# DISEÑO ÓPTIMO DE PRUEBAS

# DISEÑO ÓPTIMO DE PRUEBAS

Benjamin D. Wright
Mark H. Stone
Universidad de Chicago

MESA PRESS
Chicago
1979

MESA PRESS
Avenida Kimbark 5835
Chicago, IL
60637

Copyright © 1979 por Benjamin D. Wright y Mark H. Stone
Todos los derechos reservados

Impreso en los Estados Unidos de América
Biblioteca del Congreso, Número de tarjeta del catálogo 79-88489

Diseño de portada: Andrew W. Wright
Gráficos y composición tipográfica: Betty Stonecipher

Traducción sin fines de lucro por: Agustín Tristán-López y Mayra-Melissa Sánchez-Sifer
Instituto de Evaluación e Ingeniería Avanzada
San Luis Potosí, México. Diciembre de 2018
Autorizada por Institute for Objective Measurement
Con la aprobación de Jack Stenner y J. Michael Linacre

# AGRADECIMIENTOS

Este libro le debe más a Georg Rasch de lo que las palabras pueden expresar. Fueron muchos meses de tutoría acuciosa e inspiradora durante 1960, en los cuales me mostró áreas de oportunidad y necesidades de medición, que fueron las bases para este trabajo producido por Mark Stone y yo. Igual deuda tengo con Jane Loevinger por su sentido común durante el Simposio de 1965 convocado por la Asociación Psicológica del Medio Oeste sobre "Modelos de Probabilidad Independientes de la Muestra en Medición Psicológica", a Darrell Bock por animarme a hablar sobre "calibración de pruebas y medición de personas independientes de la muestra" en la reunión de marzo de 1967 de la Psychometric Society y a Benjamin Bloom por insistir en la "calibración de pruebas y medición de personas independientes de la muestra" durante la Conferencia por Invitación sobre Problemas en Pruebas convocada por el Educational Testing Service (ETS) en 1967.

Benjamin D. Wright

David Farr inició una sesión de capacitación de la American Educational Research Association (AERA) sobre "calibración de pruebas y medición de personas independientes de la muestra en investigación educativa" en 1969. Desde entonces, gracias al apoyo de Richard Jaeger y William Russell, se llevaron a cabo varias sesiones preliminares a los congresos de AERA, verdaderos foros de debate sobre la teoría y la práctica de la medición con el modelo de Rasch, durante las cuales se fomentaron muchas colaboraciones instructivas. Estamos especialmente en deuda con Louis Bashaw, Richard Woodcock, Robert Rentz y Fred Forster, por sus enseñanzas sobre la forma de hacer mediciones útiles con el modelo de Rasch.

El largo y arduo trabajo de análisis matemático, programación informática y verificación empírica que permiten que la medición de Rasch no solo sea práctica sino también fácil de usar fue realizado por amigos del programa MESA, de la Universidad de Chicago. Esta gran aventura comenzó en 1964 con Bruce Choppin y Nargis Panchapakesan y continuó con David Andrich, Graham Douglas, Ronald Mead, Robert Draba, Susan Bell y Geoffrey Masters.
Dedicamos nuestro libro a todos ellos.

Benjamin D. Wright
Mark H. Stone
*Universidad de Chicago*
*30 de marzo de 1979*

# PREÁMBULO

## 0.1 CÓMO USAR ESTE LIBRO

Este es un manual para aprender a hacer medición con el modelo de Rasch. Brindamos algunas explicaciones teóricas pero el énfasis se centra en la práctica. Los Capítulos 2, 4, 5 y 7 usan un pequeño problema para ilustrar con todo detalle la aplicación de la medición de Rasch. A los lectores que aprenden mejor haciendo, les recomendamos ir directamente a los capítulos 2 y 4, para trabajar el problema práctico desarrollado en ellos. Después les convendrá estudiar las secciones de los capítulos 5 y 7 donde continúa el análisis del problema y solo posteriormente regresar al capítulo 1.

Los procedimientos prácticos están debidamente sustentados y respaldado con base en los aspectos metodológicos que motivan y gobiernan la medición con el modelo de Rasch, los cuales se desarrollan en los capítulos 1, 5 y 6. A los lectores a quienes les gusta comenzar estudiando la teoría, les recomendamos leer estos capítulos antes de resolver el problema contenido en los capítulos 2 y 4. Finalmente, si la medición con el modelo de Rasch es completamente nueva para usted, le convendría comenzar con la Sección 0.3 de este preámbulo, donde encontrará la introducción al tema presentado el 28 de octubre de 1967 en la Conferencia Nacional Sobre Problemas en Pruebas convocada por el ETS (Wright, 1968).

La sección 0.2 revisa la motivación e historia de las ideas que culminaron en el libro de Rasch (1960) "Modelos probabilísticos para algunas pruebas de inteligencia y de logro". Las referencias citadas en esa sección y en otras partes de este libro se centran en trabajos que (1) se relacionan directamente con el tema que se analiza, (2) están escritas en inglés y (3) están disponibles en cualquier biblioteca universitaria o pueden ser proporcionadas por nosotros (*N.T. Numerosas referencias pueden ser encontradas en Internet en el sitio del Institute for Objective Measurement y Rasch Measurement Transactions: www.rasch.org*).

## 0.2 MOTIVACIÓN E HISTORIA

Hace cincuenta años, Thorndike se quejó de que las pruebas de inteligencia contemporáneas fallaban al especificar "hasta qué punto es apropiado sumar, restar, multiplicar, dividir y calcular valores de razón obtenidos de las medidas" (Thorndike, 1926, 1). Una buena medida de la habilidad sería aquélla "en la cual el cero represente la ausencia de la habilidad en cuestión y los valores 1, 2, 3, 4 y siguientes representen cantidades que aumenten en un intervalo constante" (Thorndike, 1926, 4). Thorndike tuvo el valor de quejarse porque creyó haber resuelto el problema para la prueba de inteligencia que había desarrollado personalmente. Algo similar planteó en su oportunidad Thurstone (1925).

El método de Thurstone consistió en tomar la proporción de personas de un grupo de edad que respondía correctamente a un ítem dado, transformarla en unidades de desviación de una distribución normal y utilizar los valores como base construir una escala. Los valores comunes para diferentes grupos etarios los obtuvo asumiendo una relación lineal entre los diferentes valores de escala de los ítems compartidos por dos o más formas (o versiones) de una prueba, utilizando las medias de los grupos y sus desviaciones estándar como parámetros para hacer la transformación a una escala común. Thurstone replanteó parcialmente el trabajo de Thorndike para mostrar que su método era mejor (Thurstone, 1927). Su "escala absoluta" (1925, 1927) corresponde aproximadamente con una escala intervalar, pero depende bastante de la distribución de las habilidades de la muestra utilizada. Además del asunto de la homogeneidad, el método de Thurstone involucra la suposición de que la habilidad se distribuye normalmente dentro de los grupos de edad en la cual

existen algunos parámetros poblacionales relevantes que toman valores fijos dentro de dichas distribuciones. Basta con que la especificación de la población sea inapropiada para que igualmente lo sean los valores estimados en la escala, es decir, no son invariantes del muestreo. Por lo tanto, las muestras que difieran en sus distribuciones de las habilidades producirán diferentes valores en la escala, tanto en magnitud como en dispersión.

Thurstone usó la versión de 1925 de su método por el resto de su vida, pero la mayor parte de las calibraciones de pruebas se han basado en técnicas más sencillas de rangos percentilares y en puntajes estándar o tipificados. Lo inadecuado de estos métodos fue expuesto por el análisis de Loevinger de 1947 al construir y evaluar pruebas de habilidad (Loevinger, 1947).

Loevinger demostró que la homogeneidad y la monotonicidad de la escala son criterios esenciales para una correcta medición. Además, "un método aceptable de escalamiento debe dar como resultado una escala derivada independiente de la escala original y del grupo originalmente probado" (Loevinger, 1947, 46). Resumiendo el estado de avance de la calibración de pruebas en 1947, Loevinger dice: "Ningún sistema de escalamiento ha demostrado ser adecuado según los criterios propuestos aquí, a pesar de ser criterios que corresponden con las afirmaciones hechas por el sistema de Thurstone" (Loevinger, 1947, 43). En cuanto a las confiabilidades basadas en correlaciones: "Hasta que logremos encontrar un sistema adecuado de escalamiento, la correlación entre las pruebas de habilidades, incluso entre dos pruebas de la misma habilidad, será accidental y en grado desconocido" (Loevinger, 1947, 46).

En 1950, Gulliksen concluyó *su Teoría de las Pruebas Mentales* con esta observación:

> Se ha realizado relativamente poco trabajo experimental o teórico sobre cambios del grupo medido y su efecto sobre los parámetros de un ítem. Si suponemos que para responder un ítem determinado se requiere una cierta habilidad, la proporción del grupo que responde correctamente aumentará o disminuirá conforme cambie el nivel de habilidad del grupo... Todavía no se cuenta con un tratamiento teórico sistemático de la medida de la dificultad del ítem dirigido a determinar la naturaleza particular de sus variaciones en función de los cambios en la habilidad del grupo. El trabajo experimental tampoco se ha dirigido a analizar los ítems y determinar la invariancia relativa de sus parámetros con los cambios sistemáticos en el nivel de habilidad del grupo medido (Gulliksen, 1950, 392-393).

En la Conferencia por Invitación de 1953 organizada por el ETS sobre problemas de las pruebas, Tucker sugirió que "una prueba ideal puede concebirse como aquélla para la cual la información transmitida por cada uno de los posibles puntajes escalados representan una ubicación en un continuo unitario, de modo que las diferencias uniformes entre puntajes escalados correspondan con diferencias uniformes entre desempeños en la prueba para todos los niveles de puntaje" (Tucker, 1953, 27). También propuso que la comparación de grupos de diferentes habilidades es un método poderoso para evaluar la homogeneidad de la prueba (Tucker, 1953, 25). En ese momento los otros participantes en la conferencia menospreciaron sus propuestas como poco prácticas y utópicas.

En 1960, Angoff escribió en un artículo de su Enciclopedia sobre medición y escalamiento que:

> La mayoría de las escalas de las pruebas que están actualmente en uso derivan sus sistemas de unidades de los datos tomados de administraciones de prueba reales, y por lo tanto dependen del rendimiento de los grupos medidos. Una prueba construida de este modo solo tiene significado si el grupo está bien definido, por ello solo tendrá significado si tiene algún parecido con los siguientes grupos o individuos que responden la misma prueba y para el mismo propósito de selección, diagnóstico o evaluación grupal. Sin embargo, si se encuentra que el muestreo para el desarrollo de la escala de una prueba no es el idóneo o que se tiene

> una utilidad limitada al grupo en el cual se hizo el escalamiento de la prueba, por ejemplo debido a cambios en la población focal o en los propósitos educativos, entonces la escala misma se pone en tela de juicio. Esto es un asunto serio. Una prueba cuya utilidad sea continua debería contar con una escala que no cambie con el tiempo, lo que permitiría dar a conocer el sistema de unidades y propiciar que las personas lleguen a familiarizarse con ella, permitiendo una acumulación de datos útiles para comparaciones históricas (Angoff, 1960, 815).

Sin embargo, esos métodos fallidos, mencionados y criticados por Loevinger, Gulliksen y Angoff, todavía se siguen utilizando ampliamente en la construcción y medición de pruebas. Esto a pesar del hecho de que en los últimos veinticinco años se han acumulado considerables evidencias de que hay mejores métodos, posibles y prácticos.

Estos mejores métodos tienen sus raíces en los modelos psicofísicos del siglo XIX de Weber y Fechner. Se basan en modelos simples acerca de lo que parece razonable que sucede cuando una persona responde a un ítem en una prueba. Tradicionalmente se han utilizado dos distribuciones estadísticas para modelar el comportamiento probabilista de este evento. La distribución normal es la base para la medición de procesos mentales en la Ley de Juicio Comparativo 1920, propuesta por Thurstone. El uso de la ojiva normal como modelo de respuesta al ítem parece haber sido introducido por Lawley y Finney en la década de 1940. Lord convirtió a la ojiva normal en la piedra fundamental de su enfoque para el análisis de ítems hasta aproximadamente 1967, cuando bajo la influencia de Birnbaum se movió hacia un modelo logístico para la respuesta al ítem (Lord, 1968).

