Centro de Investigación en Matemáticas, A.C.

CIMAT

Validación y confiabilidad de pruebas de opción múltiple para la evaluación de habilidades

T E S I S

Que para obtener el grado de

Maestro en Ciencias en Estadística Oficial

P r e s e n t a :

Ernesto Macías Calvillo

Director de Tesis:

Dr. Rogelio Ramos Quiroga

Guanajuato, Gto. Junio del 2011

CIMAT

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, A.C.

Validación y confiabilidad de pruebas de opción múltiple para la evaluación de habilidades

TESIS

Que para obtener el grado de

Maestro en Ciencias en Estadística Oficial

PRESENTA:

Ernesto Macías Calvillo

Comité de Evaluación:

Dr. Jorge Domínguez Domínguez
(Presidente)

Dr. Johan Van Horebeek
(Secretario)

Dr. Rogelio Ramos Quiroga
(Vocal y Director de Tesis)

Guanajuato, Gto. Junio del 2011

AGRADECIMIENTOS

Son muchas las personas especiales a quienes me gustaría agradecer su amistad, apoyo, ánimo y compañía en las distintas etapas de mi vida. Algunas están aquí conmigo y otras están en mis recuerdos y en el corazón. Sin importar donde estén o si alguna vez llegan a leer esta dedicatoria, quiero darles las gracias por ser parte de mí, por todo lo que me han brindado y por sus bendiciones.

Agradezco a mi esposa Teresa y a mis hijos Alejandra Yarensy, Ernesto Eduardo y Rodrigo Andrés por ser motor de mi vida, siendo mi inspiración y motivación para no reservar ningún esfuerzo para este logro… ¡Por ellos y para ellos!

A mis padres y hermanos que me enseñaron a superar la adversidad con su ejemplo de persistencia y superación permanente, les agradezco sus sacrificios en aras de mi crecimiento personal y profesional.

Gracias a mis profesores de toda la vida porque de alguna manera son parte de lo que soy, a mis maestros de posgrado por su generosidad al compartir sus experiencias y amplios conocimientos y darme ejemplo de su calidad humana en todo momento. Gracias por su paciencia.

Agradezco de manera especial y sincera al Dr. Rogelio Ramos Quiroga por aceptar la dirección de esta tesis. Su apoyo y confianza en mi trabajo y su capacidad para guiar mis ideas ha sido un aporte invaluable. Gracias por sus consejos certeros.

También, quiero agradecer al Dr. Johan Van Horebeek y al Dr. Jorge Domínguez Domínguez, quienes a accedieron amablemente a ser miembros del jurado para evaluar este trabajo.

Al Instituto Nacional de Estadística y Geografía (INEGI), mi agradecimiento por brindarme la oportunidad de desarrollarme profesionalmente y muy en lo particular al Dr. Gilberto Calvillo Vives, quien tuvo la visión de llevar a cabo esta Maestría Ciencia en Estadística Oficial en el Instituto de la cual me siento orgulloso ser egresado.

A mis compañeros de grupo y de trabajo sólo tengo palabras de agradecimiento por su amistad y disponer siempre de una palabra de aliento en los momentos difíciles. Gracias por ayudarme a no flaquear.

Al más especial de todos, A ti Señor, porque haces realidad mis sueños, por todo el amor que me rodea y porque me sostienes siempre en tus manos.

DEDICATORIA

A esa maravillosa mujer,
mi madre que me enseño a superar la adversidad
con su ejemplo de persistencia.

A mis hijos, Rodrigo, Eduardo y Alejandra
por ser motor de mi vida, por el tiempo
sacrificado en aras de mi crecimiento
personal y profesional.

A amada esposa Teresa, quien a pesar de
mis extravíos está siempre a
mi lado apoyándome.

A mis asesores de tesis por su dedicación,
apoyo y ejemplo de profesionalismo
para mi desarrollo profesional.

A mis compañeros de trabajo por su paciencia y tolerancia.

CONTENIDO

CAPÍTULO 1

PROBLEMÁTICA

1.1 Introducción. 1
1.2 Antecedentes. 4
1.3 Planteamiento del problema. 9
1.4 Objetivo general. 11
1.4.1 Objetivos específicos. 12
1.5 Justificación. 12
1.6 Limitaciones. 14

CAPÍTULO 2

MARCO CONCEPTUAL Y TEÓRICO

2.1 Competencia. 15
2.2 Evaluación. 16
2.2.1 Definición de evaluación. 16
2.2.2 Tipos de evaluaciones. 16
2.3 Pruebas. 17
2.3.1 Definiciones de pruebas. 17
2.3.2 Construcción de la prueba. 17
2.4 Estándares de calidad de las pruebas. 19
2.4.1 Objetividad. 20
2.4.2 Validez. 20
2.4.3 Confiabilidad. 21
2.4.4 Error de medición. 24
2.5 Reactivo. 25
2.5.1 Definición de reactivo. 25
2.5.2 Tipos de reactivos. 25
2.5.3 Reactivo de opción múltiple. 25
2.5.4 Elaboración de reactivos. 26
2.5.5 Teoría Clásica de los Tests (TCT). 27
2.5.6 Teoría de la Respuesta al Ítem (TRI). 35
2.5.7 Función de información. 52
2.5.8 Calibración de ítems. 56
2.5.9 Análisis de alternativas. 59
2.6 Estimación de la habilidad. 59
2.7 Calificación mínima aprobatoria. 63

CAPÍTULO 3

METODOLOGÍA

3.1 Método 67
3.2 Equipamiento, software y materiales. 70

CAPÍTULO 4

RESULTADOS

4.1 Exploración y tratamiento de los datos. 72
4.2 Análisis de reactivos por la Teoría Clásica de los Tests. 76
4.3 Análisis de reactivos por la Teoría de la Respuesta al Ítem. 80
4.4 Validez de las opciones de respuestas. 87
4.5 Validación y confiabilidad de la prueba. 89
4.5.1 Índice de validez. 89
4.5.2 Índice de consistencia. 90
4.5.3 Índice de confiabilidad de la prueba. 94
4.6 Tiempo requerido y ajuste a la extensión de la prueba. 95
4.6.1 Tiempo requerido de aplicación. 95
4.6.2 Ajuste a la extensión de la prueba. 96
4.7 Estimación de la habilidad. 97
4.8 Estimación de la calificación mínima aprobatoria. 100
4.9 Validación del método, modelo y técnicas. 101

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones. 103
5.2 Recomendaciones. 106

GLOSARIO……..108

BIBLIOGRAFÍA…….. 110

ANEXOS……...112

RESUMEN

La Teoría de Respuesta al Ítem proporciona una estrategia probabilística para estimar el nivel de la habilidad de los sustentantes en el rasgo inobservable en función del patrón de respuesta a un conjunto de preguntas de opción múltiple mediante un modelo estadístico que contempla medidas invariantes respecto a los ítems.

En la presente investigación se recopilan los conceptos y los principios que sustentan las metodologías de la validez y confiabilidad de evaluaciones, así como las propiedades psicométricas que caracterizan a los reactivos. Además, se presentan los procedimientos y las técnicas estadísticas para la estimación de los parámetros de los modelos y la estimación del nivel real de la habilidad y de la calificación umbral a evaluar en los sustentantes como aptos para realizar una competencia laboral.

En un marco formal, práctico y sencillo, se ejemplifica en un caso de estudio la aplicación de las metodologías y técnicas estadísticas para la identificación de imperfecciones del test, las cuales afectan la validez y confiabilidad de las evaluaciones.

El contexto de esta investigación se ubica en el Instituto Nacional de Estadística y Geografía a través de la Dirección de Capacitación y Calidad que demanda para cumplimiento de su Misión: la identificación de áreas de oportunidad para el desarrollo y profesionalización del personal que participa en la generación de la Estadística Oficial de nuestro país.

PALABRAS CLAVE: Evaluación, validación, confiabilidad, teoría de la respuesta al ítem, preguntas de opción múltiple, curva característica del ítem, discriminación, dificultad, función de información, estimador de la habilidad.

ABSTRACT

The Theory of Item Response provide a probabilistic strategy for estimating the level of the examinees in the unobservable trait tests measure based on the pattern of responses to a set of multiple choice questions, using a statistical model that includes measures invariant with respect to the items .

In the present investigation are collected concepts and principles underlying the methodologies of the validity and reliability of assessments and psychometric properties that characterize the items. In addition, there are procedures and statistical techniques to estimate the model parameters and estimating the actual level of skill and the minimum grade to qualify the sustaining to do a competition work.

In a formal framework, practical and simple, is exemplified by a case study, the application of methodologies and statistical techniques for identification of imperfections in the test, which affect the validity and reliability of assessments.

The context of this research is at the “Instituto Nacional de Estadística y Geografía” through the “Dirección de Capacitación y Calidad” is necessary for the enforcement of its Mission: to identify opportunity areas for the development and professionalization of personnel involved in the generation of Official Statistics of our country.

KEY WORDS: Evaluation, validation, reliability, theory of item response, multiple choice questions, item characteristic curve, discrimination, difficulty, information function, the ability estimator.

CAPÍTULO 1

PROBLEMÁTICA

Las más largas caminatas comienzan con un paso.
Proverbio hidúo

1.1 *Introducción.*

Las instituciones, en general, enfocan sus esfuerzos en mejorar continuamente sus procesos de generación de productos y servicios para alcanzar los estándares de calidad deseados y ganar competitividad en los mercados.

Uno de los factores determinantes del éxito de las empresas es la capacidad del personal que lleva acabo los procesos de trabajo; por consecuencia, es considerado el recurso más importante de la empresa y es prioritario su desarrollo e inclusión en los procesos con toda certeza de su capacidad para realizar sus actividades eficazmente.

El Instituto Nacional de Estadística y Geografía, está comprometido con la profesionalización de sus colaboradores en cumplimiento con su Misión de generar y coordinar la estadística oficial del país. Su punto de partida, para su desarrollo, es la evaluación de sus habilidades.

La evaluación es un medio eficiente para emitir juicios respecto a la calidad y maestría de la habilidad, e identificar áreas de oportunidad que podrán ser atendidas mediante capacitación.

La medición es una parte básica de la evaluación, ya que sin ella no sería posible hacer una valoración; sin embargo, la medición y la evaluación no son lo mismo: medir algo sin la emisión de un juicio no es evaluación, así como tampoco lo es la emisión de un juicio sin la información obtenida desde la medición.

La prueba es el instrumento de medición. Lo que mide es el grado de conocimientos, aptitudes y actitudes que un individuo posee. Un instrumento de medición es válido si mide lo que pretende medir con objetividad y confiabilidad, aún cuando no es común imaginar que una lista reactivos constituya una regla o *vara* de medir, se tienen muchas ventajas psicométricas teóricas y prácticas observar a las pruebas como cualquier otro instrumento de medición.

Otras cualidades técnicas requeridas de los instrumentos es que permitan comparaciones tanto transversales como longitudinales y que los resultados sean procesados rápida y confiablemente, todo ello dentro de límites razonables de costo.

Existen varios modelos de evaluación que ofrecen diferentes ideas conceptuales, pero desde el punto de vista práctico se necesitan herramientas simples que ofrezcan información suficiente para evaluar a los sustentantes y, al mismo tiempo, permitan dictaminar la calidad de la prueba y aporten elementos de análisis para el perfeccionamiento de los reactivos.

Los instrumentos que reúnen las características deseadas son las *pruebas de aplicación en gran escala* -frecuentemente identificadas como exámenes de papel y lápiz-, formados exclusivamente por preguntas cerradas. Cuando estas pruebas se elaboran deficientemente dan pie a críticas que las señalan como superficiales y, en el peor de los casos, sesgadas social y culturalmente. En cambio, si se aprovechan las metodologías y técnicas es posible contar con instrumentos sofisticados muy finos, con capacidad de aplicación por computadora y adaptados a diferentes escenarios.

Para la adopción y aplicación de *pruebas en gran escala* es necesario analizar a quienes han incursionado en estas actividades, no para imitarlos ciegamente, pero sí para aprovechar su experiencia. El conocimiento de experiencias tempranas podrá ahorrar costosos errores y contribuirá al desarrollo de este campo y a la formación de especialistas, reduciendo la brecha con otros países que han estado aplicando este tipo de evaluaciones en plenitud desde inicios del siglo pasado.

En el marco teórico del presente documento se recopilan los conceptos y principio que sustentan las metodologías de evaluación, así como los estadísticos que caracterizan a los reactivos y por consecuencia, la calidad de los exámenes desde la perspectiva de la Teoría Clásica de los Tests y la Teoría de la Respuesta al Ítem. Además, se presentan los procedimiento para la estimación de las propiedades psicométricas de los ítems, así como la estimación de la calificación real de la habilidad y de la calificación umbral para la calificar los sustentantes en aptos o no para realizar una competencia laboral.

Las metodologías y técnicas estadísticas para valorar la validez y confiabilidad de las evaluaciones se ejemplifican en un caso de estudio práctico de un examen de la capacidad de generación de información geográfica, enfocándose principalmente al análisis de los reactivos.

1.2 *Antecedentes.*

La teoría básica de la medición se inició en el siglo XVIII y desde el siglo XIX se aplicaron estas ideas en el campo de la educación en países como Alemania, Inglaterra y Estados Unidos de América.

Durante la segunda mitad del siglo XX, la Teoría de Respuesta al Ítem fue concebida por Lord en Psicología y con la difusión de las computadoras se produjeron los modelos estadísticos de Birnbaum (1957) y de Rasch, en 1960. La obra que difunde una teoría ya madura es la de Lord y Novick, de 1968: Statistical theories of mental test scores. "La TRI intenta dar una fundamentación probabilística al problema de la medición de rasgos y constructos no observables (rasgos latentes), considerando al ítem, y no al puntaje global, como la unidad básica de análisis." (Martínez Arias, 1995: 237-243).

Un aspecto particular de la historia de la medición educativa en Estados Unidos de América, es el College Board (creado en 1900), el más importante organismo en cuanto al desarrollo de pruebas para el ingreso y egreso de las universidades (Donlon, 1984: 1).

Desde 1925 el College Board desarrolla pruebas de aptitud, en contraposición a conocimientos, buscando ir más allá de la memorización de datos aislados. La prueba que todavía hoy se llama *Scholastic Aptitude Test* (SAT) se gestó en 1925 y doce años después se desarrollaron pruebas íntegramente compuestas por preguntas de opción múltiple. Cabe mencionar que el número de sustentantes anuales llegó desde fines de los años sesenta al millón y medio de personas, cifra que se ha mantenido desde entonces (Donlon, 1984: 3-8).

La evaluación en gran escala en el mundo se incrementa significativamente, por ejemplo, en España, con la creación del Instituto Nacional para la Calidad de la Educación en los años noventa comenzaron a desarrollarse estas evaluaciones y

en América, la UNESCO inicia en 1995 a implantar un programa regional de evaluación de la calidad de la educación básica, (Wolff, 1998: 21).

En México, posiblemente la experiencia más antigua sea la de la Facultad de Medicina de la UNAM, que a mediados de los sesenta desarrolló un banco de reactivos de opción múltiple para el examen de titulación de la carrera de Médico Cirujano calificado por computadora. Sin embargo, seguramente la experiencia más importante fue la relativa a los exámenes de selección para ingreso a la UNAM, tanto a nivel bachillerato como licenciatura, procesos que se volvieron complejos y delicados cuando el número de aspirantes se multiplicó y los lugares disponibles cada año fueron más insuficientes.

Para atender esta problemática, según informa el Mtro. Rafael Vidal, director técnico del Centro Nacional de Evaluación para la Educación Superior (CENEVAL), la UNAM estableció en 1963 un área especializada que logró desarrollar pruebas que, sometidas *a posteriori* a evaluaciones psicométricas, mostraron excelentes niveles de confiabilidad e inclusive, grados de equivalencia de la versión de un año a los siguientes no muy inferiores a los que se alcanzan con las modernas técnicas de igualación o equiparación. Los trabajos de esta área de la UNAM, continuados luego en el marco de la Dirección General de Planeación, fueron la base para la creación del CENEVAL en 1994.

Otras instituciones de educación superior privadas, como la Universidad Iberoamericana, o públicas, como la Universidad Autónoma de Aguascalientes, desarrollaron también sus propias pruebas de ingreso durante los años setenta sin controlar sistemáticamente las propiedades psicométricas de las pruebas y, en consecuencia, sin poder hacer análisis comparativos o longitudinales.

Otras instituciones privadas como el ITESM y la Universidad de las Américas utilizaron los servicios de la oficina establecida por el College Board en Puerto

Rico en 1963, contratando la aplicación de la Prueba de Aptitud Académica (PAA), equivalente al SAT, con fines de selección para el ingreso al nivel de licenciatura.

Para ver nacer organismos dedicados a la elaboración de pruebas que atiendan en forma sistemática cuestiones técnicas, deberán esperarse hasta los años noventa, con la fundación del CENEVAL y la modernización del área de evaluación de la SEP (1995), precedidas por el lanzamiento del Examen de Conocimientos y Habilidades Básicas (EXCOHBA) de la Universidad Autónoma de Baja California (1990).

Respecto al ámbito laboral, la evaluación de capacidades se priorizan, en el 2003 se promulgó la Ley del Servicio Profesional de Carrera en la Administración Pública Federal, cuyo objetivo es establecer las bases para la organización, funcionamiento y desarrollo del Servicio Profesional de Carrera en las dependencias de la Administración Pública Federal centralizada. Ley que establece en el artículo 44 que la capacitación y certificación de competencias son los procesos mediante los cuales los servidores públicos de carrera son inducidos, preparados, actualizados y certificados para desempeñar un cargo en la Administración Pública; y en el artículo 52 indica que los servidores profesionales de carrera deberán ser sometidos a una evaluación para certificar sus capacidades profesionales por lo menos cada cinco años. Las evaluaciones deberán acreditar que el servidor público ha desarrollado y mantiene actualizado el perfil y aptitudes requeridos para el desempeño de su cargo. Esta certificación será requisito indispensable para la permanencia en su cargo, por lo anterior, es evidente la demanda de mecanismos de evaluación válidos y confiables tanto para el diagnóstico de necesidades de capacitación como para la certificación.

En un ámbito general, se dispone del Consejo Nacional de Normalización y Certificación de Competencias Laborales (CONOCER), encargado de difundir y fomentar la cultura de la competencia laboral certificada en México, promoviendo

su normalización y certificación para que trabajadores y personas en general, puedan demostrar sus conocimientos, habilidades y destrezas adquiridas durante su vida laboral, independientemente de la forma como cada quien la haya adquirido.

El certificado de competencia laboral del CONOCER significa contar con un reconocimiento formal de validez oficial y de carácter nacional para los trabajadores que hayan demostrado evidencias de sus conocimientos, habilidades y destrezas aplicadas a una función laboral, misma que se encuentra establecida en una Norma Técnica de Competencia Laboral. Esto le servirá para promover su movilidad dentro del mercado laboral y facilitar, en apego a las necesidades reales de este mercado, tanto la oferta de su educación formal para el trabajo como su capacitación.

El 13 mayo de 1998, CONOCER otorgó al CENEVAL la acreditación como organismo certificador y desde esa fecha quedó autorizado para acreditar centros de evaluación y certificar la competencia laboral de las personas conforme a las normas técnicas de competencia laboral que tienen acreditadas.

En el INEGI, el antecedente más formal en la evaluación de habilidades surge con la implementación del Sistema Integral de Profesionalización (SIP) en 1994, con el cual la contratación del personal se sustentaba en la evaluación de habilidades requeridas para ejercer las funciones del puesto vacante.

Para agilizar y transparentar el proceso de evaluación, se implementó el Sistema de Evaluación Automatizada que generaban los exámenes con base en un banco de reactivos y una tabla de especificaciones construida a partir del perfil del puesto; en esta tabla se indicaba la temática, amplitud y profundidad a valorar, así como la importancia del dominio. Los reactivos del examen se generaban aleatoriamente conforme a estas especificaciones al momento de la aplicación directa en la computadora, de tal forma que garantizaba la total confidencialidad

del examen al no ser conocido previamente por ninguna persona. Los exámenes eran calificados automáticamente por el propio sistema logrando agilidad, imparcialidad y objetividad en los resultados.

El banco de reactivos fue construido y validado por especialistas en distintas temáticas, cada reactivo se clasificó de acuerdo con la temática y el nivel taxonómico, sin embargo, el análisis del comportamiento de los reactivos se basó únicamente en la estadística descriptiva sin utilizar otra técnica formal de validación y confiabilidad. Posiblemente esto provocó el desuso de la herramienta, pues la calificación del evaluado dependía principalmente de la suerte que tuvieran en la selección aleatoria de reactivos, más que su propia habilidad. Esta situación fue agudizada al incrementarse el banco de reactivos sin una clasificación uniforme y sin considerar las propiedades psicométricas de los reactivos.

Actualmente, la Dirección de Capacitación y Calidad del INEGI dispone de una herramienta para el diseño y aplicación computarizada de evaluaciones para propósitos de diagnóstico de capacitación principalmente, la conceptualización es similar a la implementada para el desarrollo del SIP, pero la validez de las pruebas está sustentada en la metodología de diseño y es delegada la responsabilidad al experto, quien elabora los reactivos.

Es importante mencionar el reto que tiene el INEGI con su nueva personalidad jurídica, la Ley del Sistema Nacional de Información Estadística y Geográfica en su artículo 36 establece la responsabilidad de elaborar un programa permanente y actualizado de capacidades técnicas de los servidores públicos de las unidades a efecto de apoyar la capacitación y actualización, entendiéndose como unidad (artículo 1, XV) a las áreas administrativas que cuenten con atribuciones para desarrollar actividades estadísticas y geográficas o que cuenten con registro administrativo que permitan obtener información de interés nacional: las

dependencias y entidades de la Administración Pública Federal, incluyendo a las de la Presidencia de la República y de la Procuraduría General de la República; las entidades federativas y los municipios; los organismos constitucionales autónomos y los Tribunales Administrativos Federales. Por lo anterior, es eminente la necesidad de fortalecer al área de capacitación para enfrentar el reto al extenderse considerablemente la población objetivo a atender.

Al presentar la reseña histórica de los avances técnicos y metodológicos de las evaluaciones, se ha demostrado que los problemas que enfrenta nuestro contexto institucional no son inéditos: se han presentado en otros lugares, hay soluciones bien probadas y, en el último medio siglo, los avances en este campo han sido muy importantes, aunque el reto principal es la adopción práctica de estas técnicas estadísticas de validación y confiabilidad de evaluaciones.

1.3 *Planteamiento del problema.*

La imperfección de las pruebas, de los procedimientos de aplicación y de evaluación son una ineludible realidad aún con el cumplimiento de altos estándares de calidad por los elementos arbitrarios e inciertos involucrados en los procesos. El principal factor de esta limitación es el ser humano por su impredecible desempeño. La imperfección reduce el grado de confianza de la evaluación y por consiguiente, incrementa el riesgo en tomar decisiones equívocas.

La eficiencia de las evaluaciones está sustentada principalmente en reducir la imperfección de las pruebas al diseñarlas con un enfoque sistemático y metodológico para lograr que realmente midan con confiabilidad lo que se pretende medir.

Las pruebas de desempeño son usualmente utilizadas por su realismo para evaluar las habilidades basadas en la valoración del experto en el desarrollo del proceso y en las características de un servicio o producto realizado por el sustentante. No obstante, la aplicación de este tipo de pruebas es costosa y requiere de un tiempo considerable su aplicación y evaluación. Por lo general, participan varios expertos para calificar la habilidad, ya sea porque así se requiere debido a la gran cantidad de sustentantes, o bien, porque se realice en distintos momentos o lugares, siendo ésta la situación predominante en la Dirección de Capacitación y Calidad del INEGI.

Hay tres fuentes de imperfección en la evaluación de desempeño:

1. Diferencia de severidad. Los evaluadores no son igualmente estrictos. Si sometemos a evaluación a una persona por distintos evaluadores es posible que no coincida la calificación entre éstos.

2. Discrepancia al mérito. Es probable que los evaluadores tengan discrepancia sobre la importancia del rasgo a evaluar, es decir, que su ponderación sea distinta.

3. Variación temporal. Un evaluador puede cambiar la ponderación de un rasgo en el tiempo. Ésta puede ser obtenida si un evaluador califica algún rasgo más de una vez y olvida o cambia la ponderación.

Aun cuando se dispone de técnicas para manejar la subjetividad de la evaluación, no deja de ser costosa debido a la implicación de expertos evaluadores y los tiempos requeridos para su aplicación; además, limita a los sustentantes en los casos de prueba a realizar; más aún, no todo se puede experimentar ya sea por los costos involucrados en el desarrollo o los posibles daños ocasionados por errores de los sustentantes al momento de la evaluación.

Una alternativa son las pruebas con reactivos de opción múltiple y tratamiento estadístico para lograr la validez y confiabilidad esperada; las cuales ofrecen la posibilidad de contestar gran cantidad de reactivos en relativamente poco tiempo y ser calificadas por computadoras con alta confiabilidad a bajo costo.

La aplicación de técnicas estadísticas para optimizar la validez y confiabilidad de la prueba, ofrecerá el marco de referencia para el diseño y aplicación de evaluaciones, brindando respuestas a las siguientes preguntas:

1. ¿El examen realmente mide de manera consistente lo que pretende medir?
2. ¿Cuál es el grado de confianza de los resultados de un examen?
3. ¿Qué reactivos son adecuados para discernir el nivel de la habilidad?
4. ¿Cuál debe ser la ponderación adecuada para cada reactivo?
5. ¿Qué reactivos pueden eliminarse del examen para deducir el tiempo de aplicación sin que impacte significativamente en los resultados?
6. ¿Cuál es la calificación mínima aprobatoria de un examen?

La falta de tradición psicométrica en México hace que los avances importantes en la materia sean casi desconocidos; las mejores prácticas se limitan a aplicar, con limitaciones, la teoría clásica. Por lo anterior, es oportuno abordar las técnicas estadísticas de validez y confiabilidad de evaluaciones con reactivos de opción múltiple cuyo enfoque sea sencillo y práctico.

1.4 *Objetivo general.*

El propósito fundamental que tiene el presente estudio es presentar la metodología y técnicas estadísticas para identificar las debilidades de los exámenes de preguntas de opción múltiple que afectan la validez y confiabilidad de las evaluaciones de habilidades en el contexto de capacitación en el INEGI.

1.4.1 Objetivos específicos.

1. Realizar una revisión bibliográfica para comprender los conceptos, principio y métodos de las evaluaciones.

2. Identificar las técnicas estadísticas para valorar la validez y confiabilidad de exámenes con reactivos de opción múltiple.

3. Aplicar los métodos y las técnicas estadísticas en un caso de estudio y evaluar los resultados.

1.5 *Justificación.*

En la construcción de instrumentos psicométricos es usual que el experto del tema desarrolle los reactivos correspondientes a los aspectos que desea comprobar en una prueba, él escoge la cantidad de reactivos, determina la ponderación y fija la calificación mínima aprobatoria. Con base en su buen juicio no siempre está mal, sin embargo, tampoco tiene evidencia de que esté bien porque, por lo general, no confirma la validez y confiabilidad de su instrumento. Incluso cuando siga con rigor el proceso para el diseño de pruebas, no siempre se tiene garantizada la calidad. Si agregamos que regularmente quien realiza la evaluación desconoce las técnicas psicométricas o aun conociéndolas, no verifica los supuestos en que se sustenta la metodología, dando como resultado una débil validez y confiabilidad.

En la Dirección de Capacitación y Calidad conforme a su Misión y Visión (Anexo A) es un área de oportunidad utilizar métodos de validación y confiabilidad de los exámenes utilizados en el proceso de la capacitación e incluso de certificación de competencias.

Las evaluaciones diagnósticas son esenciales para identificar las áreas de oportunidad que se requieren fortalecer mediante capacitación contribuyendo, así,

a un trabajo de excelencia y conformar grupos homogéneos que facilite el desarrollo del curso.

Los resultados de las evaluaciones intermedias y finales de un curso aportan indicadores fundamentales para la valoración del proceso enseñanza-aprendizaje en función de logros obtenidos.

El análisis de indicadores objetivos contribuyen a identificar debilidades y áreas de oportunidad que al atenderlas promoverán la mejora continua del proceso, como son: el estudiante, el instructor, los planes y programas de estudio, recursos didácticos, entre otros elementos que componen el fenómeno de capacitación y, en general, el desarrollo humano.

Con respecto a la certificación de competencias laborales, disponer de una herramienta de validez y confiabilidad de evaluaciones es fundamental para respaldar que la persona tenga la capacidad de desarrollar un trabajo confiable y de calidad.

En lo general, la aplicación de las técnicas de validez y confiabilidad podrá ser utilizada por cualquier institución de capacitación o educación, sin la necesidad de ser estadista al utilizar metodologías, técnicas estadísticas y los programas de cómputo desarrollados en este trabajo.

Mi interés particular en el tema de validez y confiabilidad es más general, se orienta a evaluaciones de habilidades con pruebas de preguntas cerradas para acotar el alcance de la investigación. Sin embargo, la metodología planteada puede ser ampliamente utilizada en diversas áreas, puesto que por naturaleza el ser humano, de manera consciente o inconsciente, siempre está tomando decisiones con base en juicios sustentados en evaluaciones con datos observados u obtenidos mediante instrumentos. Sólo es posible mejorar lo que se mide, pero hay que medirlo bien, mediante herramientas válidas y confiables.

1.6 *Limitaciones.*

El presente estudio estará delimitado a aspectos metodológicos y técnicas estadísticas para identificar las debilidades de validez y confiabilidad de evaluaciones de habilidades mediante pruebas con preguntas de opción múltiple con un enfoque práctico en el contexto de capacitación en el INEGI.

El presente estudio no incluye metodologías ni técnicas referentes a:

- Pruebas de ejecución o desempeño.
- Comparación de pruebas e igualación de calificaciones.
- Normalización y estandarización de calificaciones.
- Técnicas para valorar el sesgo de un ítem o instrumento.
- Análisis de aspectos pedagógicos, económicos ni sociales.

CAPÍTULO 2

MARCO CONCEPTUAL Y TEÓRICO

El conocimiento teórico es un tesoro cuya llave es su práctica.
Tomas Fuller (1608 – 1661)

2.1 *Competencia.*

La competencia es la capacidad de poner en operación los diferentes conocimientos, aptitudes y actitudes (capacidad cognitiva o destreza física) de manera integral en las diferentes interacciones que tienen los seres humanos para la vida y el ámbito laboral.

En el ámbito laboral es la capacidad que una persona posee para desempeñar una función productiva usando diferentes recursos bajo ciertas condiciones que aseguren la calidad en el logro de los resultados.

