando en cuenta que el puntaje observado para cada sustentante es la realización de una variable aleatoria, es posible definir el puntaje verdadero del sustentante como el *valor esperado* de X. De esta forma, para el sustentante *j* tenemos que:

$$V_j = E(X_j)$$

donde:

V_j = Puntuación verdadera del sustentante *j*

$E(X_j)$ = Valor esperado de la variable aleatoria X. En el contexto de la aplicación repetida de la prueba, este valor esperado puede estimarse a partir del promedio de las puntuaciones observadas del sustentante *j*

Definición del error de medida

De acuerdo con el modelo de la Teoría Clásica, el error de medida es la diferencia entre el puntaje observado y el puntaje verdadero, de esta forma, para el sustentante *j* tenemos que:

$$E_j = X_j - V_j$$

donde:

E_j = Error de medida para el sustentante *j*

X_j = Puntaje observado del sustentante *j*

V_j = Puntaje verdadero del sustentante *j*

Debido a que E_j es la diferencia entre una variable aleatoria y una constante para el sustentante *j*, la media de E_j es el valor esperado:

$$\mu(E_j) = E(X_j - V_j)$$

Tomando en cuenta las propiedades del valor esperado, tenemos que:

$$\mu(E_j) = E(X_j) - E(V_j)$$

Al ser V_j una constante y tomando en cuenta la definición de puntaje verdadero, entonces:

$$\mu(E_j) = E(X_j) - V_j = V_j - V_j = 0$$

En resumen, el promedio del error de medida para un sustentante en la aplicación repetida de una prueba es igual a cero.

Supuestos de la Teoría Clásica de los Tests

De las definiciones anteriores es posible derivar algunos principios básicos que permitirán aplicar la Teórica Clásica de los Tests al estudio de la confiabilidad (más adelante se utilizarán estos supuestos para deducir uno de los estimadores de la confiabilidad de la prueba más utilizados en la psicometría, el coeficiente Alfa de Cronbach).

1. El error de medida promedio para una población de examinados es cero ($\mu(E) = 0$).

2. La correlación entre los puntajes verdaderos y los errores de medida para una población de examinados es cero ($\rho_{VE} = 0$).
3. La correlación entre los errores de medida para una población de examinados en dos aplicaciones de la misma versión es igual a cero ($\rho_{E_1E_2} = 0$)

Análisis de reactivos basado en la Teoría Clásica de los Tests

Aun cuando el desarrollo de la TCT se centra en el análisis de la prueba, esta teoría hace uso de algunos indicadores de la dificultad y la discriminación de los reactivos. A continuación se describen algunos de los índices más utilizados.

Indicadores de dificultad

Evaluar la dificultad de los reactivos es importante por dos razones básicas:

- Para tener una idea de qué tan difícil o fácil será para los sustentantes contestar cada pregunta. Esto es útil en la medida en la que se puede diseñar con ello un examen, dependiendo del objetivo que se busque.

- Conocer la distribución de las dificultades de los reactivos permite diseñar pruebas que se parezcan entre sí. Se pueden lograr versiones de una misma prueba con una dificultad "similar", eligiendo los reactivos con cuidado para que sus dificultades se distribuyan de una manera predeterminada en cada una de ellas.

Existen varias formas de representar la dificultad de un reactivo. La de uso más común es simplemente la proporción de sustentantes que aciertan al reactivo[5] y algunas transformaciones simples de este valor. Quizá la transformación más conocida es el grado de dificultad.

El *Grado de Dificultad* es el porcentaje de sustentantes que contestan correctamente un reactivo. La única diferencia entre la proporción de sustentantes que aciertan y el grado

[5] A esta proporción se le conoce como "valor-p" o "p-value" en inglés. No confundir con el valor p que se utiliza en las pruebas de hipótesis estadísticas.

de dificultad es que el grado de dificultad toma valores desde 0% (cuando ningún sustentante contestó correctamente el reactivo) hasta 100% (cuando el reactivo fue contestado correctamente por todos los sustentantes), mientras que el valor-p toma valores de 0 a 1.

La interpretación de este indicador puede ser confusa debido a que valores altos del mismo indican menor dificultad (más sustentantes contestan ese reactivo), mientras que valores bajos nos hablan de una mayor dificultad. Por eso hay quienes se refieren al indicador como el "grado de facilidad del reactivo".

Indicadores de discriminación

Son diversas las maneras de estimar la discriminación de un reactivo. En la mayor parte de las pruebas, la discriminación del reactivo normalmente se mide tomando como referencia de la habilidad de un sustentante la puntuación que obtiene en el total de la prueba. Así, un reactivo discriminará de manera eficaz si lo responden correctamente más sujetos con puntuaciones altas que sujetos con puntuaciones bajas. Por el contrario, un reactivo será ineficaz si quienes más aciertan son los que obtienen bajas puntuaciones en la prueba. Un reactivo que funcione así, que tenga baja discriminación o, mucho peor, que discrimine de manera inversa será siempre sospechoso de estar equivocado; puede tener la opción correcta mal codificada o incluso evaluar una cosa distinta al resto de la prueba.

Además de la anterior –que compara el comportamiento del reactivo con el desempeño del sustentante en la prueba–, es posible abordar el problema tomando como referencia algún criterio externo. Así, por ejemplo, es posible que se tenga a la mano una clasificación previa de una muestra de sustentantes que sirva para establecer, de antemano, quiénes son competentes y que pudiera servir como un referente adicional. Un reactivo discrimina si quienes lo contestan más son aquellos de cierta categoría determinada previamente, tal vez a partir de la opinión de expertos o con otro examen.

Índice de discriminación

Al definir la discriminación como la capacidad del reactivo para distinguir entre sustentantes de alto y de bajo rendimiento, resulta lógico obtener un indicador de la diferencia del porcentaje de respuestas correctas entre un grupo de rendimiento alto y otro de bajo rendimiento. De esta manera, el índice de discriminación se define como (Crocker y Algina, 1986):

$$ID = P_S - P_I$$

donde:

ID = Índice de discriminación

P_S = Proporción de respuestas correctas en la prueba, de los sustentantes del Grupo Superior (número de respuestas correctas en el Grupo Superior entre el número de sustentantes que conforman este grupo)

P_I = Proporción de respuestas correctas en la prueba, de los sustentantes del Grupo Inferior (número de respuestas correctas en el Grupo Inferior entre el número de sustentantes que conforman este grupo)

Se pueden definir los grupos de rendimiento a partir de criterios externos (delimitados por un panel de expertos, por ejemplo) o internos (la puntuación en la prueba). Cuando se usan puntuaciones, los grupos superior e inferior pueden definirse a partir de distintos criterios. Por ejemplo, tomando la mediana de las puntuaciones totales en la prueba, de tal manera que se puede conformar el grupo superior con los sustentantes que obtuvieron un puntaje mayor o igual a la mediana y el grupo inferior con sustentantes con puntajes menores a ésta. También pueden definirse a partir de otros criterios. Siguiendo el estudio de Kelley (1939), es común utilizar los percentiles 27 y 73 para identificar a los integrantes del grupo inferior y del grupo superior, respectivamente, con lo que se logran grupos más distintos entre sí.

Correlación Punto Biserial

Existen índices de discriminación que se basan en la correlación entre las puntuaciones de un reactivo y las puntuaciones totales en la prueba. La correlación indica la fuerza y la dirección de una relación lineal entre dos variables aleatorias. Se considera que dos

variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores de la otra. De esta manera, la correlación entre las variables X e Y será positiva si al aumentar los valores de X también aumentan los valores de Y y será negativa, si al aumentar los valores de X, los de Y disminuyen.

La correlación de Pearson es un número entre -1 y 1 que mide el grado de asociación entre dos variables continuas[6]. Un valor positivo de esta correlación implica una asociación positiva, mientras que un valor negativo implica una asociación negativa.

El cálculo de la correlación de Pearson se realiza utilizando la siguiente fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{(n-1)\hat{\sigma}_X\hat{\sigma}_Y}$$

$$\rho = \frac{COV(X,Y)}{\hat{\sigma}_X\hat{\sigma}_Y} = \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{\frac{1}{(n-1)}\sqrt{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \hat{\mu}_Y)^2}} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{\sqrt{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \hat{\mu}_Y)^2}}$$

donde:

ρ = Correlación de Pearson

$COV(X,Y) = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)$ (covarianza muestral)

$\hat{\sigma}_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2}{n-1}}$ (estimador insesgado de la desviación estándar de X)

$\hat{\sigma}_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}_Y)^2}{n-1}}$ (estimador insesgado de la desviación estándar de Y)

[6] La correlación de Pearson supone que las variables tienen una distribución normal bivariada, lo que garantiza que la única relación que puede existir entre las variables es una relación lineal. También supone independencia, esto es, que los valores que alcanzan las variables en un caso particular son independientes a los valores de las variables para el resto de los casos.

X_i = Observación *i* de la variable X

Y_i = Observación *i* de la variable Y

$\hat{\mu}_X$ = Media de X

$\hat{\mu}_Y$ = Media de Y

n = Número total de observaciones

Un caso particular de la correlación de Pearson se da cuando una de las dos variables implicadas es dicotómica y la otra puede considerarse como continua. Un ejemplo de esto es cuando se quiere calcular la asociación que existe entre los resultados obtenidos en un reactivo en particular (en escala dicotómica) y la calificación global en la prueba (en escala continua). A este tipo de correlación se le conoce como *correlación punto biserial* o *correlación biserial puntual*, y puede calcularse por medio de la siguiente fórmula simplificada de la correlación de Pearson:

$$\rho_{pbis} = \frac{(\hat{\mu}_+ - \hat{\mu}_X)}{\hat{\sigma}_X}\sqrt{p/q}$$

donde:

ρ_{pbis} = Correlación punto biserial

$\hat{\mu}_+$ = Puntaje total promedio de los sustentantes que contestaron el reactivo correctamente

$\hat{\mu}_X$ = Puntaje total promedio de todos los sustentantes

$\hat{\sigma}_X$ = Desviación estándar de las puntuaciones de todos los sustentantes

p = Proporción de respuestas correctas del reactivo

$q = 1 - p$

Un criterio que generalmente se utiliza para la selección de reactivos por correlación punto biserial es aceptar sólo los que presenten un valor mayor a 0.20 en este indicador. De acuerdo con Crocker y Algina (1986), este valor garantiza una correlación significativamente mayor a cero incluso para muestras pequeñas de sustentantes (100 observaciones). Una vez definido este criterio, los reactivos que muestren una correlación menor no podrán ingresar al banco como candidatos para conformar versiones finales de examen.

Correlación Biserial

Otro índice utilizado con frecuencia para describir el poder de discriminación de un reactivo es la Correlación Biserial. A diferencia de la correlación punto biserial, en esta se asume que la variable latente (Crocker y Algina, 1986) que subyace a las respuestas de los sustentantes al ítem analizado tiene una distribución normal.

El cálculo de la correlación biserial se realiza con la siguiente fórmula:

$$\rho_{bis} = \frac{(\hat{\mu}_{+} - \hat{\mu}_{X})}{\hat{\sigma}_{X}}(p / Y)$$

donde:

ρ_{bis} = Correlación biserial

$\hat{\mu}_{+}$ = Puntaje total promedio de los sustentantes que contestaron el reactivo correctamente

$\hat{\mu}_{X}$ = Puntaje total promedio de todos los sustentantes

$\hat{\sigma}_{X}$ = Desviación estándar de las puntuaciones de todos los sustentantes

p = Proporción de respuestas correctas del reactivo

Y = Ordenada de la curva normal estándar en el valor del puntaje *z* asociado con el valor de *p* para este reactivo

La única modificación en el cálculo de esta correlación con respecto a la correlación punto biserial es la inclusión del término *Y*, que es el valor que toma curva de distribución de probabilidad de una normal estándar para un valor de *p* determinado. Por ejemplo, si el reactivo tiene una proporción de respuestas correctas de 0.5, el valor de *Y* será muy cercano a 0.4.

El valor de la correlación biserial es siempre mayor que el de la correlación punto biserial, sobre todo para reactivos en los extremos de la escala de dificultad (grado de dificultad menor a 25% o mayor a 75%). En estos casos la correlación biserial puede llegar a ser hasta cuatro veces mayor a la correlación punto biserial. Por este motivo, cuando se requiere seleccionar reactivos en los extremos de la escala de dificultad, se recomienda utilizar la correlación biserial. Para reactivos con dificultad moderada no existe mayor diferencia entre este índice y la correlación punto biserial.

Indicadores ajustados por correlación espuria

Los indicadores de discriminación descritos anteriormente, pueden arrojar valores artificialmente elevados debido a que toman en cuenta el puntaje en la prueba y el puntaje en el reactivo, siendo el primero una variable que depende directamente del segundo. Cuando la prueba cuenta con un número reducido de reactivos, el peso de cada reactivo en el número total de aciertos en la prueba puede ser muy importante, provocando correlaciones espurias entre las dos variables de interés.

Para evitar el problema de la correlación espuria entre el puntaje en el reactivo y el puntaje en la prueba, se recomienda sustraer el puntaje del reactivo bajo análisis del puntaje global antes de calcular cualquiera de los indicadores de discriminación basados en correlaciones.

Indicadores de confiabilidad

De acuerdo con la TCT, la medida resultante de un ejercicio de evaluación está afectada por una cierta cantidad de error aleatorio. Si se le aplicara una prueba a un sustentante en varias ocasiones, suponiendo que no recordara sus respuestas anteriores, sus puntajes observados en cada una de las aplicaciones difícilmente serían iguales.

Sin embargo, las mediciones repetidas generalmente muestran ciertas consistencias, y a esta tendencia, a la consistencia de un conjunto de medidas, se le denomina *confiabilidad*. Es decir, aun sabiendo que no obtendríamos los mismos puntajes de manera exacta, sí esperaríamos observar puntuaciones muy cercanas entre sí. Si una prueba es confiable –definida como tal en términos de su consistencia–, entonces la aplicación repetida de la misma prueba o de varias pruebas paralelas[7] debería de dar un resultado similar. Determinar el grado de confiabilidad de un instrumento es uno de los objetivos más importantes de la psicometría.

[7] De acuerdo con la TCT, dos pruebas son paralelas cuando (1) cada sustentante tiene el mismo puntaje verdadero en ambas (miden exactamente el mismo constructo pero con reactivos diferentes), y (2) la varianzas de sus errores de medida son las mismas.

La determinación de la confiabilidad de un instrumento se reduce a establecer qué parte de la variación observada de las puntuaciones se debe a verdaderas diferencias en el constructo medido y qué parte de debe a los errores de medida.

$$\rho_{X_1X_2} = \frac{\sigma_V^2}{\sigma_X^2}$$

donde:

$\rho_{X_1X_2}$ = Coeficiente de confiabilidad

σ_V^2 = Varianza de las puntuaciones verdaderas

σ_X^2 = Varianza de las puntuaciones observadas

El coeficiente de confiabilidad es un concepto teórico, al requerir la varianza de las puntuaciones verdaderas no permite el cálculo empírico de la confiabilidad de la prueba. Para poder estimar la confiabilidad será necesario recolectar datos a partir de la aplicación de dos pruebas paralelas; la correlación entre los puntajes observados en estas dos pruebas se conoce como *coeficiente de equivalencia*. Este método tiene el inconveniente de requerir la construcción de pruebas paralelas lo cual puede resultar muy difícil y costoso, por tal motivo, otra manera de estimar la confiabilidad consiste en aplicar la misma prueba en dos ocasiones a la misma muestra de sustentantes (test-retest). La correlación entre los puntajes obtenidos en la aplicación repetida de una misma prueba se conoce como *coeficiente de estabilidad*. Uno de los principales inconvenientes de este método radica en que, al tratarse de una misma prueba, los sustentantes podrían adquirir cierta experiencia para la segunda aplicación, lo que introduciría un sesgo en las puntuaciones observadas y por lo tanto una violación importante a la definición de pruebas paralelas.

Tanto la aplicación repetida de una prueba como la aplicación de pruebas paralelas resultan diseños experimentales poco prácticos. En una situación real de evaluación, generalmente la prueba es aplicada en una sola ocasión. Las aproximaciones más populares a la estimación de la confiabilidad surgen a partir de esta restricción. Los *coeficientes de consistencia interna* permiten calcular la confiabilidad analizando la estabilidad de la prueba en su interior, por ejemplo, obteniendo la correlación entre los

puntajes observados en dos mitades del examen, o la matriz de covarianzas de los puntajes obtenidos a partir de la segmentación de la prueba en “n” partes.

Otro método de consistencia interna para estimar la confiabilidad de una prueba es el coeficiente Alfa de Cronbach. Este coeficiente es casi con seguridad el método más utilizado en la psicometría para estimar la confiabilidad de una prueba. Propuesto por Cronbach en 1951, es una extensión de la fórmula Kuder-Richardson 20, que la precedió y que es equivalente cuando los reactivos son dicotómicos.
El coeficiente Alfa se puede calcular de acuerdo con la siguiente fórmula (ver *anexo 2* para la deducción de la fórmula):

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right)$$

donde:

$\hat{\alpha}$ = Coeficiente alfa de Cronbach
k = Número de reactivos en la prueba
$\hat{\sigma}_i^2$ = Varianza del puntaje en el reactivo *i*
$\hat{\sigma}_X^2$ = Varianza del puntaje total

El valor del Coeficiente Alfa de Cronbach suele considerarse como una aproximación al límite inferior de la confiabilidad de la prueba. En otras palabras, se espera que la confiabilidad sea siempre igual o superior a la que estima este indicador.

La fórmula para variables dicotómicas –c*oeficiente Kuder Richardson 20*– se obtiene al estimar la varianza del reactivo ($\hat{\sigma}_i^2$) a partir de pq en la fórmula general de Alfa.

El Alfa de Cronbach (representado por la letra griega alfa, α) se ofrece en la mayor parte de los paquetes de estadística más comunes, con la ventaja de que (además del indicador único) generalmente señalan el efecto que tendría eliminar un reactivo sobre el indicador total. Conocer si eliminar un reactivo determinado mejora o empeora la confiabilidad estimada para la prueba puede ser una información importante en la selección de los reactivos, debido a que parece natural que entre más reactivos

contenga la prueba, mayor será la precisión con la que ésta mide el constructo evaluado. Spearman (1910) y Brown (1910) propusieron, de manera independiente, la siguiente fórmula que relaciona la longitud de la prueba con el coeficiente de confiabilidad.

$$\rho_{2X_1X_2} = n\frac{\rho_{1X_1X_2}}{1+(n-1)\rho_{1X_1X_2}}$$

donde:

$\rho_{2X_1X_2}$ = Coeficiente de confiabilidad pronosticado al variar la longitud de la prueba

$\rho_{1X_1X_2}$ = Coeficiente de confiabilidad previo a la variación de la longitud de la prueba

n = Número de veces que aumenta el tamaño de la prueba

Esta fórmula funciona siempre y cuando los reactivos añadidos tengan características psicométricas similares a las de los reactivos originales, es por esto que en las situaciones en las que se elimina un reactivo y el valor del Coeficiente Alfa aumenta o permanece constante, lo que se está detectando es que el reactivo en cuestión no aporta a la precisión de la medida y debe ser eliminado.

Una de las confusiones más comunes se da al pensar que el Alfa de Cronbach es un indicador de la unidimensionalidad de la prueba[8]. Es cierto que la consistencia interna es una condición necesaria para la unidimensionalidad, sin embargo no es una condición suficiente. De acuerdo con Schmitt (1996), es posible alcanzar covarianzas altas entre los reactivos de pruebas que miden más de una dimensión, lo que implica que no debe de utilizarse el Alfa de Cronbach como una medida de unidimensionalidad. Para definir si una prueba es unidimensional es recomendable recurrir al análisis factorial.