La distribución logística fue utilizada por los biometristas para estudiar el crecimiento y las tasas de mortalidad en la década de 1920, siendo desde entonces Berkson el defensor de sus ventajas prácticas sobre la distribución normal. Estas aplicaciones biométricas finalmente fueron recogidas, probablemente por el trabajo de Bradley y Terry en la década de 1950, y formulado en un modelo logístico de respuesta para el análisis de ítems por Birnbaum (1968) y Balcer (1961). Baker desarrolló programas informáticos para aplicar el análisis lógito y probit del ítem y estudió su rendimiento con datos empíricos y simulados (Baker, 1959, 1963).

Sin embargo, en todos estos enfoques de análisis de ítems se buscan al menos dos parámetros para cada ítem. Se han realizado numerosos intentos para estimar la dificultad del ítem como la abscisa (o intercepto horizontal) de la ojiva de respuesta ante una probabilidad de un medio, así como también la discriminación del ítem tomando la pendiente de la ojiva en dicho intercepto. Desafortunadamente esta elaboración aparentemente razonable para resolver el problema introduce una dificultad insuperable para aplicar estas ideas en la práctica. Ha habido un debate continuo durante al menos quince años sobre si hay alguna manera útil que permita estimar otros parámetros como la discriminación del ítem o la "adivinación".

La inevitable solución de este debate ha estado implícita desde la invención de Fisher del estimador suficiente en la década de 1920 y el trabajo de Neymann y Scott en la coherencia de los estimadores condicionales en la década de 1940. Rasch (1968), Andersen (1973, 1977) y Barndorff-Nielsen (1978) demuestran, cada uno de forma decisiva, que la dificultad del ítem se puede estimar de manera consistente y suficiente a partir de los datos disponibles de respuesta correcta/incorrecta. Estas pruebas dejan en claro que los datos de respuesta dicotómica disponible para el análisis de ítems solo puede admitir la estimación de la dificultad del ítem y los intentos de estimar cualquier otro parámetro de un ítem individual están necesariamente condenados al fracaso.

Las matemáticas de estas pruebas no necesitan ser dominadas para convencerse de sus implicaciones prácticas. Cualquiera que realmente examine el funcionamiento interno de los

diversos programas de computadora disponibles para estimar la discriminación de los ítems e intenta aplicarlos a datos reales, encontrarán que las estimaciones resultantes dependen en gran medida de la muestra. Los intentos que se hagan en estos programas de computadora para iterar a una convergencia aparente, solo obtienen una "convergencia" que se "alcanza" al interferir arbitrariamente con la tendencia inevitable de que al menos una de las estimaciones de la discriminación del ítem será divergente al infinito. En la mayoría de los programas este problema insuperable se evita ya sea por no iterar en absoluto o impidiendo que cualquier estimación particular de discriminación exceda algún límite superior completamente arbitrario como por ejemplo 2.0.

Por lo que podemos decir, fue el matemático danés Georg Rasch quien comprendió por primera vez las posibilidades de una medición verdaderamente objetiva que reside en un modelo de respuesta logístico simple. Aparentemente, también fue Rasch quien aplicó por primera vez la función logística al análisis real de datos de pruebas mentales para el propósito práctico de construir pruebas. Rasch comenzó su trabajo de medición psicológica en 1945 cuando estandarizó una prueba de inteligencia grupal para el Departamento de Defensa danés. Fue al llevar a cabo ese análisis de ítems que por primera vez "se dio cuenta del problema de definir la dificultad de un ítem independientemente de la población y la habilidad de un individuo independientemente de cuáles ítems responde" (Rasch, 1960, viii). En 1952 ya había establecido el fundamento básico para una nueva psicometría y resolvió dos modelos de probabilidad para el análisis de pruebas de lectura oral. En 1953 volvió a analizar los datos de una prueba de inteligencia y desarrolló los elementos esenciales de un modelo de probabilidad logística para el análisis de ítems.

Rasch publicó por primera vez su preocupación sobre el problema de las estimaciones dependientes de la muestra en su artículo de 1953 sobre el análisis factorial simultáneo en varias poblaciones (Rasch, 1953). Pero su trabajo sobre análisis de ítems fue desconocido en Estados Unidos hasta la primavera de 1960 cuando visitó Chicago durante tres meses, presentó un trabajo en el Simposio de Berkeley sobre Estadística Matemática (Rasch, 1961), y publicó “Modelos Probabilísticos para algunas Pruebas de Inteligencia y Logro” (Rasch, 1960).

En su revisión de 1965 sobre la persona y la población como conceptos psicométricos, Loevinger escribió:

> Rasch (1960) ha ideado un enfoque verdaderamente nuevo para los problemas psicométricos... No utiliza ninguna de las psicometrías clásicas, sino que renueva la aplicación del álgebra a un modelo probabilístico. Supone que la probabilidad de que una persona responda un ítem correctamente es el producto de un parámetro de habilidad perteneciente solo para la persona y un parámetro de dificultad que pertenece solo al ítem. Más allá de especificar a una persona como el estándar de habilidad o a un ítem como el estándar de dificultad, la habilidad asignada a un individuo es independiente de la de otros miembros del grupo y de los ítems particulares con los que está siendo medido; del mismo modo que ocurre con la dificultad del ítem... De hecho, estas dos propiedades fueron sugeridas alguna vez como criterio para la escala absoluta (Loevinger, 1947); en ese tiempo los esquemas propuestos no parecían satisfacer los criterios para la escala absoluta, así como tampoco lo hacía el escalamiento de Guttman. Por lo tanto, se debe acreditar a Rasch por su contribución sobresaliente a uno de los dos problemas psicométricos centrales: conseguir medidas no arbitrarias. Rasch está preocupado por un tipo de generalización más rigurosa que el propuesto por Cronbach, Rajaratnarn y Gleser. Cuando su modelo ajusta, los resultados son independientes de la muestra de personas y de los ítems particulares dentro de límites bastante amplios, en los cuales la generalidad puede decirse que es completa (Loevinger, 1965, 151).

## 0.3 UNA INTRODUCCIÓN PARA EL PROBLEMA DE LA MEDICIÓN

El tema que me interesa es un problema en la medición. Es un viejo problema en las pruebas educativas. Alfred Binet tenía la misma preocupación hace 60 años. Louis Thurstone estaba preocupado por eso hace más de 40 años. El problema aún no está resuelto. Para algunos puede parecer un pequeño punto. Pero cuando se considera cuidadosamente, creo que se encuentra que este pequeño punto es una cuestión de vida o muerte para la ciencia de la medición mental. ¡La verdad es que las llamadas mediciones que hacemos en pruebas educativas no son muy buenas!

Desde que tuve la edad suficiente para discutir con mis amigos sobre quién tenía el mejor cociente intelectual (digo "mejor" porque algunos pensaron que 100 era perfecto y 60 estaba apenas aprobado), me he quedado perplejo por la medición psicológica. Estábamos confundidos sobre la escala. Las unidades de CI eran diferentes a cualquiera de las medidas de altura, peso y riqueza con las que estábamos aprendiendo a construir ciencias de la vida. Incluso ese logro noble, 100 por ciento, era ambiguo. Cien podría significar la buena noticia de que éramos inteligentes. O podría significar que la prueba fue fácil. A veces rezábamos por pruebas más sencillas para hacernos más inteligentes.

Más tarde aprendí la manera aproximada de usar un puntaje de una prueba. Si estuviera dispuesto a aceptar en su conjunto el conjunto de ítems que componen una prueba estandarizada, podría obtener una medida relativa de habilidad. Si mi desempeño me colocaba en el percentil ochenta entre los hombres de la universidad, sabría dónde estaba parado. ¿Será verdad? El mismo puntaje también me pondría en el percentil ochenta y cinco entre las mujeres universitarias, en el percentil nonagésimo entre los estudiantes de último año de la carrera, y por encima del percentil noventa y nueve entre los estudiantes de secundaria. ¡Mi habilidad dependía no solo de los ítems que tomaba, sino de quién era y los compañeros que tuviera!

La verdad es que un estudio científico de los cambios en la habilidad del desarrollo mental va mucho más allá de nuestras débiles habilidades para hacer mediciones. Podemos obtener respuestas cuantitativas a preguntas como: ¿Cuánto aumenta la comprensión de lectura en los primeros tres años de la escuela? ¿Qué proporción de habilidad es innata y qué otra es aprendida? ¿Qué proporción de habilidad se logra por la madurez asociada a cada año de la infancia?

Espero recordarles algunos problemas que afligen la práctica actual en la medición mental. Las escalas en las que se mide la habilidad son incómodamente resbaladizas. No cuentan con una unidad regular. Su significado y calidad estimada dependen del conjunto específico de ítems realmente estandarizados y la distribución de las habilidades particulares de los niños que se incluyeron en la muestra con la que se hizo la estandarización.

Si un niño ha respondido todo un conjunto específico de ítems que desea medir, entonces puedes obtener su posición percentilar entre los grupos de niños con los que se estandarizó la prueba. Pero, ¿cómo interpretar esta medida más allá de los límites de ese conjunto de ítems y de niños? Al cambiar a los niños se tiene un nuevo criterio. Al cambiar los ítems se tiene también un nuevo criterio de medición. Cada colección de ítems mide una habilidad propia. Cada medida depende de su significado en la cohorte aplicada. ¿Cómo podemos hacer mediciones mentales objetivas y construir una ciencia del desarrollo mental cuando trabajamos con patrones elásticos?

El crecimiento de la ciencia depende del desarrollo de métodos objetivos para transformar observaciones en medida. Las ciencias físicas son un buen ejemplo. Su base es el desarrollo de métodos de medición que son específicos del propósito previsto a medir e independiente de la variación en las otras características de los objetos medidos o los instrumentos de medida usados.

Cuando queremos una medición física, rara vez nos preocupamos por la identidad individual del instrumento de medición. Nosotros mismos no nos preocupamos por otros objetos distintos al que queremos medir y que sean o hayan sido medidos con el mismo instrumento. Es suficiente saber que el instrumento está aceptado dentro de la clase de instrumentos apropiados para el trabajo.

Cuando un hombre dice que está en el percentil nonagésimo en habilidad matemática, necesitamos saber en qué grupo y en qué prueba antes de poder darle un sentido a su afirmación. Pero cuando él dice que mide cinco pies y once pulgadas de alto, ¿pedimos ver su criterio? Sabemos que las cintas de medir pueden diferir en color, temperatura, material, peso e incluso tamaño. Sin embargo, suponemos que comparten una escala de longitud de manera suficientemente independiente de estos caracteres secundarios, lo que permite dar un significado objetivo a una medida de cinco pies once pulgadas. Esperamos que otro hombre de la misma altura medirá aproximadamente cinco pies once pulgada, incluso con otra cinta diferente. Puedo estar en un percentil de habilidad diferente respecto de cada grupo con el que me compare. Pero yo tendré el mismo peso de 175 libras en todos ellos.

Llamaremos "objetiva" a la medición que posee esta propiedad. Dos condiciones son necesarios para lograrla. Primero, la calibración de los instrumentos de medición debe ser independiente de aquellos objetos que se usan para la calibración. En segundo lugar, la medida del objeto debe ser independiente del instrumento que se utiliza para medir. En la práctica, estas condiciones solo se pueden aproximar. Pero su aproximación es lo que hace que la medición sea objetiva.

La calibración del instrumento independiente del objeto y la medición del objeto independiente del instrumento, son las condiciones que hacen posible generalizar la medición más allá de las particularidades del instrumento utilizado, con lo cual se pueden comparar objetos medidos con instrumentos similares – aunque no idénticos – y para combinar o dividir instrumentos que se adapten a nuevos requisitos de medición.

La estrella guía hacia la cual deberían apuntar los modelos de medición mental es a este tipo de objetividad. De lo contrario, ¿cómo podemos lograr una comprensión cuantitativa de las habilidades mentales o cómo podemos tratar de construir una ciencia del desarrollo mental? La calibración de la dificultad del ítem debe ser independiente de las personas particulares utilizadas para la calibración. Así como la medición de la habilidad de la persona debe ser independiente de los ítems de prueba utilizados para una medición particular.

Cuando comparamos un ítem con otro para calibrar una prueba, no debería importar quiénes responden los ítems que usamos para la comparación. Nuestro método de calibración debería darnos los mismos resultados independientemente de a quién administremos la prueba. Esta es la única forma en que podamos llegar a construir pruebas que tengan un significado uniforme independientemente de a quienes elegimos para dichos ítems.