Habilidad es el grado de competencia de un sujeto frente a un objetivo determinado, el cual es la variable latente a estimar mediante procedimientos de evaluación.

2.2 *Evaluación.*

2.2.1 Definición de evaluación.

La evaluación es un proceso sistemático de obtención de datos válidos y confiables que permite emitir un juicio de valor del nivel o capacidad para realizar un trabajo con base en una prueba que pone en manifiesto los conocimientos, aptitudes y actitudes previamente establecidos.

La medición es una parte básica de la evaluación, ya que sin ella no sería posible hacer una valoración; sin embargo, medición y evaluación no son lo mismo: medir algo (entendiendo la medición como la asignación de números a propiedades o fenómenos) sin la emisión de un juicio no es evaluar, como tampoco lo es la emisión de un juicio sin la información obtenida desde la medición.

La determinación y elaboración de los medios para evaluar dependerán, en gran medida, de la intención y de los rasgos a evaluar. Los medios para evaluar el desempeño son las pruebas de ejecución, los exámenes escritos y orales.

2.2.2 Tipos de evaluaciones.

En el contexto de la evaluación de las capacidades de un cierto grupo de personas, se pueden distinguir dos principales tipos de evaluación basadas en la forma de interpretar los resultados:

- *Evaluaciones referidas a la norma:* La información obtenida se interpreta comparando el resultado de cada sujeto con los demás miembros del grupo. Informa si un individuo sabe o puede hacer más o menos que los demás.

- *Evaluaciones referidas a criterio:* Se basan en la comparación de cada sujeto con un parámetro esperado correspondiente al nivel de dominio que garantiza la capacidad mínima.

2.3 *Pruebas.*

2.3.1 Definiciones de pruebas.

Una prueba es fundamentalmente un instrumento de medición. Lo que mide o debe medir es el grado de conocimientos, aptitudes y actitudes para realizar un trabajo. Una prueba es un instrumento que recoge información estructurada acerca de lo que se quiere medir para evaluar.

Para la mayoría de las personas y aun para la mayoría de los especialistas en psicometría concebir a la prueba como una regla es un concepto nuevo. sin embargo, tiene ventajas psicométricas teóricas y prácticas observar a las pruebas como se observa cualquier otro instrumento de medición.

2.3.2 Construcción de la prueba.

La evaluación parte de la obtención de datos para llegar a la valoración, por lo que es sustancial la construcción de la prueba sistemáticamente en tres grandes etapas básicas:

1. *Diseño.* Está orientada a la planeación del instrumento, implica: definir lo que se va a medir; señalar el propósito de la medición; establecer las

decisiones que se tomarán con los resultados y determinar las especificaciones técnicas de la prueba.

2. *Elaboración.* Es la construcción de la prueba con base en la tabla de especificaciones correspondiente al perfil referencial establecidas en el diseño y que se detallan más adelante.

3. *Validación.* Consiste en analizar los resultados de aplicaciones preliminares de la prueba para confirmar estadísticamente su validez y su confiabilidad, así como establecer las condiciones definitivas para la administración del instrumento con los sustentantes.

2.3.2.1 Perfil referencial.

El principal referente para conformar el contenido de una prueba (constructo) es la definición de lo que se quiere medir, debe ser claramente preestablecida y acotada. Una vez determinado el perfil referencial, éste queda como la fuente de información fundamental para saber el tipo de contenidos que deben especificarse en la prueba.

El perfil referencial debe fundamentarse en esquemas teóricos o conceptuales y ser susceptible de ser definido de manera mensurable.

2.3.2.2 Tabla *de* especificaciones.

La tabla de especificaciones presenta la organización lógica de los componentes (contenidos) definidos operacionalmente y el nivel cognoscitivo (amplitud y profundidad) con que deberán ser medidos en la prueba; además, indica el peso específico que se dará a cada uno de ellos a partir del señalamiento del número de reactivos que deberán elaborarse.

El propósito de la tabla de especificaciones es asegurar que la prueba incorpore una muestra representativa y adecuada de las capacidades que comprende el perfil referencial. Por lo tanto, es un modo de garantizar la completez y el balanceo del contenido.

Además, la tabla sirve como plano de referencia, ya que especifica la naturaleza de cada pregunta a integrar en la prueba.

La calidad de la prueba dependerá en gran medida de la fidelidad con que el evaluador pueda igualar las especificaciones contenidas en la tabla, es decir, de la habilidad para generar reactivos que pongan de manifiesto la conducta específica descrita en cada uno de sus enunciados.

La tabla de especificaciones es un documento que determina las reglas de armado o ensamble de la prueba; incluye las áreas que estarán incluidas en la prueba, los componentes específicos por cada área, el número de reactivos a elaborar, el nivel taxonómico y el tipo de conocimiento asignado a cada componente; permite que los componentes que se incluyen en la prueba representen fielmente el perfil referencial y las distintas versiones de la prueba sean equivalentes.

2.4 *Estándares de calidad de las pruebas.*

Para que cumplan su función adecuadamente los instrumentos de medición deben poseer características indispensables como objetividad, validez y confiabilidad; otras no son indispensables, depende de algunas particularidades del instrumento y como va ser utilizado, éstas son: la adaptabilidad y la estandarización de calificaciones, Tristán Agustín (1996).

2.4.1 Objetividad.

Se habla de objetividad cuando una evaluación es independiente de factores externos, el resultado depende exclusivamente de dominio de lo que se pretende medir y no permite diferentes.

En general, una evaluación es objetiva en la medida en la que no depende de la subjetividad de la persona que diseña, califica o interpreta los resultados.

2.4.2 Validez.

La validez es la eficiencia con que mide lo que pretende medir; por lo tanto, un instrumento de medición es válido si cumple satisfactoriamente el propósito para el que fue diseñado. Sin embargo, estrictamente hablando, uno no valida un instrumento de medición, sino el uso específico que se le da a las puntuaciones o resultados obtenidos (Cronbach, 1971).

Existen dos tipos de validez: teórica (constructo y contenido) y empírica (congruencia y predictiva):

Validez teórica de constructo. Entendida como la existencia de evidencias en torno a la consistencia entre el perfil referencial y la prueba. Aquí, el énfasis está en sustentar el grado en el que los puntajes de la prueba representan la medida de la característica o atributo psicológico que se supone evalúa la prueba, es decir, el constructo teórico; por lo que es necesario establecer procedimientos para la revisión lógica de tal adecuación, el análisis de su estructura interna o el de la relación de la prueba con variables externas. En algunos casos es posible reafirmarlo estadísticamente mediante el análisis factorial.

Validez teórica de contenido*.* Consiste en que la prueba contenga una muestra representativa, adecuada y congruente del total de los conocimientos o habilidades correspondientes al dominio disciplinario que se pretende evaluar. Este tipo de validez dependerá principalmente de tres aspectos: la claridad en la

definición del dominio que se pretenda evaluar (perfil referencial), la tabla de especificaciones utilizada para construir la prueba y el nivel de verificación técnica de los reactivos.

Validez empírica de congruencia. Representa la validación de una prueba mediante otra de la que ya se reconoce su validez. Se realiza obteniendo una correlación estadística entre ambas medidas. La congruencia es el grado de correlación de lo evaluado con otras evaluaciones hechas con el mismo significado teórico.

Validez empírica predictiva: Sirve para pronosticar el desempeño del sustentante a partir de lo demostrado en el resultado de la prueba. Se usa el término *predicción* en un sentido general para referirse a la relación entre el resultado del prueba y los eventos que ocurrieron antes, durante y después de su aplicación, por ejemplo: pronosticar el futuro rendimiento del estudiante. Podemos decir que por sí solas las calificaciones funcionan como predictivas. La validez de una función predictiva se determina al correlacionar estadísticamente las puntuaciones de la prueba con el resultado evento.

2.4.3 Confiabilidad.

En una evaluación, la confiabilidad es sinónimo de seguridad o consistencia y denota estabilidad del resultado. Es el grado de concordancia de la calificación de un sustentante en distintos escenarios y puede ser medido en términos de la probabilidad de que no cambie el resultado de la evaluación. Se refiere a la precisión con que se mide el rasgo, independientemente de lo que mida; está directamente relacionada con los errores en la medición, que afectan la precisión en la medida. Existen básicamente tres formas de confiabilidad:

- La confiabilidad en la calificación: en ocasiones se ve disminuida debido a que depende demasiado del juicio de la persona calificadora y se controla notablemente cuando la prueba es objetiva.

- La confiabilidad del contenido: surge de inadecuados y no representativos muestreos del conocimiento de la materia.

- La confiabilidad temporal: se refiere a que la medida sea estable y constante a través del tiempo. En el caso de las pruebas de rendimiento escolar esta temporalidad es relativa, sin embargo, resulta deseable en otro tipo de instrumentos; por ejemplo, cuando medimos la inteligencia.

Si una prueba no es confiable, es decir, no está midiendo aquéllos que pretende medir de forma consistente y con precisión, entonces no es confiable para ningún uso. Por otra parte, puede tener un grado perfecto de confiabilidad pero no ser válida para un uso determinado.

Existen varios métodos para obtener la confiabilidad de una prueba, algunos de ellos tradicionales que han ido cayendo en desuso principalmente por la nueva tecnología que facilita los procesos de recolección y manejo estadístico de datos. Entre los métodos más comunes están:

- Formas alternativas de construcción de dos pruebas tratando de satisfacer de la mejor manera posible condiciones de paralelismo en el contenido. Se calcula con una correlación de los puntajes de ambas formas.

- División por mitad del número de preguntas, ya sea al azar, o bien, pares a una versión y nones a otro. Se calcula con una correlación entre los puntajes de las dos mitades.

- Test-retest requiere de la aplicación de la prueba dos veces, debería resultar las mismas calificaciones sin que medie algún aprendizaje extra. Se calcula con una correlación entre los puntajes de ambas aplicaciones.

La forma más estable de medir la confiabilidad de un instrumento es mediante la consistencia interna manifestada en la relación de la calificación por tema o las preguntas con la calificación total.

El Coeficiente Alpha de Cronbach es un modelo de consistencia interna (a mayor valor de Alpha, mayor confiabilidad) basado en el promedio de las correlaciones entre los ítems y se refiere a que tan bien los ítems miden un simple constructo unidimensional. El mayor valor teórico de Alpha es 1, y afirma Cronbach (1984) que en lo general 0.65 se considera un valor aceptable, la fórmula para puntuaciones estandarizadas es:

$$\alpha = \frac{k\overline{P}}{1+(k-1)\overline{P}} \qquad (2.4\text{-}1)$$

Donde k es el número de ítems y $\overline{P}$ es la media de correlaciones entre las respuestas de los ítems.

El coeficiente es directamente proporcional al número de ítems y al promedio de correlación de los ítems.

Las pruebas multidimensionales presentan el coeficiente bajo por lo que se recomienda realizar previamente un análisis factorial para identificar las dimensiones y las preguntas de cada una de éstas y posteriormente calcular el coeficiente para cada subconjunto de ítems por separado.

Existe la posibilidad de elevar la confiabilidad del test al excluir reactivos que contribuyen menos al Coeficiente de Alfa, se recomienda un proceso iterativo de exclusión de ítems de uno a uno hasta que ya no se producen ganancias en el valor de Alpha, Es razonable definir la prueba con el número menor de reactivos para reducir la extensión y el costo.

2.4.4 Error de medición.

El error en la medición se presenta siempre que se realiza una medición. El error total está compuesto de los dos tipos de errores:

- El error constante: está relacionado con la validez del instrumento y se refiere al error que tiene a partir de la metodología empleada para la construcción de la prueba, este error es constante en cualquier medición que se realice con él, así al conocerse y ser constante es más fácil controlarlo.

- El error variable: está relacionado con la confiabilidad del instrumento, va a depender del modo en que sea aplicado, disminuimos este tipo de error en la medida que los manuales especifican su modo de empleo y sea respetado por los aplicadores, siendo su control más complicado.

Sin duda alguna, es mejor hacer todo lo posible para prevenir el error de medición que tener que evaluar o calcular sus efectos una vez que éste se ha presentado. Ayuda reducir el error de medición al escribir claramente los reactivos, plantear las instrucciones de la prueba de tal forma que se comprendan fácilmente y apegarse fielmente a las condiciones prescritas de administración del instrumento y evitar la subjetividad de las calificaciones, explicando al máximo las reglas de calificación. La mejor forma de hacer más confiables las pruebas es hacerlas más extensas.

2.5 *Reactivo.*

2.5.1 Definición de reactivo.

Los términos de pregunta, ítem y reactivo son utilizados con frecuencia como sinónimos, siendo el reactivo el más amplio, pues no se limita únicamente a preguntas. Por lo tanto, un reactivo es un planteamiento de un problema o cuestionamiento para conocer el resultado del aprendizaje o en general, el desempeño de una habilidad.

Un reactivo busca obtener información acotada y estructurada acerca de algún aspecto de lo que se quiere medir y siempre estará integrado a una prueba junto con otros reactivos.

2.5.2 Tipos de reactivos.

- Reactivo de respuesta no estructurada: Son reactivos cuya respuesta esperada puede ser variable y no está previamente estipulada. La valoración de las respuestas del sustentante con este tipo de reactivos implica un análisis de expertos bajo criterios más generales de calificación. Algunos ejemplos son: de respuesta breve, frases incompletas y ensayo.

- Reactivo de respuesta estructurada: La respuesta esperada es única y está establecida desde el diseño mismo del reactivo. De esta manera, al momento de calificar la prueba se identifica fácilmente si el sustentante tiene o no el dominio con base en la respuesta. Algunos ejemplos de este tipo de reactivos son: de opción múltiple, de identificación, de jerarquización, lista de comprensión y de correspondencia.

2.5.3 Reactivo de opción múltiple.

Un reactivo de opción múltiple se define como la formulación de una pregunta o problema que consta de un *enunciado* y una serie de respuestas, llamadas

opciones. Entre éstas hay una correcta llamada *solución*, y otras incorrectas conocidas como *distractores.*

A continuación se presenta una breve justificación del porqué los reactivos de opción múltiple son adecuados:

- Pueden ser aplicados a grandes cantidades de sustentantes a la vez.
- Son calificados de manera expedita a través de medios informáticos.
- Son validados de manera teórica y empírica.
- Se construyen con base en la tabla de especificaciones.
- Disminuye la probabilidad de que el sustentante conteste al azar entre más distractores utilice.
- Permiten diferenciar (discriminar) al sustentante que sabe del que no sabe.

2.5.4 Elaboración de reactivos.

La construcción de reactivo es el proceso mediante el cual un especialista elabora preguntas basadas en el contenido de diversas fuentes de información, siendo la *tabla de especificaciones* el referente que se debe seguir para la elaboración de éstos, ya que la tabla procede directamente del perfil referencial.

Cada reactivo representa por sí mismo un problema que, al ser resuelto, permita demostrar que se ha alcanzado un objetivo determinado, el cual está relacionado con la habilidad.

Los reactivos deberán ser independientes entre sí, sin que la información contenida en uno sugiera la solución de otro, ni la respuesta de un reactivo antecedente sea necesaria para poder contestar un reactivo posterior, sin traslaparse en contenido o vocabulario, y la elección de una respuesta debe estar basada, fundamentalmente, en la comprensión de la estructura de dicho reactivo y no de otros.

2.5.5 Teoría Clásica de los Tests (TCT).

La teoría de tests proporciona modelos de calificación de las pruebas a partir de matrices de datos que contienen las respuestas a los reactivos. El problema central de la teoría de los tests es la relación que existe entre el grado de capacidad del sujeto (variable latente) y la puntuación observada en el test. Dicho de otro modo, el objetivo de cualquier teoría de tests es realizar inferencias acerca del nivel de la habilidad a partir de las respuestas.

Por tanto, para estimar la habilidad es necesario relacionar ésta con los resultados observables en la prueba, esta relación debe ser adecuadamente descrita por una función matemática. Las distintas teorías de tests difieren justamente en la función que utilizan para relacionar la actuación observable en la prueba con el nivel del sujeto en la variable inobservable.

En cualquier caso, esta función da siempre cuenta de la información contenida en la matriz de entrada a partir de las características de interés que supuestamente está midiendo la prueba y el error que de forma inevitable se introduce siempre en cualquier proceso de medición, ya sea de características físicas o psicológicas.

El modelo estadístico general asume que las puntuaciones observadas X de una prueba dependen de la habilidad real θ y el error de medición e, formalmente dicho de la siguiente manera:

$$X = \theta + e \qquad (2.5\text{-}1)$$

Bajo el supuesto que el nivel real de la habilidad del sujeto, en la característica de interés, es la media de los valores que se obtendrían de forma empírica en caso de administrar la misma prueba al sujeto en idénticas condiciones, un número infinito de veces. El nivel de la habilidad es la esperanza de las puntuaciones

(θ=*E(X)*) y el valor esperado del error de medida es cero (*E(e)=0)* y presenta varianza del error homogénea para cada reactivo *(*$(\sigma_{e_i}^2 = \sigma_{e_j}^2)$.

Las principales teorías de tests que han surgido en el campo de la psicometría son: Teoría Clásica de los Tests, y Teoría de Respuesta a los Ítems. Esta última es el principal interés de esta investigación. Sin embargo, exploremos un poco sobre la teoría clásica por su gran tradición.

La TCT se sintetiza diciendo que el valor esperado de la puntuación verdadera θ es igual al valor esperado de la puntuación empírica *X:*

$$E(\theta) = E(X)$$

La ecuación de regresión de la puntuación empírica sobre la puntuación verdadera es la ecuación lineal que pasa por el origen y que tienen el valor unidad como pendiente de la recta:

$$E\left(X \mid \theta = \theta_k\right) = \theta_k$$

La varianza de las puntuaciones empíricas en una prueba es igual a la suma de la varianza de las puntuaciones verdaderas más la varianza de los errores de medida:

$$\sigma_x^2 = \sigma_\theta^2 + \sigma_e^2$$

La covarianza entre las puntuaciones empíricas y las puntuaciones verdaderas es igual a la varianza de las puntuaciones verdaderas:

$$COV\left(X, \theta\right) = \sigma_\theta^2$$

La razón de la varianza de las puntuaciones verdaderas con respecto a la varianza de las puntuaciones empíricas es igual al cuadrado del coeficiente de correlación

entre las puntuaciones empíricas y sus correspondientes puntuaciones verdaderas:

$$\rho_{x\theta}^2 = \frac{\sigma_\theta^2}{\sigma_x^2}$$

El modelo así formulado, con los supuestos y derivaciones, resulta inoperante porque todos contienen elementos no observables. Spearman introduce el concepto de tests paralelos, que le permitirá operar empíricamente con las puntuaciones obtenidas por los sujetos en los tests.

Los test paralelos implica que $X = \theta + e$ y $X' = \theta + e'$ tal que $X \neq X'$ y $e \neq e'$:

$$\left.\begin{array}{l} x = \lambda_x \xi + \delta_x \\ x' = \lambda_x \xi + \delta_{x'} \end{array}\right\} \Rightarrow \begin{array}{l} \lambda_x = \lambda_{x'} \\ \theta_{\delta_x} = \theta_{\delta_{x'}} \end{array}$$

$$\hat{\rho}_{xx'} = \lambda_x . \lambda_{x'}$$

La estimación del nivel de la habilidad basada en la distribución normal de los errores es: $\hat{\theta} = X$ y el error típico de la medida es $\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}}$ y el intervalo de confianza se estima en base en $P\left(X - Z\,\sigma_e \leq \theta \leq X + Z\,\sigma_e\right) \leq \alpha$.

Con respecto a las características cuantitativas de los reactivos, la TCT proporciona dos indicadores clásicos para determinar su calidad: Índice de dificultad e Índice de discriminación, los cuales son ampliamente utilizados por su facilidad de cálculo e interpretación intuitiva.

El **índice de dificultad** es la proporción de sujetos con respuesta incorrecta en un determinado reactivo (algunos autores consideran para el cálculo la proporción de respuesta correcta); es válido para pruebas de potencia, donde se espera que los reactivos no contestados reflejen desconocimiento por parte del sujeto. Si la prueba es de velocidad, la proporción no tiene que ver con la aptitud, sino que es

una característica de la posición del reactivo en la prueba, por lo tanto este índice no es adecuado para pruebas de velocidad (Gulliksen, 1950, p. 367).

En general, "un ítem de calidad se espera que sea acertado por la mitad de los estudiantes y fallados por la otra mitad restante, siempre y cuando la mitad de los estudiantes que acierten, sea la de mejor aprovechamiento y viceversa", Rosado (2003).

El **índice de discriminación** permite determinar si el reactivo lo aciertan más frecuentemente los sustentantes de alto rendimiento comparados con los de bajo rendimiento. El poder de discriminación es la capacidad del reactivo para diferenciar a los sustentantes que saben de los que no saben.

En la mayor parte de las pruebas convencionales de inteligencia, aptitudes y rendimiento académico, la discriminación del reactivo normalmente se valora con respecto a la puntuación total (criterio interno), enfatizando la validez de constructo. El ítem que discriminan frente a un criterio externo que la prueba pretende predecir maximiza la validez, mientras que el reactivo que discriminan según puntuación total maximiza la homogeneidad o la consistencia interna de la prueba. Ambos objetivos son deseados, aunque la aplicación de ambos criterios puede llevar a resultados diferentes e incluso opuestos. Purificar la prueba de cara a la consistencia interna puede reducir la validez externa.

El procedimiento para el cálculo de los índices de dificultad y de discriminación consiste en identificar al grupo de alto y bajo rendimiento conformado por los sustentantes que están por arriba y por debajo de la puntuación mediana o bien, una forma de que los indicadores sean más sensibles se conforman los grupos con el 33% de los sustentantes extremos.

Se aplica la siguiente fórmula para obtener el grado de dificultad del reactivo *i*:

$$DIF_i = (E_a + E_b) / N \qquad \textbf{(2.5-2)}$$

Donde: E_a y E_b. es el número de errores del reactivo *i* de la población del grupo de rendimiento alto y bajo respectivamente y *N* es el número de sujetos en ambos grupos.

El criterio para valorar cada reactivo y aceptarlo es que tenga un valor entre el 0.20 a 0.80 con un valor óptimo del 0.50 de acierto (Rosales, 2003).

El grado de discriminación del reactivo *i* se estima:

$$DIS_i = (A_a - A_b) / (N/2) \qquad \textbf{(2.5-3)}$$

Donde: A_a y A_b es el número de aciertos de la población del grupo alto y bajo rendimiento respectivamente en el reactivo *i* y *N* es el número de sujetos en la muestra de ambos grupos.

El estadístico DIS varía dentro de un inérvalo de -1 a 1, valores positivos indican que el reactivo discrimina a favor del grupo superior, valores negativos indican que el reactivo tiene una discriminación inversa a los esperado.

Cuando se utiliza la puntuación total para conformar los grupos, se recomienda eliminar el efecto del propio reactivo, excluyente el puntaje que este proporciona al puntaje total.

Una limitación de estos índices es que no posee una distribución muestral conocida y no podemos establecer significancia estadística ni intervalo de confianza.

El criterio para valorar cada reactivo y aceptarlo es que tenga un valor arriba de cero, el CENEVAL adopta los criterios definidos por Abel (1965) para la interpretación del indicador basado en la práctica presentada en la siguiente tabla.

Tabla1.0.1 Calidad del reactivo de acuerdo con el poder de discriminación.

Discriminación	Calidad	Recomendaciones
Más de 0.39	Excelente	Conservar
0.30 – 0.39	Buena	Posibilidades de mejorar
0.20 – 0.29	Regular	Necesidad de revisar
0.00 – 0.20	Pobre	Descartar o revisar profundamente
Menos de 0.0	Pésima	Descartar definitivamente

Es recomendable analizar ambos indicadores de forma integral para un mejor resultado. Tristán (1996) ha hecho notar que el grado de dificultad y el poder de discriminación no son independientes. Un reactivo muy fácil o muy difícil no discrimina porque lo resuelven casi todos o casi nadie. En la siguiente tabla presenta los valores máximos que pueden alcanzar la discriminación para reactivos con diferentes porcentajes de respuestas incorrectas.

Tabla 0.2 Relación entre el valor máximo de discriminación y la dificultad.

Dificultad	1	0.9	0.7	0.5	0.3	0.1	0
Máxima discriminación	0	0.2	0.6	1	0.6	0.2	0

Esta idea establece que el poder de discriminación, en vez de tener un valor fijo, satisfaga una norma discriminativa (*ND*) que dependa del grado de dificultad. Los valores propuestos son:

$$ND = 0.3DIF \text{ ; Si } 0\% \leq DIF < 79.92\% \qquad \textbf{(2.5-4)}$$

$$ND = 100 - DIF \text{ ; Si } 76.92\% \leq DIF \leq 100\%$$

Para verificar si un reactivo cumple con la norma, se define la relación discriminativa (*RD*), que es el cociente del poder de discriminación y la norma discriminativa.

Si la relación discriminativa es mayor que 1, el poder de discriminación es mayor que la norma y el reactivo se acepta automáticamente, si es menor, se recomienda analizarlo en cuanto a contenido y redacción.

Otro indicador interesante es el coeficiente de correlación biserial (r_{bis}) que se calcula para determinar el grado en el que la capacidad que mide la prueba también las mide el reactivo. La correlación biserial proporciona una estimación de la correlación producto-momento de Pearson entre la calificación total de la prueba y el hipotético del reactivo, cuando éste se dicotomiza en respuestas correctas e incorrectas (Henrysson, 1971).

La correlación del punto biserial (r_{pbis}) se utiliza para saber si las personas "adecuadas" son las que obtienen las respuestas correctas, qué tanto poder predictivo tiene el reactivo y cómo puede contribuir a las predicciones. Henrysson (1971) sugiere que la correlación del punto biserial nos dice más sobre la validez predictiva del test que el coeficiente de correlación biserial, ya que éste tiende a favorecer los reactivos de dificultad media.

Ebel y Frisbie (1986) afirman que la correlación del punto biserial simplemente describe la relación entre las respuestas a un reactivo (0 o 1) y las calificaciones en la prueba de todas las personas. La ecuación para obtener este indicador, de acuerdo a Glass y Stanley (1986), es la siguiente:

$$r_{pbis} = \frac{x_1 - x_0}{s_x}\sqrt{\frac{n_1 n_0}{n(n-1)}} \qquad \textbf{(2.5-5)}$$

Donde: x_1 y x_0 es la media de las puntuaciones totales de aquellos que respondieron correctamente e incorrectamente el reactivo *i respectivamente*; S_x es la desviación estándar de las puntuaciones totales; n_1 y n_0 es el número de casos que respondieron correctamente e incorrectamente el reactivo respectivamente y n es la suma de n_1 y n_0.

Para maximizar la confiabilidad se seleccionarán ítems con altas inter correlaciones entre sí; es decir, que correlacionen alto con el puntaje total en el test (criterio interno) y para maximizar la validez se seleccionarán ítems que correlacionen alto con el criterio externo que se desea predecir.

En casos de desequilibrio se corre un riesgo: la maximización de la confiabilidad puede producir una prueba muy homogénea que tiende a evaluar una sóla variable, pero con poca relación con un criterio externo complejo o heterogéneo. En contraposición, la selección de ítems en función del criterio externo, puede producir un test que correlaciona alto con el criterio que se desea predecir, pero de baja consistencia interna, lo que lo hace de difícil interpretación.

Una forma de evitar ambos riesgos es seleccionar los ítems que correlacionan significativamente con los criterios interno y externo.

Las ecuaciones (2) y (3) permiten determinar ambos índices, donde rbis(ci) es la correlación biserial y el criterio interno, rbis (ce) es la correlación biserial con el criterio externo y Y_i es la ordenada en la distribución normal en la que se divide el área bajo la curva en las proporciones p y 1-p.

$$\begin{aligned} &(2)\ \text{rbis (ci)}\ Y_i \\ &(3)\ \text{rbis(ce)}\ Y_i \end{aligned} \qquad \textbf{(2.5-6)}$$

El valor obtenido en (2) es una medida de la contribución del ítem a la varianza total y por ello a la confiabilidad; la suma de estos valores para todos los ítems

permite establecer la desviación típica de la prueba total. El valor obtenido en (3) representa la contribución del ítem individual a la validez de la prueba total. La ecuación (2.5-7) permite deducir que el valor numérico del coeficiente de validez es la proporción entre la suma de los índices de validez de los ítems y la suma de los coeficientes de confiabilidad.

$$r = \frac{\sum r_{bis(ce)} Y_i}{\sum r_{bis(ci)} Y_i} \qquad \textbf{(2.5-7)}$$

Afirma Magnusson: "Para obtener, un criterio determinado, la más alta validez posible en la prueba, debemos, obviamente, escoger ítems que tengan una alta correlación con el criterio externo y bajas correlaciones con los puntajes en la prueba. Un ítem cuyo índice de validez es igual a su índice de confiabilidad contribuye así en igual forma a la validez del test y a su confiabilidad".

El modelo clásico se produce la paradoja de que un ítem es fácil o difícil según la aptitud de los examinados y la aptitud de los examinados depende de que los ítems del test sean fáciles o difíciles. Esta deficiencia se debe en primer lugar a las propiedades de los ítems cuyos valores dependen de la habilidad de muestra normativa y, en segundo lugar, las estimaciones de la habilidad dependen de los ítems del test (Nunnaly y Bernstein, 1995).

2.5.6 Teoría de la Respuesta al Ítem (TRI).

El origen de estos modelos puede encontrarse en Lazarfeld (1950), pero es en la década del sesenta cuando la obra de George Rasch marca la pauta aunque quienes han divulgado mejor estos modelos han sido Lord y Birnbaum en la obra de Lord y Novick (1968) y Lord (1980).

El modelo TRI especifica que la probabilidad de que una persona *j* responda el ítem *i* correctamente $P(X_{ij}=1)$ depende de la habilidad de la persona usualmente denominada θ_j y las propiedades psicométricas del ítem *i*.

Los tres supuestos básicos son:

1. Unidimensionalidad de la variable latente. Se refiere a la naturaleza del rasgo a medir que está conformado por un único componente para describir adecuadamente la habilidad.

2. Independencia. Las respuestas de un examinado a cualquier par de ítem son independientes y no existe relación entre las respuestas de un examinado a diferentes ítems.