Error estándar de medida y estimación de la puntuación verdadera

Aunque no es posible determinar la cantidad exacta de error para un puntaje dado, la TCT postula un método para describir la variación esperada de los puntajes observados

[8] Se dice que una prueba es unidimensional, cuando el conjunto de reactivos que la componen miden una sola habilidad (la unidimensionalidad es uno de los supuestos de la TRI).

de cada sustentante, alrededor de su puntaje verdadero. Dado que definimos la puntuación verdadera como la media de los puntajes observados (obtenidos de varias aplicaciones repetidas de una prueba), los puntajes observados para cada sustentante se distribuirán de cierta forma alrededor de su puntaje verdadero. Cuando se promedian los errores estándar de todos los sustentantes, se obtiene el *error estándar de medida* (ver la deducción de la fórmula en el anexo 3).

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{X_1 X_2}}$$

donde:

$\sigma_E =$ Error estándar de medida
$\sigma_X =$ Desviación estándar de las puntuaciones obtenidas en la prueba
$\rho_{X_1 X_2} =$ Coeficiente de confiabilidad (en la práctica generalmente se utiliza el coeficiente Alfa)

Asumiendo que los errores de medición se distribuyen normalmente, es posible encontrar un intervalo de confianza con un nivel de confianza, $(1-\alpha)\times 100\%$, para las puntuaciones observadas de un sustentante cuya puntuación verdadera es *V*, como se indica en la siguiente expresión:

$$V \pm z_{1-\alpha/2}\sigma_E$$

donde:

$V =$ Puntaje verdadero del sustentante
$z_{1-\alpha/2} =$ Percentil $1-\alpha/2$ de la distribución normal estándar (1.96 para $\alpha = 0.05$)
$\sigma_E =$ Error estándar de medida

El intervalo de confianza calculado indica la precisión con la que, en promedio, la prueba está midiendo el constructo evaluado. Sin embargo, resulta poco práctico ya que en la mayoría de las situaciones cada sustentante es evaluado en una sola ocasión y por lo tanto sólo contamos con un puntaje observado por cada uno de ellos y no conocemos su puntaje verdadero *V*, ni podemos estimarlo. Además se asume que el tamaño del intervalo es igual para cualquier puntaje a lo largo del continuo de puntajes

verdaderos, este supuesto no parece especialmente realista ya que la precisión de la medida a lo largo de la escala de medición, dependerá de las características psicométricas de los reactivos que componen la prueba.

En la práctica nos gustaría estimar el intervalo alrededor del puntaje observado que contenga el puntaje verdadero tomando en cuenta un nivel de significancia α, con esto podríamos probar, por ejemplo, si los puntajes observados de dos sustentantes son significativamente distintos. Algunos autores sustituyen el puntaje observado por el puntaje verdadero en la formula anterior y cambian la interpretación del intervalo de la siguiente manera: Si un sustentante que obtuvo un puntaje observado X en una prueba, fuera evaluado un número infinito de veces con pruebas paralelas y construyéramos un intervalo de confianza con un nivel de confianza, $(1-\alpha)\times 100\%$, en cada ocasión, en el $(1-\alpha)\times 100\%$ de las veces el intervalo incluiría la puntuación verdadera de la persona.

Existen algunas aproximaciones diferentes para estimar intervalos de confianza con un sustento teórico y una interpretación más directa. Tomando en cuenta que las puntuaciones observadas guardan una relación lineal con las puntuaciones verdaderas, es posible estimar la puntuación verdadera que corresponde a una puntuación observada aplicando la ecuación de regresión:

$$\hat{V} = \hat{\rho}_{X_1X_2}(X - \overline{X}) + \overline{X}$$

Como en todo modelo de regresión lineal, la predicción tiene asociado un error que en este contexto es conocido como *Error Estándar de Estimación* y está dado por la siguiente ecuación:

$$\sigma_{VX} = \sigma_X \sqrt{\rho_{X_1X_2}(1-\rho_{X_1X_2})}$$

De acuerdo con lo anterior, el intervalo de confianza para la puntuación verdadera a un nivel de confianza, $(1-\alpha)\times 100\%$, se obtiene mediante la expresión,

$$\hat{V} \pm z_{1-\alpha/2}\sigma_{VX}$$

Debe notarse que, a diferencia del intervalo anteriormente descrito, éste sólo estará centrado en los puntajes observados que coincidan con el puntaje promedio del grupo de sustentantes evaluados (población objetivo).

Teoría de Respuesta al Ítem

Las pruebas se utilizan para tratar de medir lo que alguien sabe, lo que sabe hacer o lo que le gusta. Una prueba de matemáticas, por ejemplo, trata de aportar información sobre lo que el sustentante sabe o es capaz de resolver de esta materia. Para ello, se le hacen varias preguntas y si acierta (o si falla) sabremos un poco más sobre su capacidad.

El conocimiento en matemáticas o en física, la inteligencia o el gusto por la lectura son "variables latentes", que existen pero sólo pueden conocerse a partir de inferencias sobre sus manifestaciones externas (Bartholomew, 2011). No es posible saber cuan inteligente es alguien o lo que sabe de física, más que por algunas evidencias indirectas que aportan algunas pistas.

Los modelos de la TRI parten, para su desarrollo teórico, de la existencia de variables latentes que se manifiestan a través de las respuestas de los sustentantes a un grupo de reactivos; ejemplos de variables latentes que se evalúan en el Ceneval son la habilidad verbal y la habilidad matemática. A diferencia de la TCT que se basa en la idea de estimar un puntaje verdadero a partir de la aplicación repetida de una prueba, en la TRI se buscará estimar la habilidad[9] de los sustentantes.

La Curva Característica del Ítem (CCI)

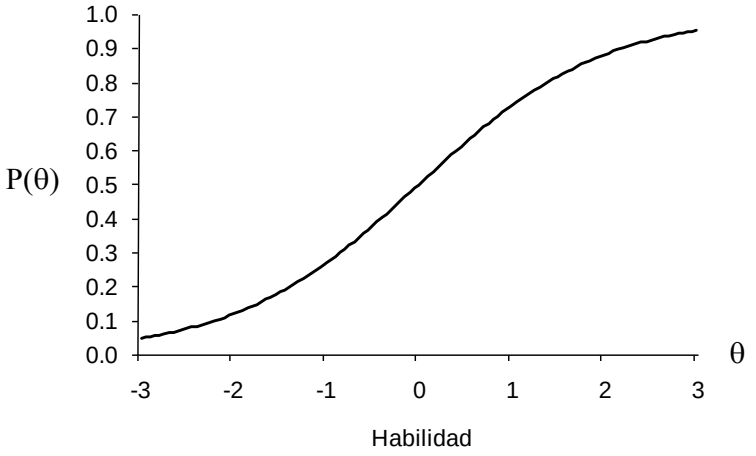
El objetivo primordial de la TRI es modelar el comportamiento de las variables latentes y determinar la probabilidad que tiene cada sustentante de acertar cuando se enfrenta a cada uno de los reactivos. La TRI utiliza una curva de distribución logística acumulada para modelar dicha probabilidad.

[9] En el marco de la TRI, se denomina *habilidad* a la hipotética variable no observable o variable latente que se quiere medir, lo que puede incluir una gran variedad de capacidades cognitivas (Baker y Kim, 2004).

Los modelos de la TRI postulan que la relación entre el rendimiento de un sustentante en determinado reactivo y la variable latente que subyace a dicho rendimiento, puede describirse por una función monótona creciente. Esto es, para cada nivel de habilidad (generalmente denominado por la letra griega theta: θ_j), existe una probabilidad, $P(X_i = 1 | \theta_j)$, de que un examinado con dicha habilidad conteste correctamente el reactivo X_i. Se esperaría que esta probabilidad crezca mientras la habilidad se incremente. La función matemática que describe la relación entre los diferentes niveles de habilidad y la probabilidad de contestar correctamente determinado reactivo es conocida como *Curva Característica del Ítem* (CCI) y constituye la base de la TRI[10].

En general, la CCI tiene una forma de "S", como se muestra en la figura 3. Las principales diferencias en los modelos más populares de la TRI radican en matices de la función matemática elegida para describir la CCI y, por lo tanto, en la forma concreta que pueda tomar esta curva.

Figura 3. Curva Característica del Ítem (CCI)



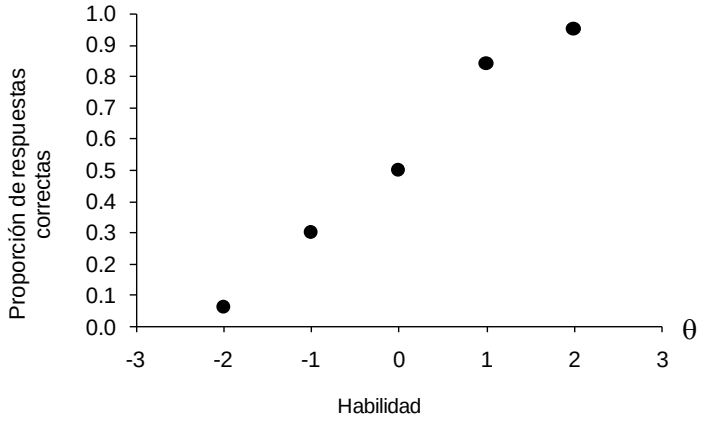
[10] Con la intención de simplificar la notación, en el resto del documento se identificará a la probabilidad condicional $P(X_i = 1 | \theta_j)$ como: $P_i(\theta_j)$, y a la función general que describe la CCI del reactivo *i* como: $P_i(\theta)$.

Interpretación de la probabilidad

Tomando en cuenta que el objetivo de estos modelos de probabilidad es expresar y predecir el comportamiento de un sustentante al enfrentarse a un reactivo específico, es importante tener una interpretación clara de la probabilidad, $P(\theta)$, en el sentido de saber qué tipo de datos reales se están tratando de modelar.

En una situación estándar de evaluación lo que se puede observar es si un sustentante con determinado nivel de habilidad, digamos θ_j, tuvo éxito o no en un reactivo en particular. Entonces, el concepto de probabilidad surge de la idea de que un número grande de examinados con el mismo nivel de habilidad (θ_j) también contestaron el reactivo, por lo que es posible calcular la proporción de respuestas correctas que obtuvo este grupo de sustentantes en el reactivo. En la figura 4 se muestra la proporción de respuestas correctas a un reactivo, por parte de siete grupos de diferente habilidad.

Figura 4. Proporción de respuestas correctas para diferentes grupos de habilidad



La tarea principal ahora será encontrar una Curva Característica del Ítem que se ajuste lo mejor posible a estas proporciones de respuesta correcta, tal y como se muestra en la figura 5.

Figura 5. Ajuste de una CCI a la proporción de respuestas correctas

Modelos logísticos de la TRI

En este apartado se definirán tres de los modelos más populares de la TRI. Una primera distinción entre ellos es el número de parámetros que utilizan para describir los reactivos. Los valores numéricos de los parámetros empleados por cada modelo definirán la forma de la CCI y describirán diferentes propiedades del comportamiento estadístico de los reactivos.

En la TRI, el modelo matemático estándar para la CCI es la función de distribución logística acumulada. Esta función es parecida a la función de distribución normal, pero tiene la ventaja de ser de más fácil manejo al no involucrar integración. La facilidad para su cálculo no es ya un factor de gran importancia en la actualidad –con la existencia de un poder de cómputo que hace todos estos cálculos triviales–, pero en la época en la que se comenzó a utilizar representaba ventajas significativas. Además, las curvas logística y normal pueden hacerse virtualmente idénticas para efectos prácticos empleando sencillos ajustes en las fórmulas.

- *Modelo logístico de un parámetro (modelo de Rasch)*

La ecuación de la CCI para el modelo logístico de un parámetro (Rasch) es la siguiente:

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \qquad i = 1, 2, ..., n$$

donde:

$P_i(\theta)$ = Probabilidad de que un examinado con habilidad θ responda correctamente el reactivo *i*

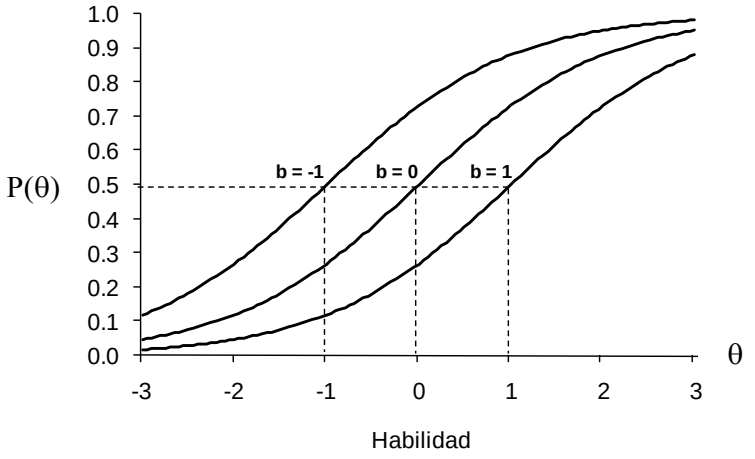
b_i = Dificultad del reactivo *i*

n = Número de reactivos en la prueba

e = Base de los logaritmos neperianos

El parámetro b_i , definido como la dificultad del reactivo *i*, indica la posición de la CCI en la escala de habilidad y se define como el punto de la escala en donde la probabilidad de respuesta correcta es igual a 0.5. En la figura 6 se muestran las CCI de tres reactivos con dificultades diferentes.

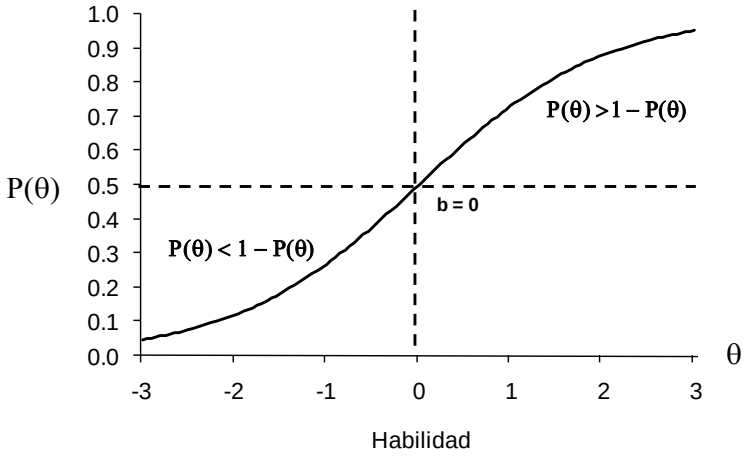
Figura 6. Reactivos con diferente parámetro de dificultad "b"



Nótese que las curvas difieren sólo en su localización a lo largo de la escala de habilidad. El reactivo más fácil tiene un parámetro de dificultad igual a -1 y se encuentra ubicado al lado izquierdo de la escala. El reactivo más difícil tiene un parámetro de dificultad igual a 1 y está ubicado hacia la derecha de la escala.

La dificultad del reactivo define un punto en la escala de habilidad justo en el que la probabilidad de éxito, $P(\theta)$, es igual a la probabilidad de fracaso, $1-P(\theta)$, esto es, un sustentante con un nivel de habilidad θ igual a la dificultad del reactivo *i* (b_i), tendrá una probabilidad de 0.5 de contestar correctamente este reactivo y, por lo tanto, una probabilidad de 0.5 de responderlo de manera incorrecta. Este punto en la escala determina un umbral que divide la escala de habilidad en dos partes; los sustentantes con un nivel de habilidad menor a la dificultad del reactivo tendrán obligadamente una probabilidad de éxito menor a 0.5 y, por lo tanto, menor a la probabilidad de fracaso. Los sustentantes con una habilidad mayor a la dificultad del reactivo tendrán una probabilidad de éxito mayor a 0.5 y consecuentemente mayor a la probabilidad de fracaso. Esta idea de umbral define adecuadamente la lógica de los modelos de probabilidad (ver figura 7).

Figura 7. El parámetro de dificultad "b" define un umbral para la probabilidad

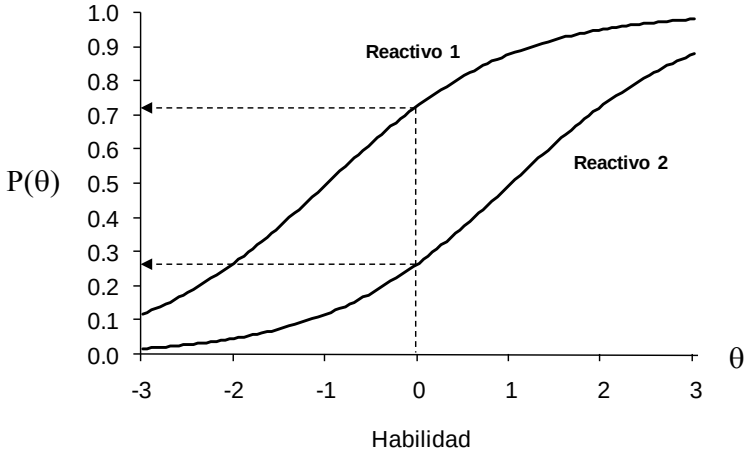


El modelo TRI de un parámetro supone que la dificultad es la única característica del reactivo que influye en la respuesta del sustentante y, por lo tanto, fija el nivel de discriminación en un valor fijo para todos los reactivos[11]. Estos supuestos originan una propiedad importante de este modelo. Para cualquier nivel de habilidad θ dado, si la

[11] En particular, el modelo de Rasch fija el valor de la discriminación de todos los reactivos en un valor de uno.

dificultad del reactivo 1 es menor a la dificultad del reactivo 2, la probabilidad de respuesta correcta del reactivo 1 será siempre mayor que la correspondiente al reactivo 2. Esto equivale a decir que las curvas características de dos reactivos con dificultades diferentes no se cruzan en ningún punto de la escala. En la figura 8 el reactivo 1 es más fácil que el reactivo 2 ($b_1 < b_2$) y la probabilidad de responder correctamente el reactivo 1 es mayor a la probabilidad de responder correctamente el reactivo 2 ($P_1(\theta) > P_2(\theta)$), en cualquier punto de la escala de habilidad.

Figura 8. En el modelo de un parámetro no hay cruzamientos de las CCI



- *Modelo logístico de dos parámetros*

El modelo logístico de dos parámetros incorpora un elemento adicional con relación al modelo anterior. A diferencia del modelo de un parámetro, este modelo permite analizar el comportamiento de los reactivos, además de por su nivel de dificultad, por su poder de discriminación.

La ecuación de modelo logístico de dos parámetros es la siguiente:

$$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \qquad i = 1, 2, ..., n$$

donde:

$P_i(\theta)$ = Probabilidad de que un examinado con habilidad θ responda correctamente el reactivo *i*

b_i = Dificultad del reactivo *i*

a_i = Discriminación del reactivo *i*

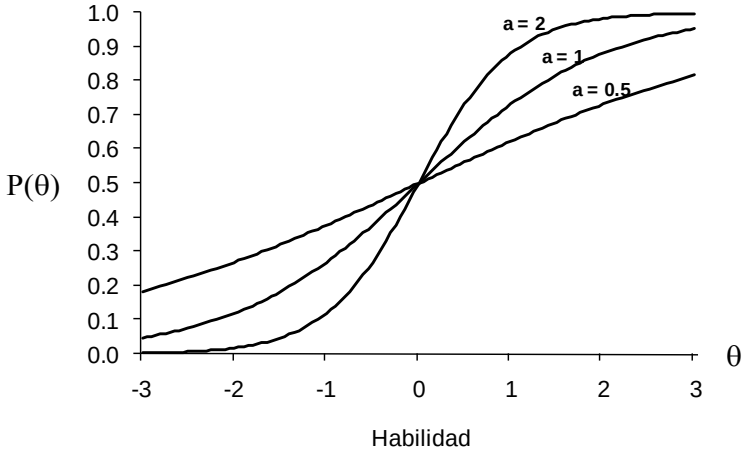
n = Número de reactivos en la prueba

e = Base de los logaritmos neperianos

El parámetro a_i es el poder de discriminación del reactivo; por lo tanto, indica qué tan bien distingue a los sustentantes de habilidad alta de los de habilidad baja. Este parámetro es proporcional a la pendiente de la CCI en el punto b_i de la escala de habilidad. Entre mayor sea el valor de a_i la CCI será más inclinada y entre menor sea el valor de este parámetro, la CCI será más plana.