Cuando exponemos a personas a una selección de ítems de prueba para medir su habilidad, no debería importar qué selección de ítems usemos o qué ítems respondan. Deberíamos ser capaces de comparar personas para llegar a medidas estadísticamente equivalentes de habilidad, cualquiera que sea la selección de ítems que se hayan utilizado, incluso cuando han sido medidos con pruebas completamente diferentes.

Es fácil exhortar sobre la objetividad y hacer sarcasmo a expensas de las prácticas actuales. Pero ¿se puede hacer algo sobre este problema? ¿Hay una mejor manera? En el viejo modo de hacer cosas, calibramos un ítem de prueba al observar cuántas personas de una muestra

estándar tienen éxito en ese ítem. El concepto tradicional de "dificultad" es la proporción de respuestas correctas en una muestra de estandarización. La calidad del ítem se juzga a partir de la correlación entre las respuestas en ese ítem y los puntajes en la prueba. La habilidad de la persona es una posición percentilar en la misma muestra "estándar". Obviamente, este enfoque se basa mucho en suposiciones sobre la idoneidad de la muestra estandarizada de personas.

Es posible un enfoque bastante diferente, en el que no se necesitan estas suposiciones sobre la distribución de la habilidad de las personas utilizadas. Este nuevo enfoque asume en cambio un modelo muy simple de lo que ocurre cuando una persona responde a un ítem. El modelo dice simplemente que el resultado del evento está enteramente gobernado por la diferencia entre la habilidad de la persona y la dificultad del ítem. Nada más. Cuanto más capaz sea la persona, mejores serán sus posibilidades de éxito con cualquier ítem. Si un ítem es más fácil, será más probable que cualquier persona lo resuelva. Es tan simple como eso.

Pero este modelo simple tiene consecuencias sorprendentes. Cuando la medición se rige por este modelo, es posible tener en cuenta las habilidades de las personas en la muestra de ítems calibrados y libera su estimación de las dificultades particulares. Los puntajes que obtienen las personas en la prueba se pueden usar para eliminar la influencia de sus habilidades al hacer la estimación de la dificultad del ítem. El resultado es una calibración de ítems libres – o lo que ese denomina independientes – de la muestra.

Lo mismo puede suceder cuando medimos personas. Los puntajes que reciben los ítems en cualquier muestra que se use para calcular sus calibraciones se pueden usar a su vez para eliminar la influencia de la dificultad del ítem a partir de la estimación de la habilidad de la persona. El resultado es un examen libre – o independiente – de las medidas de las personas[1].

[1] Adaptado de las Memorias de la Conferencia por Invitación de 1967 sobre Problemas de las Pruebas. Copyright © 1968 por Educational Testing Service. Todos los derechos reservados. Reimpreso con permiso.

# CONTENIDO

AGRADECIMIENTOS ..... v
PREÁMBULO ..... vii
0.1 Cómo usar este libro ..... vii
0.2 Motivación e historia ..... vii
0.3 Una introducción para el problema de la medición ..... xi
1 EL MODELO DE MEDICIÓN ..... 1
1.1 Cómo se mide por medio de pruebas ..... 1
1.2 Cómo se utilizan los puntajes ..... 4
1.3 Qué sucede cuando una persona contesta un ítem ..... 9
1.4 El modelo de rasch ..... 15
1.5 Uso del modelo de rasch para calibrar y medir ..... 17
1.6 Un procedimiento de estimación simple y práctico ..... 21
1.7 Cómo aparecen los valores estadísticos tradicionales en la medición de Rasch ..... 24
2 CALIBRACIÓN DE ÍTEMS A MANO ..... 28
2.1 Introducción ..... 28
2.2 La prueba de los cubos de KNOX ..... 28
2.3 Datos para el análisis de ítems ..... 29
2.4 Calibrando ítems y midiendo personas ..... 30
2.5 Discusión ..... 44
3. CALIBRACIÓN DE ÍTEMS POR COMPUTADORA ..... 46
3.1 Introducción ..... 46
3.2 Salida de bical del análisis con PROX de los datos de la prueba de los cubos de KNOX. ..... 46
3.3 Comparación de PROX a mano contra PROX por computadora ..... 55
3.4 Análisis de la PCK con el procedimiento UCON ..... 56
3.5 Comparación de UCON contra PROX con los datos de la PCK. ..... 60
3.6 Algoritmo informático para PROX ..... 61
3.7 Procedimiento incondicional UCON ..... 62
4 EL ANÁLISIS DE AJUSTE ..... 66
4.1 Introducción ..... 66
4.2. La matriz de respuesta de la PCK ..... 66
4.3 El análisis de ajuste a mano ..... 69
4.4 Registros desajustados en las personas ..... 76
4.5 Registros de ítems desajustados ..... 77
4.6 Breve resumen del análisis de ajuste ..... 79
4.7 Análisis de ajuste por computadora ..... 80
5 CONSTRUCCIÓN DE UNA VARIABLE ..... 83
5.1 Generalizando la definición de una variable ..... 83
5.2 Definición de la variable de la PCK. ..... 83
5.3 Intensificar y extender la variable de la PCK ..... 87

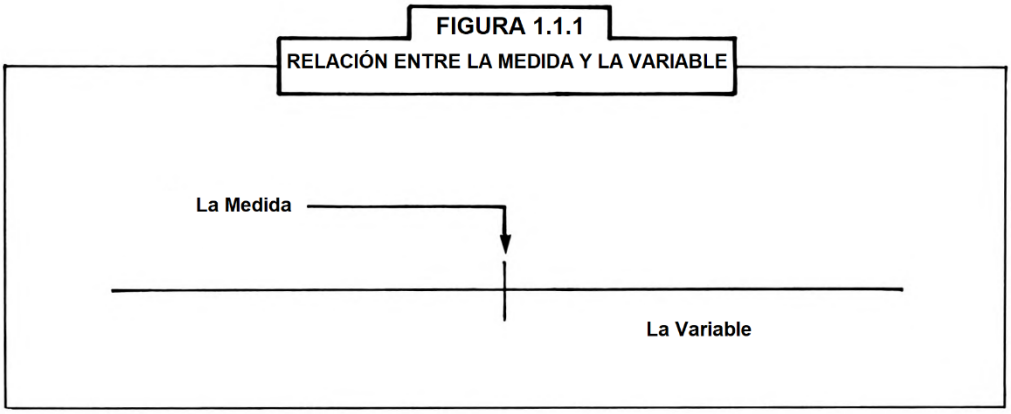
5.4 Líneas de control para gráficas de identidad ... 94
5.5 Conectando dos pruebas ... 96
5.6 Construcción de bancos de ítems ... 98
5.7 Transferencia al banco de datos de la PCKB ... 106
5.8 Equiparación con personas comunes en la PCKB ... 109
5.9 Equiparación con ítems comunes en la PCKB ... 112
5.10 La variable de la PCK como prueba con referencia a criterio ... 118
5.11 Control de calidad de la calibración de los ítems ... 121
5.12 La variable de la PCKcomo prueba con referencia a norma ... 126
6 DISEÑO DE PRUEBAS ... 129
6.1 Introducción ... 129
6.2 El objetivo de la medición ... 129
6.3 Caracteristicas métricas de la prueba ... 131
6.4 La forma de una mejor prueba ... 133
6.5 La precisión de una mejor prueba ... 134
6.6 El coeficiente de error ... 135
6.7 El diseño de una mejor prueba ... 137
6.8 Las reglas completas para el diseño óptimo de pruebas ... 139
7 CONSTRUYENDO MEDIDAS ... 141
7.1 El uso de una variable para construir medidas ... 141
7.2 Conversión de puntajes con medidas por UCON, PROX y UFORM ... 142
7.3 Medidas de mejores pruebas con UFORM ... 145
7.4 Pruebas individualizadas ... 151
7.5 Confección de la prueba por estatus ... 153
7.6 Confección de la prueba por desempeño ... 156
7.7 Prueba personalizada ... 161
7.8 Ajuste de la persona y control de calidad ... 165
7.9 Diagnostico del desajuste ... 170
7.10 Corrigiendo una medida ... 181
8. ELECCIÓN DE UNA ESCALA ... 191
8.1 Introducción ... 191
8.2 Fórmulas para construir nuevas escalas ... 192
8.3 La menor diferencia medible ... 192
8.4 Definición del factor de espaciamiento ... 195
8.5. Unidades de escalamiento normativo: nits ... 198
8.6. Unidades de escalamiento sustantivo: sits ... 199
8.7 Unidades de escalamiento de probabilidad de respuesta: chips ... 201
8.8 Formularios de informes ... 205
APÉNDICES ... 211
BIBLIOGRAFÍA ... 217
ÍNDICE ... 220

# 1 EL MODELO DE MEDICIÓN

## 1.1 CÓMO SE MIDE POR MEDIO DE PRUEBAS

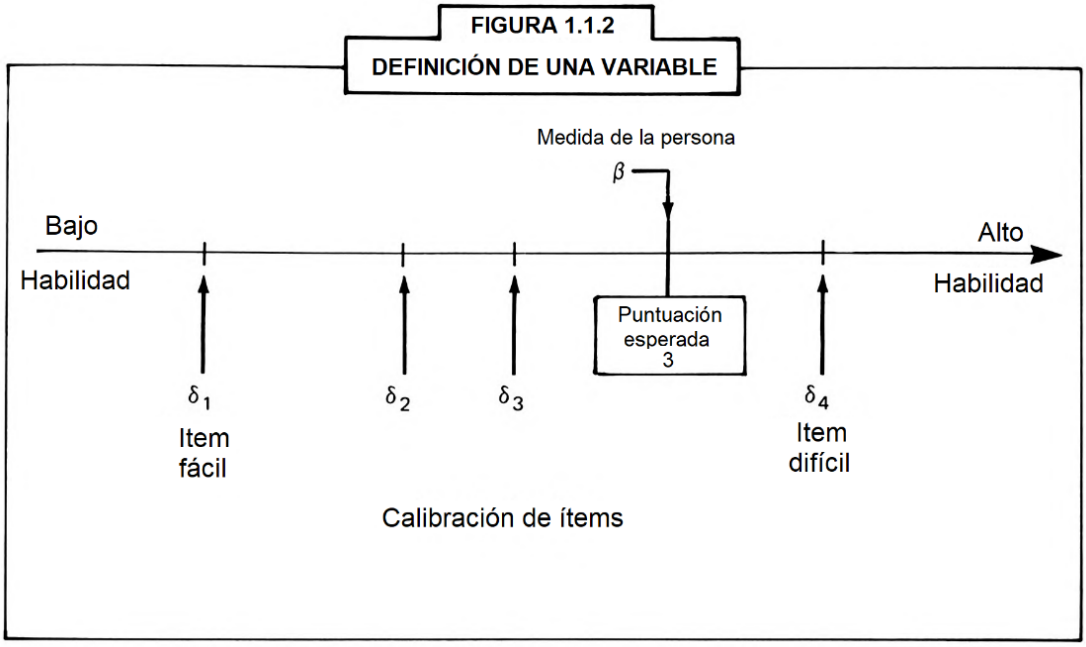
Este libro trata de cómo hacer y usar pruebas mentales. Para hacerlo exitosamente debemos tener un método que transforme las observaciones de desempeño en una prueba a medidas de habilidad mental. La idea de una medida requiere tener una idea sobre la variable en la cual se ubica la medida. Si se visualizan las variables como una recta, la medida se puede dibujar como un punto en dicha línea. La relación entre la medida y la variable se presenta en la Figura 1.1.1.

FIGURA 1.1.1
RELACIÓN ENTRE LA MEDIDA Y LA VARIABLE




Cuando aplicamos una prueba a una persona, el propósito que tenemos es estimar su ubicación en la línea implicada en la prueba. Previamente debimos construir una prueba que defina la línea. Debemos también tener una forma para transformar el desempeño de la persona en una posición en la línea. Este libro presenta cómo usar los ítems de una prueba para definir líneas y cómo usar las respuestas a los ítems para ubicar a las personas en esas líneas.