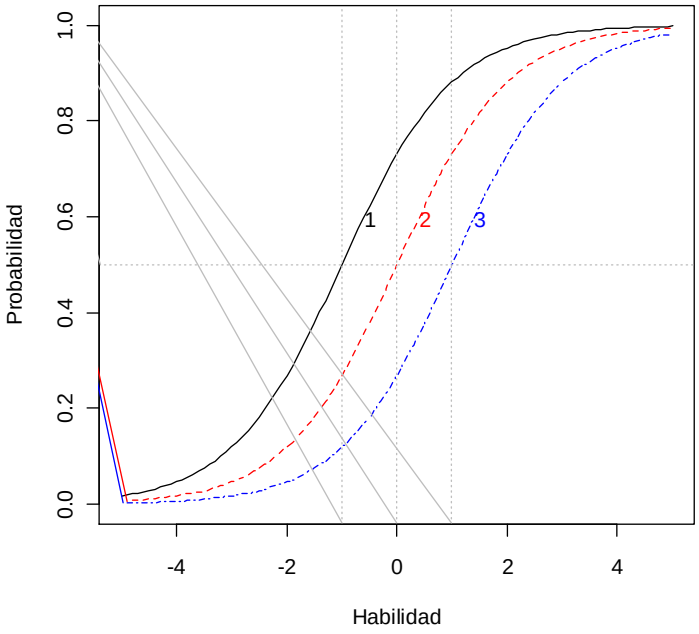
3. La relación entre el patrón de respuestas de una persona y el nivel de la habilidad puede ser descrita por una función monótonamente creciente llamada Curva Característica del Ítem (CCI).

La CCI tiene la forma de S semejante a la función de distribución acumulada de una variable aleatoria, por lo que es practico utilizar para la regresión del modelo dicotómico la función logística (modelo logit) y la normal (modelo probit).

La forma de la CCI describe cómo los cambios relativos del nivel de la habilidad cambian la probabilidad de la respuesta correcta.

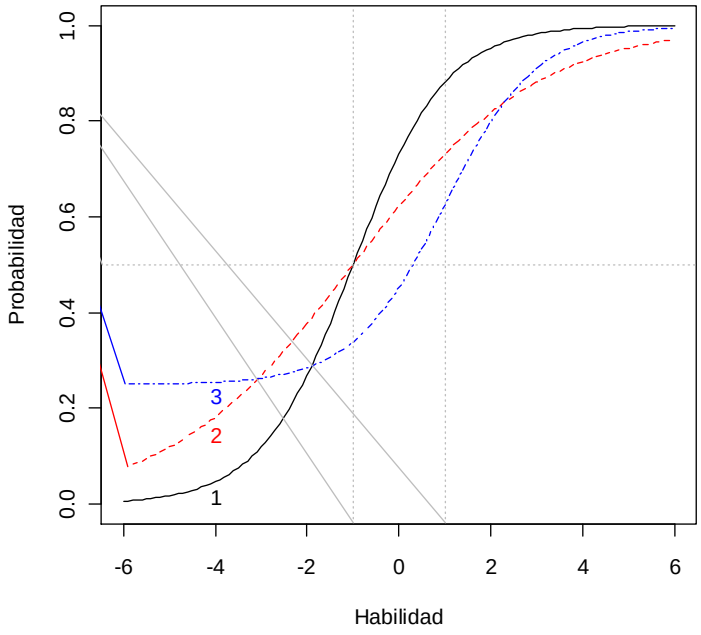
En el punto medio de la curva, los cambios pequeños en la habilidad implican grades cambios en la probabilidad de resolver el ítem. Sin embargo, los grandes cambios en los extremos de la habilidad producen muy pequeños cambios en la probabilidad.

Figura 0.1. Curva característica del ítem con distintos grados de dificultad.



La forma de las tres curvas son iguales pero difieren en la localización, la habilidad correspondiente a la probabilidad de 0.50 es mucho más baja habilidad en el ítem 1 que en el 2 y que del 3; así que el ítem 1 es más fácil, por lo tanto, la localización corresponde a la dificultad del ítem (β_i), mientras más grande sea β_i con respecto a la media de la habilidad (μ_θ) más difícil es el ítem, a diferencia de la Teoría Clásica de los Tests que se refiere a la proporción de respuestas correctas del ítem en toda la población.

Figura 0.2. Curva característica del ítem con distintos parámetros.



El parámetro de escala (α_i) indica la potencia de discriminación del ítem, siempre es positivo y típicamente menor 2.5, éste describe que tan rápido cambia la probabilidad con respecto al nivel de la habilidad. Por ejemplo, los cambios en el ítem 2 son mucho más lentos que el ítem 1; por lo tanto, el ítem 2 es menos discriminante que el ítem 1. Para el ítem 3, la probabilidad de ser resuelto nunca cae a 0 como es el caso de los ítems 1 y 2; el ítem 3 modela los casos cuando el sustentante de menor capacidad responde al azar o adivina al no conocer la respuesta.

Respecto al supuesto de la independencia se refiere a la probabilidad de respuesta correcta de un ítem depende únicamente del nivel de la habilidad y de las características del ítem γ, es decir, es independiente a la probabilidad de respuesta de cualquier otro ítem.

$$P(X_{ij}=1 \mid X_{kj}=1,\gamma_i,\gamma_k,\theta_j)=P(X_{ij}=1 \mid X_{kj}=1,\gamma_i,\theta_j)$$

El conocimiento de la respuesta de otro ítem *k* (*k≠i*) no cambia la probabilidad de la respuesta correcta del ítem *i*.

La ecuación (2.5-8) corresponde a la función logística acumulada, P_{ij} es la probabilidad correspondiente al valor esperado de responder correctamente condicionada a la características del ítem y la habilidad, por lo tanto, el rango está entre 0 y 1, y Z_{ij} está entre –∞ y +∞, P_{ij} están relacionadas en forma no lineal con Z_{ij} (es decir con θ_j)

$$P_{ij}=E(X_{ij}=1 \mid \gamma_i,\theta_j)=\frac{1}{1+e^{-Z_{ij}}} \quad \text{Donde } Z_{ij}=\xi_i+\lambda_i\theta_j \qquad (2.5\text{-}8)$$

Entonces, considerando que los momios (2.5-10) es la razón de probabilidad de acierto P_{ij} entre la probabilidad de erros (1- P_{ij}).

2.5-9

$$\frac{P_{ij}}{1-P_{ij}}=\frac{1+e^{Z_{ij}}}{1+e^{-Z_{ij}}}=e^{z_{ij}} \qquad (2.5\text{-}10)$$

El logaritmo natural de momio de acertar no solamente es lineal con la habilidad, sino también con los parámetros y es posible estimar los parámetros con mínimos cuadrado ordinario.

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=z_{ij} \qquad (2.5\text{-}11)$$

En la distribución logística, la media y la mediana coinciden ($\beta_i=\mu_i$), y la desviación σ_i es una medida de extensión de la distribución logística, de tal forma que la discriminación α_i es inversamente proporcional a la desviación σ_i.

$$Z_{ij} = (\frac{\theta_j - \mu_i}{\sigma_i}) = \alpha_i(\theta_j - \beta_i) \text{ dado que } \alpha_i = 1/\sigma_i \qquad (2.5\text{-}12)$$

El índice de discriminación α_i en el modelo Rasch es igual a 1, por lo que se conoce como el modelo logístico de un sólo parámetro (dificultad). El logaritmo natural de momios de acierto son moderados por una simple diferencia de la habilidad θ_j y la dificultad β_i,

$$\ln(\frac{P_{ij}}{1-P_{ij}}) = \log it(P_{ij}) = \theta_j - \beta_i \qquad (2.5\text{-}13)$$

Si el nivel de la habilidad es igual a la dificultad de ítem ($\theta_s = \beta_i$), entonces el logaritmo de momios de acierto será 0 y el exponente de 0 es un 1, el cual significa que es igualmente probable el acierto que el error en el ítem.

$$\ln(\frac{P_{ij}}{1-P_{ij}}) = 0 \qquad \frac{P_{ij}}{1-P_{ij}} = 1 \qquad P_{ij} = 1 - P_{ij} \qquad P_{ij} = .5$$

De lo anterior, se deduce que a medida que la habilidad es mayor que la dificultad la probabilidad de acierto es mayor que la de equivocarse y viceversa.

El modelo logístico de un parámetro predice las probabilidades de acertar condicionada la dificultad del ítem y la habilidad.

$$P_{ij} = P(X_{ij} = 1 | \beta_i, \theta_j) = \frac{1}{1 + e^{-(\theta_j - \beta_i)}} \qquad (2.5\text{-}14)$$

Las características de la CCI del modelo Rasch se muestran Figura 0.1 con tres ítems con grados de dificultad β_i de -1, 0 y 1, respectivamente, observándose lo siguiente:

1. La probabilidad incrementa de acuerdo al nivel de la habilidad para cada ítem.

2. La forma de las curvas son iguales, únicamente difieren en la localización (dificultad).

3. Las pendientes de las curvas son iguales y por lo tanto, no se cruzan.

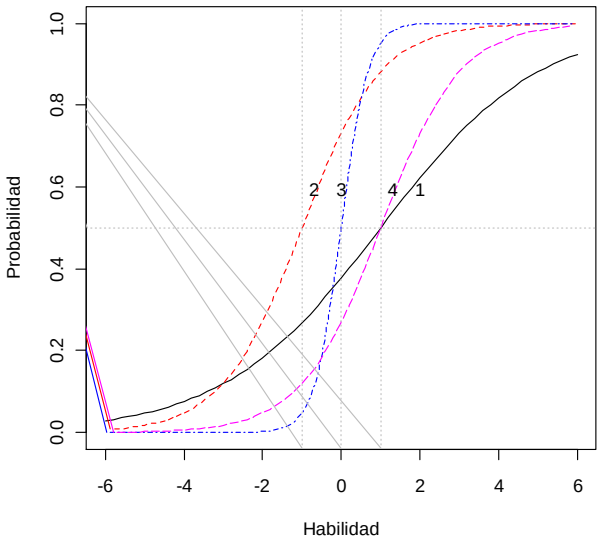
El punto de inflexión de la CCI ocurre cuando la probabilidad de acertar un ítem es 0.50 y este punto correspondiente en la escala de la habilidad al grado de dificultad. 1

La primera generalización del modelo Rasch es el modelo de dos parámetros que caracterizan al ítem: la discriminación α y la dificultad β.

$$P_{ij} = P(X_{ij} = 1 | \beta_i, \alpha_i, \theta_j) = \frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \qquad (2.5\text{-}15)$$

Note que la discriminación del ítem es un multiplicador de la diferencia entre el nivel de la habilidad y la dificultad del ítem.

Figura 0.3. Curva característica del ítem con distintos grados de discriminación.



En la grafica anterior muestra la CCI con el modelo logístico con dos parámetros, los ítem tienen la discriminación de 0 .5, 1, 3 y 1 y la dificultad de 1, -1, 0 y 1 respectivamente.

1. Los ítems 2 y 4 tienen la misma discriminación (pendiente) y únicamente difieren en la dificultad (localización), el ítem 2 es más fácil que el ítem 4.

2. El valor de discriminación del ítem 1 es relativamente bajo (0.5), por lo tanto, cambios en el nivel de la habilidad tiene menos impactos en la probabilidad, por lo que provee menos información porque el nivel de la habilidad es menos relevante a cambios.

3. El ítem 3 tiene un alto valor de discriminación (3), así que el ítem es más eficiente al alrededor del nivel de la habilidad (-1 y 0).

4. Note que la dificultad del ítem en el modelo de dos parámetros aun corresponde al nivel de la habilidad en la cual la probabilidad de acertar es de 0.50.

Los ítems de opción múltiple ocasionalmente presentan un comportamiento asintótico mayor que cero en la cola inferior de la CCI. Birnbaum (1968) señala que los sustentantes menos diestros seleccionan la respuesta correcta por azar (o adivinando), Lord (1974b) menciona que este valor asintótico es inferior de 1/m, donde m es el número de respuestas alternativas.

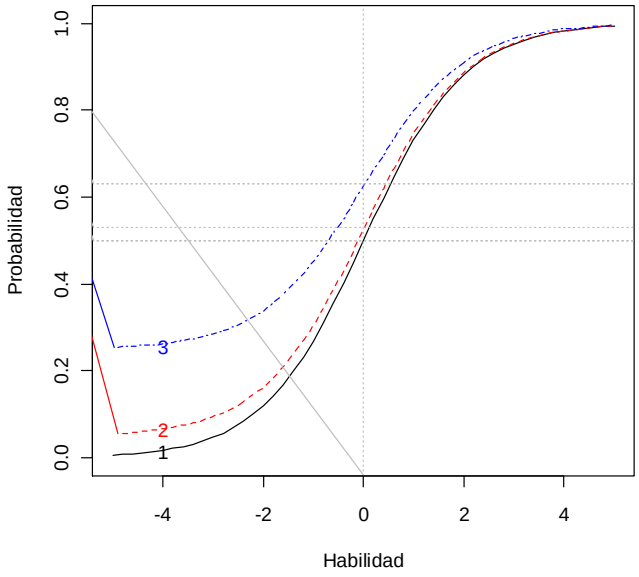
Birnbaum (1968) propuso un modelo de tres parámetros para CCI, con el cual, la probabilidad de respuesta a un ítem es:

$$P_{ij} = P\,(X_{ij} = 1 \mid \alpha_i, \beta_i, C_i, \theta_j) = C_i + (1 - C_i)\left(\frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}}\right) \qquad (2.5\text{-}16)$$

Donde C_i es la probabilidad asintótica de respuesta correcta, es decir, es el límite inferior de la curva característica.

La siguiente gráfica muestra tres ítems con igual grado de dificultad y discriminación pero con diferente grado de respuesta correcta por adivinación (0, 0.05 y 0.25).

Figura 0.4. Curva característica del ítem con distintos grados de respuesta aleatoria.



La dificultad de los ítems aún ocurre en el punto de inflexión de la CCI, pero éste es afectado por la respuesta aleatoria (*P(θ)= (1+c)/2),* por lo tanto, corresponde a 0.50, 0.53 y 0.63 respectivamente a los ítems 1, 2 y 3.

Lord y Novick (1968) señalan que la utilización de distribución acumulada de la normal en el modelado de las curvas características de ítems es ventajosa desde el punto de vista teórico, pues las probabilidades de acertar asintóticamente tienden a una distribución normal.

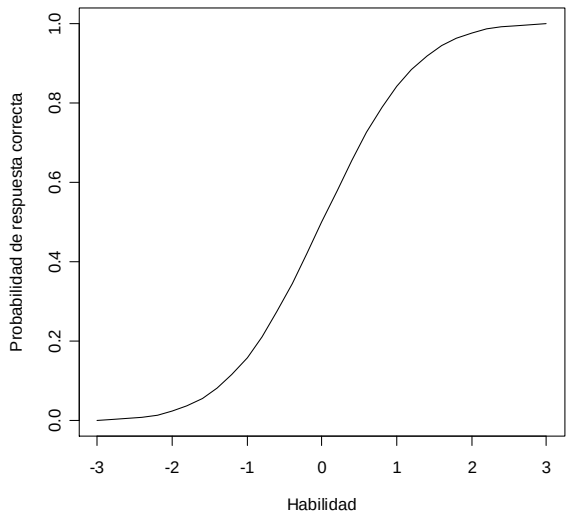
$$P_{ij} = P(\mu_i, \sigma_i, \theta_j) = \phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_{i,j}} e^{-\frac{1}{2}Z^2} dZ \qquad \textbf{(2.5-17)}$$

Donde μ_i es la media, σ_i es la desviación estándar, θ_j es la habilidad y Z_{ij} es la normal transformada.

En la distribución normal, la media y la mediana coinciden ($\beta_i = \mu_i$) y la desviación σ_i es una medida de extensión de la distribución normal, de tal forma que la discriminación α_i es inversamente proporcional a la desviación σ_i.

$$\alpha_i = 1/\sigma_i \qquad \textbf{(2.5-18)}$$

Figura 0.5. Curva característica del ítem con el modelo de la curva normal, β_i=.6 y α_i=1.2.



Para graficar la curva característica del ítem con los parámetros de βi=0.6 y αi=1.2 y considerando $\beta_i=\mu_i$ y $\alpha_i=1/\sigma_i$, entonces la normal estandarizada es:

$$Z_{ij} = \left(\frac{\theta_j - \mu_i}{\sigma_i}\right) = \alpha_i(\theta_j - \beta_i) = 1.2(\theta_j - .6) \qquad \textbf{(2.5-19)}$$

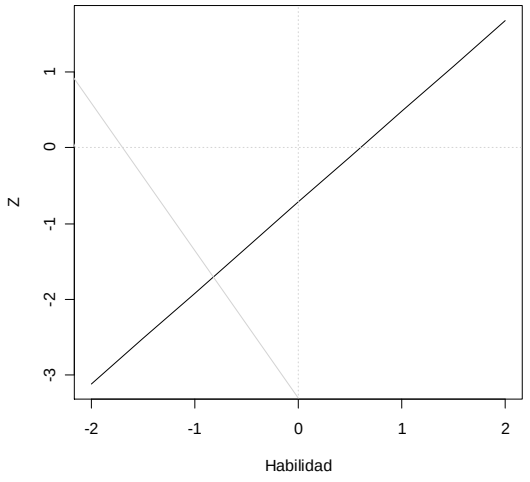
Para obtener las probabilidades de la respuesta correcta para un θ específico, se calcula la Z_{ij} con la ecuación anterior, como se muestran en la siguiente tabla:

Tabla 0.3 Probabilidad de respuesta correcta con el modelo de la curva normal, β_i=.6 y α_i=1.

Θ	θ –β	Z	$P_i(\theta)$
-2.0	-2.6	-3.12	0.0009
-1.5	-2.1	-2.52	0.0059
-1.0	-1.6	-1.92	0.0274
-0.5	-1.1	-1.32	0.0934
0.0	-0.6	-0.7	0.2358
0.5	-0.1	-0.12	0.4522
1.0	0.4	0.48	0.6844
1.5	0.9	1.08	0.8599
2.0	1.4	1.68	0.9535

La función de $P_i(\theta)$ define una relación no lineal entre la probabilidad de respuesta correcta y la habilidad θ, pero Z es una transformación de $P_i(\theta)$ de tal forma que resulta una relación lineal entre la Z y θ, para un θ_i = -0.5; la probabilidad de respuesta correcta es 0.0934 y la correspondiente Z es -1.32, graficando la Z con respecto a θ obtenemos:

Figura 0.6.Transformación de la desviación normal de la curva característica, β_i=.6 y α_i=1.2 (ζ_i=-.72 y λ_i=1.2)



Donde la ecuación de esta línea es:

$$Z'_{ij} = \xi_i + \lambda_i \theta_j \qquad \textbf{(2.5-20)}$$

Donde la ξ_i es el intercepto, λ_i es la pendiente y θ_i la habilidad, por lo tanto los parámetros son:

$$Z'_{ij} = \alpha_i(\theta_j - \beta_i) = \xi_i + \lambda_i \theta_j \qquad \textbf{(2.5-21)}$$

$$\alpha_i \theta_j - \alpha_i \beta_i = \xi_i + \lambda_i \theta_j$$

$$\alpha_i \theta_j = \lambda_i \theta_j \qquad \textbf{(2.5-22)}$$
$$\alpha_i = \lambda_i$$

El parámetro de discriminación es la pendiente de la regresión lineal de la relación de la transformación de la desviación normal de la probabilidad de respuesta correcta y la habilidad.

$$-\alpha_i \beta_i = \xi_i$$

Y sustituyendo α_i

$$-\lambda_i \beta_i = \xi_i \qquad \textbf{(2.5-23)}$$
$$\beta_i = \frac{-\xi_i}{\lambda_i}$$

Y el parámetro de la dificultad es el negativo de la razón de la intercepción y la pendiente de la regresión lineal.

La estimación de los parámetros de la curva puede ser estimados con los datos observados ajustando la regresión lineal mediante mínimos cuadrados o máxima verosimilitud para obtener el intercepto y la pendiente, y estos utilizando en las ecuaciones (2.5-22) y (2.5-23) para estimar $\hat{\alpha}_i$ y $\hat{\beta}_i$.

El modelo logístico es muy similar al a curva normal, sólo que utiliza la función logística para describir la CCI.

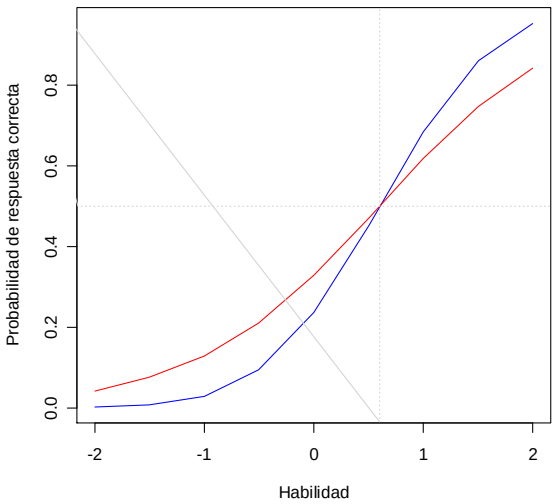
$$P_i(\theta_j) = P_i(\alpha_i', \beta_i, \theta_j) = \psi(Z) = \frac{e^{z_{ij}}}{1+e^{z_{ij}}} = \frac{1}{1+e^{-z_{ij}}} \qquad (2.5\text{-}24)$$

Donde $Z_{ij} = \alpha_i'(\theta_j - \beta_i)$ y α_i' es el parámetro de discriminación.

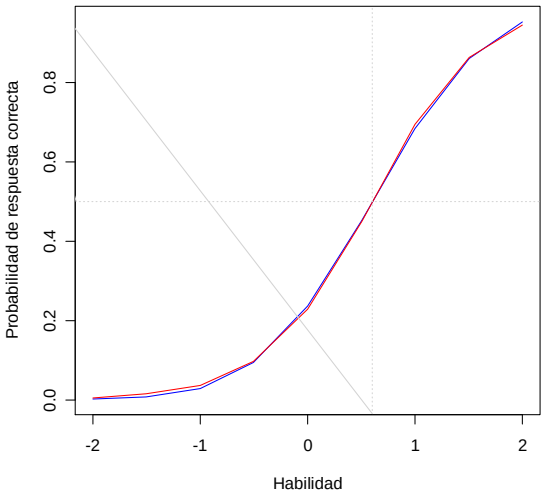
En la Figura 0.7 se grafican ambas curvas características con los parámetros de β_i=0.6 y α_i.=1.2 con el modelo de normal y logístico, La inspección general de las dos curvas revela que ambos comparten el mismo parámetro de dificultad. Sin embargo, la pendientes de las curvas no son iguales debido a la diferencias de las varianzas de la función normal y logística, Haley (1952) demostró que si a la $Z_{ij} = \alpha_i'(\theta_j - \beta_i)$ es multiplicado por 1.702 ($\alpha_i = 1.702\alpha_i'$) y se utiliza la ecuación (2.5-24), las diferencias absolutas entre la probabilidad $P_i(\theta_j)$ de ambos modelos es menos de 0.01% en todo el rango de la habilidad θ.

Figura 0.7. Comparación de curva característica con el modelo normal (azul) y logístico (rojo) con y sin ajuste.

Z sin ajustar



Z ajustada por 1.702



Afirma Lord y Novick (1968) que actualmente se utiliza más el modelo logístico por su forma cerrada y rápida evaluación, el parámetro de discriminación se usa directamente sin ajustar.

La función logística está relacionada con el logaritmo de los momios de contestar correctamente el ítem, como se mencionó anteriormente, y es posible estimar los parámetros de la siguiente forma:

$$\log\left(\frac{P_i(\theta_j)}{Q_i(\theta_j)}\right) = \log(P_i(\theta_j)) - \log(Q_i(\theta_j)) = Z \qquad \textbf{(2.5-25)}$$

Y de acuerdo con el modelo logístico:

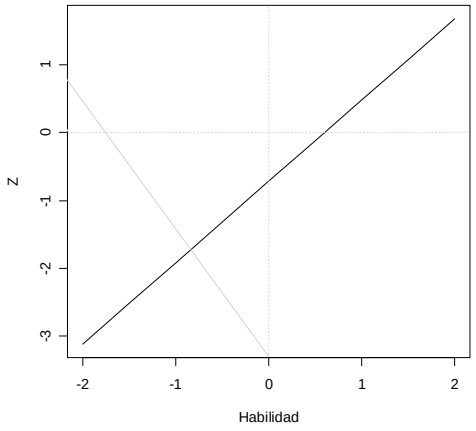
$$P_i(\theta_j) = \frac{1}{1+e^{-z_{ij}}}$$

$$e^{-z_{ij}} = \frac{1-P_i(\theta_j)}{P_i(\theta_j)} = \frac{P_i(\theta_j)}{Q_i(\theta_j)}$$

$$Z_{ij} = \log\left(\frac{P_i(\theta_j)}{Q_i(\theta_j)}\right) = \alpha_i(\theta_j - \beta_i) = \xi_i + \lambda_i\theta_j \qquad \textbf{(2.5-26)}$$

Para estimar los parámetros del ítem es conveniente usar la transformación logística para representar linealmente la relación entre θ_i y Z_{ij}, (Figura 0.8) que este ultimo corresponde a $P_i(\theta_i)$.

Figura 0.8 Transformación de la desviación logística de la curva característica, β_i=.6 y α_i=1.2.



La regresión lineal pareciera ser idéntica a la curva normal presentada en la Figura 0.6, pero la arista vertical tiene diferentes unidades de medida, para el modelo logístico es $\log(P_i(\theta_j)/Q_i(\theta_j))$ y en el caso del modelo normal, las unidades es la transformación de desviaciones de la normal de $P_i(\theta_j)$.

La estimación de los parámetros de modelo logístico de Birnbaum es similar a la anterior, transformando la CCI a una regresión lineal:

$$P_i(\theta_j) = C_i + (1 - C_i)\psi[\alpha_i(\theta_j - \beta_i)]$$

Donde $Z_{ij} = \alpha_i'(\theta_j - \beta_i)$, entonces:

$$P_i(\theta_j) = C_i + (1 - C_i)\frac{1}{1 + e^{-z_{ij}}}$$

$$e^{-z_{ij}} = \frac{1 - C_i}{P_i(\theta_j)} - 1 = \frac{Q_i(\theta_j)}{P_i(\theta_j) - C_i}$$

$$Z_{ij} = \log\left(\frac{P_i(\theta_j) - C_i}{Q_i(\theta_j)}\right) \qquad \textbf{(2.5-27)}$$

La transformación no es la misma para los dos parámetros, la arista vertical está en unidades de $\log[(P_i(\theta_j) - C_i)/Q_i(\theta_j)]$.

Tabla 0.4 Ecuaciones para el modelo logístico.

Modelo	Probabilidad de acertar
Un parámetro (Rasch)	$P_{ij} = P[X_{ij} = 1 \mid \theta_j, \beta_i] = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}$
Dos parámetros	$P_{ij} = P[X_{ij} = 1 \mid \theta_j, \beta_i, \alpha_i] = \frac{\exp(\alpha_i(\theta_j - \beta_i))}{1 + \exp(\alpha_i(\theta_j - \beta_i))}$
Tres parámetros	$P_{ij} = P[X_{ij} = 1 \mid \theta_j, \beta_i, \alpha_i, C_i] = C_i + (1 - C_i)\frac{\exp(\alpha_i(\theta_j - \beta_i))}{1 + \exp(\alpha_i(\theta_j - \beta_i))}$

Donde: θ_j es la habilidad del la persona *j* y para el ítem *i*, β_i es la dificultad, α_i es la discriminación y C_i es el parámetro de respuesta aleatoria.

Para aproximar sus valores a la curva normal acumulada es recomendable multiplicar a Z por 1.702 (D), a esta expresión se le conoce frecuentemente función logística normalizada.

La característica principal de los modelos TRI es conseguir medidas invariantes respecto al parámetro de aptitud y de los ítems, es decir, las estimaciones de la aptitud obtenidas con distintos ítem serían iguales y las estimaciones de los parámetros de los ítems obtenidos con distintas muestras de examinados serán iguales. Esta propiedad de invariancia se obtiene incorporando los parámetros de los ítems al proceso de estimación de la aptitud e incluyendo el parámetro de la aptitud en el proceso de estimación de los parámetros de los ítems.

2.5.7 Función de información.

La función de información expresa la cantidad de información que proporciona el ítem respecto al rasgo latente θ.

El estimador de la habilidad $\hat{\theta}$ es un estimador de máxima verosimilitud con distribución asintótica normal con media θ y varianza $1/I(\theta)$, donde $I(\theta)$ es la cantidad de información, así que a mayor varianza menor precisión del estimador $\hat{\theta}$ y menos información se tiene del grado la habilidad desconocida. Birnbaum (1968) ha definido la función de información de la prueba como:

$$I_t(\theta) = \sum_{i=1}^{n} \frac{[P_i^{'}(\theta)]^2}{P_i\ (\theta) Q_i(\theta)} \qquad \textbf{(2.5-28)}$$

Donde $P_i\ (\theta)$ es obtenido por la evaluación del modelo de CCI en θ y $P_i^{'}(\theta) = dP/d\theta$.

El valor de la función de información dependerá de la pendiente de la curva (cuando mayor más información) y del error estándar de medición (cuanto más pequeño mayor información). Es decir, cuanto mayor sea la pendiente y menor la variancia de un ítem, mayor será la información.

La función de información para una prueba condicionada a una habilidad θ es simplemente la suma de la función de información de los ítems en θ.

$$I_t(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad \textbf{(2.5-29)}$$

La cantidad de información de una prueba para una determinada habilidad θ es inversamente proporcional a la precisión, con la cual es estimada con la ecuación **(2.5-30**).

$$\sigma_{\hat{\theta}/\theta} = \sigma_e = \frac{1}{\sqrt{I(\theta)}} = \frac{1}{\sqrt{\sum_{i=1}^{k} D^2 \cdot P_i(\theta) \cdot Q_i(\theta)}} \qquad \textbf{(2.5-30)}$$

$\sigma_{\hat{\theta}/\theta}$ es el error estándar de la estimación y D es el valor escalable a la normalidad.

Por lo tanto, la función de información del test es:

$$I_t(\theta) = \sum_{i=1}^{n} \frac{\left(P'(u_i = 1/\theta)\right)^2}{P(u_i = 1/\theta)P(u_i = 0/\theta)} = \sum_{i=1}^{n} \frac{[D \cdot P_i(\theta) \cdot Q_i(\theta)]^2}{P(\theta)Q(\theta)} = \sum_{i=1}^{n} D^2 \cdot P_i(\theta) \cdot Q_i(\theta) \qquad \textbf{(2.5-31)}$$

La función de información del test en la TRI sustituye la confiabilidad del TCT, la ventaja conceptual es que ésta indica que tan bien es estimada la habilidad para cada nivel de habilidad, mientras que la confiabilidad de TCT es una medida de consistencia para toda la escala. De acuerdo con lo anterior, el intervalo de confianza del 95% para la habilidad estimada es $\hat{\theta} \pm 1.96\sigma_{\hat{\theta}/\theta}$.