En la figura 9 se muestran las curvas características de tres reactivos con la misma dificultad ($b_i = 0$), pero con diferente discriminación.

Figura 9. Reactivos con diferente parámetro de discriminación "a"

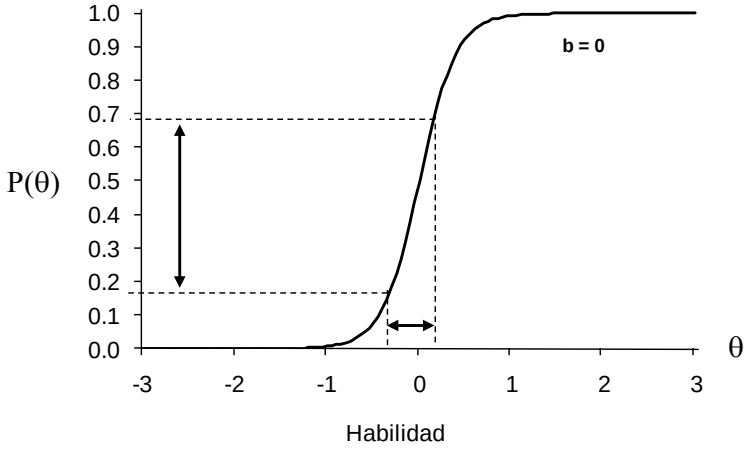


El reactivo con parámetro de discriminación igual a 0.5 tiene una CCI menos inclinada que las curvas de los dos reactivos con una mayor discriminación; de hecho, esta curva se asemeja a una recta en el intervalo de la escala visualizado.

Como puede notarse en las gráficas anteriores, las curvas características de los reactivos tienen diferente pendiente a lo largo de la escala de habilidad; sin embargo, alcanzan su mayor inclinación justamente en el punto $\theta = b_i$ (punto de inflexión de la curva). En ese punto se dice que el reactivo "da en el blanco" debido a que es el lugar en que mejor distingue entre sustentantes con baja y con alta habilidad.

La pregunta que surge en este momento es: ¿Cómo se relaciona la pendiente de la CCI con la discriminación del reactivo? Para un reactivo con una discriminación alta existe un intervalo relativamente corto en la escala de habilidad para el cual la probabilidad de respuesta correcta, $P_i(\theta)$, cambia de un valor cercano a 0 a un valor cercano a 1 (ver figura 10).

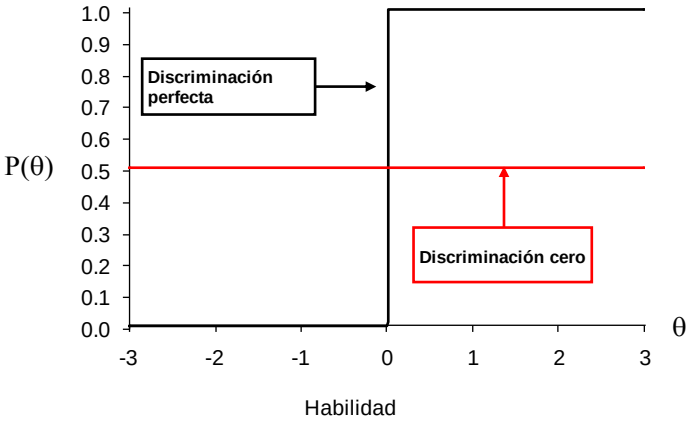
Figura 10. Interpretación de la discriminación



En la figura 11 se muestran dos reactivos con discriminación extrema. Cuando el valor del parámetro a_i se acerca a cero, la CCI tiende a una recta horizontal y cuando el parámetro de discriminación tiende infinito positivo ($+\infty$), la CCI toma una forma de

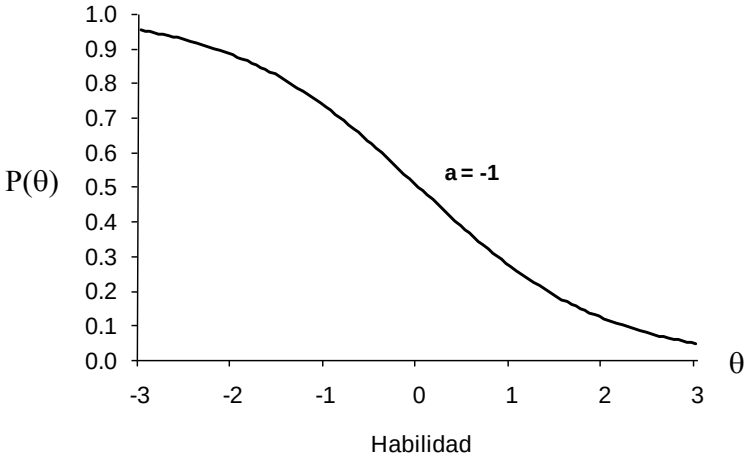
escalón. Este último caso es conocido como “patrón ideal de respuesta” o “patrón de Guttman” (Ayala, 2009. P. 36) y se establece como un modelo determinístico.

Figura 11. Reactivos con discriminación extrema



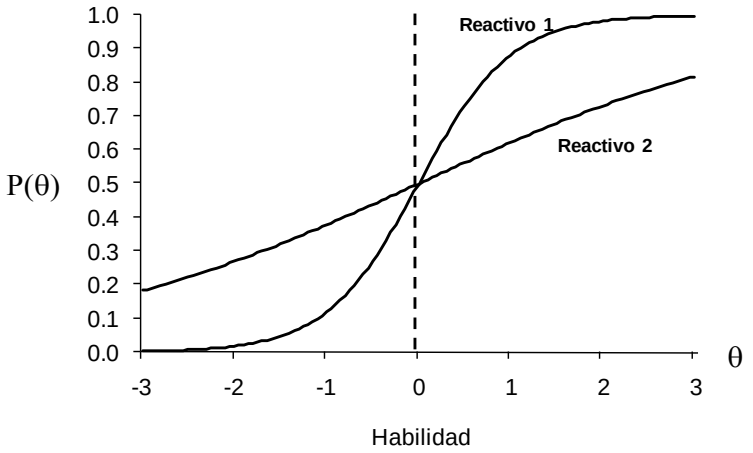
Cuando el valor del parámetro a_i es negativo, la pendiente de la curva característica será negativa, indicando un comportamiento contrario al esperado; los sustentantes de habilidad baja tendrán una mayor probabilidad de responder correctamente el reactivo que los sustentantes de habilidad alta (ver figura 12).

Figura 12. Reactivo con discriminación negativa



El hecho de que cada reactivo pueda tener distinta discriminación, hace que las curvas características puedan cruzarse, a diferencia de lo que sucedía en el modelo de un parámetro, lo que da lugar a casos que pudieran parecer paradójicos. En la figura 13, los sustentantes con una habilidad menor a cero tienen una probabilidad de respuesta correcta mayor para el reactivo 2 que para el reactivo 1. Sin embargo, esta situación se invierte para el grupo de sustentantes con habilidad mayor a cero.

Figura 13. Cruce de CCI



- *Modelo logístico de tres parámetros*

Una de las características de los reactivos de opción múltiple es que, por definición, tienen una probabilidad asociada de ser respondidos correctamente por azar. Por ejemplo, un reactivo con cuatro opciones de respuesta tiene una probabilidad de 0.25 de ser contestado por adivinación simple; un reactivo con cinco opciones de respuesta reducirá esta probabilidad de azar a 0.20, etcétera.

El modelo logístico de tres parámetros incorpora la posibilidad de que la respuesta correcta haya sido adivinada, agregando el parámetro *c*, al modelo de dos parámetros de acuerdo a la siguiente expresión.

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}} \qquad i = 1, 2, ..., n$$

donde:

$P_i(\theta)$ = Probabilidad de que un examinado con habilidad θ responda correctamente el reactivo *i*

b_i = Dificultad del reactivo *i*

a_i = Discriminación del reactivo *i*

c_i = Pseudo-adivinación del reactivo *i*

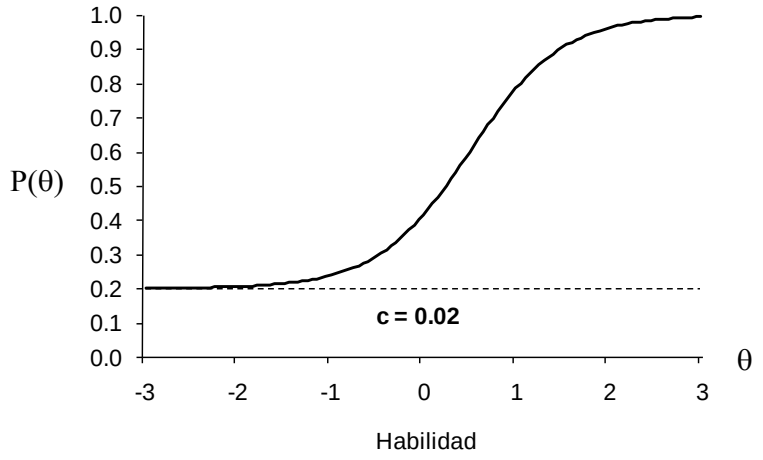
n = Número de reactivos en la prueba

e = Base de los logaritmos neperianos

La fórmula indica que existe una probabilidad fija igual a c_i, que es independiente al nivel de habilidad del sustentante. Normalmente el parámetro c_i toma valores menores a la probabilidad estimada para la adivinación aleatoria. Esto se atribuye a que los sustentantes con habilidad baja normalmente son atraídos a seleccionar alguna de las respuestas incorrectas (distractores), por lo que podrían obtener un puntaje mayor contestando aleatoriamente el reactivo. Esta es la razón por la que el parámetro c_i es denominado “pseudo-adivinación” y no simplemente “adivinación”.

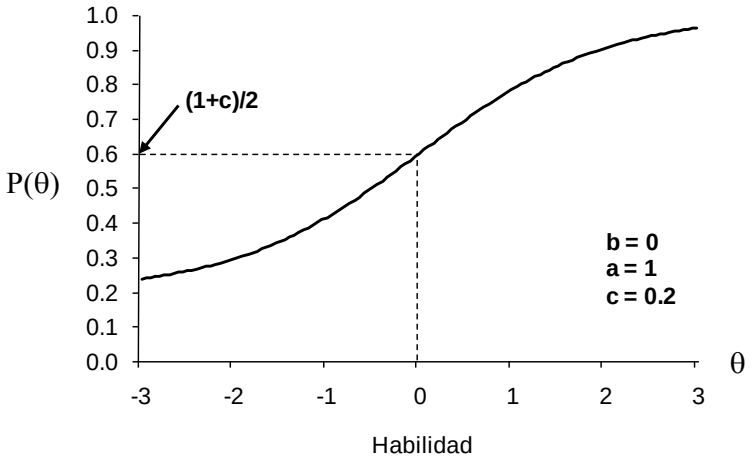
El valor del parámetro de pseudo-adivinación se refleja en la asíntota inferior de la CCI. En los modelos anteriores, la probabilidad de respuesta correcta se aproxima a cero cuando la habilidad del sustentante tiende a valores muy bajos ($-\infty$). Para este modelo, la probabilidad de respuesta correcta se aproximará al valor de c_i para valores bajos de θ, tal y como puede verse en la figura 14.

Figura 14. Parámetro de pseudo-adivinación



La incorporación del parámetro de pseudo-adivinación tiene implicaciones importantes en las propiedades matemáticas y en la interpretación del modelo. Por ejemplo, en el modelo de tres parámetros la probabilidad de respuesta correcta para un nivel de habilidad igual a b_i ya no será igual a 0.5; el desplazamiento de la asíntota inferior de la CCI aumentará esta probabilidad a un valor de $(1+c_i)/2$. La figura 15 ejemplifica esta situación.

Figura 15. Probabilidad asociada a "b" en el modelo de tres parámetros



El modelo logístico de tres parámetros es el más general de los tres. A partir de éste pueden desprenderse el modelo de un parámetro y el modelo de dos parámetros. Cuando el parámetro de pseudo-adivinación se fija en un valor de cero, el modelo de tres parámetros se convierte en el modelo de dos parámetros.

Sea $c_i = 0$, entonces:

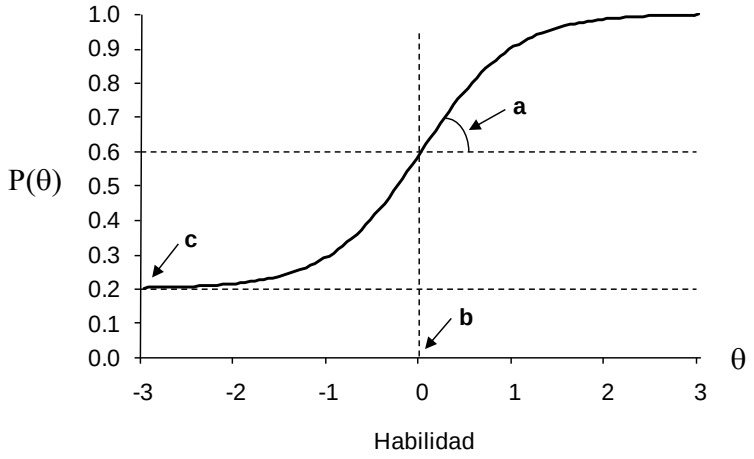
$$P_i(\theta) = 0 + (1-0)\frac{1}{1+e^{-a_i(\theta-b_i)}} = \frac{1}{1+e^{-a_i(\theta-b_i)}}$$

Si además se fija el parámetro de discriminación en un valor de 1, por ejemplo, el modelo de tres parámetros converge al modelo de Rasch. En la figura 16 se muestra la interpretación gráfica de los tres parámetros.

Sea $c_i = 0$ y $a_i = 1$, entonces:

$$P_i(\theta) = 0 + (1-0)\frac{1}{1+e^{-1(\theta-b_i)}} = \frac{1}{1+e^{-(\theta-b_i)}}$$

Figura 16. Parámetros de los modelos de la TRI



Supuestos de la TRI

Los modelos de la Teoría de la Respuesta al Ítem requieren de supuestos acerca de los datos con los que se aplicarán.

- *Unidimensionalidad*

Este supuesto indica que el constructo que se quiere medir es una variable latente continua, esto es, el conjunto de reactivos de la prueba está diseñado para medir una sola habilidad latente. De esta forma (y a manera de ejemplo), no sería válido aplicar alguno de los modelos de la TRI para analizar un examen que incluye preguntas sobre varias materias, digamos Matemáticas, Física y Español; en este caso se recurriría a analizar cada una de estas tres áreas de manera independiente, esto es, estimar un modelo para cada una de las áreas que componen el examen.

En la realidad, la unidimensionalidad perfecta no existe. Sin embargo, es suficiente la presencia de un factor o componente “dominante” que refleje la variable latente medida para poder aplicar los modelos de la TRI.

- *Independencia local*

Dado que las herramientas de la TRI se enfocan en el reactivo y no en la prueba, resulta sumamente importante que las respuestas de los sustentantes a cualquier par de reactivos sean estadísticamente independientes, esto es, que la probabilidad de responder correctamente a determinado reactivo no esté influida por la probabilidad de respuesta correcta de otros reactivos de la prueba. Por ejemplo, el contenido de un reactivo no debe dar pistas para contestar cualquier otro reactivo de la prueba.

La curva característica de la prueba

La Curva Característica del Ítem define la probabilidad de respuesta correcta a un reactivo para cada nivel de habilidad θ. Conociendo la curva característica de cada uno de los reactivos que conforman la prueba, es posible estimar el número de respuestas correctas esperado para un sustentante con habilidad θ. Este puntaje estimado representa la transformación más importante de la escala de habilidad, y es conocido como puntaje verdadero (se debe tener cuidado de no interpretar este término en el mismo sentido en el que se interpretó en el marco de la TCT).

Si U_i es la respuesta de un sustentante con habilidad θ_j al reactivo *i* (0 o 1), de acuerdo con el modelo de probabilidad de la TRI, el valor esperado de dicha respuesta está dado por la siguiente expresión:

$$E(U_i \mid \theta_j) = 1 * P_i(\theta) + 0 * (1 - P_i(\theta)) = P_i(\theta)$$

De esta manera, tomando en cuenta que todos los reactivos de la prueba son independientes y haciendo uso de la linealidad del operador de esperanza, el sustentante mencionado tiene un puntaje esperado τ (puntaje verdadero) en la prueba, igual a la suma de las probabilidades de todos los reactivos que la componen.

$$\tau(\theta) = \sum_{i=1}^{n} P_i(\theta)$$

Ejemplo:

Supongamos que una prueba compuesta por 10 reactivos es analizada a través del modelo de Rasch y se estiman los parámetros de dificultad para los reactivos expresados en la tabla 5.

Tabla 5. Parámetro de dificultad de 10 reactivos de una prueba

Reactivo	1	2	3	4	5	6	7	8	9	10
Dificultad	-2.3	-2	-1.5	-0.3	0	0.7	1.2	1.9	2.3	2.5

Para calcular el puntaje verdadero para un sustentante con habilidad $\theta = 1$ se debe seguir el siguiente procedimiento:

El primer paso es calcular la probabilidad de respuesta correcta para cada reactivo tomando en cuenta que el sustentante tiene una habilidad igual a 1 (ver tabla 6).

Tabla 6. Probabilidad de respuesta correcta de 10 reactivos de una prueba, suponiendo un nivel de habilidad de 1

$$P_i(1) = \frac{1}{1 + e^{-(1-b_i)}}$$

Reactivo	1	2	3	4	5	6	7	8	9	10
$P(\theta)$	0.96	0.95	0.92	0.79	0.73	0.57	0.45	0.29	0.21	0.18

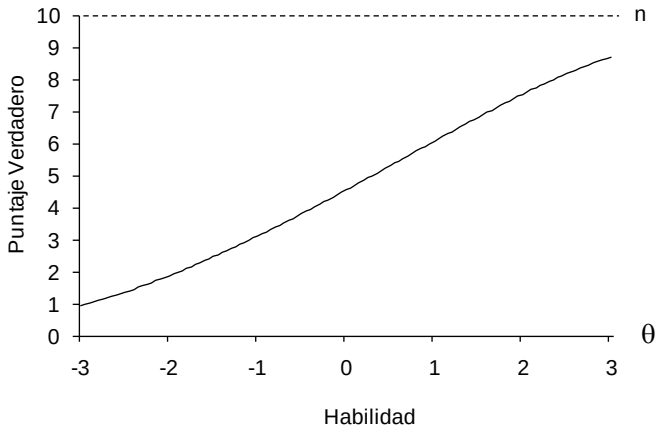
Ahora es posible calcular el puntaje verdadero sumando todas las probabilidades calculadas:

$$\tau(1) = \sum_{i=1}^{10} P_i(1) = 0.96 + 0.95 + 0.92 + 0.79 + 0.73 + 0.57 + 0.45 + 0.29 + 0.21 + 0.18 = 6.07$$

Por lo tanto, de acuerdo con el modelo de Rasch, un sustentante con habilidad $\theta = 1$, tendrá un puntaje esperado en esta prueba de *6.07*, esto quiere decir que se espera que responda aproximadamente seis de las 10 preguntas de la prueba.