Para que una prueba defina una variable de una habilidad mental, los ítems con los que se produce la prueba deben compartir una línea de exploración. Esta línea en común en dirección creciente con la habilidad puede ser dibujada como una flecha apuntando a la derecha, hacia la habilidad más alta, dejando la habilidad baja a la izquierda. El significado de la flecha lo proporcionan los ítems que la definen. Si usamos los símbolos $\delta_1, \delta_2 \ldots \delta_i \ldots$para representar los niveles de dificultad de los ítems, entonces cada $\delta_i$ señala la posición del ítem en la línea. Los valores $\delta$ son las calibraciones de los ítems a lo largo de la variable y todos esos ítems calibrados son la definición operacionalizada de lo que mide la variable. Los ítems difíciles que retan a las personas más aptas se dirigen hacia el extremo superior a la derecha de la línea. Los ítems fáciles que pueden ser contestados hasta por las personas menos aptas, generalmente definen el extremo inferior a la izquierda de la línea. La figura 1.1.2 presenta una variable definida por cuatro ítems distribuidos a lo largo de su longitud.




Una variable inicia como una idea general de lo que deseamos medir. Esta idea general se concreta al escribir ítems encargados de extraer signos de la variable esperada en función del comportamiento de las personas. Los ítems de la prueba se vuelven la definición operacional de la variable. Pero una intuición del constructor de la prueba y una construcción cuidadosa de los supuestos ítems de la prueba no son suficientes. Necesitamos contar con evidencias de que la prueba realmente se plasma en los ítems de la prueba. Debemos suministrar los ítems a un conjunto apropiado de personas y analizar sus patrones de respuesta para verificar que los ítems corresponden a un conjunto de respuestas que entre sí definen una variable.

Para localizar a una persona en esta variable debemos aplicarle la prueba con ítems que definen a la variable y después determinar si sus respuestas pueden sumarse para obtener una posición en la línea. Si usamos el símbolo $\beta$ para representar el nivel de habilidad de la persona, entonces $\beta$ señala su ubicación en la línea.

La medida de la persona $\beta$ en la Figura 1.1.2 ubica a esta persona sobre los tres ítems más fáciles y por debajo del más difícil. Si la persona responde una prueba con estos cuatro ítems su puntaje más probable sería de tres y por ello esperaríamos que conteste correctamente los tres ítems más fáciles y que el cuarto ítem lo responda de forma incorrecta. Esta observación es más importante de lo que parece, porque es la base de todos nuestros métodos para estimar las medidas de las personas con base en los aciertos en una prueba. Si queremos saber dónde se localiza una persona en una variable, obtenemos sus respuestas a algunos ítems que definen la variable. El único lugar razonable para estimar su ubicación a partir de esos datos es la región donde sus respuestas pasan de los ítems fáciles correctos a los ítems más difíciles que contesta incorrectamente.

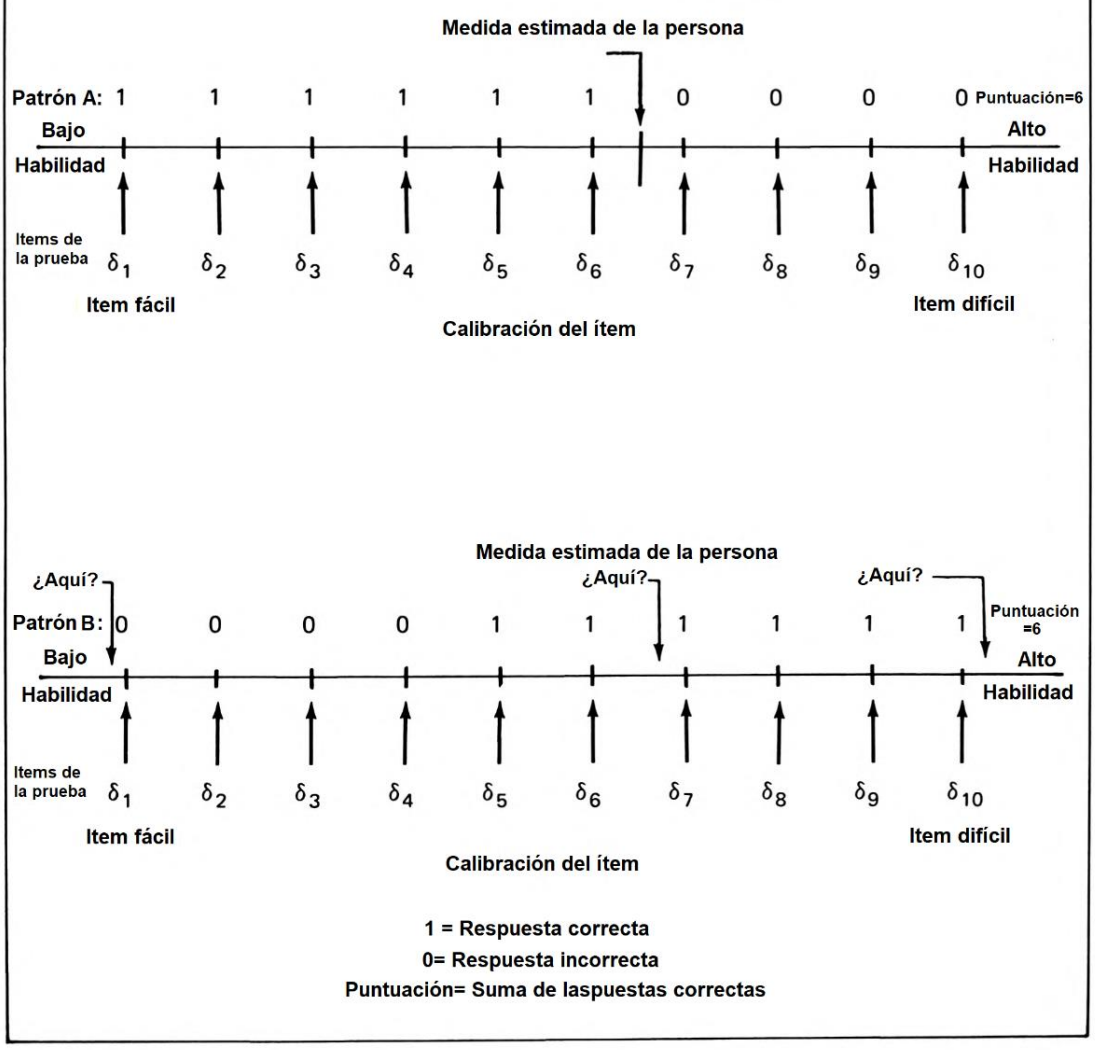
Antes de poder estimar la medida de una persona a partir de su puntaje, debemos examinar su patrón de respuestas. Debemos ver si ese patrón es consistente con la forma en que esperamos que responda a los ítems. Cuando los ítems con los que se examina a una persona han sido calibrados a lo largo de una variable de fácil a difícil, entonces podríamos esperar que el patrón de respuestas de la persona sea más o menos consistente con el orden de dificultad de estos ítems

a lo largo de la variable. Esperamos que la persona tenga éxito en los ítems que deberían ser fáciles para ella y que falle en los ítems que le deberían ser difíciles.

La Figura 1.1.3 muestra dos patrones de respuesta para una misma prueba de diez ítems. Los diez ítems se encuentran ordenados por dificultad a lo largo de la variable. Cada patrón de respuestas se registra arriba de la línea de la variable. Los 1 representan las respuestas correctas, en tanto que los 0 representan las respuestas incorrectas. Ambos patrones producen un puntaje de seis.

**FIGURA 1.1.3**

**VALIDACIÓN DE PATRONES DE RESPUESTA**




En el Patrón A, los seis ítems más fáciles son correctos y los cuatro más difíciles son incorrectos. Parece inconcebible ubicar a esta persona en cualquier lugar, excepto en la región por encima de $\delta_6$, donde está el ítem de dificultad mayor que responde correctamente, pero por debajo de $\delta_7$, que es el menos difícil del resto de ítems que se respondieron incorrectamente.

El patrón B, sin embargo, es muy difícil de conciliar con las implicaciones de un puntaje de seis. ¡Esta persona obtiene los seis ítems más difíciles correctos y los cuatro más fáciles incorrectos! Si tratamos de localizar a esta persona por encima de $\delta_{10}$, que es el más difícil de los ítems que responde correctamente, tendríamos que explicar cómo es que falló en los cuatro ítems más fáciles. ¿Podría alguien ser tan descuidado? Por otro lado, si tratamos de ubicarlos por debajo de $\delta_1$, el ítem más fácil que respondió incorrectamente, entonces, ¿cómo explicamos que los seis ítems más difíciles sean incorrectos? Cualquier otra ubicación a lo largo de la variable, como entre $\delta_6$ y $\delta_7$ para un puntaje de seis, es igualmente insatisfactorio como "medida" para la persona que produjo el Patrón B. Este patrón de respuestas no es coherente con ninguna ubicación en la variable definida por estos ítems. Estamos obligados a concluir que algo anda mal, o los ítems usados están mal calibrados o que esta persona no los respondió de la manera que pretendíamos. En cualquier caso, ninguna medida razonable puede ser derivada del Patrón B.

El ejemplo del Patrón B es importante porque nos muestra que incluso cuando hemos construido ítems que pueden definir una variable válida, también tenemos que validar cada patrón de respuesta de la persona antes de proceder a usar su puntaje como base para estimar su medida. Cuando las calibraciones de los ítems han sido validadas por un número suficiente de personas adecuadas, entonces la mayoría de los patrones de respuesta que encontraremos entre dichas personas se aproximarán al Patrón A. Sin embargo, la posibilidad de que ocurran patrones similares al Patrón B nos obliga a examinar y validar rutinariamente el patrón de respuesta de cada persona examinada, antes de hacer cualquier hipótesis para estimar una medida a partir de su puntaje de prueba.

Para usar una prueba con el propósito de medir a una persona se deben considerar cuatro pasos. Primero, debemos tener una idea clara de la variable sobre la que pretendemos tomar medidas. En segundo lugar, debemos construir ítems que sean enunciados creíbles de esta idea y que puedan conllevarnos a obtener signos de comportamiento de las personas que queremos medir. En tercer lugar, debemos demostrar que los ítems conducen a resultados consistentes con nuestras intenciones, si son aplicados a personas idóneas en la población focal. Finalmente, antes de que usemos el puntaje de la persona como base para su medida, debemos determinar si su patrón particular de respuestas es, de hecho, consistente con nuestras expectativas.