La función de información del ítem para los modelos logísticos de la CCI es:

Un parámetro (Rasch)

$$I_i(\theta) = P_i(\theta)Q_i(\theta) \qquad \textbf{(2.5-32)}$$

Dos parámetros

$$I_i(\theta) = \frac{[P_i\ (\theta)]^2}{P_i\ (\theta)Q_i(\theta)} = \alpha_i^2\{P_i(\theta)Q_i(\theta)\} \qquad \textbf{(2.5-33)}$$

Tres parámetros

$$I_i(\theta) = \alpha_i^2\,\frac{Q_i(\theta)}{P_i(\theta)}\left[\frac{P_i\ (\theta) - c_i}{1 - c_i}\right]^2 = \alpha_i^2 P_i(\theta)Q_i(\theta)\left[\frac{P_i^*(\theta)}{P_i\ (\theta)}\right]^2 \qquad \textbf{(2.5-34)}$$

La máxima información del ítem ocurre cuando θ=β para el modelo de uno y dos parámetros, en el modelo Rasch es 0.25 y para el segundo modelo es $0.25\alpha_i^2$. El modelo de tres parámetros maximiza la función información cuando el nivel de la habilidad es:

$$\theta_{\max} = \beta_i + \frac{1}{\alpha_i}\log\left[\frac{1+\sqrt{1+8c_i}}{2}\right] \qquad \textbf{(2.5-35)}$$

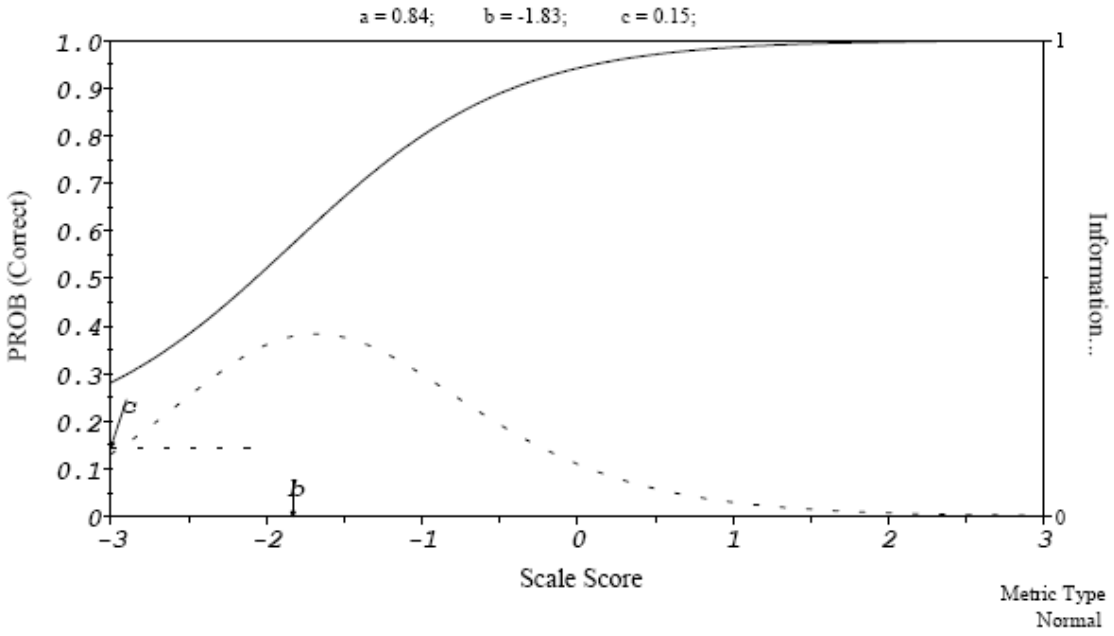
Todos los ítems tienen una función de información determinada para cada nivel θ y hay un punto de la escala donde la cantidad de información del ítem es máxima. Pues bien, para ese punto de θ es para el que el ítem es más adecuado.

En la siguiente gráfica se presenta con la línea sólida la curva característica y la punteada la función de información para un ítem con parámetros de discriminante de 0.84, con un grado dificultad de -1.83 (fácil) y con una probabilidad de acierto por azar de 0.15.

Analizando la curva característica se aprecia que la probabilidad de acierto para alumnos poco competentes es relativamente alta (aproximadamente 0.3 para θ = -3) y muy alta (más de 0.9) para estudiantes con conocimientos medios θ = 0. Por su parte, la línea de la función de información es más alta para el alumnado con puntuaciones θ entre –3 y –1. Sin embargo, no es un buen ítem para discriminar

entre rendimientos medios y altos, ya que una vez que θ toma valores positivos la altura de la curva que representa la función de información del ítem se vuelve despreciable.

Figura 0.9 Curva característica y función de información de un ítem fácil.



La función de información presenta implicaciones prácticas muy importantes para los evaluadores y constructores de pruebas objetivas. Con un banco de ítems calibrado es posible seleccionar aquellos que mejor se ajusten al nivel de los evaluados. Centrándose en la evaluación de la competencia se elegirán ítems fáciles o muy fáciles para estimar el nivel de las personas con menos talento. Por el contrario, con sustentantes más capaces se seleccionarían aquellos ítems que ofrecen máxima información en los niveles θ más altos, de esta forma se tendrían test con la precisión deseada en el rango de la habilidad de interés.

El banco de ítems calibrados es el punto de partida para la construcción de pruebas de evaluación rápidas adaptadas a diferentes niveles de rendimiento. Incluso, el mercado ya ofrece programas como BILOG y MICROCAT que, además de calibrar ítems y estimar niveles θ, permite realizar este tipo de evaluaciones de forma automatizada.

2.5.8 Calibración de ítems.

El proceso de estimación de los parámetros del modelo se denomina calibración de ítems.

La curva característica de un ítem es representada cuantitativamente por sus parámetros y éstos pueden ser estimados por máxima verosimilitud de la siguiente marera: se aplica el ítem sin presion de tiempo a una muestra de la población objetivo a quienes se les conoce previamente el grado de la habilidad θ_j, para cada nivel de habilidad (j =1, 2, 3,.., k) se contabilizan las personas f_j y el número de personas r_j que contestan correctamente el ítem. La proporción observada de respuestas correctas de la habilidad θ_j es:

$$p_j(\theta_j) = p_j = \frac{r_j}{f_j} \qquad \textbf{(2.5-36)}$$

$$q_j(\theta_j) = q_j = \frac{f_j - r_j}{f_j}$$

Asumimos que la distribución de r_j corresponde a una binomial con parámetros f_j y P_j, donde P_j es la probabilidad real de responder correctamente el ítem, por lo tanto, $E(r_j)]=f_jP_j$ y la varianza de r_j es $f_jP_jQ_j$.

$$E(p_j) = E(\frac{r_j}{f_j}) = \frac{1}{f_j} E(r_j) = P_j \qquad \textbf{(2.5-37)}$$

Considerando las respuestas correctas de k grupos de habilidades, la probabilidad del vector R=(r_1, r_2, r_3,.., r_k) está determinado por la función de verosimilitud.

$$P(R) = \prod_{j=1}^{k} \frac{f_j!}{r_j!(f_j - r_j)!} P_j^{r_j} Q_j^{f_j - r_j} \qquad \textbf{(2.5-38)}$$

Y la log verosimilitud es:

$$L = \log P(R) = const + \sum_{j=1}^{k} \log P_j + \sum_{j=1}^{k} (f_j - r_j) \log Q_j \qquad \textbf{(2.5-39)}$$

Los parámetros de la CCI se estiman al maximizar la log verosimilitud *L* considerando la función de distribución logística acumulada para el modelo de dos parámetros:

$$P_j(\theta_j) = P_i(\alpha_i', \beta_i, \theta_j) = \psi(Z) = \frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \qquad \textbf{(2.5-40)}$$

$$Q_j = Q_j(\theta_j) = 1 - P_j(\theta_j)$$

Y para el modelo de tres parámetros es:

$$P_j = c + (1 - c)\psi(Z) \qquad \textbf{(2.5-41)}$$

$$Q_j = [1 - \psi(Z)](1 - c)$$

Donde:

$$\psi(Z_j) = P_j^* = P^*(\theta_j) = \frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \qquad \textbf{(2.5-42)}$$

Una vez estimados los parámetros, se requieren evidencias que permitan sostener que el modelo es apropiado y se ajusta a los datos. Evidencias que permitan hacer un juicio global de la adecuación del modelo (Hambleton y Rogers, 1991).

Los aspectos que hay que analizar son:

- Cumplimiento de los supuestos de la unidimensionalidad y la independencia local.
- Ajuste del modelo mediante el análisis de residuales, bondad de ajuste basado en ji-cuadrado o la comparación de las distribuciones de puntuaciones predichas y observadas.
- Invarianza de los parámetros de los sujetos a través de diferentes grupos de ítems y la invarianza de los parámetros de los ítems en diferentes muestras de sujetos.

Es importante puntualizar que para la calibrar los modelos se requiere de muestras grandes (más de 300), para pequeñas el mejor modelo es el de Rasch y por esto ha sido el más popularizado.

2.5.9 Análisis de alternativas.

Las deficiencias de las preguntas de opción múltiple son frecuentemente consecuencias del comportamiento inesperado de la respuesta correcta y de los distractores, se recomienda verificar mediante técnicas estadísticas, lo siguiente:

- Todos los distractores deben de ser plausibles de ser seleccionados.
- La respuesta correcta debe ser seleccionada con mayor frecuencia por los sustentantes con más altas calificaciones.
- Los distractores deben ser seleccionadas con mayor frecuencia por los sustentantes con más bajas calificaciones.

2.6 *Estimación de la habilidad.*

El objetivo principal es obtener la media de la habilidad real para cada examinado en base la probabilidad del patrón de las respuestas de los ítems. Para obtener esta estimación de máxima verosimilitud se beben cumplir tres supuestos:

1. Los parámetros de los ítems deben ser conocidos.
2. Los sustentantes son independientes y puede ser estimada la habilidad, sustentante por sustentante.
3. Todos los ítems son moderados por la CCI de la misma familia.

Suponga: un determinado sustentante responde n ítems de una prueba y las respuestas son calificadas en correcta o incorrecta, $u_{ij}=0,1$, donde i indica el ítem ($1<=i<=n$) y j el sustentante ($1<=j<=N$) produciendo un vector de respuesta denominado $U_{ij}=(u_{1j},u_{2j},u_{3j},..u_{nj}|\theta_j)$. Con el supuesto de independencia local, las u_{ij}

son estadísticamente independientes, de tal forma que la probabilidad del vector de respuestas para el examinado está determinada por la función de verosimilitud:

$$P(U_j \mid \theta_j) = \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \qquad \textbf{(2.6-1)}$$

Y aplicando el logaritmo natural a la función de verosimilitud:

$$\mathrm{L} = \log P(U_j \mid \theta_j) = \sum_{i=1}^{n} [u_{ij} \log P_{ij} + (1-u_{ij}) Q_{ij}] \qquad \textbf{(2.6-2)}$$

La P_{ij} es función de la CCI que está expresada en los parámetros conocimos y de acuerdo al modelo se utilizarán las ecuaciones (2.5-40) y (2.5-41) para encontrar el nivel de habilidad que maximiza la log verosimilitud.

Para ilustrar el procedimiento asumimos una prueba de cinco preguntas (de un sólo parámetro) con dificultad de -2, -1, 0, 1 y 2 respectivamente y deseamos estimar el nivel de la habilidad del sustentante que contestó correctamente las cuatro primeras preguntas.

Si el nivel de la habilidad y la dificultad de los ítems son conocidos es posible calcular la probabilidad de contestar correctamente cada ítem P_{ij} , y con éstas, la probabilidad del patrón de respuestas. La Tabla 0.5 presenta las probabilidades P_{ij} para el sustentante j con dos niveles hipotéticos de la habilidad correspondientes a -1 y 2.

Tabla 0.5 Probabilidades del patrón de respuestas de un sustentante.

Ítem	Dificultad	U_j	Término	Nivel de la habilidad	
				θ_j=-1	θ_j=2
1	-2	1	P_{1j}	0.73	0.98
2	-1	1	P_{2j}	0.50	0.95
3	0	1	P_{3j}	0.27	0.88
4	1	1	P_{4j}	0.12	0.73
5	2	0	$1-P_{5j}$	0.95	0.50

Por ejemplo, La probabilidad de acertar la primera pregunta (β=-2) con el supuesto de habilidad de θ=2 es 0.98.

$$P_{1j} = P\,(U_{ij} = 1 \mid \theta_j = 2, \beta_1 = -2) = \frac{e^{\theta_j - \beta_i}}{1 + e^{\theta_j - \beta_i}} = \frac{e^{2+2}}{1 + e^{2+2}} = 0.98$$

Por lo tanto, la verosimilitud de la persona *j* dado el patrón de respuesta (1, 1, 1, 1, 0) indica que es más probable el nivel de la habilidad sea 2 que -1:

$$P(1,1,1,1,0|\theta_j=-1)=0.73*0.5*0.27*0.12*0.95=0.01$$

$$P(1,1,1,1,0|\theta_j=2)=0.98*0.95*0.88*0.73*0.5=0.30$$

El nivel de la habilidad correspondiente a cada sustentante es aquel donde la función de log verosimilitud se maximiza. Baker (1992) sugiere utilizar la técnica de Newton Raphson para obtener la estimación de la habilidad vía un procedimiento interactivo que emplea la primera y segunda derivada con respecto a θ de la log verosimilitud.

Cuando el grado de habilidad y los parámetros de los ítems son desconocidos y no es posible realizar una prueba piloto para estimarlos, las respuestas de todos los examinados de todos los ítems deben ser consideradas simultáneamente, la estimación de máxima verosimilitud deberá ser realizada en dos pasos:

- Primero. Se estiman los parámetros de los ítems con base en un parámetro de una habilidad inicial (éste puede ser el puntaje estandarizado).

- Segundo. Los parámetros de los ítems resultantes son tratados como conocidos y en función de éstos se estima la habilidad.

Este procedimiento es repetido hasta que el valor de la estimación no cambie entre dos sucesivas iteraciones de estimación.

La desventaja de este último procedimiento es que las estimaciones no son consistentes en el modelo de dos y tres parámetros. Además, las personas con puntaje máximo o cero deben ser excluidas del proceso, ya que estas habilidades no existen, al igual que los reactivos que son contestados todos correcta o incorrectamente.

El estimador $\hat{\theta}$ está sujeto a errores de estimación producidos por el propio modelo, para determinar el nivel real en que un sujeto posee el rasgo latente que mide el test, se utiliza la siguiente formula general:

$$P\left(\hat{\theta} - Z\sigma_e. \leq \theta \leq \hat{\theta} + Z\sigma_e\right) < \alpha \qquad \textbf{(2.6-3)}$$

Donde el error de medida a nivel del ítem es *e=u-P(u=1|θ)* y a nivel test es $e = X - \hat{\theta} = \sum_{i=1}^{n} u_i - \sum_{i=1}^{n} P(u_i = 1 \mid \theta)$.

El error típico de medida σ_e para un determinado nivel de $\theta = \theta_j$ puede calcularse como la raíz cuadrada de la varianza muestral de la estimación del parámetro θ ofrecida por el test:

$$\sigma_e = \sigma_{\hat{\theta}|\theta} = \sqrt{\sum_{i=1}^{n} P_i(\theta_j) Q(\theta_j)} \qquad \textbf{(2.6-4)}$$

Donde n es el número de ítems del test, $P_i(\theta_j)$ es el valor de las *CCI para* $\theta=\theta_j$ *y* $Q_i(\theta_j)$ es igual a 1- $P(\theta_j)$.

A su vez, la inversa de la varianza muestral de la estimación del parámetro θ ofrecida por el test nos da información de la precisión de la medida correspondiente a la función de información del test. Por lo tanto la estimación del error es: $\sigma_e = \sigma_{\hat{\theta}|\theta} = \frac{1}{\sqrt{I(\theta)}}$

2.7 *Calificación mínima aprobatoria.*

Las evaluaciones referidas a criterios como ya se mencionó, clasifican a los sustentantes en acreditados o no en base a un umbral correspondiente al nivel elemental de dominio de la habilidad, la cual se denomina calificación mínima aprobatoria.

Considerando la clasificación en una tabla de contingencia de 2x2, donde la clasificación de los expertos está representada por y con base en el desempeño observado y la clasificación estimada $\hat{y}$ en función del estimador de la calificación real. La calificación umbral consiste en determinar aquella que minimice los errores de clasificación representados por los falsos positivos (n_{01}) y negativos (n_{10}), para tal efecto, una técnica es utilizar la regresión logística.

Tabla 0.6 Tabla de contingencia correspondiente a la clasificación de expertos y examen.

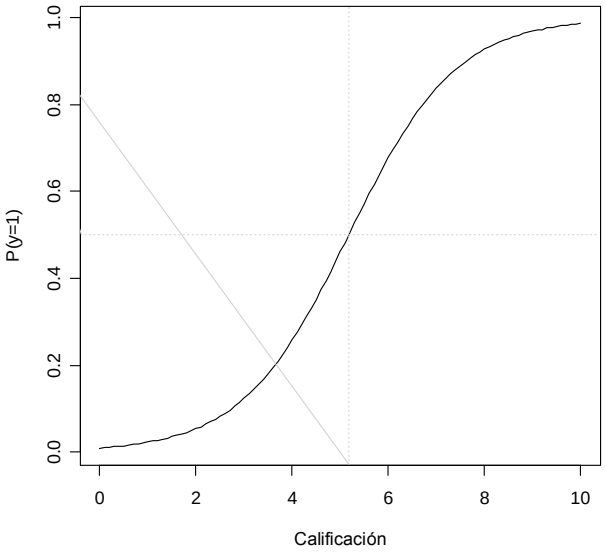
Clasificación de expertos (y)	Clasificación del examen ($\hat{y}$)	
	No aprobado	**Aprobado**
No apto	n_{00}	n_{01}
Apto	n_{10}	n_{11}

Este modelo estima la probabilidad de clasificar a un individuo como apto con base en las calificaciones obtenidas, expresado de la siguiente forma:

$$P(y_i = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \qquad \textbf{(2.7-1)}$$

Donde β_0 y β_1 son coeficientes específicos del experto evaluador y x_i es la calificación de sustentante *i*.

Figura 0.10 Probabilidad de aprobar en función de la calificación.



La Figura 0.10 muestra la gráfica de probabilidad de ser clasificado como apto con base en la calificación. La calificación umbral CU se determina cuando la probabilidad es 0.5 y se estima mediante la ecuación:

$$CU = \frac{-\hat{\beta}_0}{\hat{\beta}_1} \qquad \textbf{(2.7-2)}$$

La pendiente de la curva β_1 corresponde al poder discriminación de la prueba, un valor de cero no es deseado porque la oportunidad de pasar no dependería de la calificación, es decir, que la calificación no es buen reflejo del criterio de clasificación del examen.

Asumiendo calificaciones válidas, la alta discriminación es preferible porque corresponde a la probabilidad de aprobar cercanos cero o a uno. Valores extremos implica mayor certeza en la clasificación.

El intercepto β_0 puede ser interpretado como el grado de severidad o complejidad de la prueba.

Los parámetros β_0 y β_1 son estimados por máxima verosimilitud, la distribución de cada observación binaria corresponde a una distribución bernoulli y la verosimilitud es:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \qquad \textbf{(2.7-3)}$$

Y la log-verosimilitud es:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \log p_i + (1 - y_i)(1 - p_i) \qquad \textbf{(2.7-4)}$$

Los estimadores de máxima verosimilitud se pueden encontrar explícitamente derivando la log-verosimilitud con respecto a β_0 y β_1 igualando cada derivada a cero y resolviendo las dos ecuaciones resultantes.

Una alternativa para la estimación de los parámetros es β_0 y β_1 con base en la tabla de contingencia 2x2 es:

$$\hat{\beta}_0 = \log(\frac{n_{01}}{n_{00}}) \text{ y } \hat{\beta}_1 = \log(\frac{n_{00}n_{11}}{n_{01}n_{10}}) \qquad \textbf{(2.7-5)}$$

CAPÍTULO 3

METODOLOGÍA

El que trabaja con diligencia, pero sin método, arroja con una mano lo que gana con la otra.

C.C. Colton (1780 – 1832)

3.1 *Método*

Con el propósito de dar respuesta a las preguntas formuladas en el planteamiento de problema y cumplimiento a los objetivos de la investigación mediante un caso práctico, se analizó un examen de Geografía de México.

El objetivo de la evaluación es identificar las personas que requieren capacitación con respecto al conocimiento y compresión de términos, principios y procedimientos utilizados en el proceso de generación de información de geografía, representando ésta una competencia básica en el personal de la Dirección General de Geografía y Medio Ambiente del INEGI.

El examen fue diseñado con 66 reactivos de opción múltiple en dos secciones que evalúan ambas un único constructo unidimensional con el mismo grado de dificultad, es decir, son equivalentes. La primera sección consta de 33 reactivos aplicados a todos los sustentantes, mientras que la segunda sección, los 33 reactivos restantes fueron seleccionados aleatoriamente de un conjunto de 85 reactivos.

La calificación de la prueba total fue analizada por los jefes inmediatos y éstos manifestaron la validez empírica que corresponde los resultados a la valoración del desempeño laboral, a partir de criterios previamente establecidos sobre la ejecución del proceso de trabajo.

No obstante a lo anterior, se requiere calibrar la primera sección de la prueba para ser la única a aplicar a los 1018 empleados restantes, reduciendo el tiempo y costo implícitos, sin perjuicio de la validez y confiabilidad de la evaluación.

Para nuestro caso de estudio, se consideró la primera sección como el examen a analizar y la calificación de ambas sesiones, estandarizada en escala de 0 a 100, como la calificación real (criterio externo), es decir el parámetro a estimar con la primera sección.

Los datos disponibles para cada sustentante fueron: el criterio externo, el puntaje (número de respuestas correctas; tiempo en minutos en que se requirió para contestar el examen y para cada pregunta, la respuesta esperada y la respuesta seleccionada por el sustentante.

La muestra normativa de 200 sustentantes para la calibración del examen fue seleccionada por muestreo aleatorio, dado que no se tuvo evidencia de variabilidad por ningún rasgo de la población. Se excluyeron los registros correspondientes a preguntas que todos las contestaron correctamente o

incorrectamente; así también, los registros de las personas que presentaron 0 ó 100 puntos de calificación.

Los procedimientos que se utilizaron en ensayos experimentales fueron:

1. Organización, tratamiento y exploración de datos. En primera instancia, se organizaron los datos para facilitar su manejo, se les dio tratamiento para la estandarización de las variables, se exploraron de manera general las variables y sea eliminaron las observaciones atípica.

2. Análisis de reactivos por TCT. Con el propósito de identificar los reactivos deficientes con TCT, a cada reactivo se le calculó el grado de dificultad, la capacidad de discriminación y la norma de discriminación. Con base en estos indicadores se identificaron los reactivos aparentemente con deficiencias.

3. Análisis de reactivos por TRI. Con el propósito de identificar los reactivos deficientes con TRI, a cada reactivo se le estimó por máxima verosimilitud los parámetros de modelo logístico de tres parámetros considerando el supuesto de la habilidad conocida representada por el criterio externo. Además, se graficó la curva características de los ítems para facilitar su interpretación.

4. Análisis de alternativas de los reactivos. Para complementar los resultados de análisis de reactivos, se verificó la validez de las opciones mediante pruebas de hipótesis sobre su comportamiento esperado.

5. Validación y confiabilidad de la prueba. Para valorar la validez se compararon los resultados de la prueba con el criterio externo, valorando la capacidad de predecir el criterio externo mediante la puntación observadas de la prueba; así también, se valoró la consistencia interna; y se identificaron los reactivos que menos contribuyen a la validez y

consistencia. Respecto a la confiabilidad de la prueba se estimó la función de información para valorar en que rango de la habilidad es más eficiente el examen.

6. Tiempo de evaluación. Se identificó la distribución del tiempo utilizado para resolver la prueba por los sustentantes más sobresalientes y se ajustó la longitud de la prueba a ser contestada en un tiempo máximo de una hora, identificando y excluyendo los reactivos que menos impactan al resultado final.

7. Estimación de la habilidad. Con base en los parámetros estimados con TRI, se estimó el grado de habilidad de los sustentantes y se comparó con el criterio externo para valorar la validez.

8. Calificación mínima aprobatoria. Mediante la regresión logística se estimó el umbral para clasificar a los sustentantes en acuerdo con el dominio de la competencia en apto o no, y se valoró la calidad del estimador con la probabilidad del error de clasificación.

9. Validación del modelo, método y técnicas. En este apartado se recopiló la información que sustenta los procedimientos y las técnicas utilizadas en el caso de estudio, así como la comprobación de los supuestos empleados.

3.2 *Equipamiento, software y materiales.*

El *software* utilizado para el diseño, aplicación y evaluación fue el Sistema de Evaluación Automática desarrollado por el propio INEGI, el cual sustenta su operación en una base de datos de SQL Server, donde se obtuvieron los datos y se exportaron a archivos en formato de Excel por su facilidad de manejo.

Tabla 0.1.Diccionario de datos.

Variable	Descripción	Dominio
Sustentante	Identificador del sustentante.	1 al 200
Puntaje	Número de respuestas correcta.	0 al 33
Criterio externo	Calificación real de la habilidad.	0 al 100
Tiempo	Tiempo (minutos) en contestar las 33 preguntas.	0 al 120
Pregunta.	Identificador de la pregunta.	1 al 33
Respuesta correcta	Respuesta esperada de la pregunta.	1 al 4
Respuesta	Respuesta seleccionada por el sustentante.	1 al 4

El *software* utilizado para el procesamiento de datos fue R versión 2.8.1. por ser una herramienta estadística que se adapta perfectamente a las exigencias y rigurosidades de la investigación y por su sencillez de uso; además, esta herramienta es libre.

CAPÍTULO 4

RESULTADOS

Las matemáticas son el alfabeto con el cual Dios ha escrito el Universo.
Galileo Galilei (1564 – 1642)

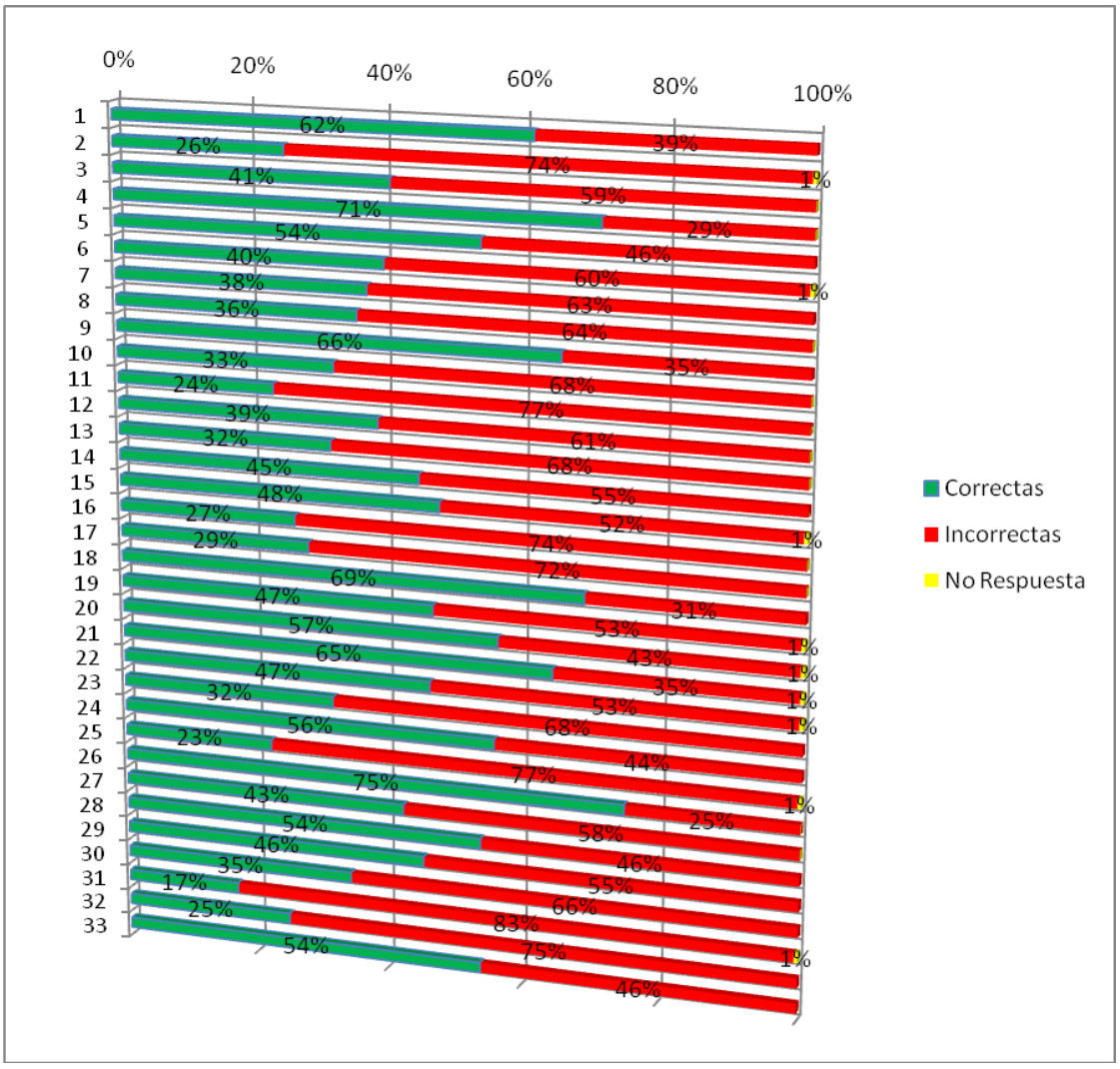
El azar es tal vez seudónimo de Dios, cuando no quiere Él poner su firma.
Anatole France (1874 - 1924)

4.1 *Exploración y tratamiento de los datos.*

Iniciamos la exploración de los datos revisando la frecuencia relativa de la respuesta de los ítems, donde observamos por cada ítem el porcentaje de sustentantes que respondieron correctamente y su complemento, los que respondieron incorrectamente o no respondieron; por premisa se deben excluir del análisis los reactivos que fueron contestados por todos los evaluados correctamente o incorrectamente, situación que no se presentó en este caso como

se muestra en la Figura 0.1. También, se observa que la no respuesta es prácticamente nula, situación que se comentará más adelante.