Si se calcula el puntaje verdadero para todos los niveles de la escala de habilidad, se obtiene lo que se conoce como *Curva característica de la prueba*. En la figura 17 se muestra la curva característica de la prueba del ejemplo anterior.

Figura 17. Curva característica de la prueba (puntaje verdadero)



La curva característica de la prueba indica la relación entre la escala de habilidad y el puntaje verdadero. Esta curva es monótona creciente, parecida a la curva característica del ítem, pero su forma estará determinada por factores como el número de reactivos en la prueba, el modelo de la TRI empleado y los valores de los parámetros de los reactivos. En todos los casos esta curva será asintótica al valor del número total de reactivos en la prueba *n*. En los modelos de uno y de dos parámetros la cola inferior de esta curva se aproximará a cero, mientras que en el modelo de tres parámetros la cola inferior de la curva característica de la prueba se aproximará a la suma de los parámetros de pseudo-adivinación de los *n* reactivos.

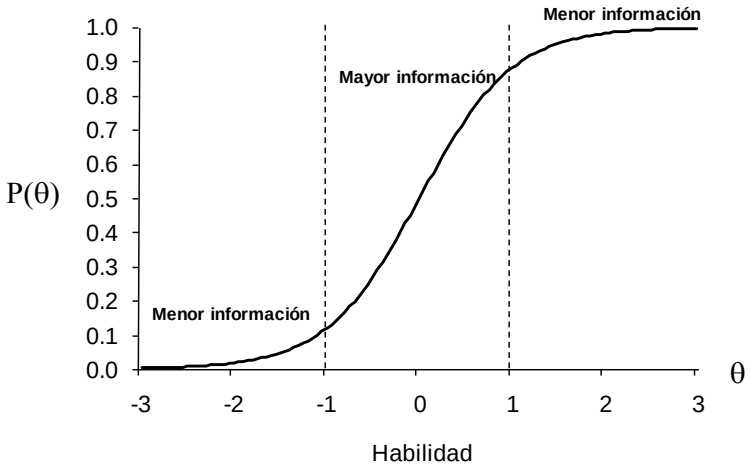
La función de información del ítem

Uno de los problemas típicos de las pruebas estandarizadas es la existencia de reactivos inadecuados para medir a determinados sustentantes. En un examen de matemáticas, por ejemplo, habrá reactivos que resulten muy difíciles para los sustentantes de habilidad baja y que en lugar de proporcionar información útil sobre

este grupo, contribuyen al error de medida de sus habilidades. Del mismo modo habrá reactivos en extremo fáciles para los sustentantes de habilidad alta, que no ayudan a discriminar entre los integrantes de este grupo.

Siguiendo con la idea anterior, un reactivo proporcionará mayor información sobre los sustentantes cuyo nivel de habilidad esté cerca de su dificultad, que de los sustentantes con una habilidad alejada de este punto[12] (ver figura 18).

Figura 18. Información del reactivo (*b = 0*)



La función de información del ítem es una noción fundamental de los modelos de la TRI, ya que permite, entre otras cosas, seleccionar reactivos para el ensamble de versiones, comparar pruebas y definir la precisión de la medida. La función de información de un reactivo se define mediante la siguiente fórmula (Birnbaum, 1968)[13]:

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \qquad i = 1, 2, ..., n$$

[12] En los modelos de uno y de dos parámetros, el punto de la escala de habilidad en que el reactivo alcanza su mayor información es b_i , en el modelo de tres parámetros este máximo se da en un nivel de habilidad ligeramente mayor al parámetro b_i .

[13] La función de información se define en términos de la varianza de la distribución condicional del estimador de máxima verosimilitud de la habilidad ($\hat{\theta}$).

donde:

$I_i(\theta)$ = Información suministrada por el reactivo *i* en el nivel de habilidad θ

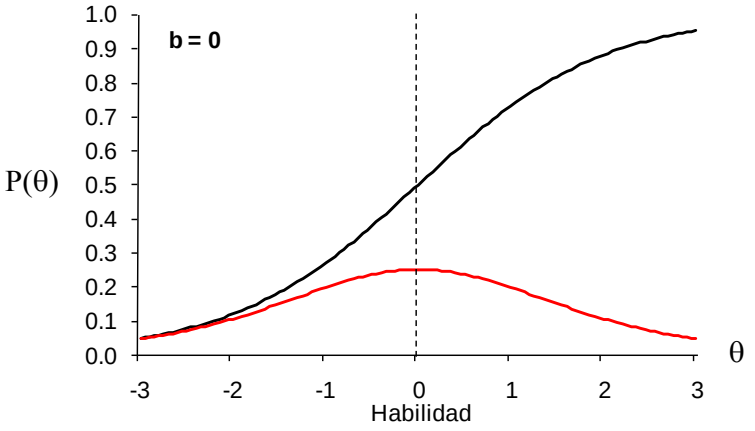
$P_i(\theta)$ = Probabilidad de que un examinado con habilidad θ responda correctamente el reactivo *i*

$P_i'(\theta)$ = Derivada de $P_i(\theta)$ con respecto a θ

$Q_i(\theta) = 1 - P_i(\theta)$

A partir de esta función puede demostrarse que la información aumenta cuando el valor del parámetro de dificultad b_i es cercano a θ, cuando el parámetro de discriminación a_i crece o cuando el parámetro de pseudo-adivinación c_i decrece. Esta propiedad se ejemplifica en las siguientes gráficas[14].

Figura 19. CCI y función de información del ítem, efecto del parámetro de dificultad

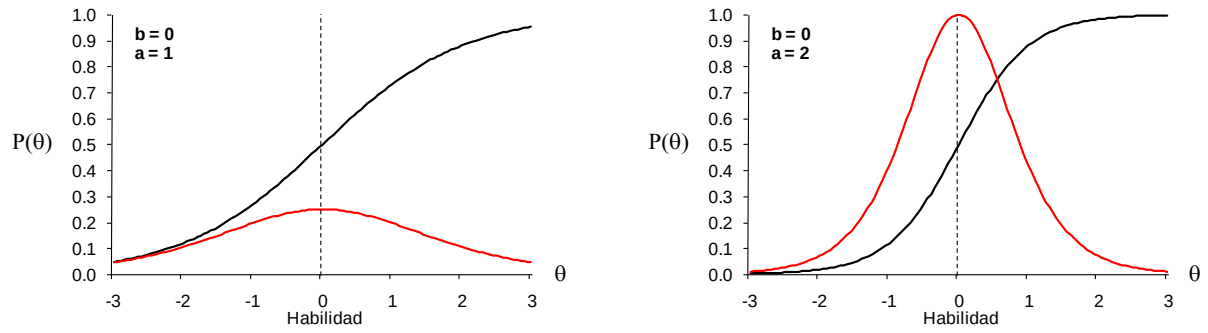


En la figura 19, la gráfica de color rojo es la función de información del reactivo y alcanza su nivel máximo cuando $\theta = b_i$; mientras θ se aleja de b_i la información disminuye[15].

[14] En las gráficas de las figuras 19, 20 y 21 el eje vertical representa la probabilidad de respuesta correcta para la CCI y la cantidad de información para la función de información del ítem.

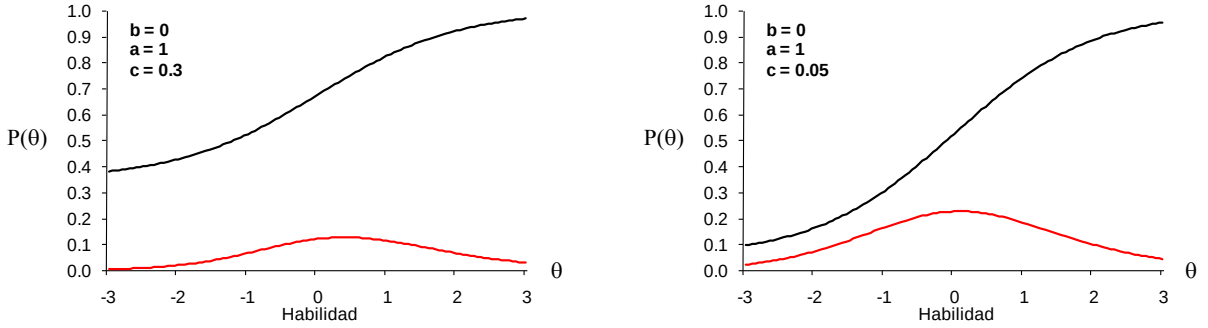
[15] Recordar que en el modelo de tres parámetros la cantidad máxima de información se da en un nivel de habilidad ligeramente mayor al parámetro b_i.

Figura 20. CCI y Función de información del ítem, efecto del parámetro de discriminación



En la figura 20 puede notarse como aumenta la información, al aumentar el valor del parámetro de discriminación del reactivo.

Figura 21. CCI y Función de información del ítem, efecto del parámetro de pseudo-adivinación



De acuerdo con la figura 21, si el valor del parámetro de pseudo-adivinación disminuye, la información aumenta.

La función de información de la prueba

Una prueba es un conjunto de reactivos; entonces, la información de la prueba en el nivel de habilidad θ se define como la suma de la información de todos los reactivos que la componen, en dicho nivel de habilidad.

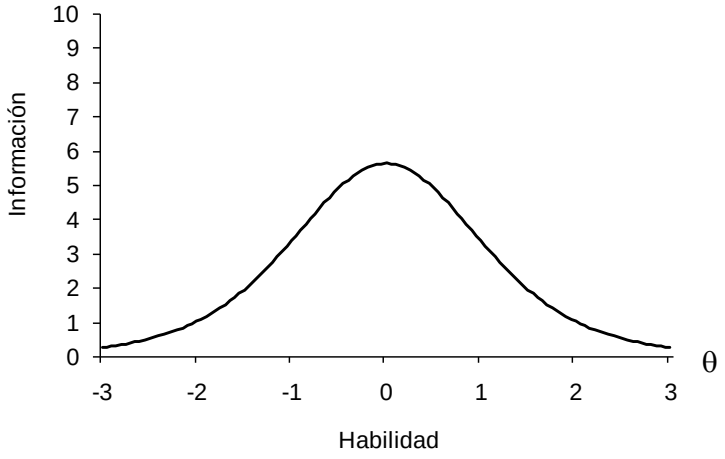
$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

donde:

$I(\theta)$ = Cantidad de información de la prueba en el nivel de habilidad θ
$I_i(\theta)$ = Información suministrada por el reactivo *i* en el nivel de habilidad θ
n = Número de reactivos en la prueba

En la figura 22 se muestra un ejemplo de función de información de la prueba.

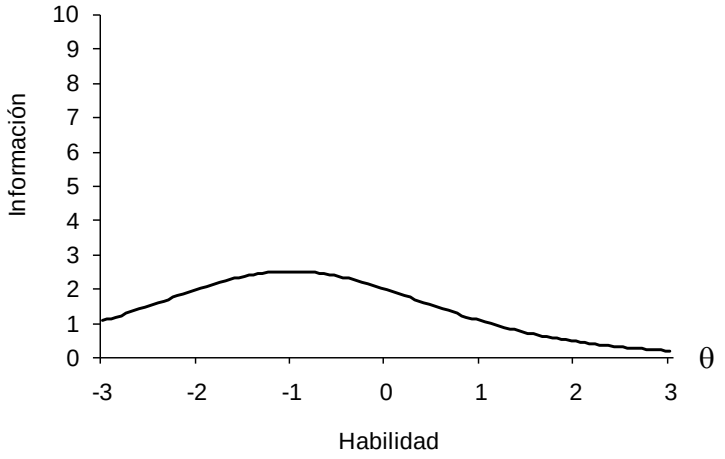
Figura 22. Función de información de la prueba (1)



Esta curva indica con qué precisión está midiendo la prueba en los diferentes puntos de la escala. En el ejemplo anterior, los sustentantes con un nivel de habilidad cercano a cero estarían siendo medidos con una mayor precisión que los sustentantes con niveles de habilidad en los extremos de la escala.

La función de información varía dependiendo de los parámetros de los reactivos que conforman la prueba; por lo tanto, la forma de esta curva puede cambiar notablemente de una prueba a otra. En la figura 23 se muestra una función de información con un nivel general de información menor a la mostrada en la figura 22. Además, esta curva alcanza su información máxima en un nivel de habilidad de -1.

Figura 23. Función de información de la prueba (2)



Nótese que el nivel de información de la prueba es más grande que el nivel mostrado por cualquiera de los reactivos. Esto se debe a que la prueba en su conjunto mide con mucha mayor precisión que un reactivo de manera individual. Por lo general, la precisión de una aprueba aumentará mientras más reactivos contenga.

Esta curva es muy útil en el ensamble de versiones. En el caso de exámenes criteriales, por ejemplo, se buscaría considerar los reactivos que promuevan que la función de información se maximice en el punto de corte de la prueba. Por otro lado, en el caso de pruebas en las que el objetivo es medir con precisión a lo largo de toda la escala, se esperaría tener una función de información de la prueba que fuera lo más parecida posible a una línea horizontal.

Error estándar de estimación

La información de la prueba está ligada a la precisión con que se está estimando el nivel de habilidad de los sustentantes; entre mayor información, mayor precisión en la estimación. El error de estimación de la prueba en el nivel de habilidad θ guarda una relación inversa con la información de la prueba en ese punto y se define por la siguiente expresión[16]:

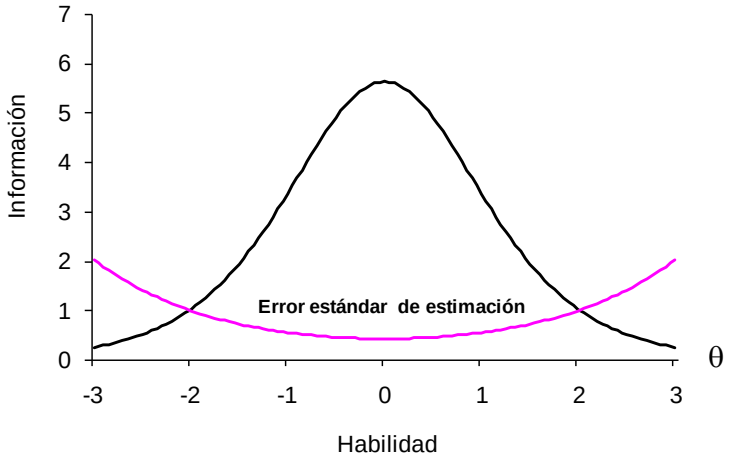
$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

donde:

$SE(\hat{\theta})$ = Error estándar de estimación de θ
$I(\theta)$ = Cantidad de información de la prueba en el nivel de habilidad θ

En la TRI el error estándar de estimación juega el mismo papel que el error estándar de medida en la Teoría Clásica de los Tests, con la gran diferencia de que en este caso el error depende del nivel de habilidad que se está estimando.

Figura 24. Función de información y error estándar de estimación de la prueba (1)

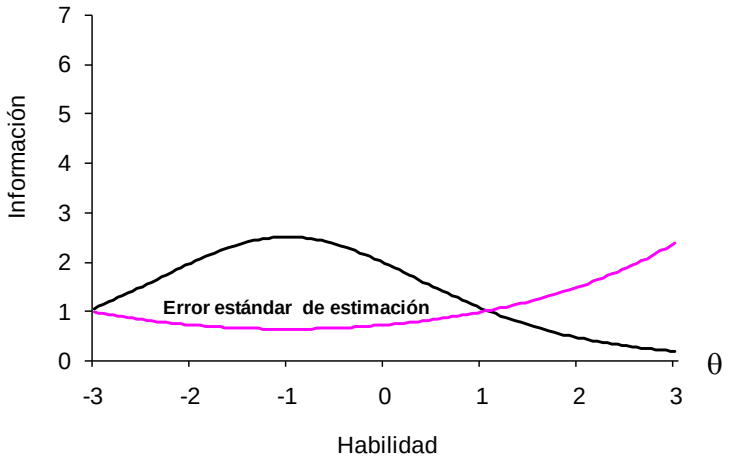


[16] Tomando en cuenta que $\hat{\theta}$ tiene una distribución normal asintótica con media θ y varianza $\sigma^2 = 1/I(\theta)$ (Baker y Kim, 2004). El estimador es eficiente al alcanzar la Cota Inferior de Cramér-Rao.

En la figura 24 se muestra la relación entre la función de información de la prueba y el error estándar de estimación (a mayor información, menor error estándar de estimación). Además, puede corroborarse que (en este ejemplo) la prueba mide con menor error a los sustentantes cuya habilidad está cercana a cero, mientras que el error es grande en los extremos de la escala de habilidad.

La cantidad de error es inversamente proporcional a la información; por lo tanto, cuando se tiene una función de información de la prueba con un nivel general bajo, el error de estimación será más grande. Esto puede verificarse en el ejemplo mostrado en la figura 25; en este caso, la mayor precisión se logra en la medición de sustentantes con un nivel de habilidad de -1.

Figura 25. Función de información y error estándar de estimación de la prueba (2)

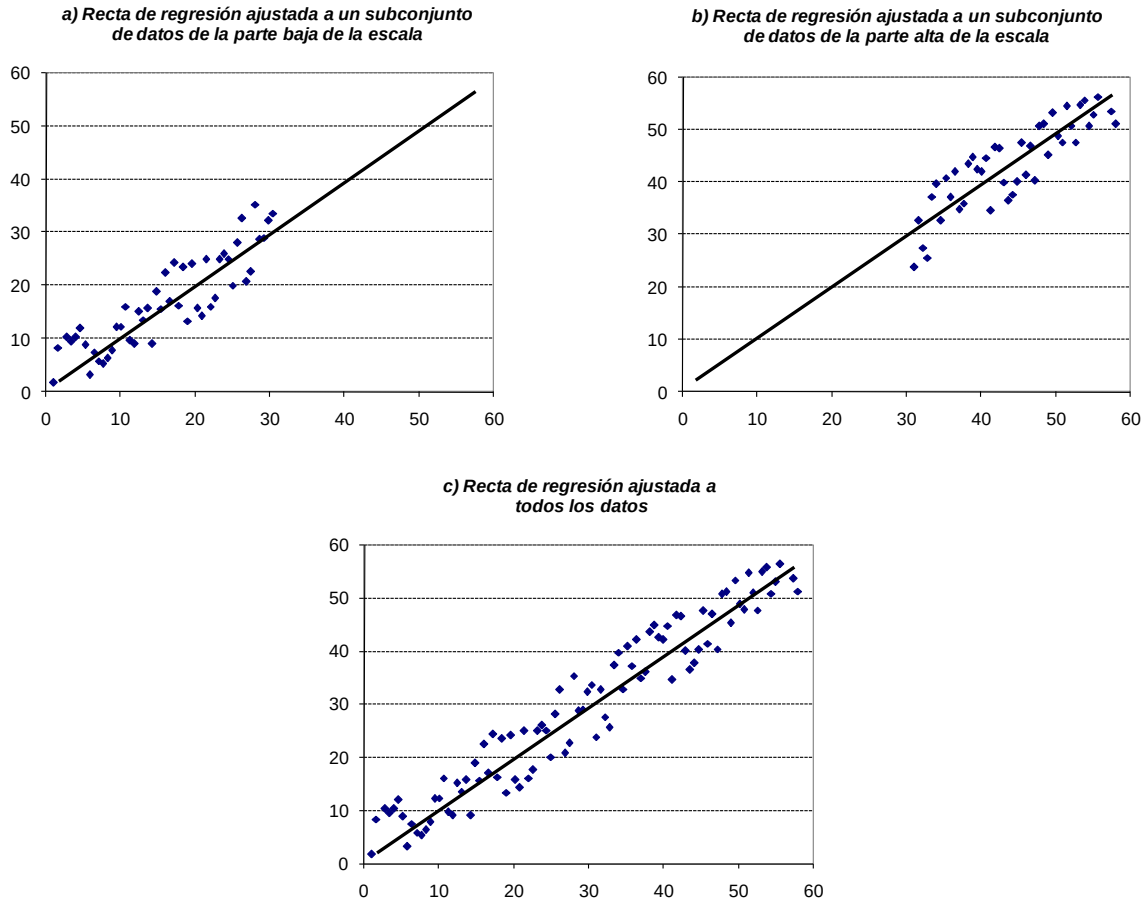


La propiedad de invarianza de los modelos de la TRI

La propiedad de invarianza es una de las principales características de los modelos de la TRI. Esta propiedad establece que los parámetros de los reactivos no dependen de la distribución de habilidad del grupo de sustentantes con el que se lleve a cabo el análisis. Además, que el parámetro de habilidad de los sustentantes no depende de la muestra de reactivos que componen la prueba.