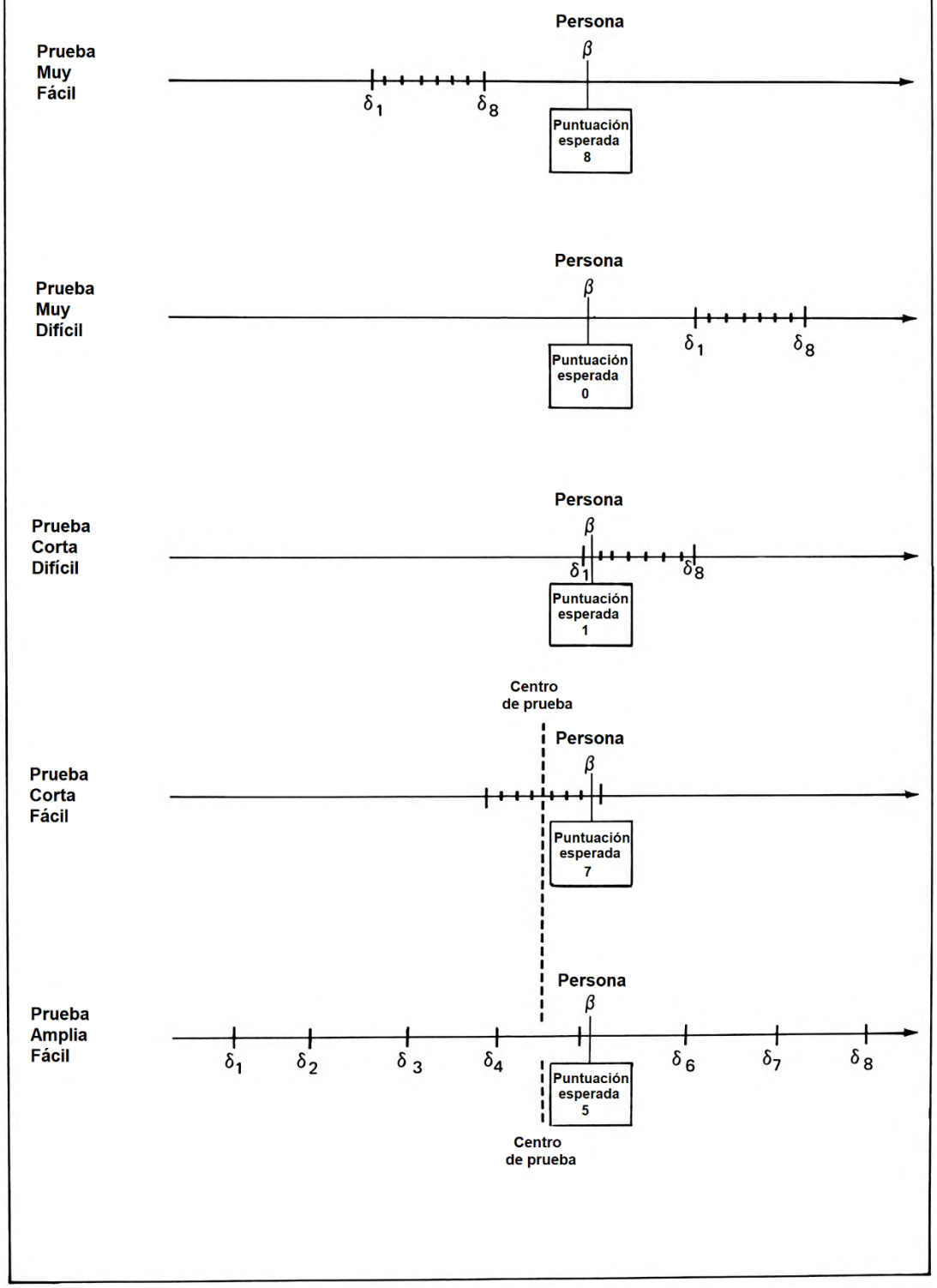
## 1.2 CÓMO SE UTILIZAN LOS PUNTAJES

Un puntaje de prueba tiene como objetivo ubicar a una persona en la variable definida por los ítems de la prueba. Casi todos los que usan puntajes de exámenes suponen que la ubicación de la persona en la variable se determina satisfactoriamente por el puntaje mismo o por alguna función lineal del puntaje, como un porcentaje de aciertos o por un valor de escala basado en la norma. Se da por sentado que el puntaje, o el valor en su escala equivalente, nos dice algo acerca de la persona examinada que va más allá del momento o de los materiales de la prueba. También se da por sentado que los puntajes son adecuados para su uso en operaciones aritméticas necesarias para estudiar el crecimiento y comparar grupos. Pero ¿realmente los puntajes tienen las propiedades necesarias para que sea razonable usarlos de todas estas formas?

Para que un puntaje en particular tenga significado, debe provenir de un patrón de respuesta consistente con los ítems que definen una variable. Pero inclusive la demostración de la validez del ítem, así como la validez de la respuesta no garantiza que el puntaje sea útil. Para poder generalizar hacia la habilidad de la persona y descubrir lo que implica un puntaje, debemos tener en cuenta las particularidades de los ítems usados y ajustar algunos detalles. ¿Cómo es, entonces, que el puntaje de la prueba de una persona depende de las características de los ítems que responde?

**FIGURA 1.2.1**

**COMO LA PUNTUACIÓN DEPENDE DEL NIVEL Y LA DISPERSIÓN DE LA DIFICULTAD DE LOS ITEMS**



La figura 1.2.1 muestra lo que puede suceder cuando una persona en un nivel de habilidad particular responde cinco pruebas diferentes, todas las cuales miden en la misma variable pero que difieren en el nivel y la distribución de sus dificultades de los ítems de fácil a difícil, con un rango estrecho a amplio. Las dificultades de los ocho ítems en cada prueba se marcan en la línea de la variable. Para ver cada prueba por separado, hemos vuelto a dibujar la línea de la variable cinco veces, una para cada prueba.

La habilidad de la persona en la medida también está marcada en cada línea, para que podamos ver cómo se encuentra esta persona con respecto a cada prueba. Si bien cada prueba tiene una posición diferente, porque la variable depende de las dificultades de sus ítems, la posición de esta persona, por supuesto, es la misma en cada línea. La figura 1.2.1 también muestra los puntajes que esperaríamos pudiera tener esta persona ante las cinco pruebas.

La primera línea, "Prueba muy fácil", tiene ítems tan fáciles para esta persona que podemos esperar que tenga un puntaje de ocho. La segunda línea, "Prueba muy difícil", tiene ítems tan difíciles que esperamos un puntaje de cero. La tercera línea, "Prueba corta difícil", tiene siete de sus ítems por encima de la habilidad de la persona y uno por debajo; en este caso el puntaje que esperaríamos ver sería uno. La cuarta Línea, “Prueba corta fácil" tiene siete de sus ítems por debajo de la habilidad de la persona y entonces esperamos un puntaje de siete. Por último, la quinta línea, "Prueba amplia fácil", tiene cinco ítems que deberían ser fáciles para la persona. Esta prueba se centra en la misma posición que la prueba corta fácil, por lo tanto tiene el mismo nivel de dificultad medio, sin embargo, debido a su mayor rango en dificultad de ítem, esperamos solo un puntaje de cinco.

¡Para una persona dada tenemos cinco puntajes esperados: cero, uno, cinco, siete y ocho! Aunque sabemos que la habilidad de la persona no cambia, los cinco puntajes diferentes sugieren cinco habilidades diferentes. Los puntajes de los exámenes obviamente dependen tanto de las características de los ítems de la prueba como de la habilidad de la persona que responde la prueba.

Si el significado de un puntaje de prueba depende de las características de los ítems de prueba, entonces antes de que podamos determinar la habilidad de una persona a partir de su puntaje de prueba, debemos "ajustar" su puntaje para los efectos de los ítems particulares utilizados en la prueba. Este ajuste debe ser capaz de traducir los puntajes obtenidos por las personas en la prueba a medidas de habilidad independientes de la prueba.

Lamentablemente, con puntajes en la prueba de cero, en las que no hay ningún caso de éxito, y de ocho, en que no hay ningún caso de falla, no hay un método satisfactorio para determinar una medida finita para la persona. Todo lo que podemos hacer en esas situaciones es observar que la persona que respondió todo incorrectamente o todo correctamente está sustancialmente por debajo o por encima del nivel operativo de la prueba aplicada. Si deseamos estimar una medida finita para esa persona, entonces tendremos que encontrar una prueba específica que sea más apropiada a su nivel de habilidad.

Podríamos sentirnos tentados a interpretar puntajes perfectos como "dominio completo". Pero a menos que la prueba en cuestión en realidad contenga los ítems más difíciles que alguna vez pudieran llegar a escribirse, siempre existiría la posibilidad de otros ítems que fueran todavía más difíciles para esta variable. Estos ítems más difíciles pueden producir respuestas incorrectas, hasta con nuestra persona que puntúa perfectamente, revelando que el dominio no estaba completo después de todo. Cuando una prueba es extremadamente fácil, por supuesto, todos reconocen que incluso un puntaje perfecto es bastante consistente con la habilidad intermedia.

La dependencia de los puntajes de las pruebas en la dificultad del ítem es un problema con el cual la mayoría de los usuarios de las pruebas están familiarizados. Casi todo el mundo se da cuenta de que el cincuenta por ciento de respuestas correctas en una prueba fácil no significa lo mismo que cincuenta por ciento de respuestas correctas en una prueba difícil. Algunos usuarios de pruebas se dan cuenta de que el setenta y cinco por ciento de lo correcto en una prueba corta no implica tanta habilidad como setenta y cinco por ciento correcto en una prueba amplia. Pero hay otro problema en el uso de puntajes brutos que a menudo se pasa por alto.

Es una práctica común calcular diferencias en los puntajes de las pruebas para medir una tendencia de crecimiento o de mejora, combinando puntajes de una prueba haciendo sumas y restas para comparar grupos, agregar y restar cuadrados, así como productos cruzados de puntajes de prueba para hacer análisis de regresión. Pero cuando estas operaciones aritméticas simples se aplican a los puntajes de las pruebas, los resultados son siempre ligeramente distorsionados y pueden ser sustancialmente engañosos. Aunque los puntajes de las pruebas generalmente permiten estimar bastante bien el orden de las habilidades de las personas, nunca estiman el espaciamiento satisfactoriamente. Los puntajes de las pruebas no son lineales en las medidas que implican y para los cuales son usados.

En el uso estadístico de puntajes de prueba, los efectos de piso y techo son ocasionalmente reconocidos. Pero casi nunca se ajustan estos extremos. Los efectos de frontera causan que cualquier diferencia de puntaje varíe en su significado sobre el rango de puntaje en la prueba. La distancia en la variable que implica una diferencia particular en los puntajes no es la misma en un extremo de la prueba que en otro. Una diferencia de cinco puntos, por ejemplo, implica un cambio más grande en la habilidad al final de una prueba que en el medio.

La figura 1.2.2 ilustra este problema con los puntajes de pruebas. Tenemos dos personas con medidas, $\beta_A$ y $\beta_B$, que están a una distancia fija en la misma variable. Se administran cinco pruebas diferentes a ambas personas, todas las pruebas miden la misma variable. Las ubicaciones de las personas y, por lo tanto, su diferencia medible en la variable, sigue siendo la misma de prueba a prueba, pero sus puntajes más probables varían ampliamente. Esto se debe a que las cinco pruebas difieren en los niveles de dificultad de los ítems, extensión y espaciado. Veamos cómo los puntajes esperados resultantes reflejan la diferencia fija entre estas dos personas.

La Prueba I está compuesta de ocho ítems, todos los cuales se encuentran entre la Persona A y la Persona B. Esperamos que la Persona A no contestará correctamente ninguno de estos ítems llegando a un puntaje de cero mientras que se podría esperar que la Persona B acierte los ocho ítems para un puntaje de ocho. En esta prueba, sus habilidades aparecerán separadas por ocho puntos. Las habilidades están tan espaciadas cómo es posible con esta prueba.

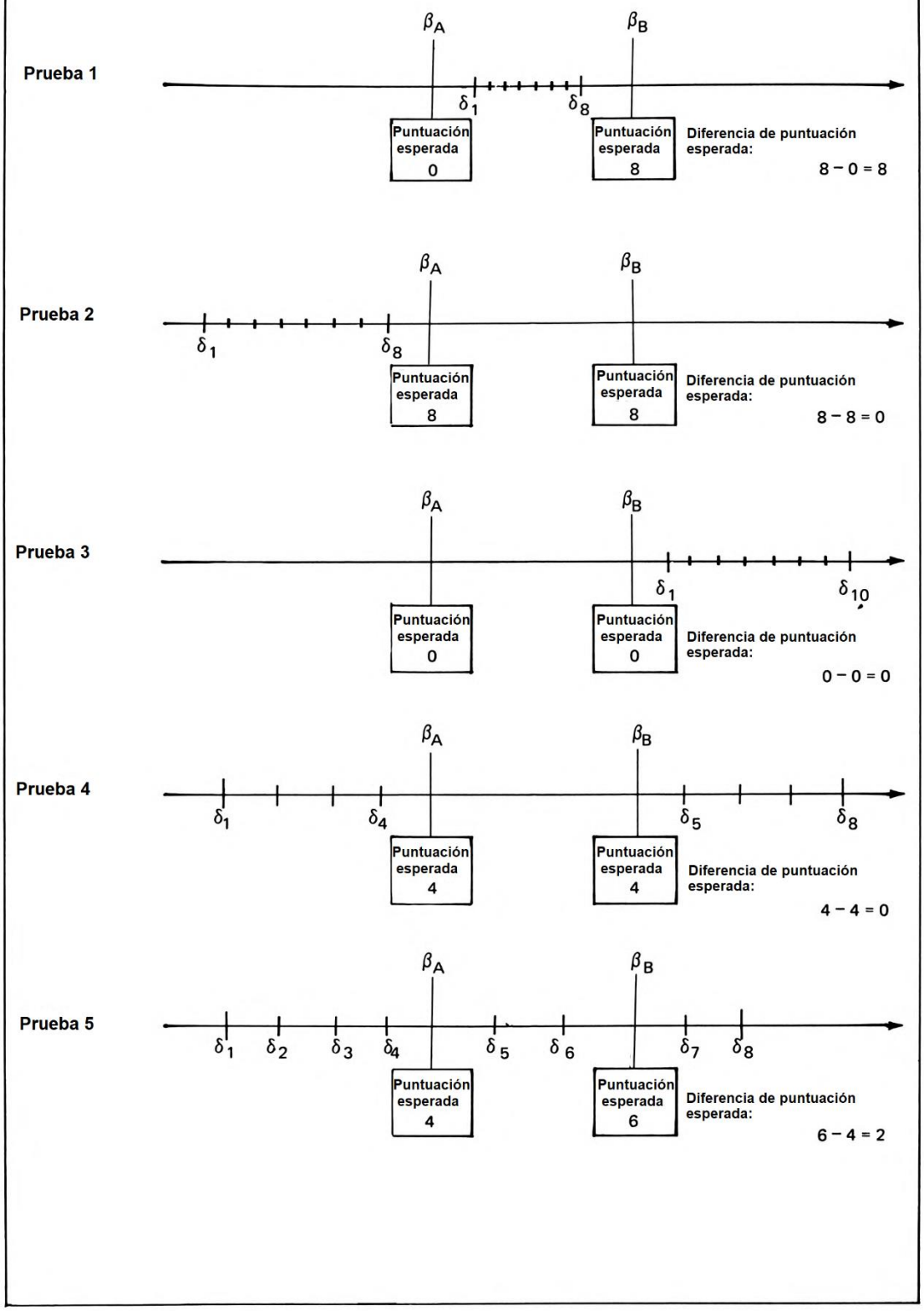
La Prueba II se compone de ocho ítems, todos ellos muy por debajo de las dos personas. Esperamos que ambas personas puedan obtener puntajes de ocho porque esta prueba es demasiado fácil para ambos. Ahora su diferencia esperada en los puntajes de prueba es cero y parecen tener la misma habilidad.