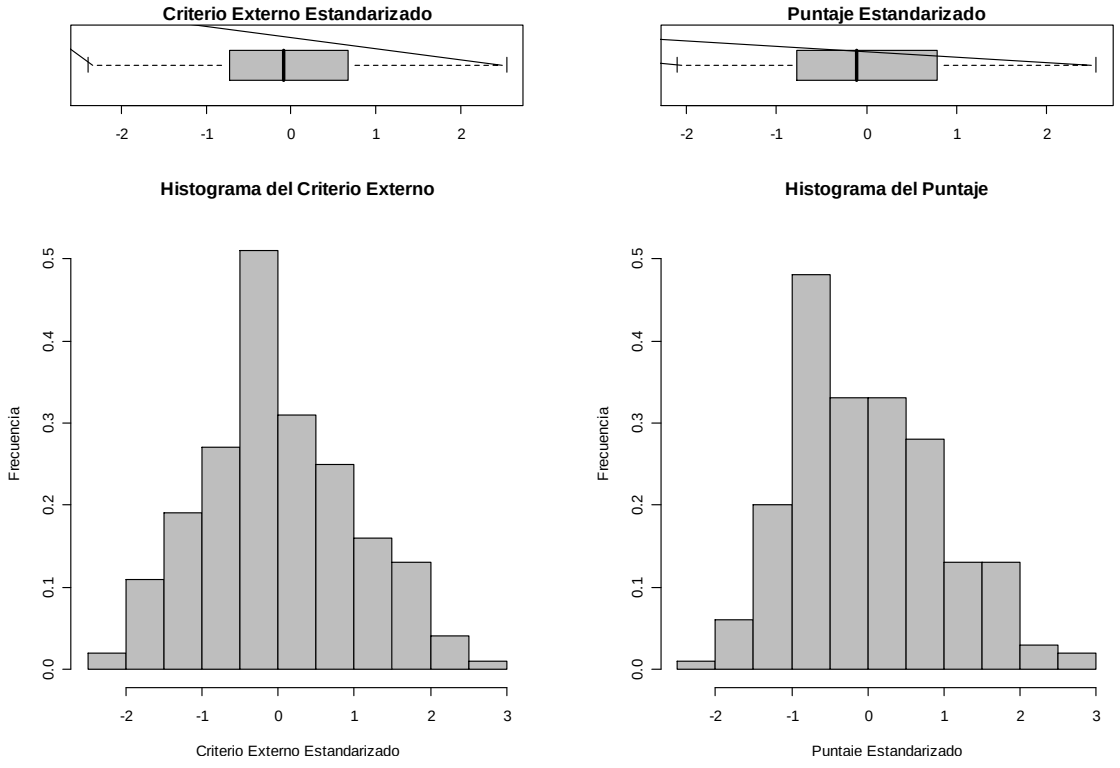
Figura 0.1. Porcentaje de respuesta por pregunta.



El criterio externo está determinado en el rango de 19 a 81 puntos, con media 49, mediana de 40 y una desviación estándar de 12.5, respecto al criterio interno presenta un rango observado de 5 a 26 puntos, con media 14.5, mediana de 14 y

una desviación estándar de 4.5. La Figura 0.2 presenta la distribución de ambos criterios ya estandarizados para facilitar su comparación.

Figura 0.2. Histogramas del criterio externo y puntaje estandarizados.

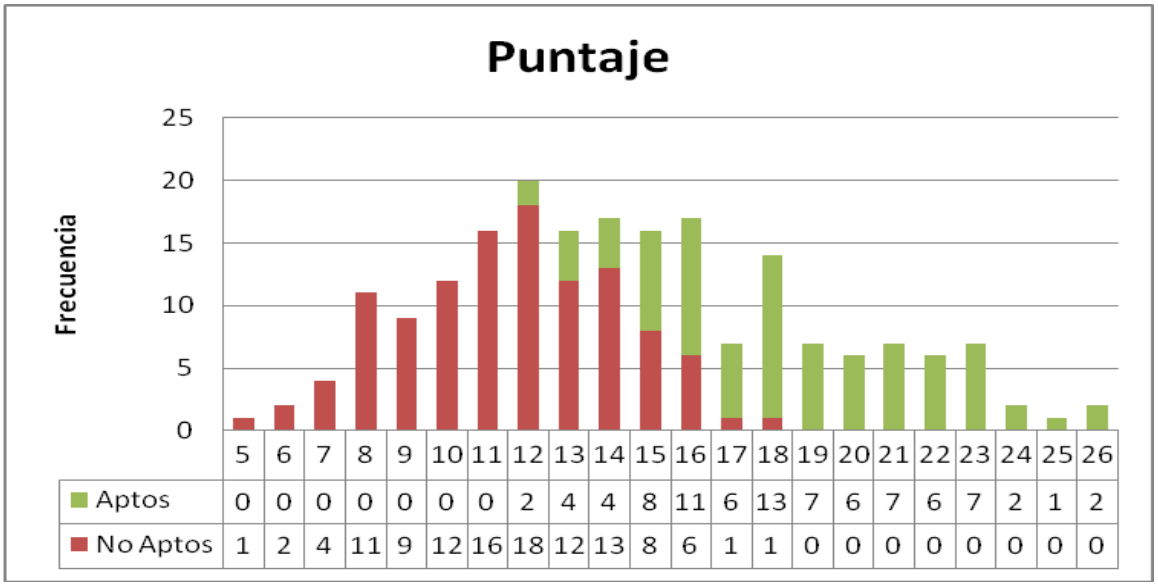


Observamos que el puntaje está sesgado a la derecha, no obstante no presentan diferencia significativa.

Los sustentantes fueron clasificados *a priori* en aptos o no en la competencia con base en el criterio externo, se consideraron a las personas aptas, a aquellas cuyo criterio externo fuera mayor a 50 puntos (0.08 estandarizado) de acuerdo con la experiencia del experto, con el cual 86 (43%) sustentantes se consideraron capaces en comprender el proceso de generación de información geográfica, el reto requiere fortalecer sus conocimientos mediante capacitación.

En la Figura 0.4 se muestra la clasificación de los sustentantes relacionada con la puntuación obtenida del examen, se observa el sesgo de obtener esta clasificación en función del puntaje y un umbral, cualquiera que sea éste, es evidente que se presentaran errores en la clasificación, ya sean falsos positivos o falsos negativos.

Figura 0.3. Distribución del puntaje con respecto a la clasificación de apto o no en la habilidad.

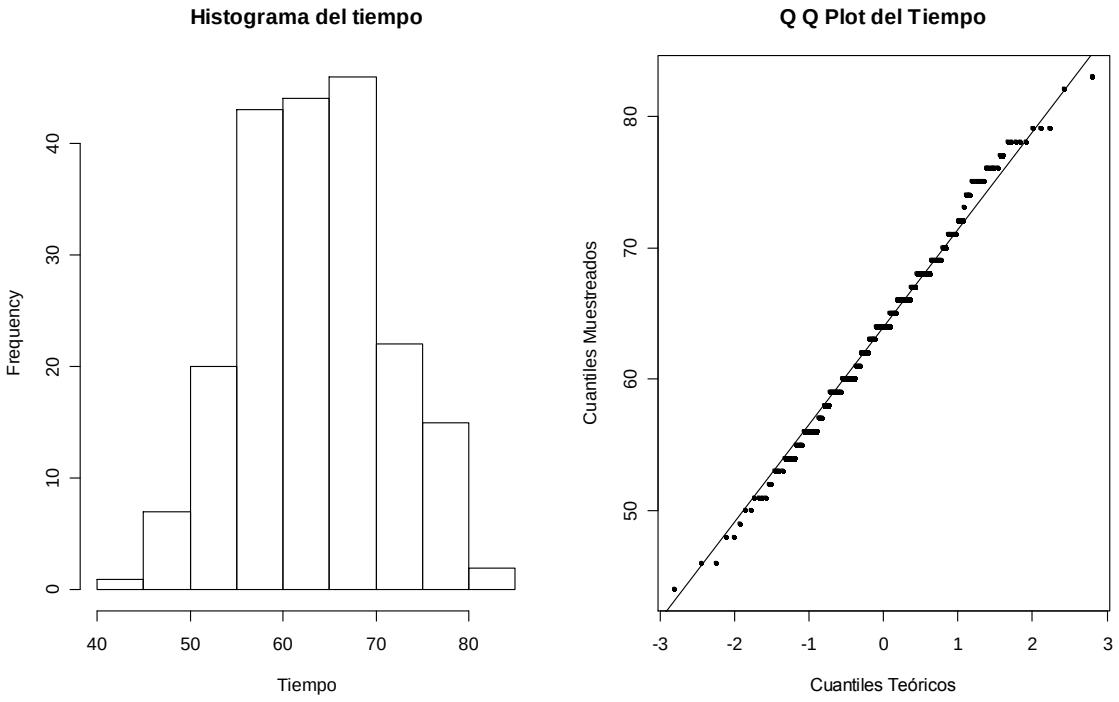


	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Aptos	0	0	0	0	0	0	0	2	4	4	8	11	6	13	7	6	7	6	7	2	1	2
No Aptos	1	2	4	11	9	12	16	18	12	13	8	6	1	1	0	0	0	0	0	0	0	0

El umbral de clasificación está entre 12 y 18 puntos y debe ser estimado de tal forma que maximice la eficiencia de clasificación bajo la premisa de igualdad de gravedad de cometer un error al reprobar injustamente o aprobar sin merecer. Sin embargo, cuando el impacto es distinto por el tipo de error, el umbral corresponde a minimizar una función de costos.

Con respecto al tiempo requerido para contestar la prueba, se observa en la Figura 0.4 la tendencia de la distribución normal con media de 64 minutos y una desviación estándar de 7.8, el rango observado está en 44 a 83 minutos.

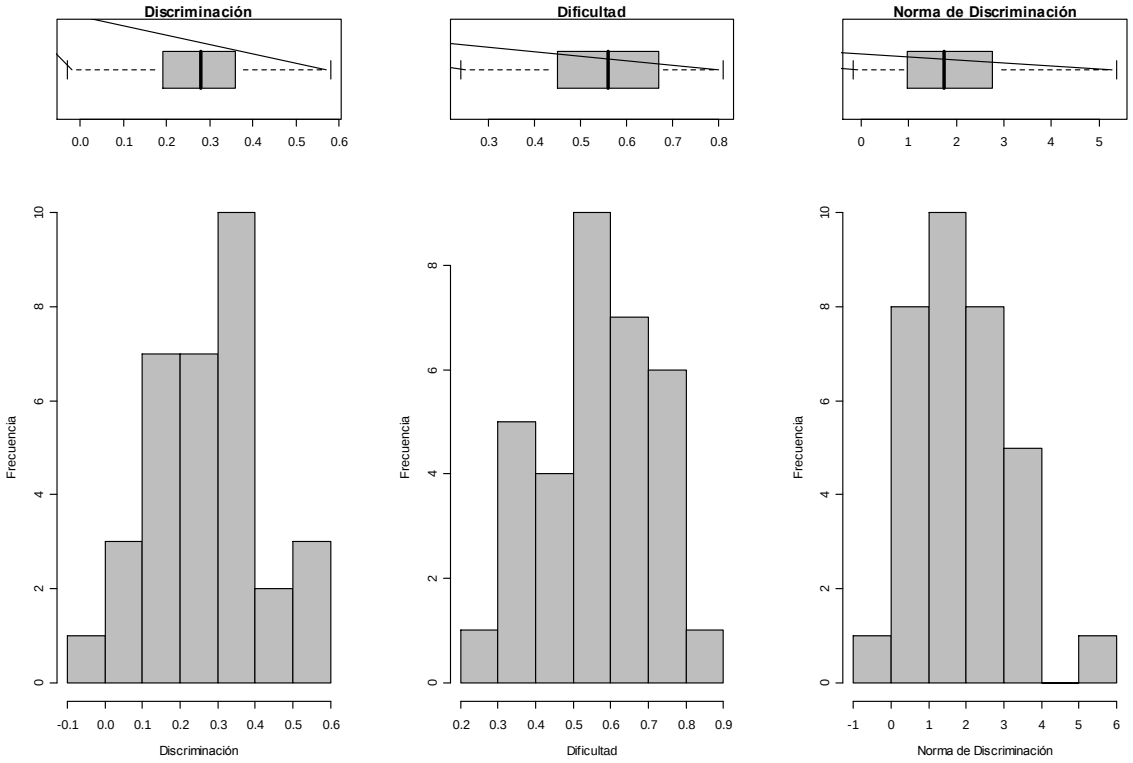
Figura 0.4. Distribución del tiempo requerido para contestar la prueba.



4.2 *Análisis de reactivos por la Teoría Clásica de los Tests.*

En general, los ítems de alta calidad se espera que sean contestados por la mitad de los sustentantes, siempre y cuando quienes los aciertan sean los de mayor dominio de la habilidad, los indicadores que permiten valorar estas características psicométricas con el enfoque clásico son los índices de poder de discriminación y del grado de dificultad, pero afirma Tristán (1996) que ambos no son independientes, un indicador que engloba a ambos es la norma de discriminación. La Figura 0.5 presenta la distribución de estos indicadores.

Figura 0.5. Distribución las propiedades de los reactivos con el enfoque a la Teoría Clásica.



En el enfoque clásico los indicadores son relativos al grado del dominio de la habilidad del grupo a quienes está dirigida la evaluación; por tal razón, es conveniente verificar y calibrar la prueba respecto a una muestra representativa a la población objetivo.

El índice de dificultad de los reactivos está en un rango de 0.24 a 0.81, con una media de 0.56 y una mediana de 0.56. Aun cuando lo ideal son los ítems con grado de dificultad alrededor de 0.50, el rango esperado es de 0.20 a 0.80; por ello, los ítems son adecuados con excepción de los reactivos 25 y 31 que presentan una dificultad muy alta.

En cuanto al índice de discriminación, hay reactivos con muy poca capacidad de discriminación (menor que 0.2), estos reactivos son: 3, 9, 10, 13, 15, 16, 17, 23, 25, 28 y 31; los reactivos con el índice de dificultad deficiente también están incluidos en estos. La capacidad media de discriminación es de 0.27, la mínima es de -0.03 y de 0.58, la máxima.

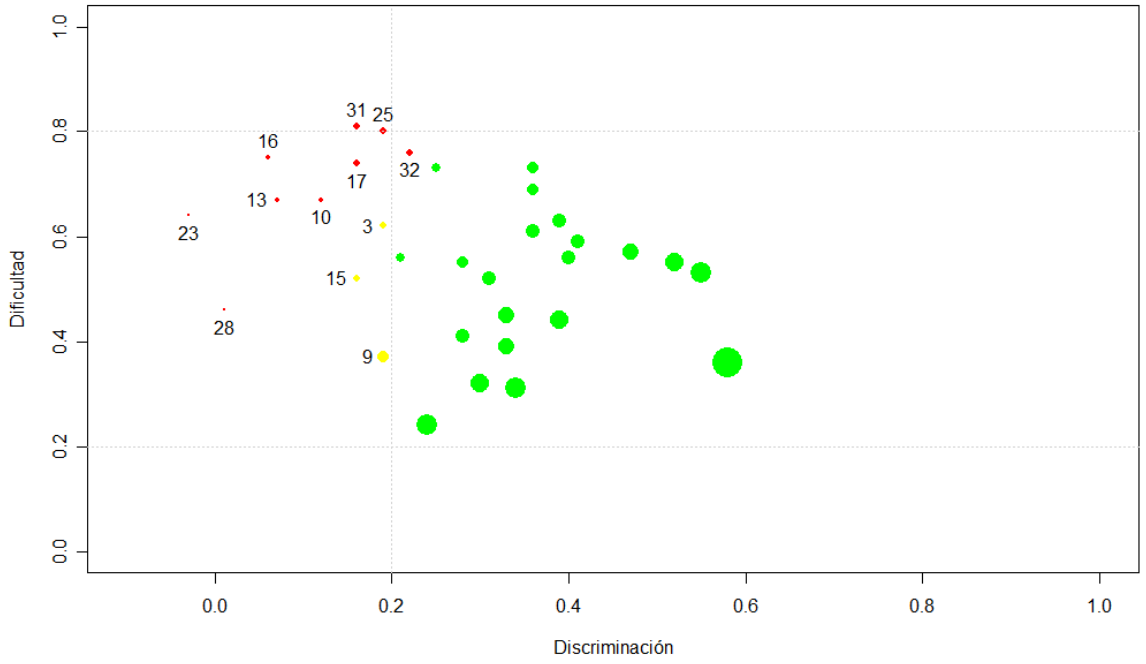
Considerando el grado de dificultad y el poder de discriminación conjuntamente con la norma de discriminación sugerida por Tristán (1996), la prueba tiene reactivos con norma menor que la unidad (10, 13, 16, 17, 23, 25, 28, 31 y 32), observe que hay un nuevo reactivo (32) que en el análisis individual no se había detectado ninguna deficiencia y otros, de acuerdo con la norma no son tan críticos (3, 9 y 15).

Tabla 0.1. Identificación de reactivos deficientes de acuerdo con el enfoque clásico.

Reactivo	Grado de dificultad	Poder de discriminación	Norma de discriminación
3		X	
9		X	
10		X	X
13		X	X
15		X	
16		X	X
17		X	X
23		X	X
25	X	X	X
28		X	X
31	X	X	X
32			X

En la Figura 0.6 se identifican los reactivos deficientes de acuerdo con el grado de dificultad y al poder de discriminación. Cada reactivo está representado por un punto, el tamaño corresponde la norma de discriminación, los puntos verdes son aquellos reactivos que no presentan deficiencias aparentes, los amarillos son los que implican deficiencia por el poder de discriminación o grado de dificultad y los rojos son los que definen una norma de discriminación menor que 1, es decir, los más deficientes de todos.

Figura 0.6. Identificación gráfica de reactivos deficientes con enfoque clásico



En la Tabla 0.2 se presenta los índices y recomendaciones para cada reactivo de acuerdo a la Teoría Clásica de los Tests.

Tabla 0.2. Índices y observaciones de los reactivos de acuerdo con la Teoría Clásica de los Tests.

Reactivo	Grado de Dificultad		Poder de Discriminación		Norma de Discriminación	
	Índice	Observaciones	Índice	Observaciones	Índice	Observaciones
1	0.39	Recomendado	0.33	Bueno	2.82	Aceptar
2	0.73	Recomendado	0.36	Bueno	1.64	Aceptar
3	0.62	Recomendado	0.19	Pobre	1.02	Aceptar
4	0.31	Recomendado	0.34	Bueno	3.66	Aceptar
5	0.45	Recomendado	0.33	Bueno	2.44	Aceptar
6	0.56	Recomendado	0.40	Excelente	2.38	Aceptar
7	0.57	Recomendado	0.47	Excelente	2.75	Aceptar
8	0.61	Recomendado	0.36	Bueno	1.97	Aceptar
9	0.37	Recomendado	0.19	Pobre	1.71	Aceptar
10	0.67	Recomendado	0.12	Pobre	0.60	Rechazar
11	0.73	Recomendado	0.25	Regular	1.14	Aceptar
12	0.63	Recomendado	0.39	Bueno	2.06	Aceptar
13	0.67	Recomendado	0.07	Pobre	0.35	Rechazar
14	0.56	Recomendado	0.21	Regular	1.25	Aceptar
15	0.52	Recomendado	0.16	Pobre	1.03	Aceptar
16	0.75	Recomendado	0.06	Pobre	0.27	Rechazar
17	0.74	Recomendado	0.16	Pobre	0.72	Rechazar

Reactivo	Grado de Dificultad		Poder de Discriminación		Norma de Discriminación	
	Índice	Observaciones	Índice	Observaciones	Índice	Observaciones
18	0.32	Recomendado	0.30	Regular	3.12	Aceptar
19	0.55	Recomendado	0.28	Regular	1.7	Aceptar
20	0.44	Recomendado	0.39	Bueno	2.95	Aceptar
21	0.36	Recomendado	0.58	Excelente	5.37	Aceptar
22	0.55	Recomendado	0.52	Excelente	3.15	Aceptar
23	0.64	Recomendado	-0.03	Pésima	-0.16	Rechazar
24	0.41	Recomendado	0.28	Regular	2.28	Aceptar
25	0.80	No Recomendado	0.19	Pobre	0.79	Rechazar
26	0.24	Recomendado	0.24	Regular	3.33	Aceptar
27	0.59	Recomendado	0.41	Excelente	2.32	Aceptar
28	0.46	Recomendado	0.01	Pobre	0.07	Rechazar
29	0.53	Recomendado	0.55	Excelente	3.46	Aceptar
30	0.69	Recomendado	0.36	Bueno	1.74	Aceptar
31	0.81	No Recomendado	0.16	Pobre	0.66	Rechazar
32	0.76	Recomendado	0.22	Regular	0.96	Rechazar
33	0.52	Recomendado	0.31	Bueno	1.99	Aceptar

Estos estadísticos no presentan una distribución muestral conocida por lo que se ha estimado el valor esperado y la varianza mediante el procedimiento de *bootstrap*, con 500 muestras aleatorias de tamaño 50. Los resultados se muestran en el anexo F, donde se observa sistemáticamente los reactivos débiles son 9, 10, 13, 14, 15, 16, 17, 23, 25, 28, 31 y 32.

4.3 *Análisis de reactivos por la Teoría de la Respuesta al Ítem.*

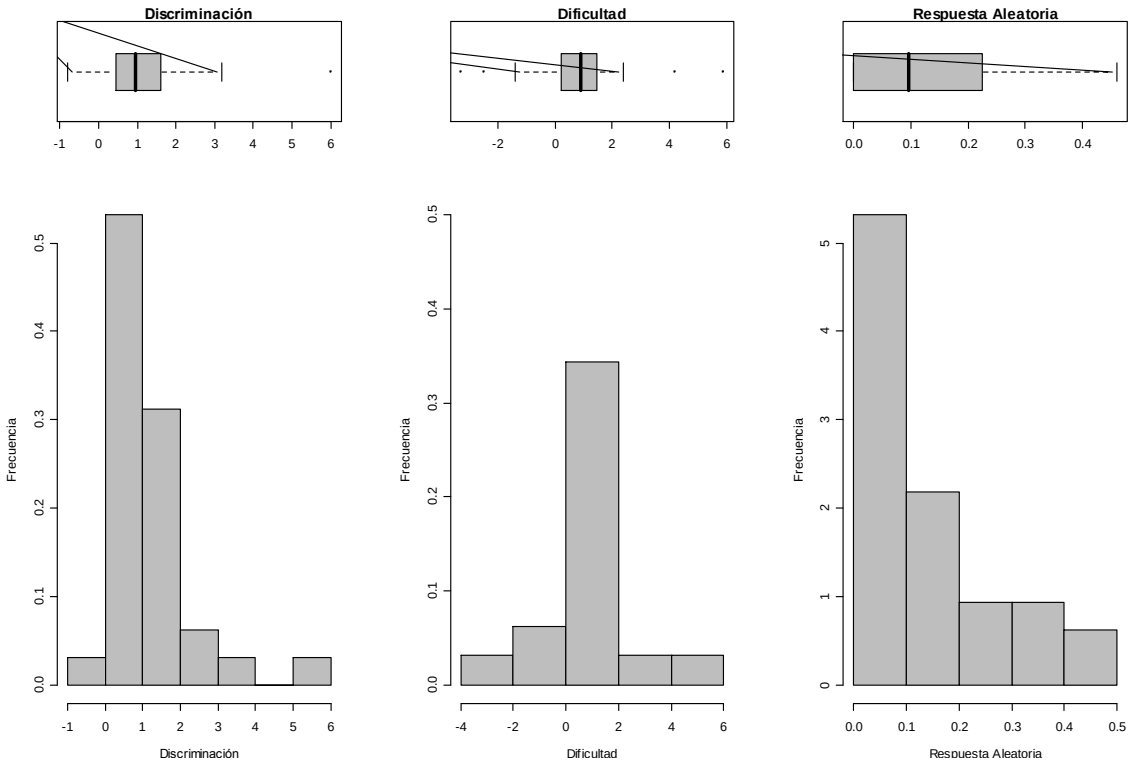
En el contexto de la Teoría de la Respuesta al Ítem, se estimó los parámetros de la curva característica de cada reactivo mediante máxima verosimilitud con el modelo logístico de tres parámetros con base en el patrón de respuestas y al criterio externo.

$$P_{ij} = P[X_{ij} = 1 \mid \theta_j, \beta_i, \alpha_i, c_i] = c_i + (1 - c_i)\frac{\exp(\alpha_i(\theta_j - \beta_i))}{1 + \exp(\alpha_i(\theta_j - \beta_i))}$$

La decisión de considerar el modelo de tres parámetros se sustenta en que éste es el modelo que más se ajusta a los datos de la muestra normativa, dado que

existe reactivos plausibles de respuesta aleatoria y con la discriminación diferente de la unidad, como se observa en la Figura 0.7 la distribución de los parámetros. Aunque, se podría enfocar la calibración de ítems para utilizar un modelo más parsimonioso, eliminando o modificando los reactivos para ajustarlos a un modelo con menos parámetros.

Figura 0.7. Distribución de las propiedades de los reactivos de acuerdo con la Teoría de la Respuesta al Ítem.



Nota: Se excluyó el ítem 16 por ser atípico.

El índice de discriminación presenta una media de 1.12 con una desviación estándar de 1.17, es de esperarse que este indicador sea mayor que cero para que curva sea una función monótonamente creciente de la habilidad, situación que no presenta los reactivos 16 y 23, pues el primero tiene un valor de cero que

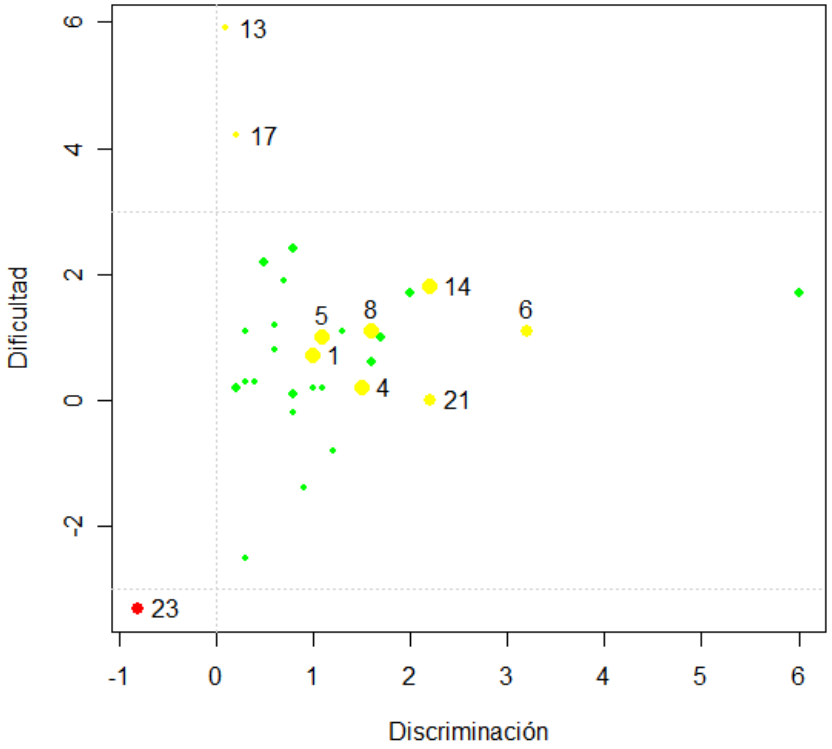
implica que la probabilidad de ser contestado correctamente es independiente a la habilidad y el segundo con un valor negativo indica que es más probable que sea contestado por quienes tienen menos habilidad, situación totalmente anormal.

En lo que se refiere al grado de dificultad, este indicador está expresado en la misma escala de medida que la habilidad y por tanto, sus valores también son de $-\infty$ a $+\infty$, aunque típicamente es esperado en un rango de -3 a 3. Sin embargo, hay preguntas que están fuera de este rango (13, 16, 17 y 23), es decir, demasiado fáciles o difíciles. El reactivo 16 al ser independiente a la habilidad no presenta un punto de inflexión para estimar la dificultad. La media de este indicador es de -0.8242 con 9.52 de desviación estándar.

La probabilidad de acertar un ítem al azar por quienes no conocen la respuesta (cuando θ atiende a $-\infty$), es en promedio de 0.1236 con una desviación de 0.15, presentándose un máximo de 0.46. Los reactivos con respuesta aleatoria superior al máximo esperado de 0.25 (puesto que cada pregunta tiene 4 opciones de respuesta) son 1, 4, 5, 6, 14, 21, 23 y 24, los cuales requieren indispensablemente analizar detalladamente el comportamiento de los distractores.

La Figura 0.8 presenta las propiedades psicométricas de los reactivos. Cada punto representa un reactivo donde la dimensión corresponde al parámetro de respuesta aleatoria, los verdes corresponden a los reactivos aparentemente eficientes, los amarillos son los que fueron identificados deficientes en algún parámetro y los rojos, en ambos parámetros. Los reactivos recomendados a ser revisados por el experto son: 1, 4, 5, 6, 13, 14, 16, 17, 21, 23 y 24, para corregir o sustituir.

Figura 0.8. Identificación de reactivos deficientes de acuerdo con la Teoría de la Respuesta al Ítem.



Nota: Se excluyó del gráfico el ítem 16 por ser atípico.

Los estimadores de los parámetros de los reactivos se muestran en la Tabla 0.3, marcado con rojo los que presentan deficiencias.

Tabla 0.3. Índices de los reactivos de acuerdo con la Teoría de Repuesta de Ítems.

Reactivo	Discriminación	Dificultad	Respuesta aleatoria	Reactivo	Discriminación	Dificultad	Respuesta aleatoria
1	1.0	0.7	0.4	17	0.2	4.2	0.0
2	1.3	1.1	0.0	18	1.2	-0.8	0.0
3	0.3	1.1	0.0	19	0.4	0.3	0.0
4	1.5	0.2	0.5	20	0.8	0.1	0.2
5	1.1	1.0	0.4	21	2.2	0.0	0.3
6	3.2	1.1	0.3	22	1.1	0.2	0.0
7	1.7	1.0	0.2	23	-0.8	-3.3	0.3
8	1.6	1.1	0.2	24	1.6	1.1	0.4
9	0.3	-2.5	0.0	25	0.7	1.9	0.0
10	0.5	2.2	0.1	26	0.9	-1.4	0.0
11	2.0	1.7	0.1	27	1.6	0.6	0.1
12	0.6	0.8	0.0	28	0.2	0.2	0.1
13	0.1	5.9	0.0	29	1.0	0.2	0.0

14	2.2	1.8	0.4
15	0.3	0.3	0.0
16	0.0	-53.1	0.0

30	0.6	1.2	0.0
31	6.0	1.7	0.1
32	0.8	2.4	0.1
33	0.8	-0.2	0.0

Nota: El valor esperado y la varianza se muestran en el Anexo F.