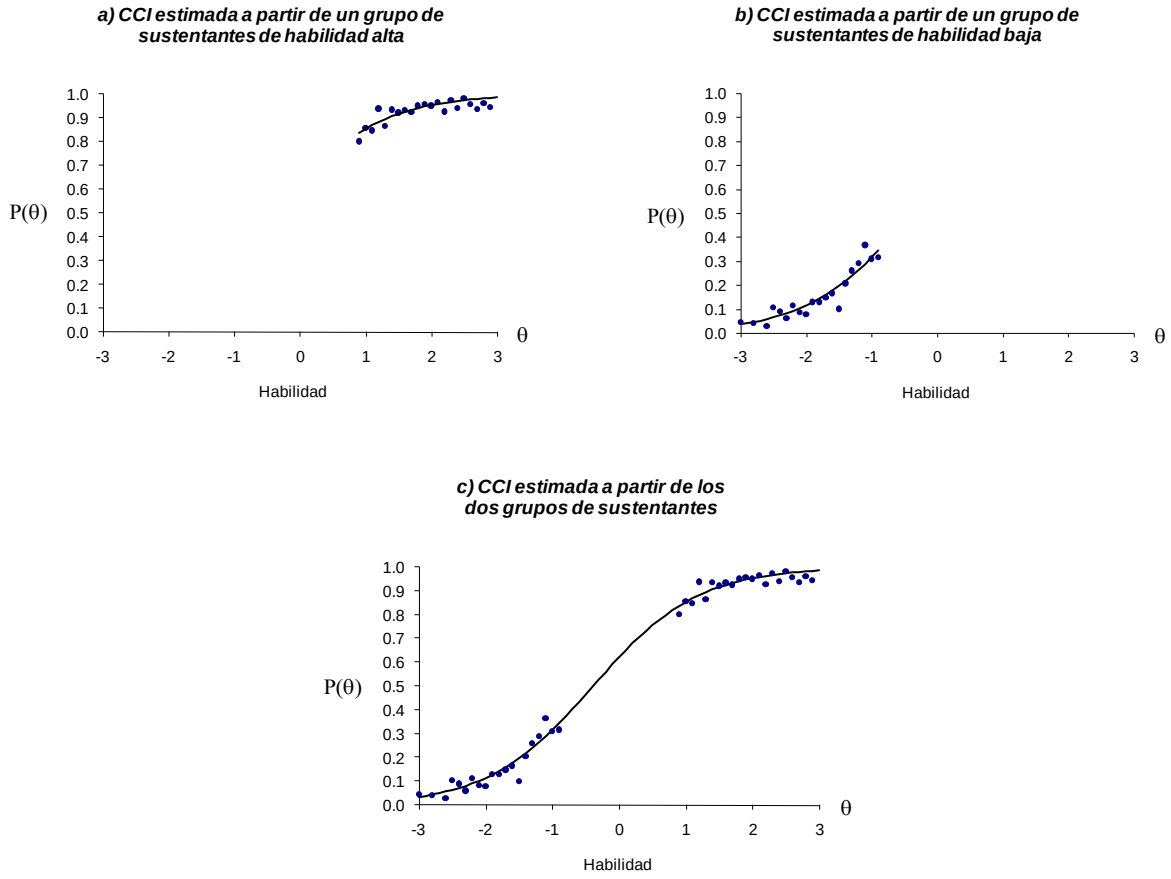
Aun cuando la propiedad de invarianza puede resultar sorprendente, es posible utilizar el modelo de regresión lineal para explicarla. Cuando el modelo de regresión se ajusta de manera adecuada a los datos, se obtendrá la misma recta de regresión (a reserva del error estándar de estimación) para cualquier subconjunto de datos observados (ver figura 26).

Figura 26. Propiedad de invarianza de grupo ejemplificada por el modelo de regresión lineal



De la misma forma, los modelos de la TRI estimarán la misma CCI a partir de muestras diferentes de sustentantes de la población objetivo (a reserva del error estándar de estimación. Ver figura 27).

Figura 27. Propiedad de invarianza de grupo de la TRI



Estimación de la habilidad y de los parámetros de los reactivos

El éxito de la aplicación de los modelos de la TRI radica en gran medida, en disponer de procedimientos que permitan estimar los parámetros de los reactivos y las habilidades de los sustentantes de manera satisfactoria.

La estimación de las habilidades cuando se conocen los parámetros de los reactivos y la estimación de los parámetros de los reactivos cuando se conocen las habilidades de los sujetos, se soluciona de manera adecuada utilizando el procedimiento de máxima verosimilitud. Sin embargo, la situación más común es que tanto los parámetros de los reactivos como las habilidades de los sustentantes sean desconocidos.

En una evaluación típica, el insumo para realizar el análisis son las respuestas de cada uno de los *N* sustentantes a las *n* preguntas que contiene la prueba. Esto lleva a tener

que estimar *N+n* parámetros de manera conjunta si se recurre al modelo TRI de un parámetro, *N+2n* para el modelo de dos parámetros y *N+3n* para el modelo de tres parámetros. Antes de comenzar la estimación es necesario resolver un problema de *identificación* que surge debido a que tanto los parámetros de los reactivos como las habilidades de los sujetos son desconocidos.

En los modelos de la TRI, la probabilidad de respuesta correcta, que está dada por la función de respuesta al ítem, $P_i(\theta)$, permanece invariante ante ciertas transformaciones lineales de los parámetros θ, *a* y *b*.

$$\theta_J = A\theta_I + B$$

$$a_J = \frac{a_I}{A}$$

$$b_J = Ab_I + B$$

$$P(\theta_J) = \frac{1}{1+e^{-a_J(\theta_J - b_J)}} = \frac{1}{1+e^{-\frac{a_I}{A}[(A\theta_I+B)-(Ab_I+B)]}} = \frac{1}{1+e^{-a_I(\theta_I - b_I)}} = P(\theta_I)$$

Para eliminar esta indeterminación, en el modelo de un parámetro es necesario fijar la media de la escala de habilidad (o de la escala de dificultad), mientras que en los modelos de dos y tres parámetros, además se requiere fijar la desviación estándar de la escala. Por este motivo, la interpretación de la escala depende de cómo está definida en cada uno de los programas que se utilizan para estimar los modelos de la TRI. Sin embargo, una vez definida la escala, las diferencias entre sustentantes o reactivos tendrán el mismo significado. Generalmente los programas de cómputo definen la escala de habilidad-dificultad estandarizando la dificultad de los reactivos o la habilidad de los sujetos. Por tal motivo, de manera general, los resultados son expresados en una escala que toma valores que van de -3 a 3.

Una vez resuelto el problema de identificación es posible proseguir con la estimación de los parámetros. Una de las opciones es la estimación de máxima verosimilitud conjunta (*JMLE* por sus siglas en inglés). Esta estimación debe hacerse en dos etapas. En la primera se eligen valores iniciales para la habilidad, generalmente se utiliza el logaritmo

del cociente del número de respuestas correctas dividido entre el número de respuestas incorrectas de cada sustentante. Estos valores se estandarizan (para evitar el problema de identificación) y se estiman los parámetros de los reactivos utilizando el método de máxima verosimilitud. En la segunda etapa, se toman los valores estimados de los parámetros de los reactivos para estimar la habilidad de los sujetos (nuevamente a partir del método de máxima verosimilitud). Este procedimiento se repite hasta que los valores estimados no cambien significativamente de una etapa a otra (Hambleton y otros, 1991).

Baker y Kim (2004) presentan una serie de técnicas para la estimación de de los modelos TRI, entre las que se encuentran la *estimación por maxima verosimilitud condicional (CMLE), la estimación por maxima verosimilitud marginal (MMLE)* y el algoritmo *de maximización de expectativas (EM).*

Ajuste del modelo a los datos

Los modelos de la TRI cuentan con diversas aplicaciones para resolver determinados problemas en evaluación. Sin embargo, muchas de las propiedades de estos modelos –incluyendo la invarianza de los parámetros de los reactivos y de la habilidad de los sustentantes– sólo se alcanzarán de manera satisfactoria si los modelos se ajustan correctamente a los datos.

Las CCI tratan de modelar la proporción de respuestas correctas de diferentes grupos de habilidad; en este sentido, muchos de los estadísticos que se emplean para determinar la bondad del ajuste de los modelos de la TRI se basan en el cálculo de las diferencias entre las probabilidades predichas por el modelo y las proporciones de respuesta correcta observadas a partir de los datos.

En el programa BILOG-MG (Zimowski *et al*., 2003), por ejemplo, se utiliza un estadístico de prueba ji-cuadrada que se calcula bajo el siguiente procedimiento (Embretson y Reise, 2000):

En primer lugar, el programa estima los parámetros de los reactivos y la habilidad de los sustentantes. Con base en esta habilidad se construyen *G* intervalos a lo largo de la

escala de habilidad, y dentro de ellos se clasifica a cada uno de los sustentantes. A continuación se calcula la proporción de respuestas correctas alcanzada en cada grupo y se compara con la probabilidad estimada por el modelo seleccionado, de acuerdo con la siguiente fórmula:

$$\chi^2_G = 2\sum_{g=1}^{G}\left[R_g \log\frac{R_g}{N_g P_g(\theta_M)} + (N_g - R_g)\log\frac{R_g}{N_g(1-P_g(\theta_M))}\right]$$

donde:

G = Número de intervalos o grupos de habilidad
χ^2_G = Estadístico ji-cuadrada con G grados de libertad
R_g = Proporción de respuestas correctas en el grupo g
$P_g(\theta_M)$ = Probabilidad (o proporción de respuestas correctas) predicha por el modelo, basada en la habilidad media de los sustentantes del grupo g
N_g = Número de sustentantes en el grupo g

Este estadístico prueba la hipótesis nula de que el reactivo se ajusta de manera adecuada a los datos, por lo que en todo momento el objetivo será no rechazar esta hipótesis.

Análisis de distractores

Hasta ahora se han descrito indicadores de las características psicométricas de los reactivos y de la prueba, sin embargo, cuando se ensamblan pruebas con reactivos de opción múltiple, resulta muy importante analizar el comportamiento estadístico de cada una de las respuestas incorrectas (distractores). Como se mencionó anteriormente, existe una probabilidad de contestar este tipo de reactivos por azar y es en relación con esta idea en donde los distractores juegan un papel crucial. Como su nombre lo indica, el principal objetivo de los distractores es llamar la atención los sustentantes que tienen una habilidad baja en el constructo medido disminuyendo la probabilidad de que contesten correctamente por casualidad[17].

[17] Es por esto que en el marco de la TRI se habla de pseudo-azar o pseudo-adivinación en lugar de azar puro.

El primer paso para realizar este análisis de manera gráfica consiste en dividir a los sustentantes en grupos de habilidad (seleccionando como criterio de habilidad, el puntaje observado en la prueba). Dependiendo del tamaño de la muestra, se puede seleccionar a los sustentantes de acuerdo a los quintiles o deciles del criterio de habilidad elegido[18].

Para cada grupo de habilidad se debe calcular el porcentaje de sustentantes que eligió cada una de las posibles opciones de respuesta del reactivo (respuesta correcta y distractores). La tabla 7 ejemplifica este cálculo para un reactivo de cuatro opciones de respuesta en el que la opción correcta es la "B" y las tres restantes corresponden a distractores (puede notarse que la opción correcta está marcada con un asterisco).

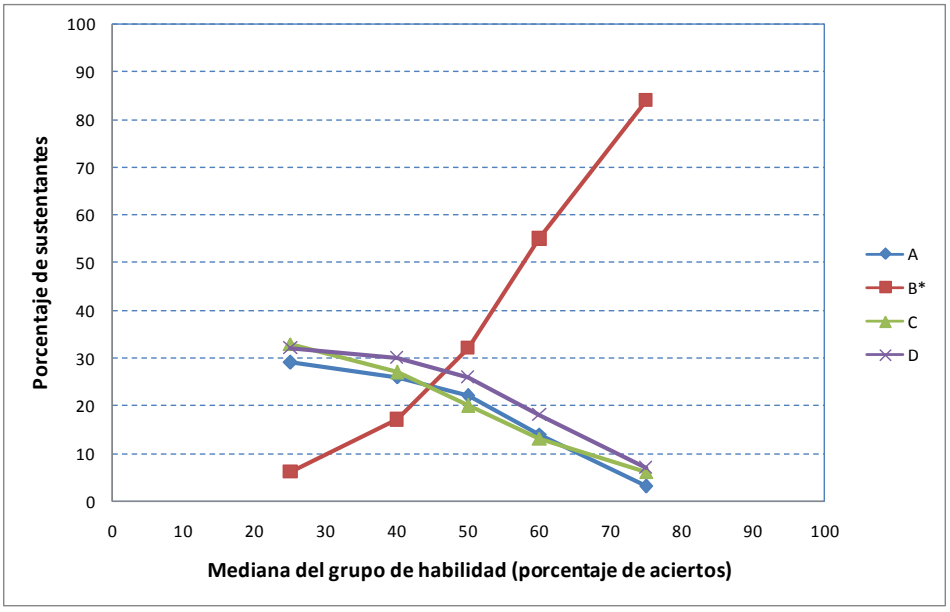
Tabla 7. Porcentaje de sustentantes que eligió cada opción de respuesta (por grupos de habilidad definidos por quintiles)

Grupo	Mediana del grupo (% de aciertos)	Opción de respuesta				Total
		A	B*	C	D	
1	25	29.6	5.6	32.6	32.2	100
2	40	26.0	17.4	26.8	29.8	100
3	50	21.8	32.4	19.5	26.3	100
4	60	14.3	54.5	12.9	18.3	100
5	75	2.6	83.5	6.5	7.4	100

Conforme se incremente la habilidad de los grupos (el grupo 1 es el de menor habilidad y el 5 es el de mayor habilidad), un reactivo con un comportamiento estadístico adecuado debe mostrar un aumento en el porcentaje de sustentantes que seleccionan la opción correcta, y una disminución en el porcentaje de sustentantes que seleccionan los distractores. Si se grafica la mediana de habilidad de cada grupo contra el porcentaje de selección de cada opción de respuesta, pueden visualizarse mejor estos patrones esperados.

[18] Esto permitirá emparejar el tamaño de los grupos e impedir hacer deducciones con un número reducido de sustentantes para alguno(s) de los grupos.

Figura 28. Frecuencia relativa de las opciones de respuesta por grupo de habilidad. Funcionamiento esperado

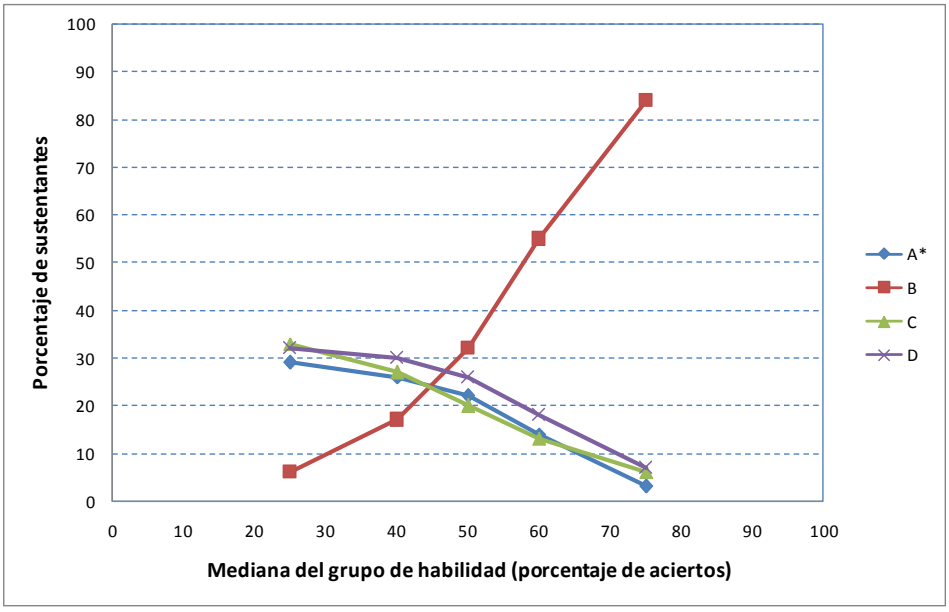


La figura 28 muestra un reactivo de cuatro opciones de respuestas en la que sólo una, la “B”, es la correcta. La curva roja representa la respuesta correcta y puede verse que se comporta como es de esperar en un buen reactivo: lo contestan menos los sustentantes de bajo desempeño (a la izquierda en la gráfica) y más los sustentantes de desempeño alto (a la derecha). Los distractores también tienen un comportamiento adecuado; todos son opciones verosímiles para los grupos de bajo desempeño (su proporción de elección disminuye conforme aumenta la habilidad del grupo) y su inclusión en el reactivo está justificada.

Los patrones irregulares en la gráfica de distractores ayudan a delimitar algunas de las posibles causas del mal funcionamiento del reactivo. A continuación se muestran algunos ejemplos:

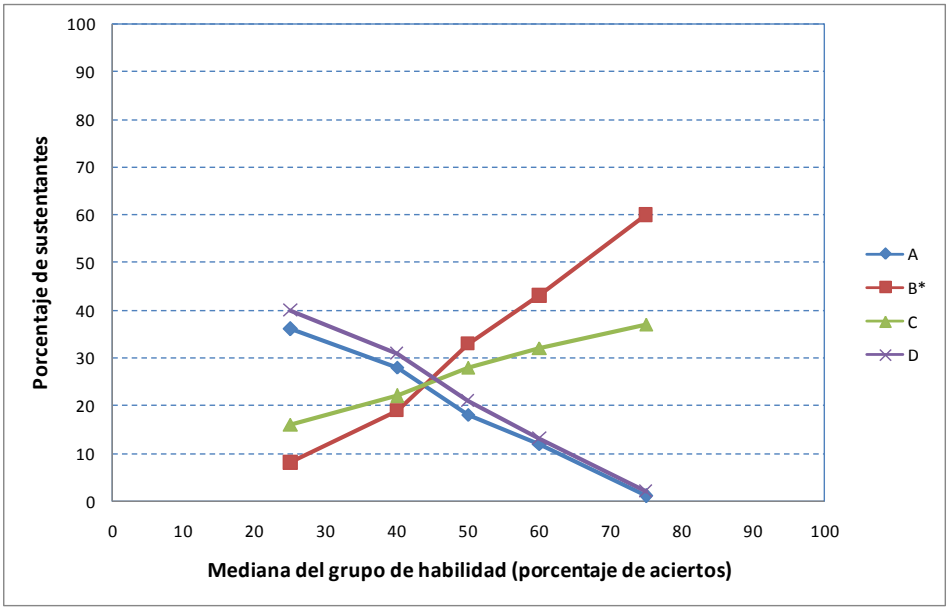
- Cuando un distractor funciona como respuesta correcta y la respuesta correcta como distractor, puede tratarse de un error de codificación en la clave de respuesta correcta (ver figura 29).