La Prueba III se compone de ocho ítems muy difíciles. Ahora esperamos que ambas personas obtengan puntajes de cero porque esta prueba es muy difícil para ellos. Una vez más, su puntaje esperado tiene una diferencia de cero y sus habilidades parecerán ser las mismas.

La prueba I es más exitosa para separar a las Personas A y B. En cambio las Pruebas II y III fallan en esta cualidad porque están demasiado lejos del objetivo. Tal vez solo sea necesario centrar una prueba correctamente para poder observar la diferencia entre dos personas.

**FIGURA 1.2.2**

**LA NO LINEALIDAD DE LOS PUNTAJES**



La Prueba IV se centra entre la Persona A y la Persona B, pero sus ítems están tan dispersos que hay una gran brecha en su centro en la cual caen la Persona A y la Persona B. Esto da un resultado esperado de cuatro para ambas personas, porque cuatro ítems son demasiado fáciles y cuatro ítems son demasiado difíciles para ambos. Incluso para esta prueba que es más o menos centrada en las dificultades de los ítems, se esperaría una diferencia de puntaje de cero y sus habilidades parecerán ser las mismas.

Para concluir, la Prueba V es amplia y bastante centrada en las personas A y B. Contiene dos ítems que caen entre sus posiciones y por lo tanto los separan. Esperamos que la Persona A responda correctamente los cuatro ítems más fáciles para un puntaje más probable de cuatro. Para la Persona B, sin embargo, esperamos que no solo obtengan los mismos cuatro ítems correctos sino también los siguientes dos más difíciles porque estos dos ítems también están por debajo de su nivel de habilidad. Por lo tanto, en la prueba V la diferencia esperada en los puntajes entre las personas A y B se convierte en dos. En esta prueba, sus habilidades por lo general parecen un poco distintas, aunque no extremadamente diferentes.

¿Qué podemos inferir sobre las diferencias en la habilidad entre las personas A y B de puntajes como estos? Las personas A y B tenderán a parecer igualmente capaces en las Pruebas II, III y IV; serán algo distintas si se utiliza la Prueba V y muy diferentes en la Prueba I. Si las diferencias entre los puntajes de las pruebas de las mismas dos personas se puede hacer que varíen tan ampliamente cambiando las dificultades de los ítems en la prueba, ¿cómo podemos usar las diferencias en puntajes de las pruebas para estudiar las diferencias de habilidad en una variable?

La respuesta es, no podemos. No como están construidos los puntajes. Para poder usar los puntajes de una prueba, que son no lineales en la variable que implican, o para analizar las diferencias, debemos encontrar una forma de transformar los puntajes de las pruebas en medidas que se aproximen a la linealidad.

Los puntajes de los exámenes siempre contienen una distorsión potencialmente engañosa. Si tenemos la intención de usar los resultados de la prueba para estudiar el avance y para comparar grupos, entonces debemos usar un método para producir medidas a partir de los puntajes de las pruebas que marquen las ubicaciones a lo largo de la variable en un intervalo o una variable lineal.

En esta sección, hemos ilustrado dos problemas serios con los puntajes de las pruebas. La primera ilustración mostró cómo los puntajes de las pruebas están limitados por la prueba y cómo tenemos que ajustarlos a las características de sus ítems, antes de que podamos usar los puntajes como base para la medición. La segunda ilustración presentó la manera en que los puntajes de las pruebas no marcan ubicaciones en sus variables de forma lineal y cómo tenemos que transformar los puntajes de las pruebas en medidas que sean lineales antes de poder usarlas para estudiar un avance o crecimiento en una habilidad o para comparar grupos.

## 1.3 QUÉ SUCEDE CUANDO UNA PERSONA CONTESTA UN ÍTEM

Las discusiones en las Secciones 1.1 y 1.2 establecen nuestra necesidad de 1) ítems que podamos demostrar que definen una variable, 2) patrones de respuesta válidos que pueden usarse para ubicar a personas en esta variable, 3) medidas independientes de la prueba que pueden usarse para caracterizar a las personas de manera general y 4) medidas lineales que pueden usarse para estudiar el avance y comparar grupos. Ahora debemos construir un método que responda a estos requisitos.

Las respuestas de personas individuales a ítems individuales son los datos brutos con los cuales debemos comenzar. El método que desarrollamos debe tomar estos datos y a partir de ellos construir las calibraciones y las medidas de las persona con las propiedades que requerimos. La figura 1.3.1 muestra una matriz de datos simple que contiene las respuestas de ocho personas a

una prueba de cinco ítems. Los cinco ítems están identificados en la parte superior de la matriz. Las ocho personas se identifican a la izquierda. La respuesta de cada persona a cada ítem se indica con "1" para una respuesta correcta y "0" para una respuesta incorrecta. Observe que a la derecha de la figura 1.3.1 aparece la suma de las respuestas para el total de ítems como puntaje de cada persona y la suma de las respuestas de todas las personas en cada columna aparecen como puntajes de los ítems.

**FIGURA 1.3.1**
**UNA MATRIZ DE DATOS DE RESPUESTAS OBSERVADAS**

| Nombre de la persona | Nombre del ítem 1 | 2 | 3 | 4 | 5 | Puntaje de la persona |
|---|---|---|---|---|---|---|
| a | 1 | 0 | 0 | 0 | 0 | 1 |
| b | 0 | 1 | 0 | 0 | 0 | 1 |
| c | 1 | 1 | 0 | 0 | 0 | 2 |
| d | 1 | 0 | 1 | 0 | 0 | 2 |
| e | 1 | 1 | 1 | 0 | 0 | 3 |
| f | 1 | 1 | 0 | 1 | 0 | 3 |
| g | 1 | 1 | 1 | 1 | 0 | 4 |
| h | 1 | 1 | 1 | 0 | 1 | 4 |
| | 7 | 6 | 4 | 2 | 1 | 20 |
| | Puntaje del ítem | | | | | |

La figura 1.3.1 muestra cómo son los datos básicos. Pero antes de que podamos poner estos datos para trabajar debemos responder una pregunta fundamental. ¿De dónde creemos que vienen estos datos? ¿Qué se supone que estos puntajes de ítems y personas nos dicen sobre los ítems y las personas? ¿Cómo creemos que se producen estos patrones de 1 y 0? Para descubrir cómo utilizar estos datos debemos establecer un modelo razonable para lo que suponemos que sucede cuando una persona intenta responder un ítem.

Quisiéramos que la habilidad $\beta_v$ de la persona n, lo cual es su ubicación en la variable, gobierne qué tan lejos debemos esperar que se produzcan respuestas correctas a los ítems a lo largo de la variable. De hecho, es la única situación en la que podemos usar las dificultades de los ítems y las respuestas a ellos como base para medir a la persona.

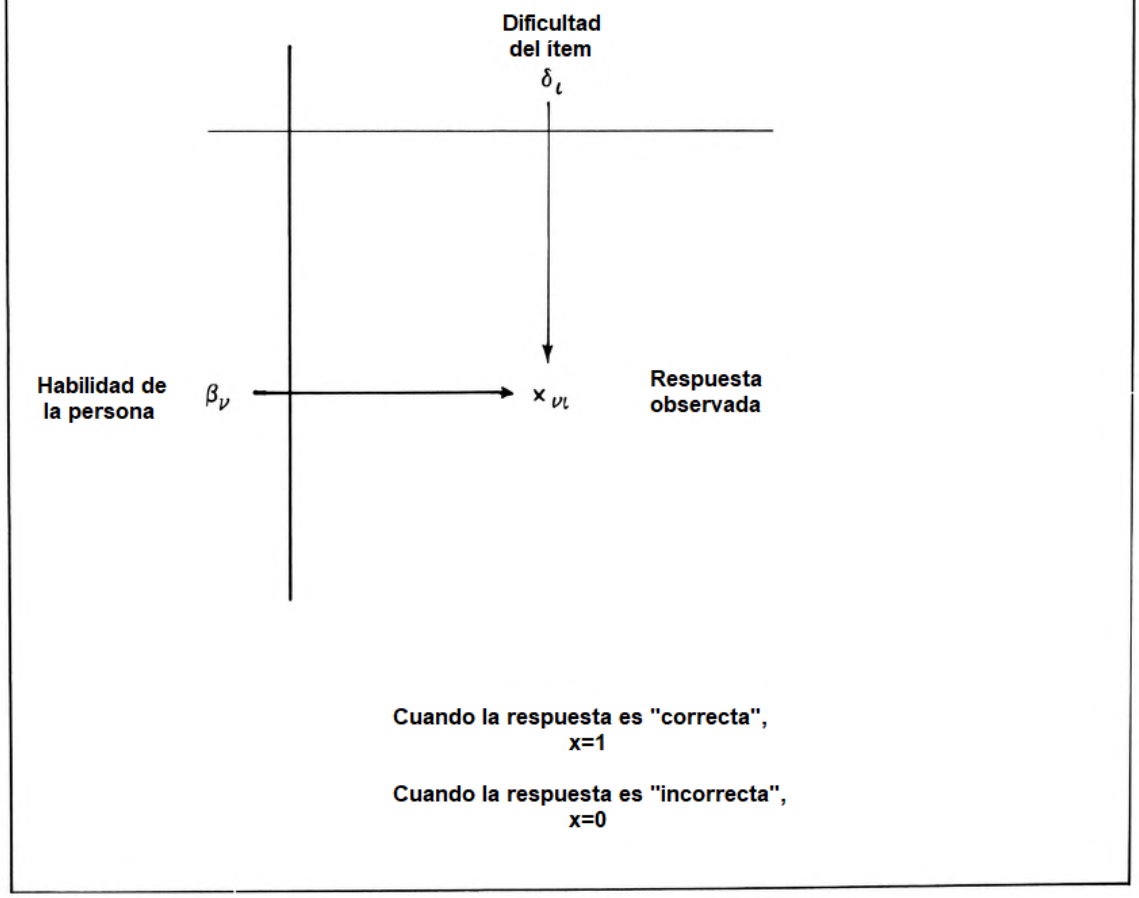
Por supuesto, podemos pensar en otros factores que pueden afectar las respuestas de una persona. Cuando se trata de ítems de opción múltiple, puede ocurrir alguna adivinación y las personas difieren en qué tanto están dispuestas a arriesgar en su adivinación. Es interminable la cantidad de acontecimientos perturbadores que interfieren con la posibilidad de disponer de

una observación clara e inequívoca. Pero, si los que esperamos medir realmente es la habilidad de la persona, entonces parece poco razonable dejar de hacer nuestro mejor esfuerzo para organizar las cosas de modo que sean las habilidades las que dominen el funcionamiento de la prueba. De hecho, para eso es la práctica administrativa que se realiza en las pruebas, por ejemplo, para controlar y minimizar la intrusión de cualquier influencia que interfiera con ella.