Las curvas características de los 33 reactivos como lo sugiere Birnbaum se presentan en el Anexo H, para ejemplificar la interpretación cualitativa de la curva característica utilizamos el reactivo No. 2, la curva azul presenta la probabilidad de responder el reactivo condicionado a la habilidad y la curva verde corresponde a la función de información que indica máxima eficiencia de discriminación.

La CCI de ítem corresponde a la ecuación:

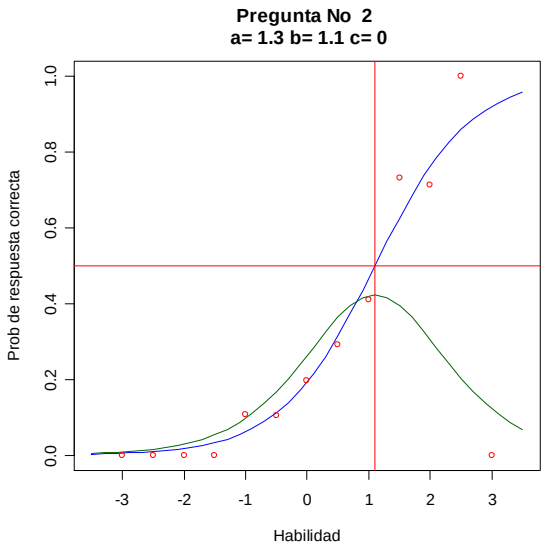
$$P_{2j} = P[X_{2j} = 1 \mid \theta_j, \beta_2, \alpha_2, c_2] = \frac{\exp(1.3(\theta_j - 1.1))}{1 + \exp(1.3(\theta_j - 1.1))}$$

El momio es la razón de probabilidad de acertar P_{2j} entre la probabilidad de no acertar Q_{2j}, entonces, el logaritmo natural de los momios de acertar es modelado por la diferencia entre la habilidad y la dificultad del ítem, donde la diferencia es ponderada por índice de discriminación.

$$\ln\left[\frac{P_{2j}}{Q_{2,j}}\right] = \alpha_2(\theta_j - \beta_2) = 1.3(\theta_j - 1.1)$$

De lo anterior, se deduce que si la habilidad es igual a la dificultad, entonces el logaritmo de momios es cero y el exponente de cero es uno, lo que significa que es tan probable el acertar como el fracasar. Sin embargo, en la medida de que la habilidad sea mayor que la dificultad la probabilidad de acertar será mayor que la de no acertar y entre mayor sea el poder discriminatorio, mayor serán las diferencias.

Figura 0.9. Curva característica del modelo logístico con tres parámetros.



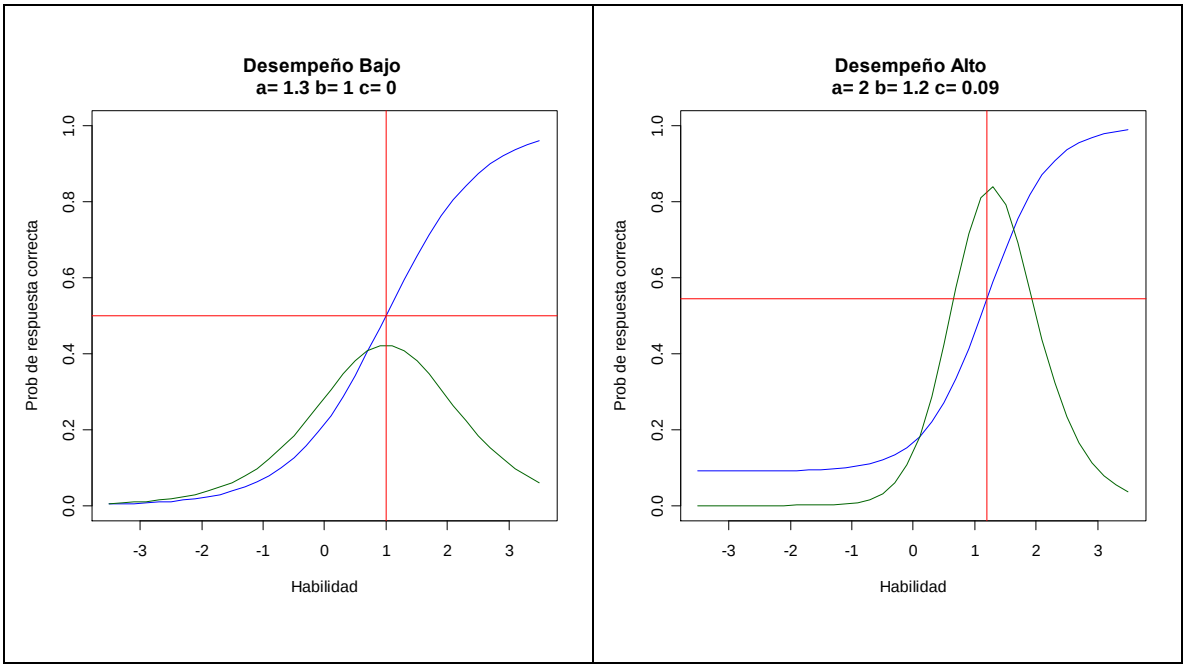
La curva característica describe como los cambios en el nivel de la habilidad cambian el nivel de probabilidad de respuesta correcta; para los sustentantes de rendimiento bajo (θ <-1), la discriminación es deficiente, la probabilidad de acierto para θ = -3 es prácticamente cero y para θ = 0 no supera el 0.2, en cambio en la habilidad media-alta y alta, incrementos pequeños en la habilidad implica grandes incrementos de probabilidad, observa la probabilidad de acertar para $\theta = \beta$ (1.3) es 0.5 y los sustentantes con θ = 2 o superior tienen una probabilidad mayor de 0.8. La función de información confirma estas impresiones, el ítem es más preciso (tiene un menor error de medida) para los valores θ positivos y más concretamente para aquellos comprendidos entre 0 y 2, donde se concentra mayor área bajo la curva.

La interpretación de las demás curvas características de los ítems es similar a la anterior, observando su forma, los reactivos identificados con deficiencias son apreciables a simple vista.

La invarianza del parámetro dificultad se comprobó al dividir la muestra en dos grupos con distinto rendimiento (40.7 y 59 puntos sin estandarizar) presentando sin cambio significativo el estadístico de dificultad en general para todos los reactivos, es decir, independencia al nivel de conocimientos de la muestras. Particularmente, el reactivo Núm. 2 se muestra gráficamente dicha propiedad en la Figura 0.10, se observa en el grupo de desempeño alto que presenta una respuesta aleatoria de 0.09 que en el grupo de desempeño bajo no ocurrió, lo cual ocasiona que el punto de inflexión de la curva ya no sea en 0.5 sino en 0.55, por lo tanto, el valor de dificultad incrementa de 1 a 1.2.

También, se observa en la función de información (línea verde) más prolongada en el grupo de desempeño alto que en el bajo, lo que confirma que este ítem es más eficiente en este grupo, esto es evidente cuantitativamente al ser mayor el índice de discriminación.

Figura 0.10 Invarianza de parámetro dificultad.



Cabe mencionar que la habilidad fue estandarizada conforme a la media y desviación típica del criterio externo de toda la muestra, si se estandariza de acuerdo con la habilidad de cada subgrupo, las dificultades no son comparables por estar en escalas distintas.

4.4 *Validez de las opciones de respuestas.*

La eficiencia de los ítems puede ser potencializada al analizar y corregir el comportamiento de la opción de la repuesta correcta y de los distractores:

- Los distractores deben ser plausibles de ser seleccionados. Los distractores que fácilmente son descartadas reducen la calidad del ítem porque aumenta la posibilidad aleatoria de ser contestados correctamente.

- Los distractores deben ser seleccionados con mayor probabilidad por los sustentantes con menor capacidad y también, se espera que la respuesta correcta sea más seleccionada por el grupo alto que el bajo. Esta característica mide la capacidad de discriminación del distractor.

- La respuesta correcta debe ser seleccionada con mayor probabilidad por el grupo de sustentantes con mayor capacidad. Si la opción correcta no es la más elegida puede ocurrir que el enunciado o las opciones no sean claras o que existan opciones incorrectas parcialmente correctas.

Cuando el tiempo de aplicación no es una limitante, la no respuesta de las preguntas por el grupo bajo es una característica deseada, es evidencia que el sustentante no responde al azar y en cambio, la no respuesta en el grupo alto sugiere el análisis de la pregunta por sospecharse de redacción confusa.

En lo particular de este caso de estudio, al ser contestadas 99.8% de las preguntas, sugiere que el sustentante contestó aleatoriamente las preguntas aun

cuando no conocía la respuesta, representando una oportunidad que no desaprovechó, la situación hubiera sido distinta si se le hubiera notificado y aplicado penalización por respuestas incorrectas.

Verificar cualitativamente la validez de los distractores de acuerdo con las frecuencias observadas como se recomienda en el apartado 2.5.9 sobre el análisis de distractores no es simple ni eficiente, es preferible aplicar pruebas estadísticas para comparar la probabilidad de ser seleccionadas y verificar el cumplimiento del comportamiento esperado.

La Tabla 0.4 que a continuación se presenta contiene las preguntas con distractores pocos probables de ser seleccionados, con mayor probabilidad de ser seleccionada que la respuesta correcta y con mayor probabilidad de ser seleccionada en el grupo alto que el bajo.

Tabla 0.4 Preguntas con distractores deficientes.

Pregunta	Distractores		
	Sin probabilidad de ser seleccionado	Con mayor probabilidad de ser seleccionada que la respuesta correcta	Con mayor probabilidad de ser seleccionada en el grupo alto que el bajo
2		C	
5	C		
8	B		
10		D	
13		D	
15			C
16		B	B
17		C	
20	C		
23	B	D	
24		C	
25		C	
26	C y D		
31		D	
32		C	D

La validez de las alternativas de respuesta para todos los ítems se presenta en detalle en el Anexo I.

4.5 *Validación y confiabilidad de la prueba.*

Toda prueba tiene un componente de varianza de error adjudicado al examen y no al sustentante, por esta razón se debe verificar si la prueba es pertinente al grupo de sustentantes, los indicadores clásicos para determinar la calidad son los índices de validez, consistencia y confiabilidad, tal como se describió en el marco teórico.

4.5.1 Índice de validez.

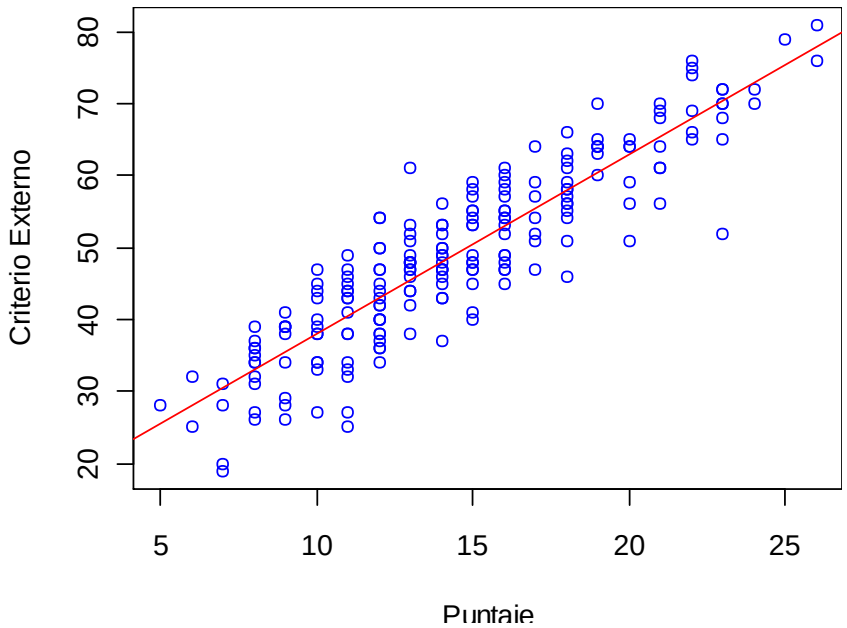
Con el supuesto que el criterio externo es el parámetro (calificación real) que deseamos estimar con la calificación de la prueba, el indicador de validez empírica corresponde al grado de correlación entre el criterio externo y los puntajes, la cual es de 0.90 con un intervalo de confianza al 95% de 0.87 a 0.92. La relación entre el puntaje y el criterio externo se muestra en la Figura 0.11 y dado que se presenta homogeneidad de varianza, es posible determinar la ecuación de equivalencia mediante regresión lineal:

$$\hat{y} = 12.93 + 2.5x \qquad \textbf{(4.5-1)}$$

Donde: $\hat{y}$ es el estimador del criterio externo que se pretende predecir a partir de los puntajes de la prueba (x).

El coeficiente de determinación de la regresión lineal o el cuadrado de la correlación entre el criterio externo y el puntaje indica que el puntaje de la prueba explica 81% de la variabilidad del criterio externo.

Figura 0.11. Correlación del criterio externo y el puntaje.



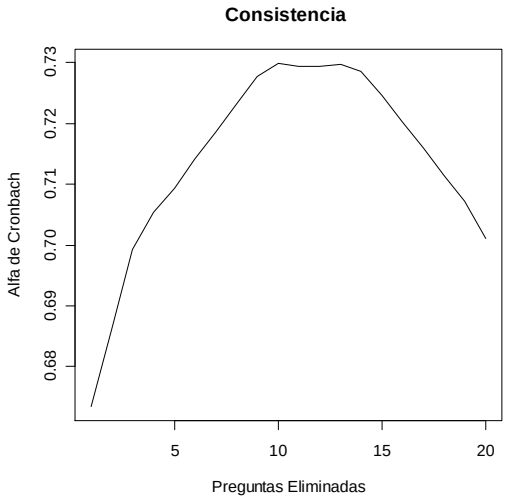
4.5.2 Índice de consistencia.

La consistencia interna es la forma más estable de medir la confiabilidad del instrumento e inclusive permite valorar la validez teórica, siempre y cuando el universo de conocimientos esté bien definido y el muestreo del conocimiento sea representativo.

Con soporte en el supuesto de unidimensionalidad de la prueba, puesto que el contenido es homogéneo, se aplicó el coeficiente Alpha de Cronbach para valorar el grado de presión del examen, obteniendo un valor de 0.66, un poco superior al mínimo aceptable (0.65) que recomienda Cronbach, valores por debajo de ése indica que las estimaciones pueden ser imprecisas por causa de ítems deficientes o porque son pocos y por supuesto, está la posibilidad de que el examen sea multidimensional. También, es posible elevar este indicador de consistencia al excluir reactivos que contribuyen menos al coeficiente de Alfa, observa la gráfica

de tendencia del indicador al excluir los reactivos, se obtiene el máximo en la décima exclusión y disminuye después del cuarto décimo.

0.12 Gráfica de tendencia de la consistencia al excluir ítems.



Los ítems que menos aportan a la consistencia interna están ordenados por relevancia en la Tabla 0.5, éstos coinciden con los reactivos de pobre potencia de discriminación del modelo TRI.

Tabla 0.5 Consistencia obtenida al excluir reactivos.

Orden	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Pregunta	23	28	16	13	15	10	17	3	9	14
Consistencia	0.67	0.69	0.70	0.71	0.71	0.71	0.72	0.72	0.73	0.73

Al elevar la consistencia interna se corre el riesgo de disminuir la validez de la prueba porque ésta puede estar midiendo varios rasgos de la habilidad, y al maximizar la consistencia interna se unifica a medir sólo uno. Por lo tanto, el procedimiento deberá aplicarse por separado a cada rasgo a evaluar y siempre tener presente la validez, pues de nada sirve disponer pruebas consistentes que no midan lo que pretenden evaluar.

La contribución de los reactivos a la validez y consistencia de la prueba fue valorada mediante la correlación del punto biserial de las respuestas con respeto al criterio externo e interno, los resultados se presentan gráficamente en la Figura 0.13 y cuantitativamente en Tabla 0.6. Los reactivos con baja validez y consistencia (menor que 0.2) son 3, 9, 10, 13, 15, 16, 17, 23 y 28, es importante observar que el reactivo 23 tiene correlación negativa en cuanto validez, es decir, la probabilidad de ser contestado correctamente es mayor por los sustentantes con menor habilidad que los examinados con mayor habilidad y el reactivo 16 simplemente no contribuye ni a la validez y a la consistencia. Esto también es apreciado en la curva característica de estos ítems.

Figura 0.13. Contribución de los reactivos a validez y confiabilidad.

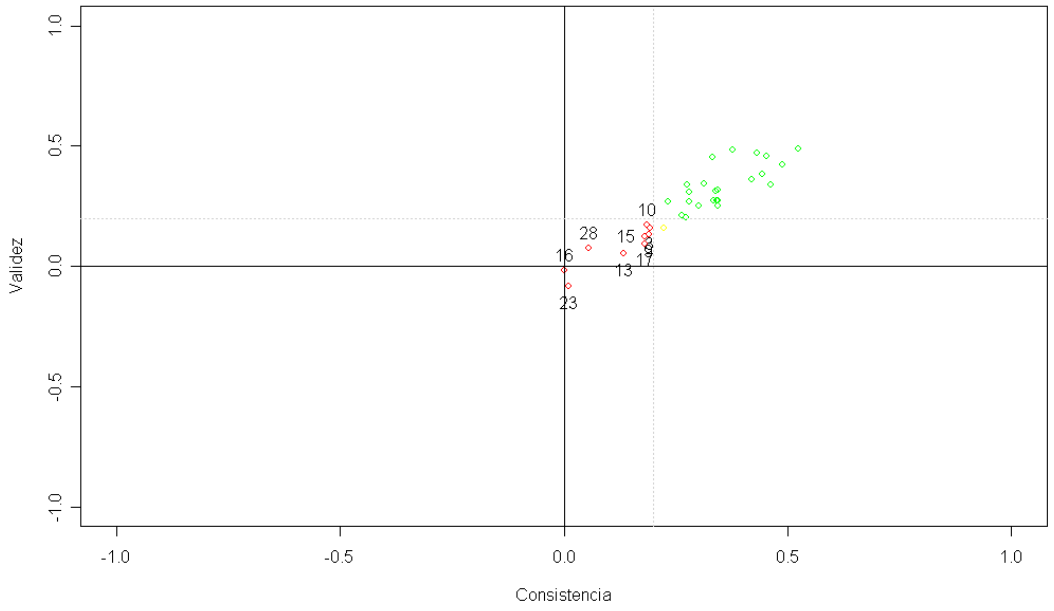


Tabla 0.6. Validación y consistencia de los reactivos.

Reactivo	Correlación de criterios	
	Externo (Validación)	Interno (Consistencia)
1	0.2485	0.3012
2	0.4718	0.4327
3	0.1598	0.1919
4	0.3185	0.3445
5	0.2734	0.3347
6	0.3400	0.4617
7	0.3807	0.4430
8	0.3601	0.4201
9	0.1318	0.1898
10	0.1735	0.1856
11	0.2744	0.3453
12	0.2745	0.3410
13	0.0510	0.1329
14	0.1582	0.2234
15	0.1226	0.1805
16	-0.0166	0.0007

Reactivo	Correlación de criterios	
	Externo (Validación)	Interno (Consistencia)
17	0.0941	0.1800
18	0.4518	0.3330
19	0.2110	0.2650
20	0.3059	0.2797
21	0.4892	0.5241
22	0.4570	0.4528
23	-0.0843	0.0118
24	0.2670	0.2804
25	0.2683	0.2339
26	0.3413	0.3127
27	0.4844	0.3781
28	0.0763	0.0547
29	0.4226	0.4890
30	0.2521	0.3437
31	0.3101	0.3406
32	0.2029	0.2730
33	0.3391	0.2769

Por lo anterior y según sea la gravedad del caso, se deben someter a una revisión detallada por el experto temático para ser corregidos o sustituidos los reactivos: 1, 3, 4, 5, 6, 9, 10, 13, 14, 15, 16, 17, 21, 23, 24, 25, 28, 31 y 32; de acuerdo a al

análisis de validez y consistencia, Índice de confiabilidad, Teoría Clásica de los Tests y Teoría de la Respuesta al Ítem.

4.5.3 Índice de confiabilidad de la prueba.

La confiabilidad de la prueba se obtiene con la acumulación de la función de información de los ítems (Figura 0.14); maximiza su precisión (error de medida más pequeño) en los sustentantes con habilidad θ positivos y, más concretamente, para aquellos comprendidos entre 1 y 2 desviaciones estándar.

Figura 0.14 Confiabilidad de la prueba.

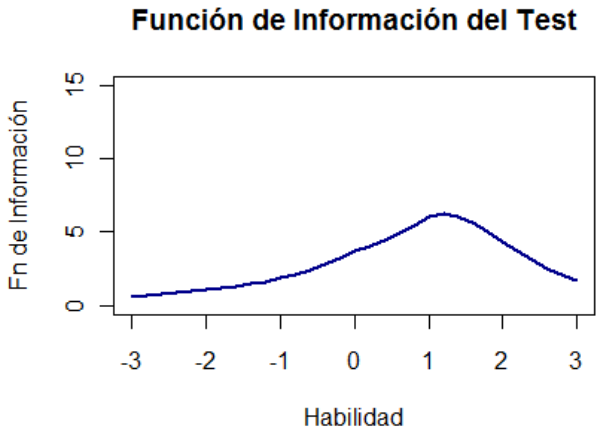
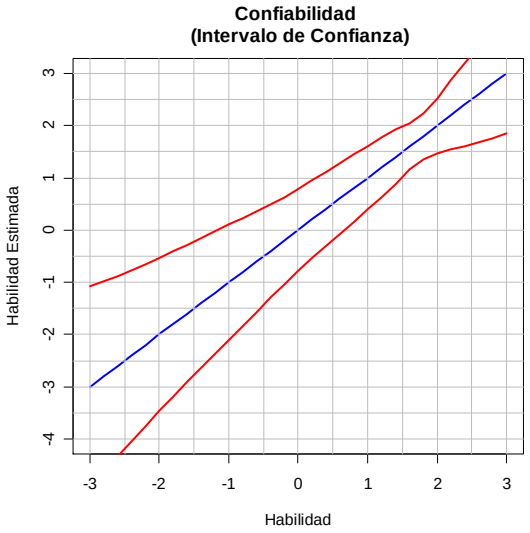


Figura 0.15 Intervalo de confianza.



Por lo anterior, para optimizar la validez y confiabilidad de la prueba, además de sustituir o corregirlos los reactivos deficientes según sea la gravedad de cada caso, se sugiere incorporar más reactivos con grados de dificultad a todo lo largo de la escala de la habilidad; de lo contrario la prueba es deficiente para los sustentantes por debajo de la media.

Conocida la función de información del examen, es posible estimar el error de la estimación de la habilidad, multiplicando a ésta por 1.702 para ajustar al modelo

logístico normal y considerando que los estimadores de máxima verosimilitud son distribuciones normales asintóticamente, los intervalos de confianza de 95% son estimados con $\hat{\theta} \pm 1.96 * SE(\hat{\theta})$, como se muestra en la Figura 0.15.

La precisión no es homogénea, el intervalo de confianza para una habilidad de 1.5 es de 1 a 2, mientras que para una habilidad de -2 el rango está -3 a -0.5, 2.5 veces más grande, lo que es mucho menos confiable.

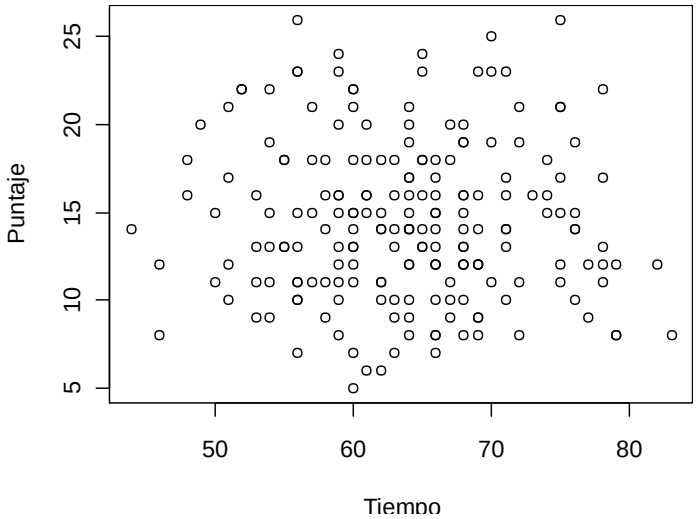
4.6 *Tiempo requerido y ajuste a la extensión de la prueba.*

Antes de aplicar la prueba se les informó a los sustentantes que no disponían de tiempo límite para responder la evaluación y también fue de su conocimiento que durante la sesión de aplicación se realizarían, de manera simultánea, otras evaluaciones distintas que requerían diferentes tiempos, de tal forma que la entrega fuera totalmente independiente, sin la influencia de otros sustentantes.

4.6.1 Tiempo requerido de aplicación.

Como se muestra en la Figura 0.4, el tiempo requerido para responder el examen presenta una distribución normal con media de 63 y una desviación estándar de 7 minutos. La calificación es independiente al tiempo requerido porque no hay evidencia de correlación (-0.0442) y visualmente, se observa en la Figura 0.16 falta de tendencia entre las variables.

Figura 0.16 Gráfica de dispersión del puntaje con respecto al tiempo.



Esta situación fue excepcional. Por lo general se presentan distribuciones con sesgo negativo. También, es de esperarse que los sustentantes que no poseen la capacidad de responder la prueba, la entreguen más pronto o más tardado de lo esperado con respecto a los que sí poseen la habilidad, situación que no sucedió.

Para estimar el tiempo requerido se utilizó el tiempo observado en los sustentantes aptos, excluyendo los valores atípicos. En este caso, el tiempo esperado para 99% de los sustentantes concluyan el examen fue de 81 minutos.

4.6.2 Ajuste a la extensión de la prueba.

Es razonable definir la prueba con el número menor de ítems para reducir el tiempo de aplicación y costos asociados, siempre y cuando se mantenga la validez teórica de contenido. Con soporte al supuesto que el tiempo requerido para cada reactivo es igual, el ajuste de la extensión de la prueba se puede realizar directamente por proporción, por ejemplo: para un tiempo máximo de 60 minutos le corresponderían 25 preguntas ($\frac{60}{81}33$).

Las ocho preguntas a excluir deberán ser las que impacten menos en la validez y confiabilidad, como es el caso de las preguntas indicadas en la Tabla 0.5, las cuales contribuyen menos a la calificación total o están correlacionadas con otras preguntas. Si todos los reactivos son igualmente importantes (validez), se deberá hacer la exclusión de forma proporcional al número de preguntas indicadas en la tabla de especificaciones como lo sugiere el CENEVAL para conservar equidad de constructo.

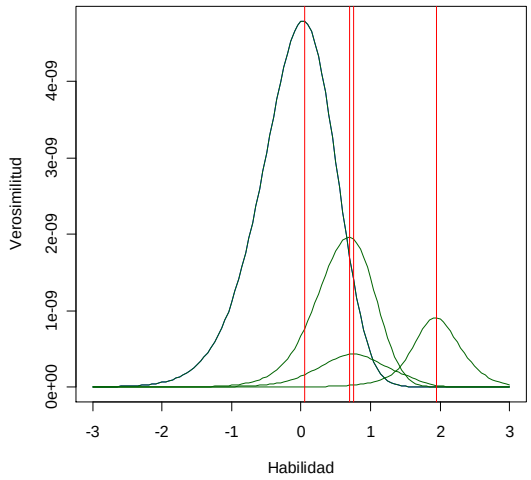
La correlación de la habilidad estimada de las pruebas con 33 y 25 reactivos fue de 0.99, prácticamente, equivalentes.

4.7 *Estimación de la habilidad.*

El puntaje bruto de la prueba es un estadístico suficiente para estimar la habilidad, en este caso particular por presentar una correlación alta con el criterio externo, tal como se mencionó anteriormente. Sin embargo, los estimadores de máxima verosimilitud es nuestro interés.

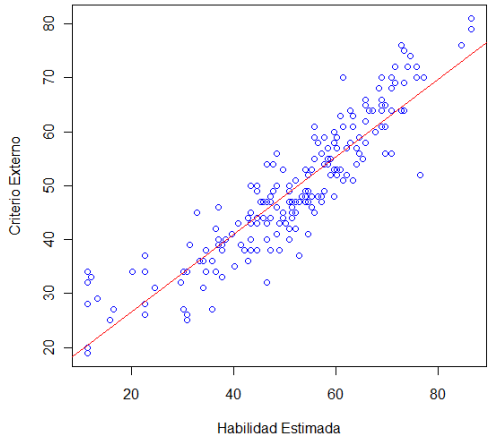
La siguiente gráfica muestra la verosimilitud de la probabilidad de cuatro sustentantes de acuerdo con los patrones de respuestas y los parámetros del modelo de los reactivos previamente calibrados con TRI. La estimación de la habilidad corresponde al punto donde la curva se maximiza (0.05, 0.70, 0.75 y 1.95). También, se observa en las gráficas que las curvas alcanzan diferentes alturas, las curvas llanas corresponden a patrones de respuestas inconsistentes o poco esperadas de acuerdo con el comportamiento general.

Figura 0.17 Gráfica de verosimilitud del patrón de respuestas de cuatro sustentantes.



Comparando la habilidad estimada respecto al criterio externo observamos una alta correlación (0.90), lo que evidencia la capacidad predictiva del estimador. La habilidad estimada explica 82% de la variabilidad del criterio externo, un poco más (1%) que el propio puntaje.

Figura 0.18 Correlación del criterio externo y la habilidad estimada.



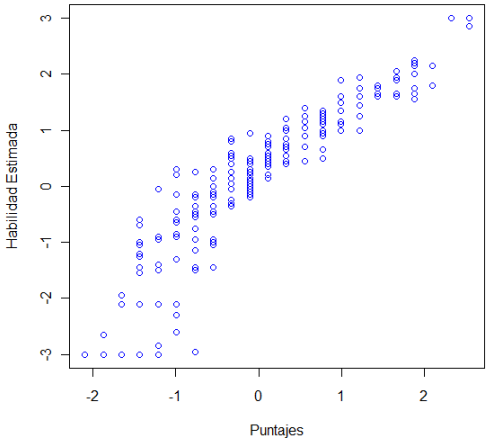
$$\hat{y} = 12 + 0.7x$$

Donde: $\hat{y}$ es el estimador del criterio externo y *x* es la habilidad estimada.