Figura 29. Frecuencia relativa de las opciones de respuesta por grupo de habilidad. Funcionamiento no esperado (1)



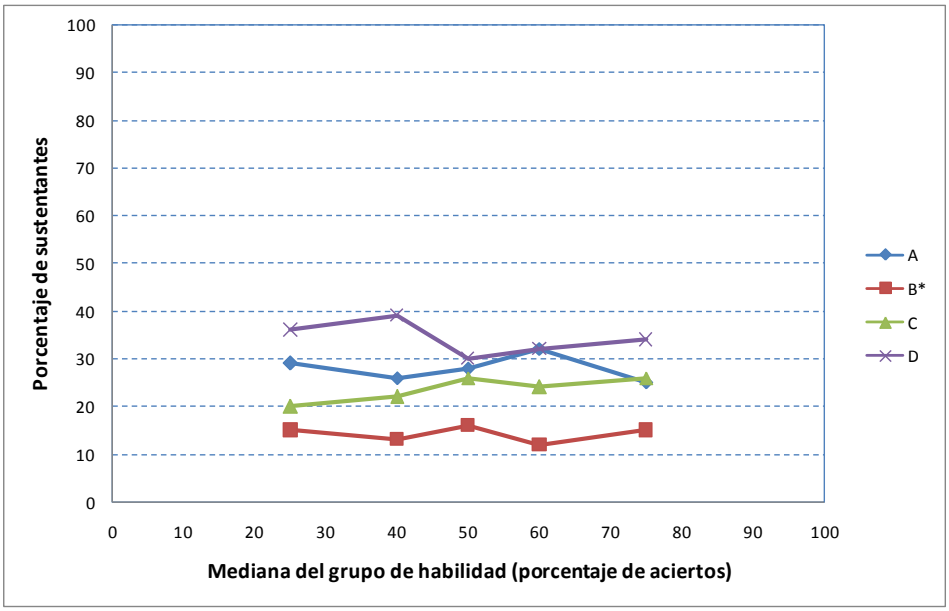
- Cuando uno de los distractores tiene un comportamiento similar al presentado por la respuesta correcta, puede ser que dicho distractor tenga elementos que confundan a los sustentantes de habilidad alta o incluso que el reactivo tenga dos opciones de respuesta correctas (ver figura 30).

Figura 30. Frecuencia relativa de las opciones de respuesta por grupo de habilidad. Funcionamiento no esperado (2)



- Cuando todas las opciones de respuesta (incluyendo la correcta) se comportan de manera similar, puede deberse a la ausencia de una opción de respuesta correcta o a problemas en la redacción del reactivo (ver figura 31).

Figura 31. Frecuencia relativa de las opciones de respuesta por grupo de habilidad. Funcionamiento no esperado (3)



El método gráfico para revisar el funcionamiento de los distractores es muy utilizado y ofrece de manera sencilla una forma de revisar de un vistazo el comportamiento de todos los distractores en todos los rangos de habilidad. Existen, sin embargo, otros métodos. El análisis de los distractores también puede hacerse a partir del cálculo de indicadores de discriminación de cada opción de respuesta, como el índice de discriminación o la correlación punto biserial.

Análisis de funcionamiento diferencial

Existen diferentes métodos estadísticos para detectar reactivos con funcionamiento diferencial. Un método general basado en puntuaciones observadas en el reactivo es el método *Mantel-Haenszel*, que compara la razón de momios de respuesta correcta de los dos grupos de interés para diferentes intervalos de habilidad y puede calcularse e interpretarse de acuerdo al siguiente procedimiento (Holland y Weiner, 1993):

- De los dos grupos de sustentantes que se quieren comparar (por ejemplo mujeres y hombres) se define uno como grupo de referencia (R) y como grupo focal (F).
- Se agrupan los sustentantes de los dos grupos en *k* intervalos te tamaño similar (entre tres y cinco), tomando percentiles del puntaje observado en la prueba.
- Para cada intervalo *i* se construye una tabla de contingencia como la siguiente:

Tabla 8. Tabla de contingencia para el intervalo *i*

		Respuesta al reactivo analizado 1	0	
Grupo	R	a_i	b_i	$a_i + b_i$
	F	c_i	d_i	$c_i + d_i$
		$a_i + c_i$	$b_i + d_i$	$T_i = a_i + b_i + c_i + d_i$

donde:

R = Grupo de referencia
F = Grupo focal
a_i = Número de sustentantes en el grupo de referencia que contestaron correctamente el reactivo
b_i = Número de sustentantes en el grupo de referencia que contestaron incorrectamente el reactivo
c_i = Número de sustentantes en el grupo focal que contestaron correctamente el reactivo
d_i = Número de sustentantes en el grupo focal que contestaron incorrectamente el reactivo

- Se calcula la razón de momios para la probabilidad de respuesta correcta entre los dos grupos (R y F), de acuerdo a la siguiente fórmula.

$$\alpha_i = \frac{p_{Ri}}{q_{Ri}} \Big/ \frac{p_{Fi}}{q_{Fi}} = \frac{\frac{a_i}{a_i + b_i}}{\frac{b_i}{a_i + b_i}} \Big/ \frac{\frac{c_i}{c_i + d_i}}{\frac{d_i}{c_i + d_i}} = \frac{a_i}{b_i} \Big/ \frac{c_i}{d_i} = \frac{a_i d_i}{b_i c_i}$$

donde:

α_i = Razón de momios para el intervalo *i*

p_{Ri} = Proporción de sustentantes en el grupo de referencia que contestaron correctamente el reactivo

p_{Fi} = Proporción de sustentantes en el grupo focal que contestaron correctamente el reactivo

q_{Ri} = Proporción de sustentantes en el grupo de referencia que contestaron incorrectamente el reactivo

q_{Fi} = Proporción de sustentantes en el grupo focal que contestaron incorrectamente el reactivo

La razón de momios indica si hay diferencias en el rendimiento de los dos grupos al contestar el reactivo. De tal forma que si $\alpha_i = 1$, no hay diferencia en el rendimiento de los dos grupos y, por lo tanto, no hay funcionamiento diferencial del reactivo en el intervalo *i*. Si $\alpha_i > 1$, el rendimiento del grupo de referencia fue mejor al mostrado por el grupo focal; cuestión que se revierte si $\alpha_i < 1$.

- Hasta el momento se calculó la razón de momios para cada intervalo de habilidad. El método Mantel-Haenszel estima un estadístico conjunto para todos los intervalos mediante una suma ponderada de las α_i .

$$\alpha = \frac{\sum_i^k \frac{a_i d_i}{T_i}}{\sum_i^k \frac{b_i c_i}{T_i}}$$

donde:

α = Estadístico de Mantel-Haenszel

T_i = Número total de sustentantes en el intervalo *i*

a_i, b_i, c_i y d_i están definidas como antes

Este estadístico toma valores que van de cero a infinito positivo. Aunque es posible aplicar una prueba estadística formal (Ji-cuadrada) para determinar la significancia

en las tablas de contingencia y por ende la presencia de funcionamiento diferencial, algunos autores han preferido transformar el estadístico Mantel-Haenszel en una escala simétrica en la que el valor de cero indica que no existe funcionamiento diferencial del reactivo. Con esto se evita obtener siempre diferencias significativas en el funcionamiento de los reactivos, cuando las muestras son lo suficientemente grandes (Holland y Weiner, 1993). A esta escala se le conoce como Delta (δ) y se define de la siguiente manera:

$$\delta = -2.35 * \ln(\alpha)$$

- La escala Delta es directamente comparable con la escala de dificultad para los reactivos que utiliza el *Educational Testing Service* (ETS). La escala revierte la interpretación del estadístico, de tal modo que valores positivos de Delta indican que el reactivo funciona a favor del grupo focal, mientras que valores negativos indican que el reactivo desfavorece a este grupo.

- El ETS interpreta los valores Delta de la siguiente manera (Ziecky, 2003):

Tabla 9. Interpretación del estadístico Delta de Mantel-Haenszel para funcionamiento diferencial

Categoría DIF del ETS	Valor de Delta
Sin importancia	$\lvert\delta\rvert < 1$
Leve a moderada	$1 \leq \lvert\delta\rvert < 1.5$
Moderada a alta	$1.5 \leq \lvert\delta\rvert$

Capítulo 5. Práctica de análisis de reactivos con Mathematica

En este capítulo se presentan los pasos necesarios para realizar un análisis básico de reactivos. El objetivo es analizar el comportamiento estadístico de 20 reactivos de opción múltiple (con cuatro opciones de respuesta en donde solo una es correcta y el resto son distractores) pertenecientes a una misma área de conocimiento. Estos reactivos se aplicaron a una muestra representativa de la población objetivo de 500 sustentantes.

En este ejercicio se calcularán y analizarán los principales estadísticos postulados por la Teoría Clásica de los Tests (TCT), utilizando un algoritmo desarrollado en el programa Mathematica 8[19] (Wolfram, 2010). Aunque el ejemplo presentado es un caso sencillo, el algoritmo permite realizar análisis de reactivos de *n* versiones con pruebas conformadas por diferentes áreas (en ese caso el programa realiza el análisis independiente de cada una de las áreas o sub-pruebas).

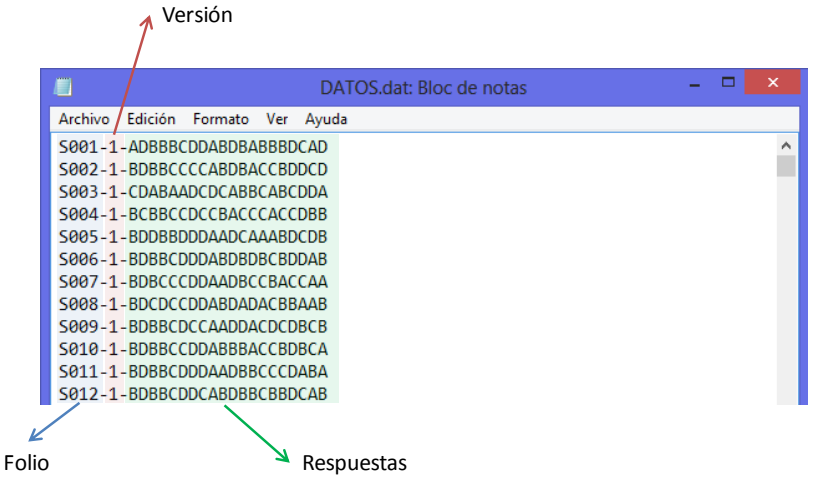
El ejercicio comienza con la descripción de los datos a utilizar (incluidos en los archivos anexos), continúa con la explicación de la configuración del programa y finaliza con la descripción de los resultados.

Datos necesarios para el análisis

Archivo de datos (respuestas de los sustentantes a la prueba)

Como datos adjuntos a este trabajo el lector encontrará tres archivos para realizar la práctica. El primero, “Datos.dat”, es un archivo de texto que contiene las respuestas de 500 sustentantes a una prueba de Matemáticas compuesta por 20 reactivos. Este archivo puede abrirse utilizando el bloc de notas de Windows o cualquier procesador de texto. A continuación se describe la estructura de los datos:

[19] El programa sólo admite pruebas conformadas por reactivos de opción múltiple con cuatro opciones de respuesta.

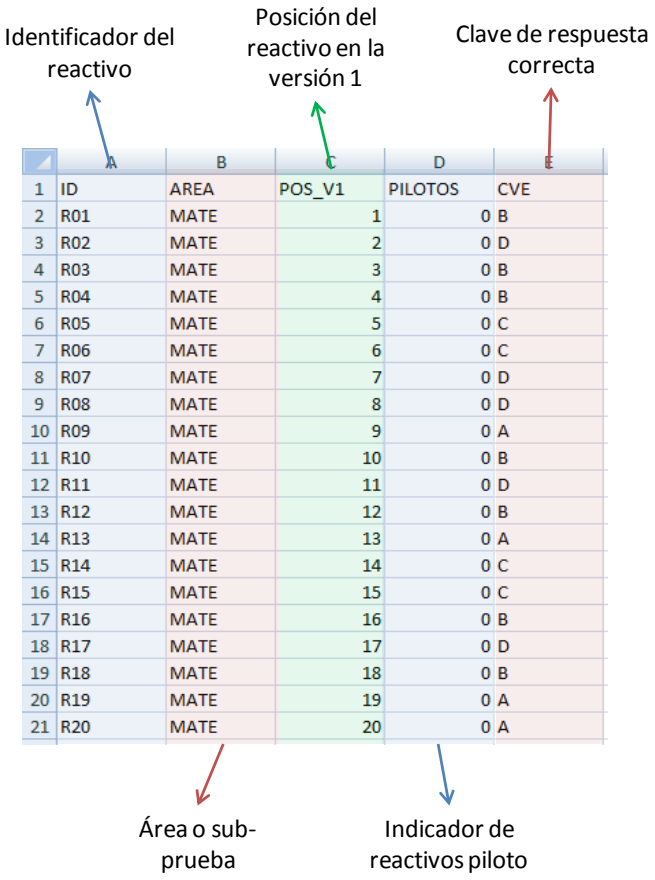


Como puede observarse, cada uno de los renglones del archivo de texto contiene la información de un sustentante. Los primeros cuatro caracteres (de la primera columna a la cuarta columna) contienen el folio del sustentante, este folio debe ser único dentro del archivo. En la quinta columna aparece un guión seguido del número de versión de examen que contestó cada sustentante. En el presente ejemplo, sólo se analizará una versión, sin embargo, es común que con el objetivo de disminuir la probabilidad de copia al momento de la aplicación de la prueba, se apliquen versiones que contienen diferentes reactivos y/o que presentan un ordenamiento diferente de los mismos reactivos.

Finalmente, de la posición ocho a la 27, se pueden visualizar las respuestas de cada sustentante a los 20 reactivos que contiene la prueba. Es importante señalar que en una situación real, además de las letras utilizadas para reconocer cada una de las opciones de respuesta del reactivo, por lo general es posible observar espacios en blanco, cuando un sustentante no seleccionó alguna de las opciones, u otros caracteres que se utilizan para codificar errores, por ejemplo, dobles o triples respuestas.

Plantilla de calificación (estructura del examen)

El siguiente archivo anexo lleva el nombre de “PLANTILLA.csv” y es un archivo de texto separado por comas en el que está expresada la estructura del examen que se aplicó a los 500 sustentantes. Es posible visualizar este archivo en la hoja de cálculo de Excel y contiene la siguiente información:



	A	B	C	D	E
1	ID	AREA	POS_V1	PILOTOS	CVE
2	R01	MATE	1	0	B
3	R02	MATE	2	0	D
4	R03	MATE	3	0	B
5	R04	MATE	4	0	B
6	R05	MATE	5	0	C
7	R06	MATE	6	0	C
8	R07	MATE	7	0	D
9	R08	MATE	8	0	D
10	R09	MATE	9	0	A
11	R10	MATE	10	0	B
12	R11	MATE	11	0	D
13	R12	MATE	12	0	B
14	R13	MATE	13	0	A
15	R14	MATE	14	0	C
16	R15	MATE	15	0	C
17	R16	MATE	16	0	B
18	R17	MATE	17	0	D
19	R18	MATE	18	0	B
20	R19	MATE	19	0	A
21	R20	MATE	20	0	A

Como puede observarse, cada renglón contiene la información de un reactivo en particular. En la primera columna aparece un identificador único del reactivo (ID), en este caso está compuesto por tres caracteres. En la segunda columna está un identificador del área o sub-prueba (AREA), en este ejemplo se está analizando una prueba de Matemáticas, usualmente las evaluaciones incorporan baterías de exámenes que pueden medir constructos diversos. La tercera columna (POS_V1) define la posición de cada reactivo en la versión uno (acomodo de los reactivos en el cuadernillo o forma); en el caso de que se estuvieran analizando *n* versiones (*n* ≥ *2)*, sería necesario incorporar *n* columnas de posición identificadas por el encabezado “POS_Vi”.

Una de las maneras más eficientes de probar (pilotear) reactivos nuevos es incorporándolos en versiones operativas de la prueba, en este sentido resulta muy relevante para el análisis y la calificación identificar los reactivos que se incorporaron con fines de verificación estadística. En la cuarta columna de la plantilla de la prueba (PILOTOS), tendrán que señalarse con un “1” los reactivos que son piloto y con un “0”

los que son operativos (estos últimos son los que se tomarán en cuenta para calificar a los sustentantes). En este ejemplo, no hay reactivos piloto. Es importante señalar que aún cuando los reactivos piloto no forman parte de la calificación de los sustentantes y por lo tanto no se toman en cuenta para estimar la confiabilidad y error de medida de la prueba, el programa sí estimará sus características psicométricas.

Para finalizar la descripción de la plantilla del examen, en la última columna se presentan las claves de respuesta correcta de cada reactivo (CVE). En este caso, al tratarse de reactivos con cuatro posibles opciones, se observan letras de la A a la D.

Para que el programa funcione de manera adecuada, la plantilla del examen deberá contener la información relacionada con las cinco variables explicadas anteriormente, identificadas con los mismos encabezados utilizados en este ejercicio.

Algoritmo de análisis (programa de Mathematica)

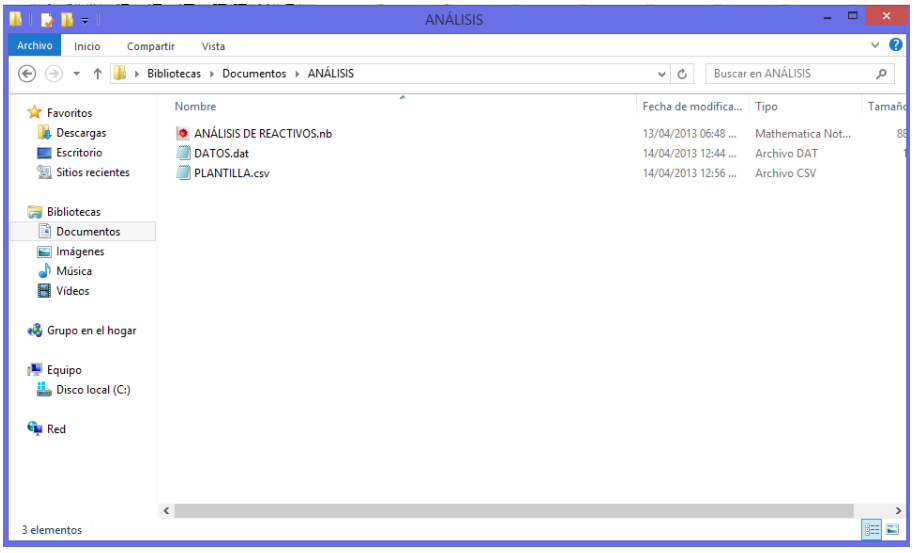
El último archivo adjunto contiene el programa computacional que tomará los dos archivos anteriormente descritos para realizar la mayoría de los análisis de la TCT revisados en este trabajo. El archivo “ANÁLISIS DE REACTIVOS.nb” sólo podrá ser ejecutado y modificado en el programa *Wolfram Mathematica*. Los algoritmos contenidos en este programa son de carácter general, por lo que puede ser utilizado para realizar el análisis de cualquier prueba compuesta por reactivos de opción múltiple (con cuatro opciones de respuesta). En la siguiente sección se explica la configuración básica del archivo, lo que permitirá realizar el análisis del examen ejemplo.