También nos gustaría conocer la ubicación en la variable del ítem ι, por medio de su dificultad $\delta_\iota$, para determinar qué tan lejos puede llegar un ítem a lo largo de la variable. Al igual que con las personas, podemos pensar en las características del ítem, como la discriminación y la vulnerabilidad a la adivinación, lo que podría modificar las respuestas de las personas. Algunos psicómetras intentan estimar estas características adicionales del ítem aunque hay buenas razones para esperar que dichas estimaciones están condenadas al fracaso. Pero, de nuevo, parecería poco razonable no hacer nuestro mejor esfuerzo para organizar las cosas de modo que sea la dificultad del ítem la que decida la forma en que personas de diversas habilidades responden a ese ítem. En todo caso, el hecho es que si utilizamos puntajes no ponderados como resultados de nuestra prueba estamos asumiendo que, para fines prácticos, son las dificultades de los ítems y las habilidades de las personas, las que dominan las respuestas de cada persona.

**FIGURA 1.3.2**

**LAS CONDICIONES ESENCIALES QUE CAUSAN UNA RESPUESTA**



Estas consideraciones nos llevan a establecer un modelo de respuesta que es la representación más simple posible. La Figura 1.3.2 coloca a la persona n de habilidad $\beta_v$ respondiendo al ítem ι de dificultad $\delta_i$, para producir la respuesta $x_{vi}$. Estos son los elementos esenciales que tomaremos en cuenta cuando tratemos de explicar los datos en la Figura 1.3.1. La figura 1.3.2 propone que la respuesta $x_{vi}$ ocurre cuando la persona $v$ responde al ítem ι que se puede considerar gobernada por la habilidad de la persona $\beta_v$, y la dificultad del ítem $\delta_i$ y nada más.

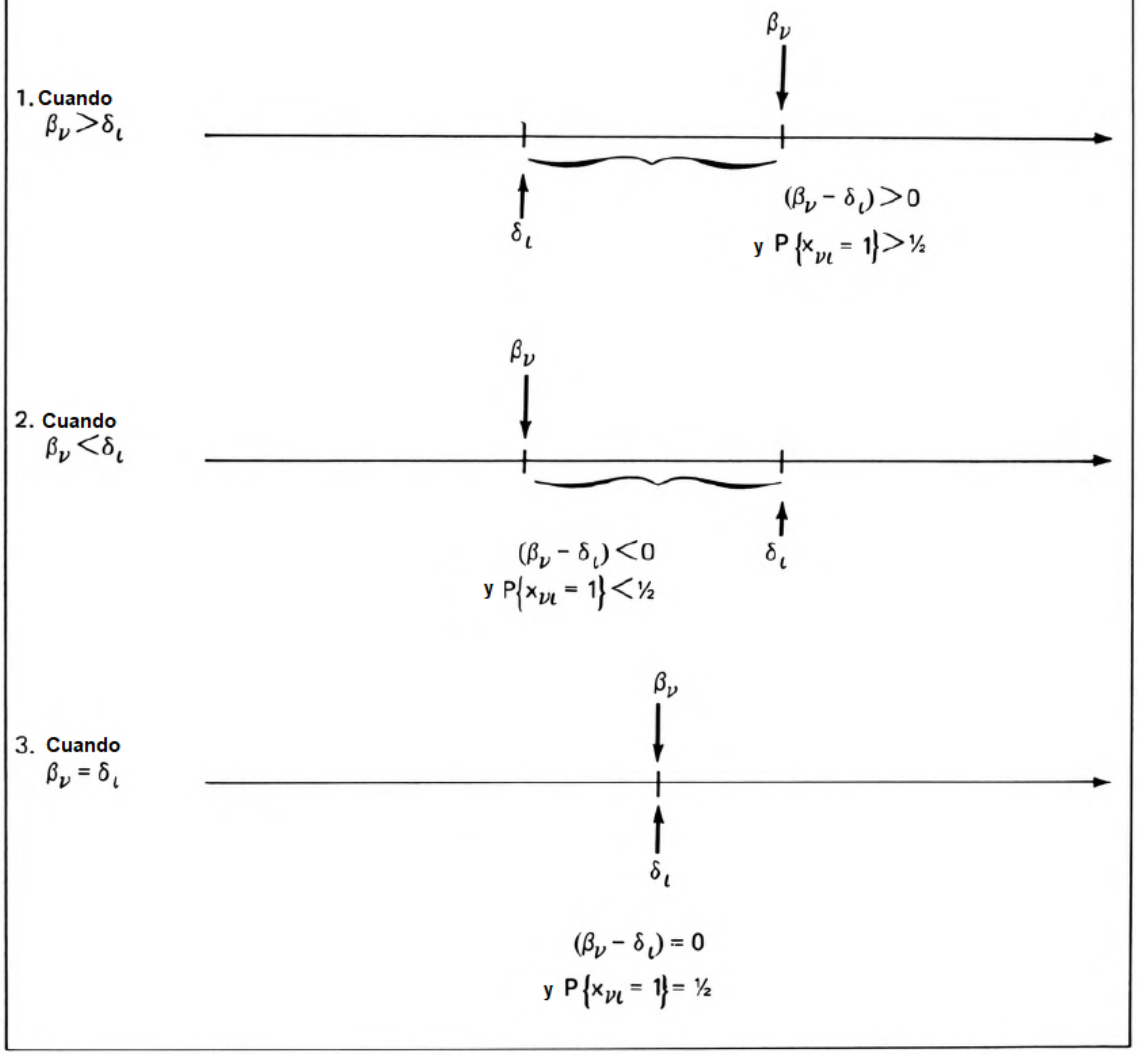
Nuestro siguiente paso es decidir cómo queremos que la habilidad de la persona $\beta_v$, y la dificultad del ítem $\delta_i$ interactúen para producir $x_{vi}$. ¿Cuál es una manera razonable y útil de establecer una relación matemática entre $\beta_v$ y $\delta_i$? Atendiendo a que requerimos que $\beta_v$ y $\delta_i$ representen ubicaciones a lo largo de una variable común que ambos comparten, la formulación más conveniente y natural de esta relación será su diferencia $(\beta_v - \delta_i)$.

Sin embargo, identificar la diferencia $(\beta_v - \delta_i)$ no termina nuestro trabajo, porque también debemos decidir cómo queremos que esta diferencia regule el valor de la respuesta $x_{vi}$. Aún en el caso de que una persona sea más capaz que la dificultad del ítem, es decir $\beta_v$ es mayor que $\delta_i$, ocasionalmente podría ocurrir que esta persona no dé una respuesta correcta a ese ítem relativamente fácil con lo que el valor resultante de $x_{vi}$ sería "0". También puede ocurrir ocasionalmente que una persona de habilidad moderada tenga éxito en un ítem muy difícil. Obviamente, se vuelve incómodo forzar una relación determinista de la forma en que $(\beta_v - \delta_i)$ rige el valor de la respuesta $x_{vi}$. Una mejor manera de lidiar con este problema es reconocer que la forma en que la diferencia $(\beta_v - \delta_i)$ influye en la respuesta $x_{vi}$ solo puede ser probabilística y configurar en consecuencia nuestro modelo de respuesta.

La figura 1.3.3 muestra cómo sería más razonable asociar la probabilidad de una respuesta correcta con la diferencia $(\beta_v - \delta_i)$. Cuando $\beta_v$ es más grande que $\delta_i$, el nivel de habilidad de la persona $v$ es mayor que el nivel de dificultad del ítem ι y su diferencia $(\beta_v - \delta_i)$ es mayor que cero, entonces queremos que la probabilidad de una respuesta correcta sea mayor que un medio (es decir >½). Cuando, por otro lado, el nivel de habilidad de la persona ν es menor que el nivel de dificultad del ítem ι, entonces su diferencia $(\beta_v - \delta_i)$, es menor que cero, con lo cual querríamos que la probabilidad de una respuesta correcta sea menos de un medio (es decir, <½). Finalmente, cuando los niveles de la habilidad de la persona y la dificultad del ítem son los mismos, de modo que su diferencia $(\beta_v - \delta_i)$ es cero, entonces la única probabilidad que parece razonable asignar a una respuesta correcta (o incorrecta) es exactamente un medio (½).

La curva de la figura 1.3.4 resume las implicaciones de la figura 1.3.3 para todas las relaciones entre las probabilidades de respuestas correctas y las diferencias entre la habilidad de la persona y la dificultad del ítem. Esta curva especifica las condiciones que debe cumplir nuestro modelo de respuesta. Las diferencias $(\beta_v - \delta_i)$ podrían surgir de dos maneras. Podrían ser el resultado de una variedad de habilidades de la persona que reaccionan a un solo ítem o pueden surgir de una variedad de dificultades de ítems que miden la habilidad de una persona. Cuando la curva que describe un ítem se dibuja tomando como variable la habilidad $\beta$, se llama "curva característica del ítem o CCI" (ICC por sus siglas en inglés) porque muestra la forma en que el ítem incide en las respuestas de las personas para cada nivel de habilidad. Cuando la curva se dibuja con la dificultad $\delta$ como su variable para que describa cómo una persona responde a una variedad de ítems, podemos llamarlo "curva característica de la persona o CCP" (PCC por sus siglas en inglés).

**FIGURA 1.3.3**
**COMO LAS DIFERENCIAS ENTRE LA HABILIDAD DE LA PERSONA Y LA DIFICULTAD DEL ITEM DEBEN AFECTAR LA CONFIABILIDAD DE UNA RESPUESTA CORRECTA**




La curva de la figura 1.3.4 es una imagen del modelo de respuesta que necesitamos para resolver el problema de cómo los parámetros, $\beta_v$ y $\delta_i$ que queremos estimar, dependen de los datos $x_{vi}$ que podemos observar. Para medir a una persona, debemos estimar $\beta_v$, y para calibrar un ítem debemos estimar $\delta_i$. Para estimar cualquiera de estos parámetros de las respuestas observadas de las personas a los ítems debemos construir una fórmula matemática que sea fiel a la relación dibujada en la figura 1.3.4 y que relacione $\beta_v$, $\delta_i$ y $x_{vi}$ de una manera útil. Esta formulación también debe ser capaz de mostrarnos cómo usar datos parecidos a los de la Figura 1.3.1, con los cuales hacer estimaciones de la habilidad de la persona que sean independientes de la prueba y estimaciones de la dificultad del ítem que sean independientes de la muestra.

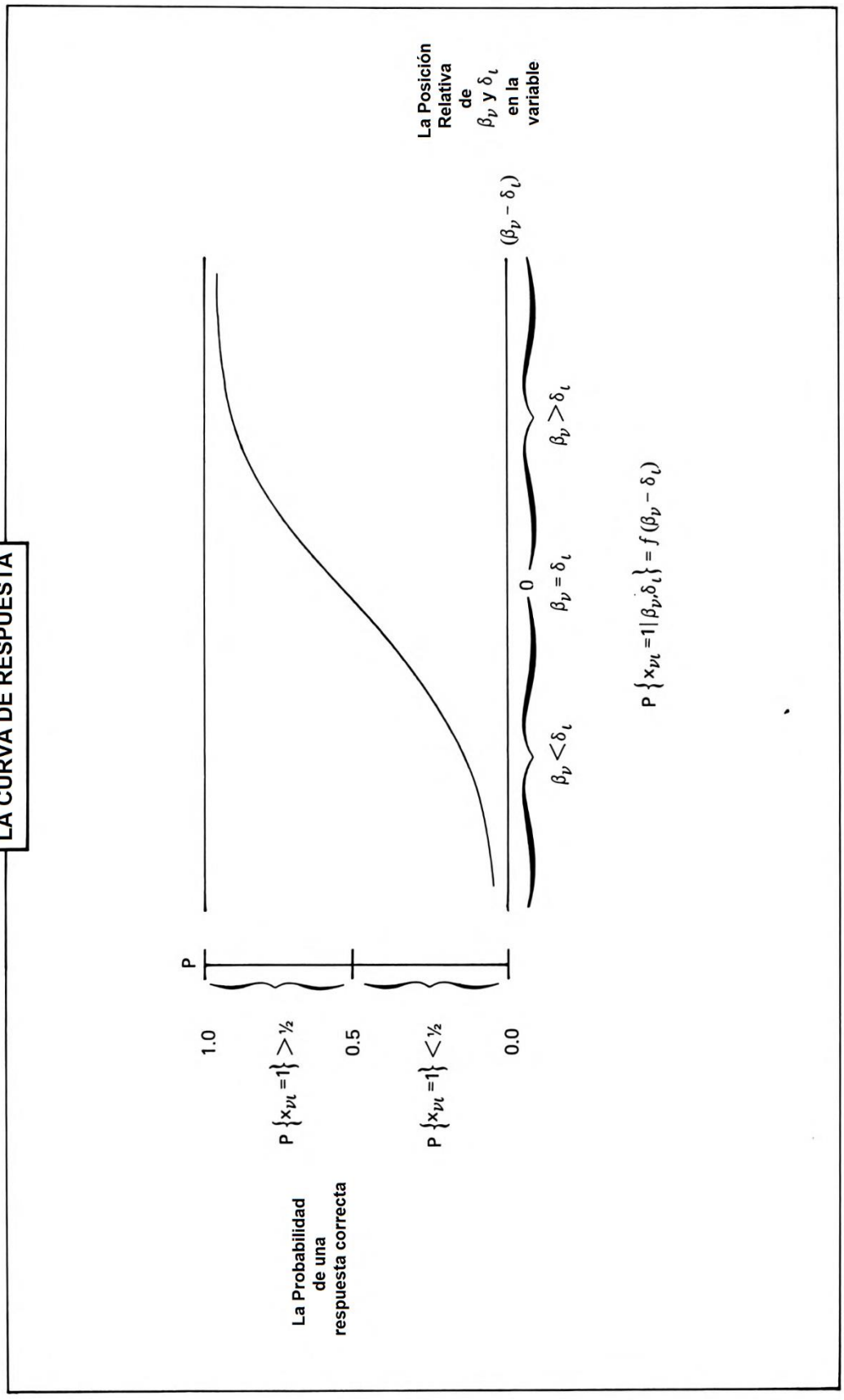


FIGURA 1.3.4
LA CURVA DE RESPUESTA
La Probabilidad de una respuesta correcta
P
1.0
0.5
0.0
P {x_νι =1} > ½
P {x_νι =1} < ½
La Posición Relativa de β_ν y δ_ι en la variable
(β_ν - δ_ι)
0
β_ν < δ_ι
β_ν = δ_ι
β_ν > δ_ι
P {x_νι =1 | β_ν, δ_ι} = f (β_ν - δ_ι)