Coeficiente de determinación: 0.82

La siguiente gráfica muestra que a sustentantes con el igual puntaje les corresponden distintos grados de habilidad estimada; esto es causado porque aun cuando tienen el mismo número de respuestas correctas, las preguntas difieren en el índice de discriminación que funciona como un ponderador que impacta a la probabilidad de respuesta correcta. En el caso del modelo Rasch, donde el índice de discriminación es uno, el puntaje es un estimador suficiente para estimar la habilidad.

Figura 0.19 Habilidad estimada con respecto al puntaje.



Además, se comprueba visualmente en la gráfica que la dispersión de la habilidad no es homogénea ya que el error de la estimación es mayor con los valores menores a cero, como se indica con la función de información del examen e intervalos de confianza anteriormente presentados.

Para verificar la invarianza del estimador de la habilidad, únicamente se consideraron los 16 reactivos más fáciles (dificultad<0.8) y comprobamos con la correlación de este estimador y el criterio externo de 0.83 que no pierde significativamente validez predictiva por ser los reactivos más fáciles ni por no

aplicar todos los ítems, aunque si pierde precisión por no cubrir éstos el rango total del parámetro. Ésta es la característica más importante de TRI.

4.8 *Estimación de la calificación mínima aprobatoria.*

Con clasificación *a priori* de los sustentantes como apto o no, se estimó la calificación mínima aprobatoria con la habilidad estimada, utilizando la regresión logística.

Los parámetros estimados de la regresión logística son -2.77 y 5.31 correspondientes al intercepto y la pendiente, con el error estándar de 0.5343 y 0.8969 respectivamente, ambos parámetros son significativos y el punto de corte es $-\hat{\beta}_0 / \hat{\beta}_1 = 0.52$, con lo cual se obtiene la siguiente Figura 0.20.

La eficiencia del modelo es de 93%, es decir, 186 de 200 clasificaciones de los evaluadores expertos mantuvieron su denominación con la prueba de escritorio y el punto de corte estimado, como se muestra en la **¡Error! No se encuentra el origen de la referencia.**.

Figura 0.20 Curva logística de la prueba.

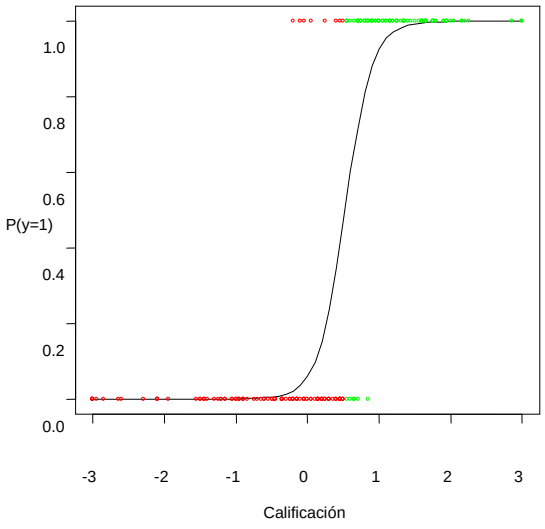


Tabla 0.7 Clasificación de sustentantes por los expertos y la prueba de escritorio.

Expertos	Prueba de escritorio		
	No apto	Apto	Total
No apto	108	6	114
Apto	8	78	86
Total	116	84	200

Valorando la capacidad predictiva del puntaje en la clasificación, el punto de corte se estimó en 0.12 presentado una eficiencia de clasificación de 87%, menor que con la habilidad estimada.

4.9 *Validación del método, modelo y técnicas.*

La bondad del ajuste de modelos de uno, dos y tres parámetros fueron analizados mediante la prueba X^2 de Pearson, con la cual se determinó que el modelo que menor discrepancia presenta entre la distribución observada y la teórica, en este caso de estudio, es el modelo de tres parámetros.

La unidimensionalidad de la habilidad se fundamentó en la tabla de especificaciones del perfil referencial y fue verificado mediante el análisis factorial.

La CCI representa la probabilidad de respuesta correcta condicionada al nivel de la habilidad, por lo que es una función monotónica creciente que todos los

reactivos del examen presentan con excepción del 16 y 32: el primero manifiesta una probabilidad de respuesta independiente a la habilidad; y el segundo, una probabilidad mayor con los que tienen menos habilidad.

La independencia local sustenta que los reactivos son independientes unos de otros, sin que la información contenida en uno sugiera la solución de otro, ni la respuesta de un reactivo antecedente sea necesaria para poder contestar un reactivo posterior.

La estabilidad de parámetro de dificultad se comprobó que mantiene una constante independientemente del nivel de conocimientos de las muestras que se elija para su calibración.

La verificación de la invarianza (alta asociación) de la estimación θ de los evaluados fue comprobado por no presentar variación significativa al ser evaluado con dos grupos de ítems de índices de dificultad diferentes, el estadístico para confirmar la invarainza fue las correlaciones entre el nivel θ estimado con ítems fáciles y difíciles.

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

Los que se enamoran de la práctica sin la teoría son como los pilotos sin timón ni brújula, que nunca podrán saber adónde van.

Leonardo de Vincí (1452- 1519)

5.1 *Conclusiones.*

El diseñar buenos exámenes de opción múltiple no es sencillo, se requiere conocimiento y experiencia en la temática, pedagogía, técnicas estadísticas y sobre todo, tiempo; pero la inversión se compensa ampliamente por la calidad de la evaluación manifestada en la objetividad, validez y confianza de los resultados, que potencializa la capacidad y facilidad de aplicación a gran escala a costos relativamente bajos.

El enfoque clásico tiene gran tradición por su sencillez; la eficacia depende, en primer instancia, del perfil referencial susceptible de ser definido de manera

mensurable, suficiencia y balanceo del constructo; y en segunda instancia, de la calidad de los reactivos que deberán ser validados y en general, debe calibrarse el test mediante una muestra representativa de la población objetivo, pues de la muestra dependen los indicadores de dificultad y discriminación, el test o los ítems resultarán fáciles si los sujetos son muy competentes; y difíciles, si los sujetos son menos aventajados.

La Teoría de Respecta de Ítem se sustenta en un modelo de probabilidad, de tal forma que ésta depende de la habilidad de la persona y las características del ítem por lo que los supuestos básicos son:

- La relación entre el patrón de respuestas y el nivel de la habilidad puede ser descrita por una función monótonamente incremental.
- La independencia de un ítem, representada por la probabilidad de responder correctamente, únicamente depende del nivel de dominio y de las características del ítem. Es decir, es independiente a otros ítems.

En caso de que la variable latente sea multidimensional, se deberán identificar los ítems para cada dimensión y para una de éstas, calibrar con su correspondiente constructo. Se puede basar en la tabla referencial y comprobarse con análisis factorial.

Las tareas fundamentales de TRI son la comprobación del modelo (verificar los supuestos y la invarianza de parámetros), la calibración de ítems, comprobar el ajuste al modelo elegido y estimar el nivel habilidad de los evaluados.

La principal aportación de la TRI frente al enfoque clásico es ofrecer puntaciones invariantes con respecto al instrumento empleado y a la muestra con la que se estandariza el test. El parámetro de dificultad se mantiene constante independientemente del grupo que se elija para su estimación. Es decir, la

estimación de la dificultad del ítem es independiente del nivel de conocimientos de la muestra.

En términos prácticos, la ventaja de la TRI frente a la TCT es que los ítems calibrados desde TRI permiten llevar a cabo evaluaciones rápidas y adaptadas a las características de los sustentantes, cuestión imposible en el enfoque clásico.

En la TCT, el núcleo del análisis es el test, la prueba objetiva y estandarizada. Cuando se evalúa la competencia de una persona con una escala ajustada desde TCT es necesario aplicar el protocolo completo para estimar su nivel de rendimiento. En cambio, desde la óptica TRI no es necesario aplicar un conjunto de ítems completo, con implementar unos pocos ítems se tiene la evidencia necesaria para estimar la habilidad de los evaluados. Esto es así porque en la TRI el protagonismo pasa del test al ítem. Ya no interesa el test en su conjunto, sino el ítem individual. La tarea clave no es estandarizar una prueba, sino estimar los parámetros de los ítems.

Así, que al disponer de un banco de ítems calibrado es posible calcular el nivel habilidad de los evaluados a partir de un conjunto pequeño y representativo de ítems de dicho banco. Pero, además de rápidas, la TRI permite realizar evaluaciones adaptadas a las características de los sujetos.

En la TCT, el intervalo de confianza del error de medida es constante a lo largo de la escala. Sin embargo, en la TRI se establece un intervalo confianza para cada uno de los valores de la habilidad. Es decir, un ítem determinado no tiene la misma precisión para todos los niveles de la escala de habilidad. Esto supone que hay ítems más adecuados para medir en ciertos niveles de la habilidad. Precisamente para aquéllos donde el error estándar es más pequeño.

Una desventaja de los exámenes de opción múltiple es que no siempre se presta para ser evaluado sobre cómo se hace algo, sólo se puede demostrar de una

forma directa, por ejemplo: el arte, la música, el lenguaje, el deporte o conducir un automóvil, para estos casos es indispensable las evaluaciones de auténticas.

Otra desventaja, el procedimiento estadístico de la estimación de la habilidad o propiamente dicho, la calificación con la TRI que percibe el sustentante puede ser que no sea lo suficientemente comprensible para los sustentantes, pues si hay desconocimiento de esta teoría por quien realiza las evaluaciones, con mayor razón por quienes son evaluados. Lo ideal es seleccionar ítems con índice de dificultad y respuesta aleatoria de uno y cero respectivamente para que la puntación sea un estimador suficiente para la habilidad.

Además, los programas de cómputo son de propósito general y quizá se requiera la participación de un estadista para la verificación de supuestos, el análisis e interpretación de resultados.

La validez y confiabilidad de las evaluaciones son funciones insustituibles del experto temático y pedagogo; las técnicas estadísticas presentadas en este contexto sólo son indicad5.1-1ores que orientan el análisis de los reactivos y en general, del examen.

5.2 *Recomendaciones.*

La elaboración de los exámenes es totalmente recomendado seguir el proceso de diseño, elaboración y validación con rigor metodológico, sin dejar de ser práctico y realista a las circunstancias. La aplicación de métodos y técnicas para valorar la validez y confiabilidad de las evaluaciones depende principalmente del grado de riesgo o consecuencias de tomar decisiones equivocadas con base en los resultados de la evaluación; cuando el riesgo es bajo, quizá sean suficienten las técnicas clásicas por su sencillez; en caso contrario, cuando se requiere de mayor precisión se justifica la aplicación de la Teoría de la Respuesta al Ítem, sobre todo, cuando la aplicación es a gran escala.

Cuando la actividad de evaluación es cotidiana, es recomendable utilizar un banco de preguntas como fuente para la elaboración de exámenes, donde las preguntas estén clasificadas no sólo por su dominio temático y taxonómico, sino también por sus propiedades psicométricas. La inversión de calibrar los ítems del banco con la teoría de respuesta es totalmente rentable por las ventajas ya mencionadas.

En la vida real, las habilidades no se expresan siempre en un binomio de aciertos y errores, en varias ocasiones no hay respuestas totalmente correctas o equivocadas, aun en los ítems tradicionales de opción múltiple hay errores más graves que otros, como sucede frecuentemente en estudios de casos donde todas las alternativas son correctas, pero unas son más eficientes que otras.

Otra situación análoga es cuando se valoran actitudes donde la respuesta a un ítem es de tipo categórica y las selecciones de éstas están relacionadas con el nivel de la habilidad. Para evaluar de manera eficaz estos casos se requiere la implementación de modelos innovadores en psicometría que consideren respuestas de ítems como una variable continúa entre 0 y 1, o incluso variables ordinales, las cuales no han sido abordadas en esta investigación.

GLOSARIO

Banco de ítems	Conjunto de ítems con propiedades psicométricas conocidas.
Bondad de ajuste	Procedimiento estadístico por el cual se determina si un modelo matemático es adecuado para describir los resultados empíricos.
Coeficiente de determinación	Es la proporción de la variación total en la variable dependiente Y que es explicada por la variación en la variable independiente X.
Coincidencias	Se considera coincidencia cuando dos personas, habiendo respondido mal el ítem, coincidan en la alternativa elegida. Es decir, hace falta que coincidan en la alternativa elegida y además que ésta sea incorrecta. En preguntas de opción múltiple, han surgido métodos para detectar si dos personas han respondido independiente el uno del otro. Para ello hay que calcular el número de coincidencias.
Consistencia interna (*alpha*)	Es un indicador indirecto de la fiabilidad del test que indica el grado en el que la gente que tiende a acertar unos ítems tiende a acertar también los demás ítems. Es decir, el grado en que los ítems ordenan a las personas de la misma manera. Oscila generalmente entre 0 y 1. Valores por encima de 0.7 son adecuados (recomendables por encima de 0.8). Valores por debajo de ese valor indican que nuestras estimaciones pueden ser imprecisas (ver análisis del test).
Correlación	Es un grupo de técnicas estadísticas usadas para medir la fuerza de la relación entre dos variables.
Correlación ítem-test ó índice de homogeneidad corregido	Es la correlación entre la puntuación en el ítem y la puntuación en el resto del test e indica el grado de relación entre sacar una puntuación alta en el test y acertar el ítem. Si todos los ítems están midiendo lo mismo, los estudiantes que aciertan un ítem tendrán mayor probabilidad de acertar el resto de las preguntas del test. Puede hablarse de correlación biserial o biserial-puntual según se asuma o no que al hecho de acertar o fallar un ítem subyazca una capacidad o variable continúa.
Dificultad de un ítem	Proporción de sujetos que escoge la opción correcta. Se suele obtener como el cociente entre el número de personas que lo aciertan y el número de personas que lo responden. Es una proporción que va desde 0 (ítem muy difícil) a 1 (ítem muy fácil). Lo ideal es que los ítems no sean ni demasiado fáciles ni demasiado difíciles.
El Perfil Referencial de	Es la descripción y delimitar cuáles serán las conductas específicas a determinar mediante la medición.
Estimación de parámetros	Procedimiento estadístico por el que se utiliza la información contenida en una muestra para asignar un valor a los parámetros de una determinada población.
Estimación puntual	Es obtener valores aproximados de parámetros poblacionales.
Estimador	Un estadístico es un estimador coherente de un parámetro de población, si al

coherente	aumentar el tamaño de la muestra se tiene casi la certeza de que el valor del estadístico se aproxima bastante al valor del parámetro de la población.
Estimador eficiente	La eficiencia se refiere al tamaño de error estándar del estadístico. Si comparamos dos estadísticos de una muestra del mismo tamaño y tratamos de decidir cuál de ellos es un estimador más eficiente, escogeríamos el que tuviera el menor error estándar.
Estimador insesgado	Es aquel para el cual la media de la distribución muestral es el parámetro estimado.
Estimador suficiente	Un estimador es suficiente si utiliza una cantidad de la información contenida de la muestra que ningún otro estimador podría extraer información adicional de la muestra sobre el parámetro de la población que se está estimando.
Hipótesis	Es una afirmación o suposición respecto al valor de un parámetro poblacional.
Ítem	Cada una de las preguntas o elementos de un examen o test.
Modelo de respuesta al ítem	Modelo matemático que proporciona la probabilidad de que un sujeto acierte un ítem, conocido el nivel de rasgo del sujeto y la dificultad (u otras características) del ítem.
Nivel de confianza	Es la probabilidad de que el parámetro poblacional se encuentre dentro del intervalo.
Punto de corte	Puntuación del test que permite clasificar a los sustentantes según superen o no dicha puntuación.
Test	Conjunto de preguntas (o ítems) administrados a un sujeto para estimar su nivel de conocimientos, inteligencia, actitudes o personalidad.

BIBLIOGRAFÍA

Algina, C. L. (1986). *Introduction to Classcal and Moderm Test Theory.* Holt Rinehart and Winston.

CENEVAL (1999). Informe de resultados 1998. México, CENEVAL.

Cronbach, L. (1971). *Test validation.* Washigton: Consejo Americano en Educación.

Donlon, Thomas F. (Ed.) (1984). The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests. New York, College Entrance Examination Board. Especialmente los capítulos I (The Admissions Testing Program: A Historical Overview, pp. 1-11) y II (Psychometric Methods Used in the Admissions Testing Program, pp. 13-36).

Ebel, R.L. y Frisbie, D.A. (1986). Essentials of Education Measurement. Englewood Cliffs, NJ: Prentice Hall.

Glass, G. y Stanley, J. (1986). Métodos estadísticos aplicados a las ciencias sociales. (Traducción: E. Galvis y E. Guzmán). México: Prentice-Hall Hispanoamericana.

Gulliksen, H. (1950/1987). *Theory of mental tests.* Hillsdale: Lawrence Erlbaum Associates.

Henrysson, S. (1971). Gathering, Analysing, and Using Data on Test Items. En R.L. Thorndike (Ed.), Educational Measurement. Washington, DC: American Council on Education.

Lord, F.M. y Novick, M.R. (1968). *Statistical theories of mental tests scores.* Reading, Mass: Addison - Wesley Pub. Co.

Martínez Arias, Rosario (1995). Psicometría: teoría de los tests psicológicos y educativos. Madrid. Síntesis.

Tristán Agustín, (1996) V. R. *Introducción al diseño de pruebas objetivas.*

WOLFF, Lawrence (1998). Las evaluaciones educacionales en América Latina: Avance actual y futuros desafíos. Santiago de Chile, PREAL.

ANEXOS

Anexo A. Misión, visión, valores y proceso.

Misión.

Brindar servicios de capacitación que satisfagan las necesidades reales del personal del INEGI y su desarrollo personal, así como promover acciones que impulsen el intercambio de conocimientos y experiencias en el ámbito nacional e internacional, que contribuyan al logro de los propósitos interinstitucionales.

Visión.

Ser altamente competitivos; orientados a satisfacer las necesidades de capacitación de nuestros usuarios, reconocidos como líderes innovadores en este campo, a nivel nacional e internacional, mediante una administración efectiva en un ambiente de trabajo en equipo.

Valores.

Aprendizaje, Compromiso, Comunicación, Honestidad, Justicia, Profesionalismo, Respeto, Responsabilidad, Servicio al cliente, Trabajo en equipo.

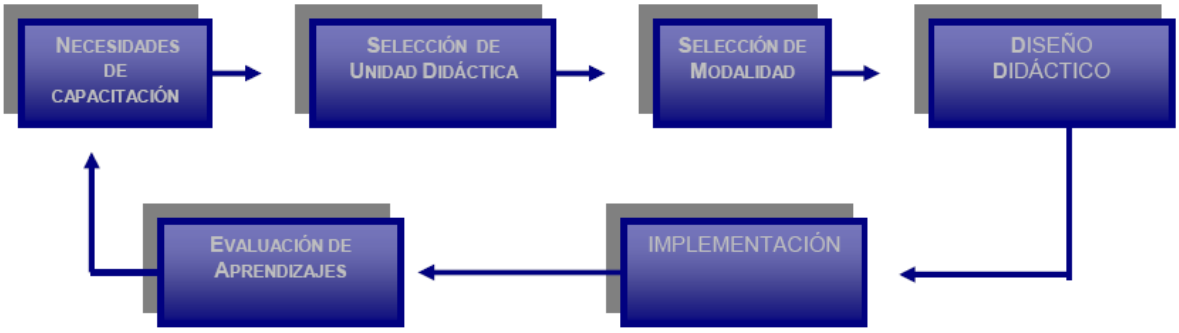
Objetivo.

Formar, actualizar y especializar al personal del Instituto para contribuir a la profesionalización del capital humano del INEGI y al desarrollo de programas de investigación y nuevas metodologías en materia estadística y geográfica.

Estrategia.

Brindar servicios de capacitación de calidad que atiendan las diversas necesidades de conocimiento del INEGI bajo un enfoque integral y flexible.

Proceso de capacitación.



Inicio: Detección de necesidades de capacitación.

Fin: Evaluación de aprendizajes.

Anexo B. Examen de Geografía de México.

Tabla de especificaciones.

Contenido del examen	Cantidad de preguntas						
	Conocimiento	Compresión	Aplicación	Análisis	Síntesis	Evaluación	Total
Origen del cosmos, formación y estructura de la Tierra.	3						3
Definición, aplicación de la Geografía y representación de la Tierra.	2	3					5
Origen histórico del territorio mexicano.	3	2					5
Referencias geográficas y extensiones territoriales de México.	2	3					5
Características edafológicas, fisiológicas, climáticas e hidrográficas.	3	2					5
Distribución de la población mexicana y su economía.	3	2					5
Regiones naturales y biogeografía.	3	2					5

Reactivos.

1 Consideraban a la Geografía como meramente descriptiva:

A) Griegos
B) Calderos
C) Chinos
D) Egipcios

2 ¿Quién es el padre de la Geografía moderna, que la definió como una ciencia a inicios del Siglo XX?

A) Emmanuel D Martone
B) Emmanuel Kant
C) Humboldt
D) Pedro S. Laplace

3 Una de las teorías nebulares acerca del origen del sistema solar fue establecida por _____, quien consideró que una enorme nebulosa en forma

incandescente formó nuestro actual sistema solar.

A) Laplace
B) Kant
C) Alfvén
D) Hoyle

4 Es una de las teorías más aceptadas con respecto a la formación del universo:

A) Bing bang
B) Universo pulsante
C) Fuerza centrípeta
D) Gravitación universal

5 Son líneas imaginarias que nos permiten situar la distancia angular entre un punto de la superficie terrestre y el plano del Ecuador.

A) Paralelos
B) Meridianos
C) Altitudes
D) Latitudes

6 El medio que nos rodea, por ejemplo las montañas, océanos, cordilleras, etc., fue formado por procesos muy prolongados de tiempo real. Esto se refiere a:

A) Hecho geográfico
B) Principio geográfico
C) Fenómeno geográfico
D) Evento geográfico

7 Montañas, océanos y bosques son:

A) Hechos geográficos
B) Fenómeno geográficos
C) Fenómenos naturales
D) Hechos naturales

8 Se presenta sobre la superficie terrestre de manera violenta y puede ser observado:

A) Fenómeno geográfico
B) Hecho geográfico
C) Hecho catastrófico
D) Evento natural

9 Se refiere específicamente al estudio de una parte de la superficie terrestre, es decir, comprende estudios exhaustivos de una comarca o región:

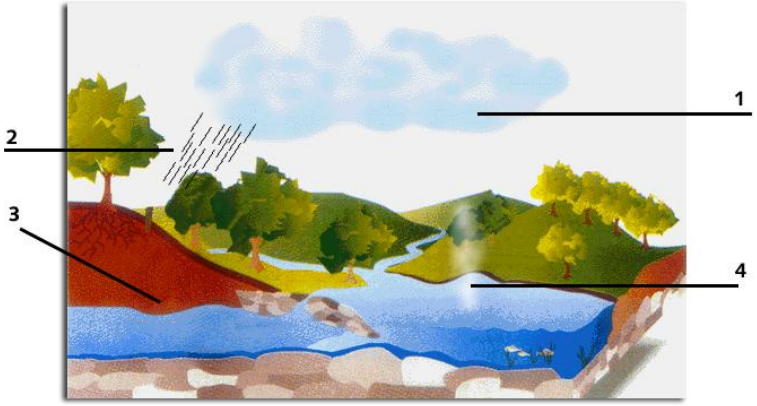
A) Geografía particular
B) Geografía general
C) Geografía humana

D) Geografía física

10 El principio geográfico que determina la duración y el alcance de un hecho o de un fenómeno geográfico, es decir, su dimensión en el tiempo y en el espacio es:

A) Extensión
B) Generalización
C) Causalidad
D) Temporalidad

11 Las etapas del ciclo del agua son:



A) 1. Condensación; 2 Precipitación; 3 Filtración; 4 Evaporación
B) 1. Precipitación; 2 Condensación; 3 Filtración; 4 Evaporación
C) 1 Filtración; 2 Precipitación; 3 Evaporación; 4 Condensación
D) 1. Evaporación; 2 Filtración; 3 Precipitación; 4 Condensación

12 El clima se define como:

A) El conjunto de fenómenos meteorológicos que caracterizan el estado medio de la atmósfera en una región de la superficie terrestre.
B) El conjunto de elementos atmosféricos como presión, viento y temperatura que se presentan en la atmósfera en una región de la superficie terrestre.
C) El conjunto de factores que alteran el comportamiento de los elementos del tiempo en un lugar de la superficie terrestre.
D) El conjunto de elementos meteorológicos como presión, viento y precipitación que se presentan durante un día.

13 El estudio puntual de una comunidad es tema de la:

A) Geografía particular

B) Geografía general
C) Geografía física
D) Geografía humana

14 Definir, ubicar, determinar la duración y extensión y conocer la correlación de las relaciones mutuas entre fenómenos iguales o semejantes, son los principios de:

A) Método geográfico
B) Determinismo geográfico
C) Fenómeno geográfico
D) Hecho geográfico

15 Explicar lo basta que fue la erupción del volcán Santa Elena en la década de los 80s. en EEUU, nos lo permite el:

A) Principio de extensión
B) Principio de localización
C) Principio de causalidad
D) Principio de generalización

16 Este principio explica cómo un gran sismo provocó el tsunami de diciembre de 2004.

A) De relación
B) De causalidad
C) De extensión
D) De localización

17 Este principio explica las condiciones análogas del medio geográfico, es decir, las posibilidades de que en otro lugar de la Tierra se presente un mismo fenómeno:

A) De generalización
B) De extensión
C) De causalidad
D) De localización

18 Selecciona el año y el nombre del documento mediante el cual México perdió más de la mitad de su territorio:

A) El tratado de Guadalupe Hidalgo en 1848
B) La Constitución de 1824
C) El tratado de Mesilla en 1853
D) La constitución de 1857

19 La superficie total del país incluyendo el territorio insular es de:

A) 1.9 millones de km^2
B) 2.0 millones de km^2
C) 1.8 millones de km^2
D) 2.1 millones de km^2

20 ¿Cuál es la isla que delimita el país en su parte más occidental?

A) Isla Guadalupe
B) Isla Mujeres
C) Isla de Cedros
D) Isla Clarión

21 Los meridianos al Oeste del Meridiano de Greenwich entre los que se encuentra la República Mexicana son:

A) 86° y 118°
B) 82° y 96°
C) 86° y 120°
D) 85° y 120°

22 México colinda con los países:

A) Estados Unidos de América; Guatemala y Belice.
B) Canadá; Estados Unidos de América y Guatemala.
C) Estados Unidos de América; Belice y Nicaragua.
D) Guatemala; El Salvador y Estados Unidos de América.

23 México tiene fronteras políticas al:

A) Norte y Sur
B) Sur, Este y Noroeste
C) Norte, Sureste y Suroeste
D) Norte, Sureste

24 Los estados limítrofes con la frontera norte son:

A) Baja California, Tamaulipas, Nuevo León, Sonora, Coahuila de Zaragoza y Chihuahua
B) Coahuila de Zaragoza, Tamaulipas, Sonora, Durango, Chihuahua y Nuevo León.
C) Chihuahua, Baja California; Sonora, Tamaulipas, Nuevo León, Sinaloa.
D) Sinaloa, Chihuahua, Baja California, Sonora, Coahuila de Zaragoza y Tamaulipas.

25 Los estados fronterizos sur y sureste del país son:

A) Campeche; Chiapas, Tabasco y Quintana Roo.
B) Yucatán, Quintana Roo, Tabasco y Chiapas.
C) Veracruz de Ignacio de la Llave, Chiapas, Campeche y Tabasco.
D) Tabasco, Chiapas, Quintana Roo y Oaxaca.

26 Los principales tipos de clima en nuestro país son:

A) Cálidos, secos, templados
B) Cálidos, secos, fríos
C) Cálidos, secos, polar

D) Secos, fríos, polares

27 Dentro de la clasificación climática, la f representa:

A) Abundantes lluvias todo el año
B) Escasas lluvias durante el verano
C) Abundantes lluvias en verano
D) Escasas lluvias todo el año

28 La población total de México según el Conteo Nacional de Población y Vivienda 2005 es:

A) Más de 100 millones
B) Más de 120 millones
C) Alrededor de 90 millones
D) Alrededor de 97 millones

29 Se considera también una actividad económica primaria:

A) La silvicultura
B) Los recursos naturales
C) El turismo
D) La agroindustria

30 El grupo étnico más numeroso en México es:

A) Náhuatl
B) Maya
C) Zapoteca
D) Otomí

31 En México se localizan dos grandes regiones biográficas que son:

A) Neártica y neotropical
B) Neártica y neodesértica
C) Neotropical y neosubtropical
D) Neodesértica y neosubtropical

32 Biodiversidad se define como:

A) La variedad de especies vegetales y animales que habitan en una región natural
B) La distribución y variedad florística de las regiones geográficas del país
C) La distribución y variedad faunística de las regiones geográficas del país
D) La variedad de especies vegetales sustentables en las regiones naturales

33 Este corte transversal presenta la estructura interna de la tierra y en orden ascendente cada uno de sus elementos:

A)
1. Núcleo interno

2. Núcleo externo
3. Discontinuidad de Gutemberg
4. Manto inferior
5. Manto superior
6. Discontinuidad de Mohorovicic
7. Corteza

B)

1. Núcleo interno
2. Núcleo externo
3. Discontinuidad de Mohorovicic
4. Manto inferior
5. Manto superior
6. Discontinuidad de Gutenberg
7. Corteza

C)

1. Núcleo externo
2. Núcleo interno
3. Discontinuidad de Morovicic
4. Corteza
5. Manto inferior
6. Manto superior
7. Discontinuidad de Gutenberg

D)

1. Manto inferior
2. Manto superior
3. Discontinuidad de Gutenberg
4. Núcleo interno
5. Núcleo externo
6. Discontinuidad de Mohorovicic
7. Corteza

Anexo C. Índice de dificultad y discriminación por la Teoría Clásica de los Tests.