Configuración del programa

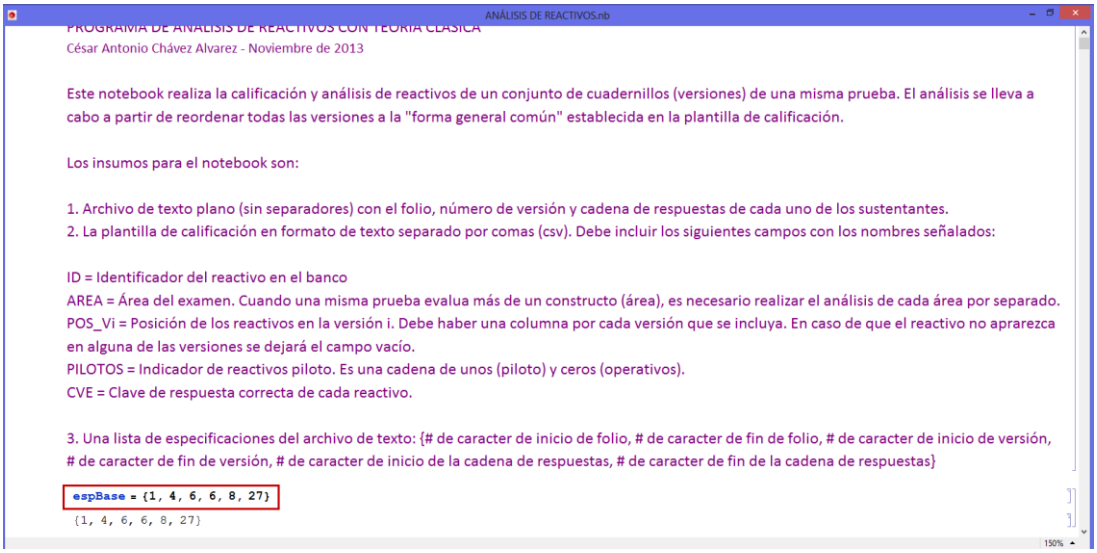
Los *notebooks* de *Mathematica* son interfaces que permiten la programación de alto nivel a partir de miles de funciones y opciones organizadas. Estos documentos están constituidos por celdas, algunas de las cuales realizan cálculos en vivo. El *notebook* desarrollado para realizar el análisis de reactivos a partir de la TCT contiene una primera sección con la explicación general del programa y de los datos requeridos, así como algunas celdas que permitirán configurar el código dependiendo de las características del análisis. Todas estas variables de configuración pueden modificarse

directamente en el *notebook*. Debe tenerse precaución de no cambiar el código contenido en otras celdas.

Antes de comenzar será necesario establecer (crear) un directorio de trabajo en el que deben copiarse los tres archivos adjuntos, en ese mismo directorio se guardarán todos los archivos de salida. Para este ejemplo se creó el directorio “ANÁLISIS” dentro de la carpeta de “Documentos”.

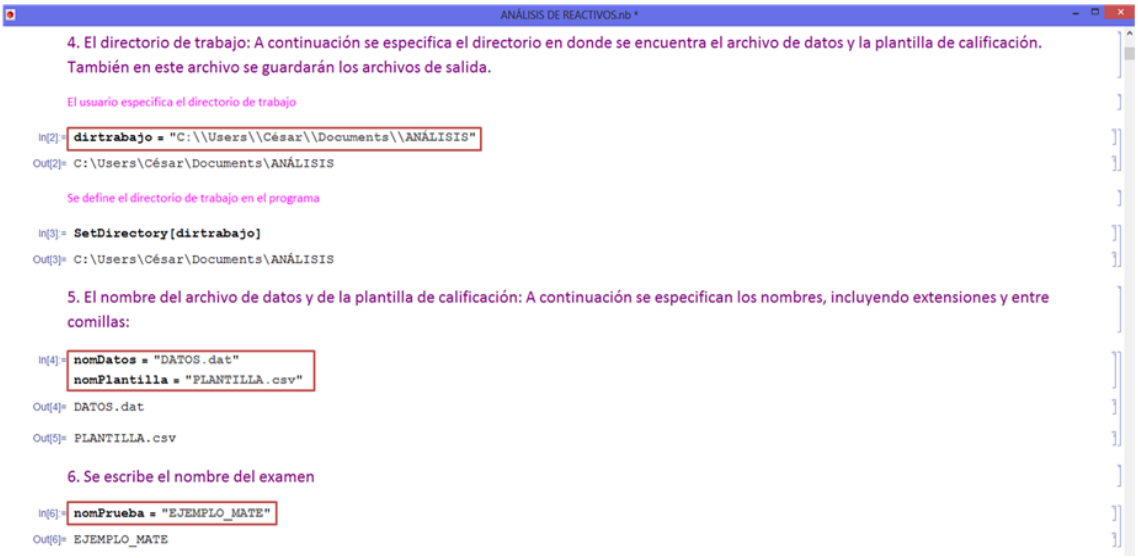


A continuación debe abrirse el archivo “ANALISIS DE REACTIVOS.nb”, en donde la primera variable que se va a configurar es “espBase”. Esta variable es una lista de seis elementos en la que se especifica la estructura del archivo de datos.

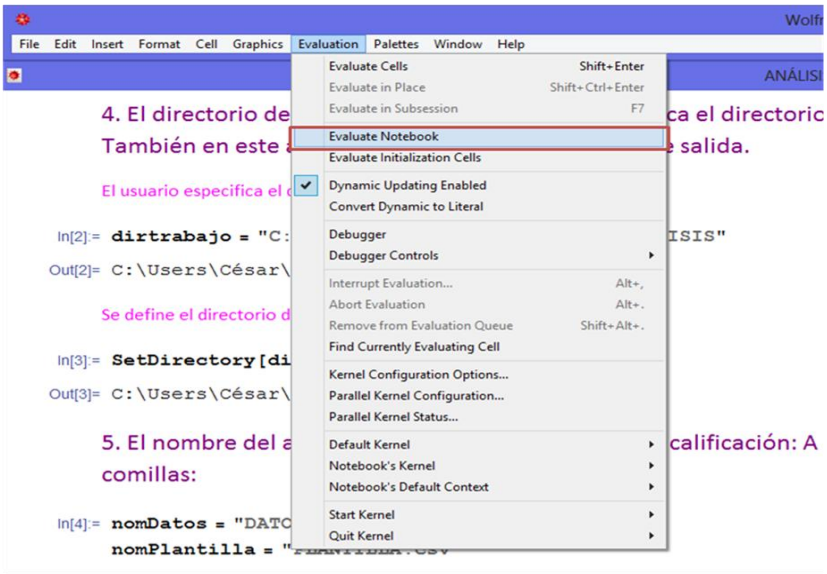


El primer número de esta lista indica la posición de inicio del folio, el segundo la finalización del folio, el tercer número especifica la posición de inicio de la versión, mientras que el cuarto el final de la misma. Los dos últimos números indican las posiciones de inicio y fin de las respuestas de los sustentantes.

A continuación debe especificarse la dirección completa del directorio de trabajo en la variable “dirtrabajo”. En la variable “nomDatos” se escribirá el nombre del archivo de datos y en la variable “nomPlantilla” el nombre del archivo de la plantilla que contiene las características de la prueba; en ambos casos debe incluirse la extensión correspondiente. Finalmente se escribe el nombre con el que el usuario quiera identificar el examen analizado, en la variable “nomPrueba”, en este case se utilizó el nombre “EJEMPLO_MATE”.



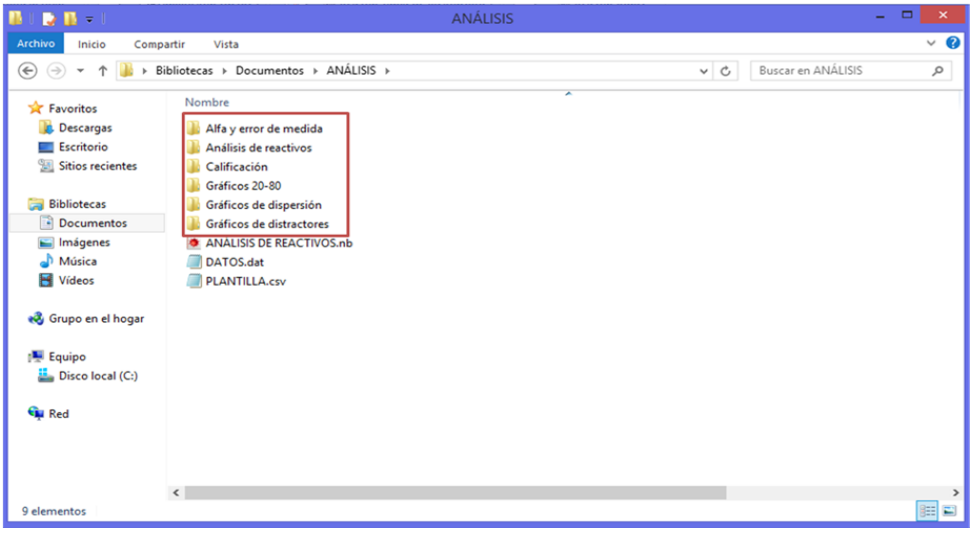
Una vez realizada la configuración anterior se procede a evaluar el *notebook* completo, esto puede realizarse utilizando la función “Evaluate Notebook” en el menú “Evaluation”.



El programa comenzará a EJECUTARSE, lo cual se indicará en el título de la ventana con la palabra “Running”. Al terminar la evaluación esta palabra desaparecerá.

Descripción de los resultados

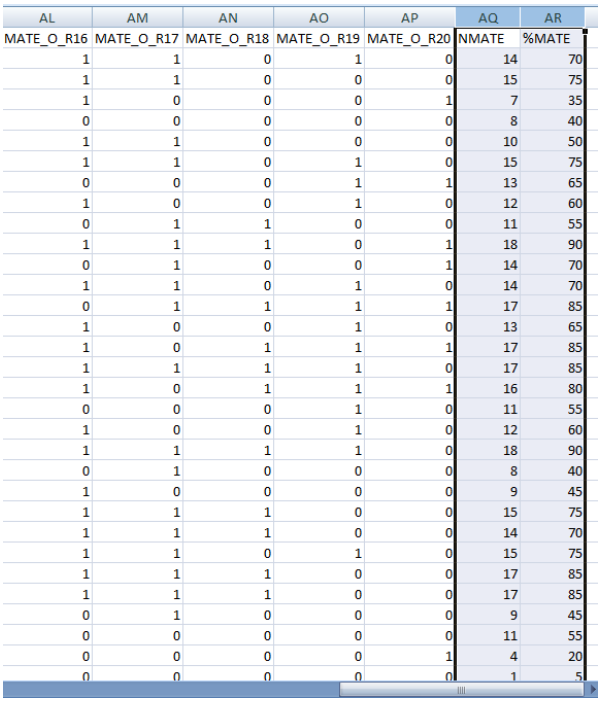
Una vez finalizada la evaluación, el usuario deberá verificar que se hayan creado una serie de carpetas y archivos de salida en el directorio de trabajo.



A continuación se describe el contenido de los seis directorios de resultados establecidos.

Calificación

Este directorio contiene el archivo “baseCalificación.csv”, el cual incluye las respuestas de los sustentantes a cada uno de los reactivos de la prueba, la calificación que obtuvieron en ellos (ceros o unos) y el número y porcentaje de aciertos alcanzado en los reactivos operativos de cada área (calificación).



AL	AM	AN	AO	AP	AQ	AR
MATE_O_R16	MATE_O_R17	MATE_O_R18	MATE_O_R19	MATE_O_R20	NMATE	%MATE
1	1	0	1	0	14	70
1	1	0	0	0	15	75
1	0	0	0	1	7	35
0	0	0	0	0	8	40
1	1	0	0	0	10	50
1	1	0	1	0	15	75
0	0	0	1	1	13	65
1	0	0	1	0	12	60
0	1	1	0	0	11	55
1	1	1	0	1	18	90
0	1	0	0	1	14	70
1	1	0	1	0	14	70
0	1	1	1	1	17	85
1	0	0	1	0	13	65
1	0	1	1	1	17	85
1	1	1	1	0	17	85
1	0	1	1	1	16	80
0	0	0	1	0	11	55
1	0	0	1	0	12	60
1	1	1	1	0	18	90
0	1	0	0	0	8	40
1	0	0	0	0	9	45
1	1	1	0	0	15	75
1	1	1	0	0	14	70
1	1	0	1	0	15	75
1	1	1	0	0	17	85
1	1	1	0	0	17	85
0	1	0	0	0	9	45
0	0	0	0	0	11	55
0	0	0	0	1	4	20
0	0	0	0	0	1	5

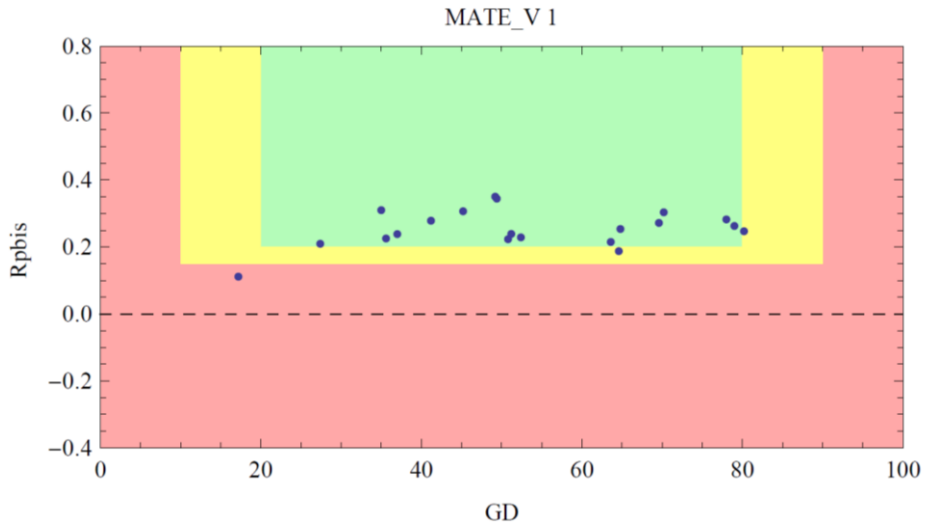
Análisis de reactivos

En este directorio se encuentra el archivo “análisisR.csv”, el cual contiene el análisis de todos los reactivos (operativos y pilotos). El resultado del análisis incluye el identificador del reactivo, el número total de sustentantes que lo respondieron, el grado de dificultad y la correlación punto biserial ajustada por correlación espuria. En este ejemplo el reactivo “R06” muestra una discriminación ligeramente menor a la aceptable (0.20), mientras que el reactivo “R20” está fuera de rango en ambos indicadores (si se considera un intervalo de 20% a 80% para el grado de dificultad). Se recomienda revisar estos dos reactivos para identificar posibles problemas en la manera en que fueron elaborados.

	A	B	C	D
1	REACTIVO	OBS	GD	Rpbis
2	MATE_O_R01	500	78	0.28231605
3	MATE_O_R02	500	79	0.26277637
4	MATE_O_R03	500	63.6	0.21502603
5	MATE_O_R04	500	69.6	0.27187483
6	MATE_O_R05	500	64.8	0.25370701
7	MATE_O_R06	500	64.6	0.18761727
8	MATE_O_R07	500	70.2	0.30356459
9	MATE_O_R08	500	52.4	0.2287517
10	MATE_O_R09	500	80.2	0.24700458
11	MATE_O_R10	500	51.2	0.23896163
12	MATE_O_R11	500	49.4	0.34430556
13	MATE_O_R12	500	49.2	0.35019532
14	MATE_O_R13	500	45.2	0.30670858
15	MATE_O_R14	500	50.8	0.22322567
16	MATE_O_R15	500	37	0.23847728
17	MATE_O_R16	500	35	0.30993979
18	MATE_O_R17	500	35.6	0.22540964
19	MATE_O_R18	500	27.4	0.20963243
20	MATE_O_R19	500	41.2	0.27845824
21	MATE_O_R20	500	17.2	0.11144284
22				

Gráficos de dispersión

El directorio “Gráficos de dispersión” contiene archivos en formato PDF, con representaciones gráficas que complementan el análisis de reactivos. En ellos se presenta la distribución del grado de dificultad y de la correlación punto biserial ajustada por correlación espuria de los reactivos incluidos en cada versión (se genera un gráfico para cada área del examen y uno más para el examen global). Al contar con una sola versión y una sola área para este ejemplo, en este caso el gráfico correspondiente al área de Matemáticas coincide con el gráfico del examen completo.

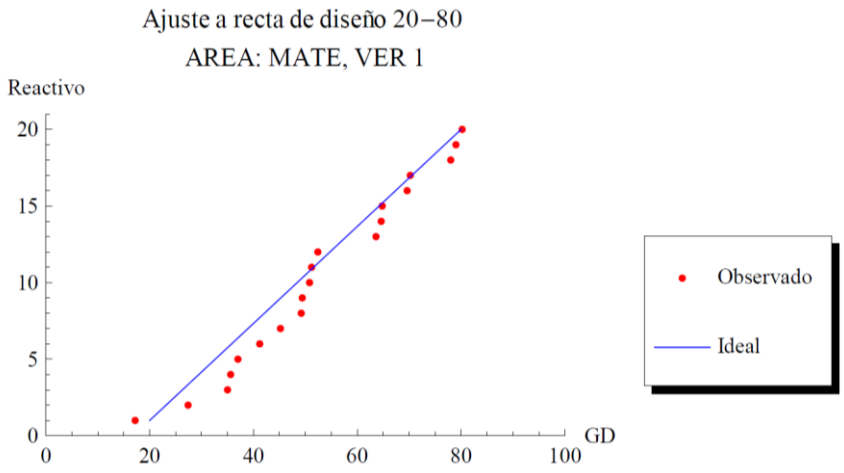


En el gráfico anterior pueden observarse los dos reactivos que tienen valores fuera de rango. En este caso se define un área verde para localizar los reactivos óptimos

($20 \leq GD \leq 80$ y $Rpbis \geq 0.20$), un área amarilla para definir reactivos que están fuera de rango pero que se encuentran cerca de los límites de aceptación ($10 \leq GD < 20$ y $Rpbis \geq 0.15$ ó $80 < GD \leq 90$ y $Rpbis \geq 0.15$ ó $20 \leq GD \leq 80$ y $0.15 \leq Rpbis < 0.20$), estos reactivos pueden definirse como sub-óptimos. Finalmente el área roja muestra los reactivos rechazados por estar lejos de los límites de aceptación para los valores correspondientes.

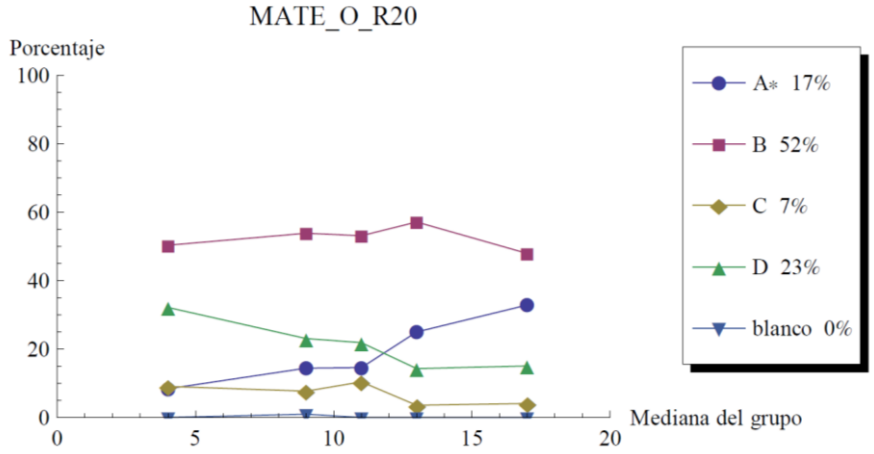
Gráficos 20-80

Uno de los modelos de ensamble de versiones utilizados en el Ceneval consiste en distribuir los reactivos de cada área del examen de manera uniforme en un rango de grado de dificultad de 20% a 80%. Este diseño permite distribuir el error estándar de medida a lo largo de la escala, por lo que es recomendado sobre todo en exámenes de asignación. El directorio “Gráficos 20-80”, contiene archivos en formato PDF en los que se presenta, de manera gráfica, la distribución del grado de dificultad de los reactivos para cada área del examen.



Gráficos de distractores

Este directorio contiene un archivo con el gráfico de distractores (en formato PDF) para cada uno de los reactivos analizados. A continuación se muestra el gráfico correspondiente al reactivo 20.



Aún cuando la proporción de elección de la respuesta correcta “A” para este reactivo va creciendo conforme aumenta la habilidad de los grupos, el distractor “B” parece estar llamando la atención tanto a los sustentantes de habilidad baja como los de habilidad alta, por lo tanto es recomendable realizar una revisión cualitativa de este reactivo.

Adicionalmente a los archivos que contienen los gráficos de distractores, en este directorio se incluye el archivo “porcResp.csv”, en el cual se presenta la proporción de sustentantes que eligió cada opción de respuesta en cada reactivo.

	A	B	C	D	E	F
1	REACTIVO	%A	%B	%C	%D	%blanco
2	MATE_O_R01	11.8	78	5.8	4.4	0
3	MATE_O_R02	8.6	5.6	6.4	79	0.2
4	MATE_O_R03	9	63.6	13.8	13.6	0
5	MATE_O_R04	7.8	69.6	6.6	16	0
6	MATE_O_R05	8	23.6	64.8	3.2	0.4
7	MATE_O_R06	7.8	8.8	64.6	18.6	0
8	MATE_O_R07	8.2	15.4	6	70.2	0.2
9	MATE_O_R08	9.2	4.6	33.4	52.4	0.4
10	MATE_O_R09	80.2	6	7.2	4.6	2
11	MATE_O_R10	23.6	51.2	16	9.2	0
12	MATE_O_R11	14	15	20.8	49.4	0.8
13	MATE_O_R12	7.2	49.2	37.2	6.2	0.2
14	MATE_O_R13	45.2	32.4	13	9.2	0
15	MATE_O_R14	15	19.6	50.8	14.2	0.4
16	MATE_O_R15	22	29.2	37	11.6	0.2
17	MATE_O_R16	20.4	35	35.2	9	0.4
18	MATE_O_R17	21.2	25.8	17.4	35.6	0
19	MATE_O_R18	22.6	27.4	19	30.2	0.8
20	MATE_O_R19	41.2	19.4	26.8	12.4	0.2
21	MATE_O_R20	17.2	52.4	7.4	22.8	0.2

Alfa y error de medida

Este directorio contiene un archivo con la estimación del Coeficiente Alfa de Cronbach y del error estándar de medida, para cada versión analizada. En este ejemplo la

confiabilidad de la única versión analizada es cercana a 0.70, esto quiere decir que el 70% de la variación en los puntajes observados se debe cambios en los puntajes verdaderos (habilidad) de los sustentantes. El examen muestra un valor de confiabilidad modesto que puede mejorar si aumenta el poder discriminativo de los reactivos o se incrementa el número de reactivos que lo componen.

	A	B	C	D
1	Ver_1_500 sustentantes.csv			
2	Área	Num. Reactivos	Alfa de Cronbach	Error Est. de Med.
3	MATE	20	0.681919509	1.989956005
4				

Reflexiones finales

En este trabajo se presentaron las principales técnicas de análisis de reactivos utilizadas en el desarrollo y en la aplicación de instrumentos de evaluación. Los modelos expuestos representan un eslabón fundamental en la cadena de procedimientos seguidos para la elaboración de pruebas educativas.

Aún cuando se sigan estrictos controles de calidad durante las fases de diseño de los instrumentos y de la elaboración de los reactivos, no es hasta que se realiza el análisis cuantitativo de los datos cuando es posible garantizar la confiabilidad de las mediciones realizadas y por lo tanto de las inferencias que se hacen a partir de estas mediciones.

A partir del análisis de reactivos es posible detectar errores que muchas veces son poco transparentes para los elaboradores de reactivos y para los expertos que validan que el reactivo es adecuado para medir lo que pretende medir. Los datos ofrecen una dosis de claridad que difícilmente se logra de otro modo.

Este texto representa solo una pequeña introducción a la gran variedad de métodos cuantitativos que son empleados en el campo de la psicometría, los cuales resultan cada vez más importantes para orientar acciones de mejora educativa. Actualmente estamos más acostumbrados a mirar información como la proveniente del *Programa para la Evaluación Internacional de Alumnos* (PISA, por sus siglas en inglés) desarrollado por la Organización para la Cooperación y Desarrollo Económicos

(OCDE), por lo que entender cada uno de los detalles de los procesos de evaluación resulta fundamental.

Anexo 1. Diseño matricial

El diseño matricial es un procedimiento de ensamble de cuadernillos que permite distribuir los reactivos de forma tal que se cubran todos los contenidos contemplados en la estructura. Bajo este diseño no es necesario que a todos los sustentantes se les aplique la prueba completa. En el siguiente ejemplo, los 240 reactivos de un examen hipotético son piloteados en 8 cuadernillos diferentes. Para garantizar la representatividad de la muestra para cada uno de los reactivos bajo análisis, es necesario aplicar todos los reactivos de manera aleatoria entre los sustentantes elegidos.

Tabla 10. Número de reactivos por área para el piloteo de la estructura completa

	Áreas			
	Matemáticas	*Español*	*Razonamiento Lógico-Matemático*	*Razonamiento Verbal*
Cuadernillo 1	**15**	**15**		
Cuadernillo 2		**15**	**15**	
Cuadernillo 3			**15**	**15**
Cuadernillo 4	**15**			**15**
Cuadernillo 5	**15**	**15**		
Cuadernillo 6		**15**	**15**	
Cuadernillo 7			**15**	**15**
Cuadernillo 8	**15**			**15**
Total	**60**	**60**	**60**	**60**

El diseño matricial comúnmente se utiliza para hacer inferencias a nivel agregado (escuela, municipio, estado, sistema educativo, etc.), sin la necesidad de que todos los sustentantes respondan la prueba completa. En el caso del piloteo de reactivos, este tipo de diseños es conveniente cuando se trata de exámenes muy largos y se puede garantizar que en cada cuadernillo habrá un grupo de reactivos que alcancen parámetros adecuados y que puedan servir como referente de calibración. En el caso del piloteo de exámenes pequeños, es necesario introducir un grupo de reactivos con parámetros psicométricos conocidos (ancla), a partir de los cuales se pueda inferir de manera inicial la habilidad de los sustentantes en el rasgo medido.

Anexo 2. Deducción de la fórmula del Coeficiente Alfa de Cronbach

El propósito de este anexo es derivar la fórmula del Coeficiente Alfa de Cronbach, demostrando que la confiabilidad de una prueba compuesta por “k” pruebas paralelas, puede expresarse como una función de la varianza de los puntajes de la prueba completa y de las covarianzas de los puntajes de las pruebas paralelas que la componen. Para entender mejor la derivación de la fórmula deberán tomarse en cuenta las siguientes definiciones:

1. Para cada pareja de pruebas *i* y *j*, la covarianza de los puntajes observados en dichas pruebas se define como:

   $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$

2. Cuando *i* y *j* son estrictamente paralelas, la varianza de los puntajes verdaderos de la prueba *i* es igual a su covarianza con los puntajes verdaderos de la prueba *j*.

   $\sigma^2_{V_i} = \sigma_{V_iV_j}$

3. Para cualquier par de pruebas *i* y *j*, la covarianza entre puntajes verdaderos es igual a la covarianza entre puntajes observados.

   $\sigma_{V_iV_j} = \sigma_{ij}$

Una vez especificado lo anterior, definimos el puntaje en la prueba “compuesta” como una suma de puntajes en *k* pruebas paralelas, $C = A + B + ... + K$. El puntaje verdadero en la prueba completa es $V_C = V_A + V_B + ... + V_K$, por lo tanto, la varianza de este puntaje verdadero puede calcularse de la siguiente manera:

$$\sigma^2_{V_C} = \sigma^2_{V_A} + \sigma^2_{V_B} + ... + \sigma^2_{V_K} + \sum_{i \neq j} \sigma_{V_iV_j}$$

Debido a que las *k* pruebas paralelas tienen varianzas iguales y covarianzas iguales a las de cualquier otra prueba,

$$\sigma^2_{V_C} = k\sigma^2_{V_i} + k(k-1)\sigma_{V_iV_j}$$

Recordando la definición 2, que se especificó al inicio de este anexo, tenemos que $\sigma^2_{V_i} = \sigma_{V_iV_j}$. Sustituyendo en la ecuación anterior,

$$\sigma^2_{V_C} = k\sigma_{V_iV_j} + k(k-1)\sigma_{V_iV_j}$$

Simplificando,

$$\sigma^2_{V_C} = k^2\sigma_{V_iV_j}$$

Recordando la definición 3 tenemos que $\sigma_{V_iV_j} = \sigma_{ij}$. Sustituyendo en la ecuación anterior,

$$\sigma^2_{V_C} = k^2\sigma_{ij}$$

Retomando la definición del coeficiente de confiabilidad de la sección correspondiente de este documento,

$$\rho_{C_1C_2} = \frac{\sigma^2_{V_C}}{\sigma^2_C} = \frac{k^2\sigma_{ij}}{\sigma^2_C}$$

Esto se cumple siempre que los *k* componentes sean paralelos. Sin embargo, en la realidad resulta prácticamente imposible dividir una prueba en *k* componentes estrictamente paralelos. En este caso, es posible utilizar la suma de las covarianzas de los componentes y la varianza de la prueba completa para estimar una *cota inferior* de la confiabilidad de la prueba. Para ello se deben establecer tres desigualdades (Lord y Novick, 1968):

1. Si los *k* componentes de la prueba compuesta no fueran estrictamente paralelos, habría al menos un componente *g* para el cual la varianza de los puntajes verdaderos es mayor o igual a su covarianza con cualquier otro componente:

$$\sigma^2_{V_g} \geq \sigma_{ig}$$

2. Para cualesquiera dos componentes que no sean estrictamente paralelos, la suma de las varianzas de sus puntajes verdaderos es mayor o igual que el doble de su covarianza:

$$\sigma_{V_i}^2 + \sigma_{V_j}^2 \geq 2\sigma_{ij}$$

3. La suma de las varianzas de los puntajes verdaderos de componentes no paralelos son mayores o iguales a la suma de las *k(k-1)* covarianzas divididas entre (*k-1*):

$$\sum \sigma_{V_i}^2 \geq \frac{\sum_{i \neq j} \sigma_{ij}}{k-1}$$

Agregando la suma de las covarianzas en cada lado de la última desigualdad tenemos,

$$\sum \sigma_{V_i}^2 + \sum_{i \neq j} \sigma_{ij} \geq \frac{\sum_{i \neq j} \sigma_{ij}}{k-1} + \sum_{i \neq j} \sigma_{ij}$$

Simplificando la expresión de la derecha,

$$\sum \sigma_{V_i}^2 + \sum_{i \neq j} \sigma_{ij} \geq \frac{k\sum_{i \neq j} \sigma_{ij}}{k-1}$$

Recordando que la expresión de la izquierda es la definición de $\sigma_{V_C}^2$,

$$\sigma_{V_C}^2 \geq \frac{k}{k-1} \sum_{i \neq j} \sigma_{ij}$$

Dividiendo ambos lados entre σ_C^2,

$$\frac{\sigma_{V_C}^2}{\sigma_C^2} \geq \frac{k}{k-1} \left(\frac{\sum_{i \neq j} \sigma_{ij}}{\sigma_C^2} \right)$$

Tomando en cuenta que $\sigma_C^2 = \sum \sigma_i - \sum_{i \neq j} \sigma_{ij}$,

$$\rho_{C_1 C_2} \geq \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_C^2}\right)$$

La expresión del lado derecho de la desigualdad es conocida como *Coeficiente Alfa de Cronbach* (Cronbach, 1951). Cuando una prueba está compuesta por componentes no paralelos, podemos estimar una cota inferior de la confiabilidad a partir de este coeficiente. Para realizar este cálculo se requiere conocer el número de sub-pruebas o componentes, la varianza de los puntajes totales y la suma de todas las covarianzas entre los puntajes de las sub-pruebas.

La mayor aplicación del Coeficiente Alfa de Cronbach se da cuando la prueba bajo análisis se define como una prueba compuesta por *n* sub-pruebas, en donde cada sub-prueba es un solo reactivo. De esta forma, para estimar el coeficiente será necesario saber el número total de reactivos, la varianza de los puntajes totales y las covarianzas entre los puntajes de los reactivos.

Anexo 3. Deducción de la fórmula para el error estándar de medida

El propósito de este anexo es derivar la fórmula del error estándar de medida en la TCT. Para ello se utilizará la definición del modelo clásico, los supuestos correspondientes y la definición del coeficiente de confiabilidad.

Recordando la definición de puntaje observado,

$$X = V + E$$

donde:

X = Puntuación observada
V = Puntuación verdadera
E = Error de medida

Obtenemos la varianza en ambos lados de la ecuación,

$$\sigma_X^2 = \sigma_V^2 + \sigma_E^2 + 2\sigma_{VE} = \sigma_V^2 + \sigma_E^2 + \frac{2\rho_{VE}}{\sigma_V \sigma_E}$$

Por el segundo supuesto de la TCT,

$$\sigma_X^2 = \sigma_V^2 + \sigma_E^2$$

Dividiendo ambos lados de la ecuación entre σ_X^2 ,

$$\frac{\sigma_V^2}{\sigma_X^2} + \frac{\sigma_E^2}{\sigma_X^2} = 1$$

El primer término del lado izquierdo es la definición del coeficiente de confiabilidad $\rho_{X_1 X_2}$,

$$\rho_{X_1X_2} + \frac{\sigma_E^2}{\sigma_X^2} = 1$$

Despejando y obteniendo la raíz cuadrada obtenemos la definición de error estándar de medida,

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{X_1X_2}}$$

Como se comentó anteriormente, en la práctica el coeficiente de confiabilidad suele estimarse a partir del Coeficiente Alfa de Cronbach, por lo que podemos escribir,

$$\hat{\sigma}_E = \hat{\sigma}_X \sqrt{1 - \hat{\alpha}}$$

Bibliografía

1. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). "*Standards for Educational and Psychological Testing"*. Washington, DC: American Educational Research Association.
2. Ayala, R. (2009). "*The Theory and Practice of Item Respobse Theory"*. New York, NY: The Guilford Press.
3. Baker, F. (2001). *"The Basics of Item response Theory"*. Washington, DC: Education Resources Information Center. Recuperado el 3 de agosto de 2008, de http://echo.edres.org:8080/irt/baker/.
4. Baker, F. y Kim, S. (2004). "*Item Response Theory, Parameter Estimation Techniques"*. New York, NY: Marcel Dekker, Inc.
5. Bartholomew, D. *et al*. (2011). "*Latent Variable Models and Factor Analysis"*. West Sussex, UK: Jonhn Wiley & Sons.
6. Brown, W. (1910). "*Some experimental results in the correlation of mental abilities"*. British Journal of Psychology, 3, p. 296-322.
7. Centro Nacional de Evaluación para la Educación Superior, A.C. (2008). "*Metodología Ceneval"*. Recuperado el 15 de mayo de 2012 de http://www.ceneval.edu.mx
8. Crocker, L. y Algina, J. (1986). "*Introduction to Classical & Modern Test Theory"*. Belmont, CA: Wadsworth.
9. Cronbach, L. J. (1951). "*Coefficient alpha and the internal structure of tests"*. Psychometrika, 16, p. 297-334.

10. Downing, S. y Haladyna, T. (2006). *"Handbook of Tests Development"*. Mahawah, NJ: Laurence Erlbaun Associates..

11. Educational Testing Service (2000). "*ETS Standards for Quality and Fairness"*. New Jersey, Princeton, NJ: Educational Testing Service.

12. Embretson, S. y Reise, S. (2000). *"Item Response Theory for Psychologist"*. Mahawah, NJ: Laurence Erlbaun Associates.

13. Hambleton, R. y Swaminathan, H. (1985). *"Item response theory : principles and applicactions"*. Boston, MA: Kluwer Academic Publishers.

14. Hambleton, R. *et al*. (1991). *"Fundamentals of Item Response Theory"*. Newbury Park, CA: SAGE Publications.

15. Holland, P y Weiner, H. (1993). *"Differential Item Functioning"*. Mahawah, NJ: Laurence Erlbaun Associates.

16. Karabatsos, G. (2002). "*A comparison of 36 person fit statistics of item response theory"*. Paper presented in a session entitled "Rash Mersin fit análisis" george Karabastos chair in the International Objetive Meassurement Workshop conference, New Orleans, Louisiana.

17. Kelley, T L. (1939). "*Selection of upper and lower groups for the validation of test items"*. Journal of Educational Psychology, 30, 17-24.

18. Linacre J. y Wright B. (1989). "*Mantel-Haenszel DIF and PROX are Equivalent! Rasch Measurement Transactions"*, 3:2 p.52-53. Recuperado el 26 de febrero de 2009 de http://www.rasch.org/rmt/rmt32a.htm

19. Lord, F. M. (1980). *"Applications of ítem response theory to practical testing problems"*. Hillsdale, New Jersey: LEA.

20. Lord, F. M. y Novick, M. R. (1968). "*Statistical theories of mental test scores"*. Addison-Wesley.

21. Martínez. R. (1996). *"Psicometría: Teoría de los tests psicológicos y educativos"*. Madrid, España: Síntesis.

22. Nunnally, J. y Bernstein, I. (1994). *"Psychometric Theory"*. New York: McGraw-Hill.

23. Schmitt, N. (1996). *"Uses and abuses of coefficient alpha",* Psychological Assessment, Vol. 8, p.350-353.

24. Spearman, C. (1904). "*General Intelligence, Objectively Determined and Measured"*. The American Journal of Psychology 15 p.201–292.

25. Spearman, C. (1907). "*Demonstration of Formulae for True Measurement of Correlation"*. The American Journal of Psychology 18 p.161–169.

26. Spearman, C. (1910). "*Correlation calculated from faulty data"*. British Journal of Psychology, 3, p. 271-295.

27. Spearman, C. (1913). *"Correlations of sums and differences".* British Journal of Psychology, 5, p.417-426.

28. Van der Linden, W. y Hambleton, R. (1997). *"Handbook of Modern Item Response Theory"*. New York: Springer.

29. Wainer, H. *et al*. (2007). *"Testlet Response Theory and Its Applications"*. New York: Cambirdge University Press

30. Wolfram, S. Mathematica (Versión 8.0.0.0.0)[Software de cómputo]. Champaign, IL, E.U: Wolfram Research, Inc.

31. Ziecky, M (2003). *"A DIF Primer"*. ETS, New Jersey. Tomado el 12 de marzo de 2009 de https://www.ets.org/Media/Tests/PRAXIS/pdf/DIF_primer.pdf

32. Zimowski, M. *et al*. (2003). BILOG-MG (Versión 3.0.2327.2) [Software de cómputo]. Lincolnwood, IL, E.U: Scientific Software International, Inc.