Núm. pregunta	Dificultad			Discriminación			Norma de discriminación		
	Media	Desviación estándar	Probabilidad de rechazo	Media	Desviación estándar	Probabilidad de rechazo	Media	Desviación estándar	Probabilidad de rechazo
1	0.42	0.0715	0.00	0.26	0.1348	0.35	2.08	1.1012	0.19
2	0.75	0.0616	0.28	0.31	0.1216	0.19	1.51	0.5814	0.14
3	0.64	0.0718	0.01	0.13	0.1546	0.67	0.69	0.8295	0.65
4	0.33	0.0623	0.03	0.26	0.1279	0.30	2.62	1.2288	0.11
5	0.47	0.0721	0.00	0.26	0.1387	0.33	1.87	1.0363	0.21
6	0.59	0.071	0.00	0.36	0.1413	0.13	2.10	0.8891	0.12
7	0.60	0.0697	0.00	0.37	0.1453	0.11	2.13	0.9043	0.11
8	0.63	0.0681	0.01	0.31	0.1379	0.21	1.71	0.8043	0.19
9	0.39	0.0701	0.00	0.10	0.1462	0.75	0.85	1.3128	0.58
10	0.70	0.0638	0.10	0.08	0.1366	0.83	0.37	0.6794	0.83
11	0.74	0.067	0.27	0.20	0.1290	0.51	1.00	0.652	0.47
12	0.64	0.0711	0.01	0.30	0.1405	0.23	1.60	0.8081	0.22
13	0.70	0.0671	0.12	0.03	0.1389	0.89	0.14	0.7092	0.89
14	0.57	0.0754	0.00	0.13	0.1397	0.70	0.75	0.8498	0.65
15	0.54	0.0754	0.00	0.09	0.1406	0.78	0.58	0.8972	0.70
16	0.77	0.0643	0.40	-0.02	0.1327	0.97	-0.12	0.6837	0.96
17	0.76	0.067	0.34	0.07	0.1279	0.84	0.37	0.6721	0.81
18	0.34	0.0659	0.02	0.21	0.1355	0.46	1.98	1.3738	0.24
19	0.56	0.0744	0.00	0.22	0.1464	0.46	1.32	0.9106	0.37
20	0.47	0.0658	0.00	0.31	0.1382	0.20	2.26	0.9907	0.12
21	0.39	0.0612	0.00	0.49	0.1120	0.00	4.25	0.7476	0.00
22	0.58	0.0656	0.00	0.44	0.1250	0.02	2.56	0.8024	0.02
23	0.67	0.0689	0.04	-0.07	0.1462	0.97	-0.34	0.7596	0.97
24	0.44	0.0699	0.00	0.19	0.1431	0.53	1.48	1.1025	0.36
25	0.81	0.0599	0.68	0.16	0.1150	0.63	0.89	0.6648	0.52
26	0.27	0.059	0.17	0.16	0.1237	0.62	1.95	1.4621	0.25
27	0.61	0.0704	0.01	0.33	0.1459	0.18	1.83	0.8676	0.17

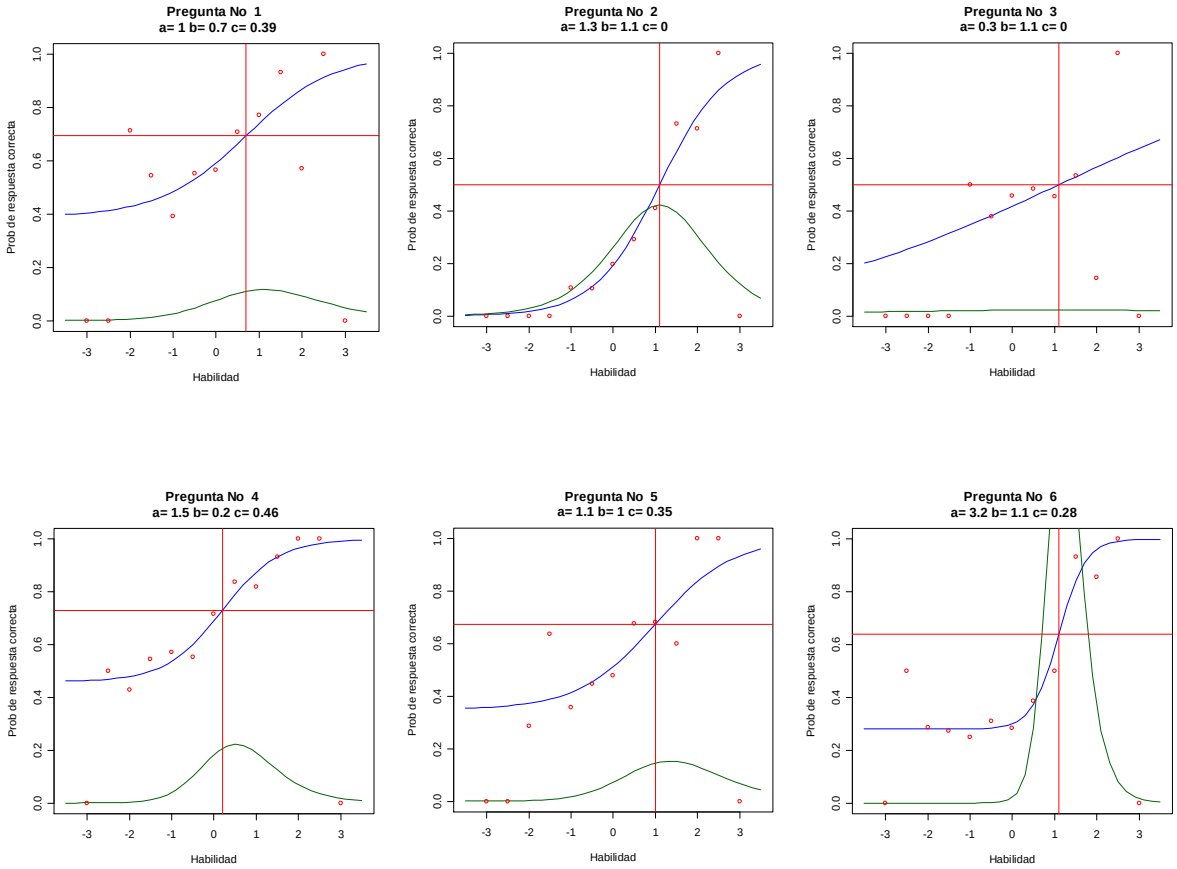
28	0.49	0.0735	0.00	-0.04	0.1447	0.95	-0.32	1.0293	0.92
29	0.57	0.065	0.00	0.48	0.1197	0.02	2.86	0.817	0.01
30	0.71	0.064	0.11	0.29	0.1278	0.24	1.40	0.6498	0.23
31	0.83	0.0621	0.77	0.14	0.1168	0.70	0.83	0.7369	0.53
32	0.78	0.064	0.46	0.16	0.1238	0.63	0.82	0.6342	0.58
33	0.52	0.0744	0.00	0.22	0.1465	0.44	1.40	0.9773	0.34

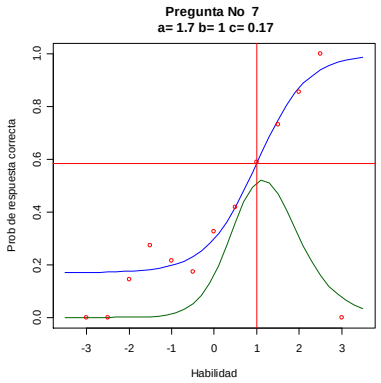
Anexo D. Índice de discriminación, dificultad y respuesta aleatoria por la Teoría de la Respuesta al Ítem.

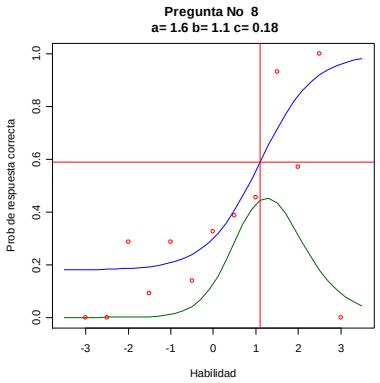
Núm. pregunta	Discriminación			Dificultada			Respuesta aleatoria		
	Media	Desviación estándar	Probabilidad de rechazo	Media	Desviación estándar	Probabilidad de rechazo	Media	Desviación estándar	Probabilidad de rechazo
1	1.14	0.6253	0.00	0.35	0.8580	0.00	0.32	0.1786	0.72
2	1.60	0.5222	0.00	1.09	0.3471	0.00	0.03	0.0398	0.00
3	0.35	0.1408	0.00	1.33	0.8714	0.07	0.00	0.0224	0.00
4	1.53	0.6059	0.00	-0.06	0.5641	0.00	0.39	0.1586	0.85
5	1.34	0.7374	0.00	0.69	0.7302	0.00	0.28	0.1808	0.67
6	2.69	0.4506	0.00	1.11	0.2150	0.00	0.26	0.0557	0.69
7	1.78	0.5917	0.00	0.97	0.2451	0.00	0.15	0.0794	0.10
8	1.74	0.6629	0.00	1.12	0.2660	0.00	0.16	0.0844	0.11
9	0.32	0.2506	0.08	-2.22	2.1441	0.37	0.10	0.1616	0.18
10	0.65	0.3937	0.00	2.21	0.8579	0.17	0.08	0.0932	0.07
11	2.09	0.5967	0.00	1.68	0.2883	0.00	0.13	0.0489	0.01
12	0.88	0.6087	0.00	1.08	0.5330	0.00	0.07	0.1091	0.12
13	0.90	1.0959	0.00	3.26	0.9129	0.61	0.09	0.1318	0.26
14	1.58	1.1568	0.02	1.58	0.9408	0.08	0.24	0.1893	0.61
15	0.52	0.5876	0.06	1.37	1.2958	0.13	0.13	0.1680	0.31
16	0.34	0.4565	0.00	3.91	0.3306	0.96	0.01	0.0446	0.00
17	0.35	0.4056	0.00	3.43	0.7710	0.70	0.01	0.0420	0.01
18	1.30	0.2632	0.00	-0.82	0.2014	0.00	0.02	0.0590	0.02
19	0.57	0.3121	0.01	0.78	0.8281	0.02	0.10	0.1309	0.18
20	0.86	0.3169	0.00	0.03	0.4945	0.00	0.14	0.1278	0.27

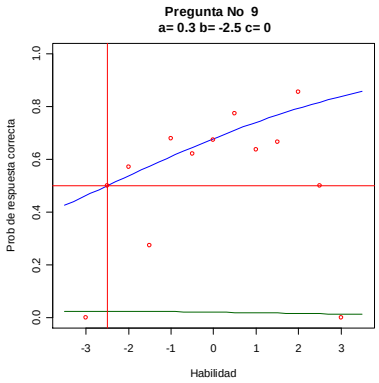
21	2.30	0.5603	0.00	-0.08	0.2809	0.00	0.24	0.1313	0.61
22	1.16	0.2563	0.00	0.24	0.1876	0.00	0.03	0.0476	0.00
23	0.20	0.2805	0.00	3.97	0.2435	0.98	0.00	0.0330	0.01
24	1.85	0.8772	0.00	0.96	0.5624	0.00	0.39	0.1333	0.87
25	0.92	0.4904	0.00	2.05	0.5722	0.06	0.03	0.0495	0.00
26	1.17	0.3610	0.00	-0.92	0.6465	0.00	0.17	0.2061	0.36
27	1.81	0.6285	0.00	0.59	0.2228	0.00	0.11	0.0907	0.06
28	0.40	0.5150	0.20	0.96	1.7803	0.14	0.18	0.1759	0.34
29	1.01	0.2124	0.00	0.25	0.1821	0.00	0.00	0.0228	0.00
30	0.78	0.5836	0.00	1.45	0.5077	0.01	0.05	0.0873	0.06
31	2.65	0.6024	0.00	1.79	0.2115	0.00	0.09	0.0358	0.00
32	1.15	0.8678	0.00	2.29	0.6832	0.13	0.08	0.0842	0.01
33	1.23	0.6601	0.00	0.19	0.5621	0.00	0.15	0.1700	0.33

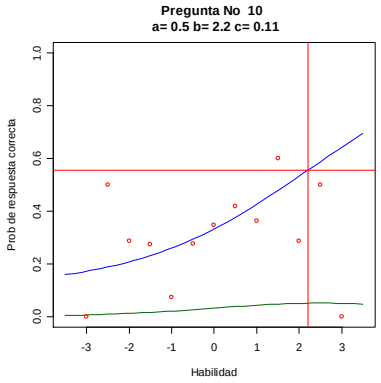
Anexo E. Curva característica de ítems del examen de generación de información geográfica.

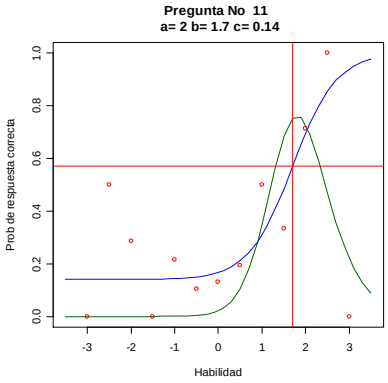


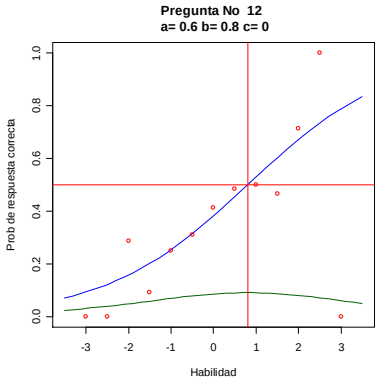

Pregunta No 7
a= 1.7 b= 1 c= 0.17
Prob de respuesta correcta
Habilidad

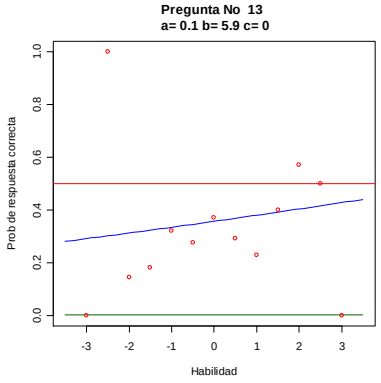

Pregunta No 8
a= 1.6 b= 1.1 c= 0.18
Prob de respuesta correcta
Habilidad

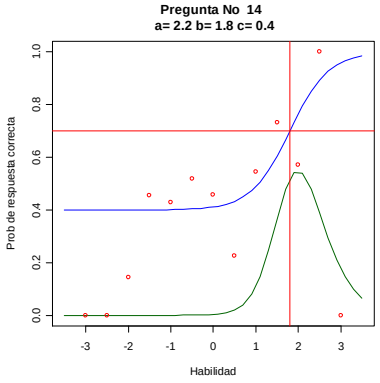

Pregunta No 9
a= 0.3 b= -2.5 c= 0
Prob de respuesta correcta
Habilidad

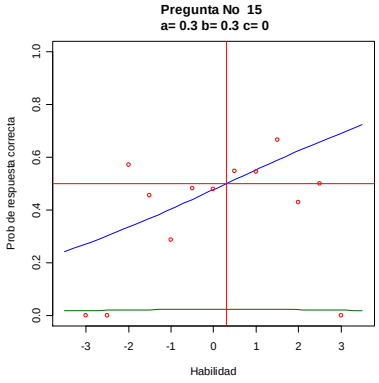

Pregunta No 10
a= 0.5 b= 2.2 c= 0.11
Prob de respuesta correcta
Habilidad

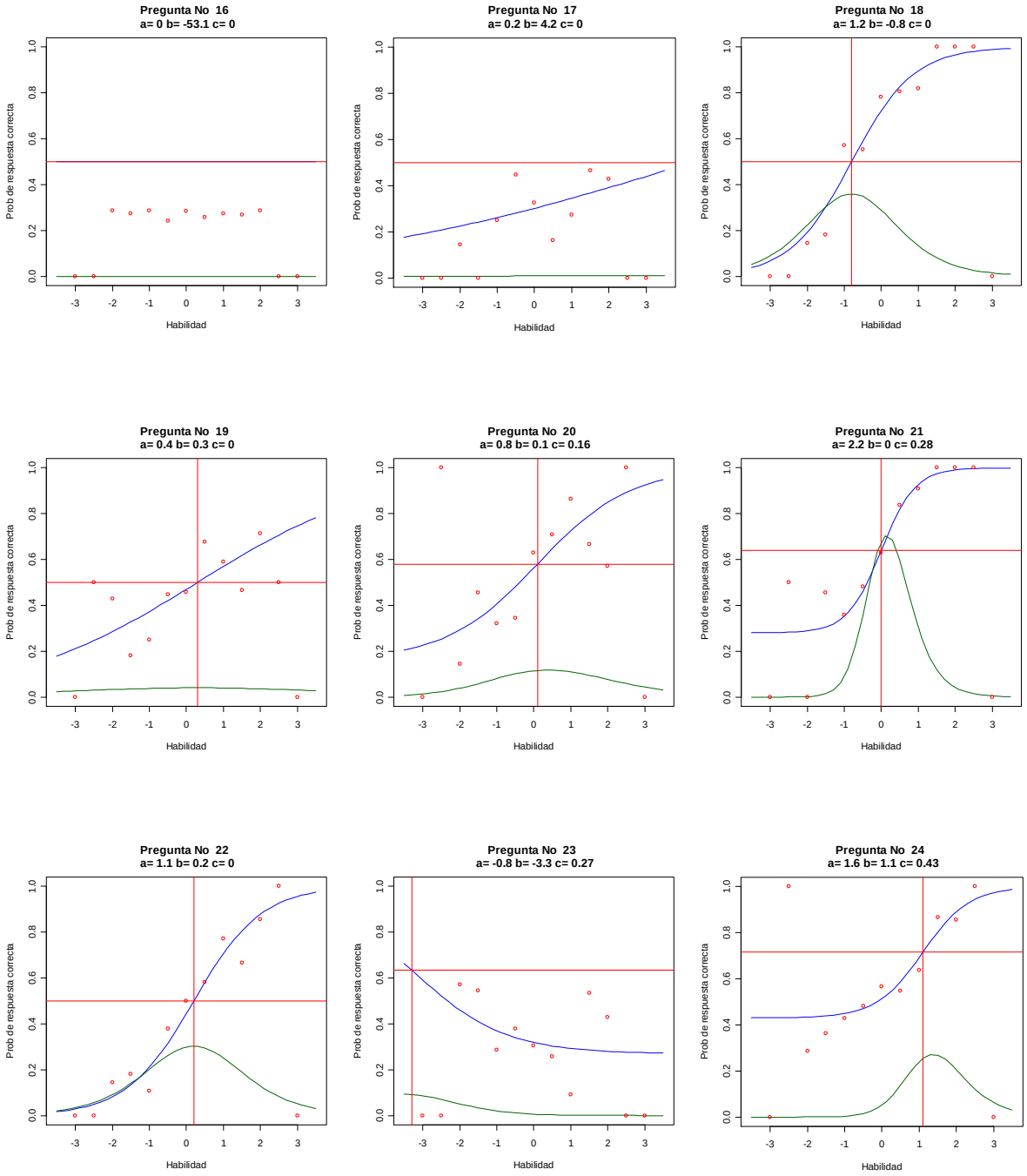

Pregunta No 11
a= 2 b= 1.7 c= 0.14
Prob de respuesta correcta
Habilidad

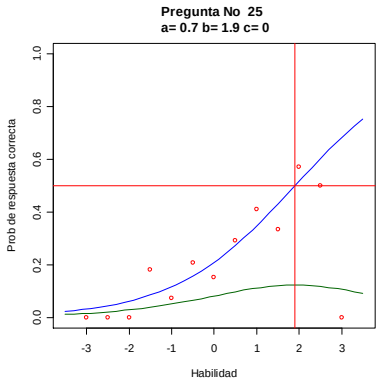

Pregunta No 12
a= 0.6 b= 0.8 c= 0
Prob de respuesta correcta
Habilidad

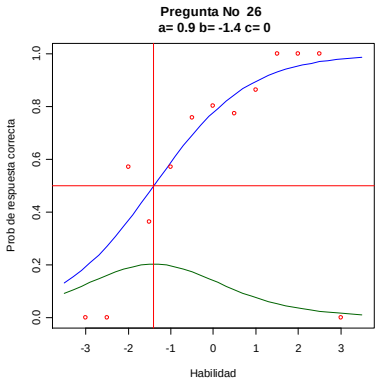

Pregunta No 13
a= 0.1 b= 5.9 c= 0
Prob de respuesta correcta
Habilidad

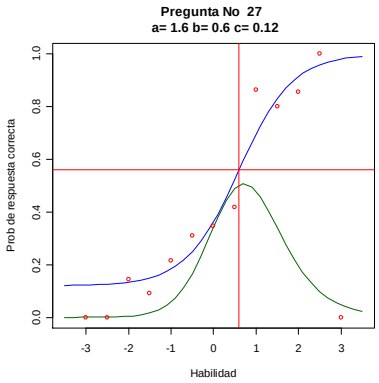

Pregunta No 14
a= 2.2 b= 1.8 c= 0.4
Prob de respuesta correcta
Habilidad

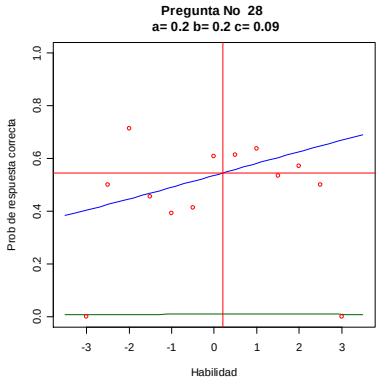

Pregunta No 15
a= 0.3 b= 0.3 c= 0
Prob de respuesta correcta
Habilidad

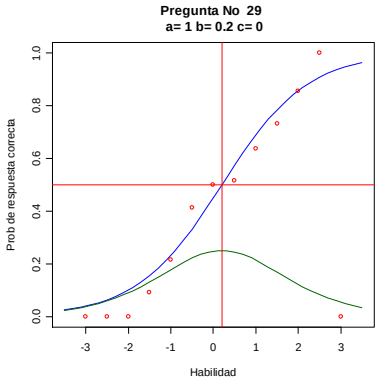

Pregunta No 16
a= 0 b= -53.1 c= 0
Pregunta No 17
a= 0.2 b= 4.2 c= 0
Pregunta No 18
a= 1.2 b= -0.8 c= 0
Pregunta No 19
a= 0.4 b= 0.3 c= 0
Pregunta No 20
a= 0.8 b= 0.1 c= 0.16
Pregunta No 21
a= 2.2 b= 0 c= 0.28
Pregunta No 22
a= 1.1 b= 0.2 c= 0
Pregunta No 23
a= -0.8 b= -3.3 c= 0.27
Pregunta No 24
a= 1.6 b= 1.1 c= 0.43
Prob de respuesta correcta
Habilidad

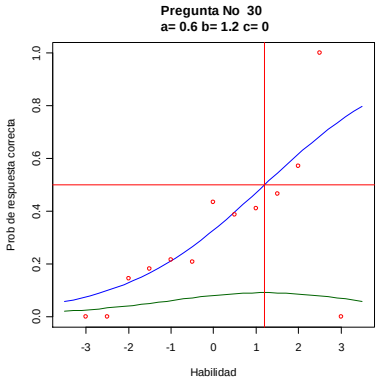

Pregunta No 25
a= 0.7 b= 1.9 c= 0
Prob de respuesta correcta
Habilidad

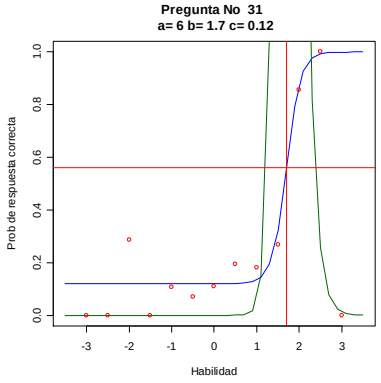

Pregunta No 26
a= 0.9 b= -1.4 c= 0
Prob de respuesta correcta
Habilidad

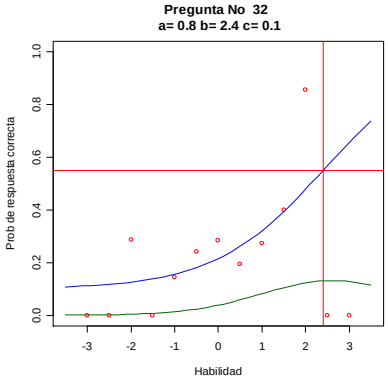

Pregunta No 27
a= 1.6 b= 0.6 c= 0.12
Prob de respuesta correcta
Habilidad

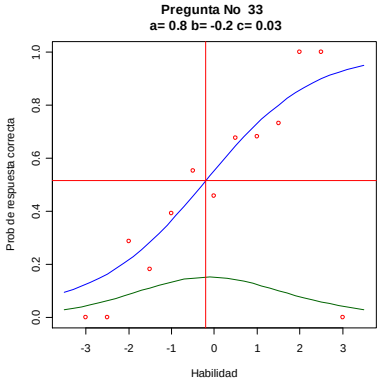

Pregunta No 28
a= 0.2 b= 0.2 c= 0.09
Prob de respuesta correcta
Habilidad


Pregunta No 29
a= 1 b= 0.2 c= 0
Prob de respuesta correcta
Habilidad


Pregunta No 30
a= 0.6 b= 1.2 c= 0
Prob de respuesta correcta
Habilidad


Pregunta No 31
a= 6 b= 1.7 c= 0.12
Prob de respuesta correcta
Habilidad


Pregunta No 32
a= 0.8 b= 2.4 c= 0.1
Prob de respuesta correcta
Habilidad


Pregunta No 33
a= 0.8 b= -0.2 c= 0.03
Prob de respuesta correcta
Habilidad

Anexo F. Validez de las opciones de respuestas.

Pregunta	Gpo.	Opciones				Observaciones
		A	B	C	D	
1	Alto	53	3	3	8	
	Bajo	30	5	10	23	
	Total	83	8	13	31	
2	Alto	31	3	25	8	El distractor C con mayor probabilidad que la respuesta correcta.
	Bajo	6	20	30	12	
	Total	37	23	**55**	20	
3	Alto	33	**21**	1	12	
	Bajo	19	**17**	4	28	
	Total	52	**38**	5	40	
4	Alto	59	0	3	5	
	Bajo	35	8	9	16	
	Total	94	8	12	21	
5	Alto	49	4	0	14	El distractor C sin probabilidad de ser seleccionado.
	Bajo	26	22	2	18	
	Total	75	26	2	32	
6	Alto	44	4	6	12	
	Bajo	16	9	30	13	
	Total	60	13	36	25	
7	Alto	46	4	1	16	
	Bajo	13	8	10	37	
	Total	59	12	11	53	
8	Alto	39	**0**	**20**	8	El distractor B sin probabilidad de ser

Pregunta	Gpo.	Opciones				Observaciones
		A	B	C	D	
	Bajo	14	**0**	**16**	38	seleccionado.
	Total	53	**0**	**36**	46	
9	Alto	50	4	2	11	
	Bajo	36	6	5	21	
	Total	86	10	7	32	
10	Alto	**27**	1	**9**	**30**	El distractor D con mayor probabilidad que la respuesta correcta.
	Bajo	18	4	**5**	41	
	Total	**45**	5	**14**	**71**	
11	Alto	28	15	3	21	
	Bajo	10	22	8	28	
	Total	38	37	11	49	
12	Alto	39	21	7	0	
	Bajo	12	36	10	10	
	Total	51	57	17	10	
13	Alto	26	1	2	**38**	El distractor D con mayor probabilidad que la respuesta correcta.
	Bajo	20	5	5	**38**	
	Total	46	6	7	**76**	
14	Alto	38	15	10	4	
	Bajo	23	19	17	9	
	Total	61	34	27	13	
15	Alto	39	8	**17**	3	El distractor C con mayor probabilidad en el grupo alto que el bajo.
	Bajo	27	18	**6**	16	
	Total	66	26	**23**	19	

Pregunta	Gpo.	Opciones				Observaciones
		A	B	C	D	
16	Alto	20	**37**	8	2	El distractor B con mayor probabilidad que la respuesta correcta y con mayor probabilidad en el grupo alto que el bajo.
	Bajo	15	**21**	15	17	
	Total	35	**58**	23	19	
17	Alto	24	6	**24**	13	El distractor C con mayor probabilidad que la respuesta correcta.
	Bajo	12	12	**34**	10	
	Total	36	18	**58**	23	
18	Alto	57	2	6	2	
	Bajo	36	5	20	7	
	Total	93	7	26	9	
19	Alto	41	16	**3**	7	
	Bajo	21	20	**3**	23	
	Total	62	36	**6**	30	
20	Alto	52	2	**0**	12	El distractor C sin probabilidad de ser seleccionado.
	Bajo	25	21	**1**	21	
	Total	77	23	**1**	33	
21	Alto	64	1	0	2	
	Bajo	24	16	10	17	
	Total	88	17	10	19	
22	Alto	49	1	12	5	
	Bajo	13	10	31	13	
	Total	62	11	43	18	
23	Alto	24	**0**	1	**42**	El distractor B sin probabilidad de ser

Pregunta	Gpo.	Opciones				Observaciones
		A	B	C	D	
	Bajo	25	**1**	4	**38**	seleccionado y el distractor D con mayor probabilidad que la respuesta correcta.
	Total	49	**1**	5	**80**	
24	Alto	50	12	**0**	5	El distractor C sin probabilidad de ser seleccionado.
	Bajo	30	25	**1**	12	
	Total	80	37	**1**	17	
25	Alto	21	**8**	**33**	4	El distractor C con mayor probabilidad que la respuesta correcta.
	Bajo	7	**7**	**37**	17	
	Total	28	**15**	**70**	21	
26	Alto	60	5	**1**	1	Los distractores C y D sin probabilidad de ser seleccionados.
	Bajo	43	25	**0**	0	
	Total	103	30	**1**	**1**	
27	Alto	43	5	6	13	
	Bajo	14	16	14	24	
	Total	57	21	20	37	
28	Alto	38	10	12	7	
	Bajo	36	12	12	8	
	Total	74	22	24	15	
29	Alto	51	1	4	11	
	Bajo	13	17	24	14	
	Total	64	18	28	25	
30	Alto	34	12	8	13	
	Bajo	9	25	17	17	
	Total	43	37	25	30	

Pregunta	Gpo.	Opciones				Observaciones
		A	B	C	D	
31	Alto	19	2	8	**38**	El distractor D con mayor probabilidad que la respuesta correcta.
	Bajo	7	3	17	**40**	
	Total	26	5	25	**78**	
32	Alto	25	6	**29**	**7**	El distractor C con mayor probabilidad que la respuesta correcta y el distractor D con mayor probabilidad en el grupo alto que el bajo.
	Bajo	9	17	**41**	**1**	
	Total	34	23	**70**	**8**	
33	Alto	44	19	2	2	
	Bajo	22	36	3	7	
	Total	66	55	5	9	

Nota: La respuesta correcta es la indicada con la letra A.