## 1.4 EL MODELO DE RASCH

Para construir una forma matemática factible para la curva en la figura 1.3.4 comenzamos combinando los parámetros, $\beta_v$ para la habilidad de la persona y $\delta_i$ para la dificultad del ítem, a través de su diferencia ($\beta_v - \delta_i$). Queremos que esta diferencia regule la probabilidad de lo que se supone debe suceder cuando la persona $v$ usa su habilidad $\beta_v$ contra la dificultad $\delta_i$ del ítem ι. Pero la diferencia ($\beta_v - \delta_i$) puede variar de menos infinito a más infinito mientras que la probabilidad de una respuesta exitosa debe permanecer entre cero y uno. Para abordar esta situación utilizamos la diferencia ($\beta_v - \delta_i$) como un exponente del número e = 2.71828…, que es la base de los logaritmos naturales y escribimos el resultado como

$$e(\beta_v - \delta_\iota) = \exp(\beta_v - \delta_\iota)$$

Esta expresión exponencial varía entre cero y más infinito y podemos acotarlo en el intervalo entre cero y uno por medio de la relación

$$\exp(\beta_v - \delta_\iota)/[1 + \exp(\beta_v - \delta_\iota)].$$

Esta formulación corresponde con una curva que sigue bastante bien la ojiva de la figura 1.3.4. Puede ser utilizada para especificar la probabilidad de una respuesta exitosa como

$$P\{x_{v\iota} = 1|\beta_v,\delta_\iota\} = \exp(\beta_v - \delta_\iota)/[1 + \exp(\beta_v - \delta_\iota)] \qquad \textbf{[1.4.1]}$$

que es el modelo de Rasch.

Cualquier otra expresión matemática que describa una ojiva de la forma de la Figura 1.3.4 puede ser una solución al problema de linealidad, mediante la transformación de puntajes que están restringidos entre 0 y 100 por ciento, a "medidas" que van desde menos infinito a más infinito.

Cualquier forma matemática que relacione la probabilidad de $x_{v\iota}$ con la diferencia entre $\beta_v$ y $\delta_\iota$ y que tenga parámetros que puedan estimarse nos podría permitir estudiar la validez del ítem y su respuesta. Todo lo que tenemos que hacer es especificar un modelo viable de cómo ($\beta_v - \delta_\iota$) gobierna la probabilidad de $x_{v\iota}$, usar este modelo para estimar $\beta_v$ y $\delta_\iota$ a partir de algunos datos y examinar la forma en que estos datos se ajustan a las predicciones calculadas a partir del modelo.

Sin embargo, no cualquier ojiva ni cualquier formulación resuelven el problema. De hecho, solo la formulación de la Ecuación 1.4.1, que es el modelo de Rasch, nos permite estimar $\beta_v$ y $\delta_\iota$ de forma independiente el uno del otro, para conseguir que las estimaciones $\beta_v$ se liberen de los efectos de $\delta_\iota$ y que las estimaciones de $d_\iota$ se liberen de los efectos de $\beta_v$.

La función logística de la Ecuación 1.4.1 proporciona un modelo de respuesta simple y útil que hace posible tanto la linealidad de la escala como la generalidad de la medida. Aunque los biometristas han utilizado la función logística desde 1920, fue el matemático danés Georg Rasch (1960) quien apreció por primera vez su significado psicométrico. Rasch denomina "objetividad específica" a la característica especial de esta función logística simple que hace posible la generalidad en la medición. Tanto él como otros autores han demostrado que no hay otra alternativa matemática a la formulación de la ojiva de la figura 1.3.4 que permite la estimación de la medida de la persona $\beta_v$ y la calibración del ítem $\delta_\iota$, de forma independiente una de la otra (Rasch, 1961, 1967; Andersen, 1973, 1977; Barndorff-Nielsen, 1978), en particular cuando se derivan estimadores $\beta_v$ y $\delta_\iota$ por la maximización de una función condicional de verosimilitud, son

imparciales, consistentes, eficientes y suficientes (Andersen, 1970, 1971, 1972a, 1973, 1977, Haberman, 1977). Otras aproximaciones simples para estos estimadores condicionales de máxima verosimilitud que son lo suficientemente precisas para casi todos los propósitos prácticos se describen en Wright y Panchapakesan (1969), Wright y Douglas (1975a, 1975b, 1977a, 1977b) y Wright y Mead (1976). Estos procedimientos han sido útiles en una amplia variedad de aplicaciones (Connolly, Nachtman y Pritchett, 1971; Woodcock, 1974; Willmott y Fowles, 1974; Rentz y Bashaw, 1975, 1977; Andrich, 1975; Mead, 1975; Wright y Mead, 1977; Cornish y Wines, 1977; Draba, 1978; Elliott, Murray y Pearson, 1977.

**TABLA 1.4.1**

**HABILIDAD DE LA PERSONA Y DIFICULTAD DEL ÍTEM EN LÓGITOS Y LA PROBABILIDAD DE RASCH DE UNA RESPUESTA CORRECTA**

| Habilidad de la persona | Dificultad del ítem | Diferencia | Respuesta Correcta | | Información en una respuesta |
|---|---|---|---|---|---|
| | | | Posibilidades | Probabilidad | |
| $\beta_v$ | $\delta_i$ | $(\beta_v - \delta_i)$ | $\exp(\beta_v - \delta_i)$ | $\pi_{vi}$ | $I_{vi}$ |
| 5 | 0 | 5 | 148. | .99 | .01 |
| 4 | 0 | 4 | 54.6 | .98 | .02 |
| 3 | 0 | 3 | 20.1 | .95 | .05 |
| 2 | 0 | 2 | 7.39 | .88 | .11 |
| 1 | 0 | 1 | 2.72 | .73 | .20 |
| 0 | 0 | 0 | 1.00 | .50 | .25 |
| 0 | 1 | − 1 | 0.368 | .27 | .20 |
| 0 | 2 | − 2 | 0.135 | .12 | .11 |
| 0 | 3 | − 3 | 0.050 | .05 | .05 |
| 0 | 4 | − 4 | 0.018 | .02 | .02 |
| 0 | 5 | − 5 | 0.007 | .01 | .01 |

$$\pi_{vi} = \exp(\beta_v - \delta_i) / [1 + \exp(\beta_v - \delta_i)]$$

$$I_{vi} = \pi_{vi}(1 - \pi_{vi})$$

Podemos ver en la Ecuación 1.4.1 que cuando la persona n es más lista que la dificultad del ítem ι, entonces $\beta_v$ es mayor que $\delta_i$, por lo que su diferencia es positiva y la probabilidad de éxito en el ítem ι es mayor que un medio. Cuanto más supere la habilidad de la persona la dificultad del ítem mayor será la diferencia positiva y estará más cerca de 1 que es la probabilidad de éxito. Pero cuando el ítem es demasiado difícil para la persona, entonces $\beta_v$ es menor que $\delta_i$, su diferencia será negativa y la probabilidad de éxito de la persona menor que un medio. Cuando un ítem exija más de una persona, mayor será la diferencia negativa y la probabilidad de tener éxito se reducirá a cero.

Las unidades matemáticas para $\beta_v$ y $\delta_i$ definidas por este modelo se llaman "lógitos". La habilidad de la persona en lógitos es el logaritmo natural del momio de tener éxito en ítems del tipo elegido para definir el punto "cero" en la escala. La dificultad del ítem de lógitos es el logaritmo natural del momio de inducir una falla de personas con habilidad "cero".

La Tabla 1.4.1 proporciona ejemplos de varias habilidades de personas y dificultades de ítems en lógitos, sus diferencias ($\beta_v - \delta_i$) y las probabilidades de éxito resultantes. Los primeros seis renglones ilustran las habilidades de varias personas y sus probabilidades de éxito relativas a ítems de dificultad cero. Los últimos seis renglones dan ejemplos de varias dificultades de ítems y la probabilidades de éxito en ellos por personas con habilidad nula.

El origen y la escala de los lógitos utilizados en la Tabla 1.4.1 son arbitrarios. Podemos agregar cualquier valor constante a todas las habilidades y a todas las dificultades sin cambiar la diferencia ($\beta_v - \delta_i$). Esto significa que podemos colocar el punto cero en la escala para que no se tengan habilidades ni dificultades negativas. También podemos introducir factores de escala que consideremos convenientes, tan grandes como queramos para eliminar la necesidad de utilizar fracciones decimales. El Capítulo 8 investiga estas posibilidades en detalle.

La última columna de la Tabla 1.4.1 proporciona la información relativa $I_{vi} = \pi_{vi}(1 - \pi_{vi})$ disponible en una respuesta observada para cada ($\beta_v - \delta_i$). Cuando la dificultad del ítem $\delta_i$ está a una distancia de un lógito de la habilidad de la persona $\beta_v$, la información sobre $\delta_i$ o $\beta_v$ en una observación es mayor que 0.20. Pero cuando la dificultad del ítem es mayor a dos lógitos fuera del objetivo esperado, la información es menor a 0.11 y para $|\beta_v - \delta_i| > 3$ es menor que 0.05. Las implicaciones para el muestreo eficiente en una calibración y el diseño óptimo de una prueba son que las respuestas dentro de la región donde $|\beta_v - \delta_i| < 1$ son más del doble de útiles de lo que se tiene con ítems calibrados y personas fuera de la región $|\beta_v - \delta_i| > 2$ más de cuatro veces que cuando se está fuera de la región $|\beta_v - \delta_i| > 3$.

## 1.5 USO DEL MODELO DE RASCH PARA CALIBRAR Y MEDIR

Hemos establecido la necesidad de un enfoque específico para la medición y mostrado cómo se pueden abordar los problemas de medición con un modelo que explica lo que sucede cuando la persona responde un ítem. Ahora estamos listos para trabajar este modelo a través de las matemáticas y averiguar cómo usarlo para calibrar ítems y medir personas. El modelo especifica que la persona ν con habilidad $\beta_v$ que produce una respuesta $x_{vi}$ al ítem ι con dificultad $\delta_i$ tiene una probabilidad de acierto dada por:

$$P\{x_{vi}|\beta_v,\delta_i\} = \exp[x_{vi}(\beta_v - \delta_i)]/[1 + \exp(\beta_v - \delta_i)] \qquad \textbf{[1.5.1]}$$

La respuesta $x_{vi}$ puede tomar solo dos valores,

$x_{vi} = 0$ cuando la respuesta es incorrecta y

$x_{vi} = 1$ cuando la respuesta es correcta.

Cuando insertamos cada uno de estos valores de $x_{vi}$ en la Ecuación 1.5.1, encontramos que conduce a las expresiones complementarias

$$P\{x_{vi} = 1|\beta_v,\delta_i\} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)] \qquad \textbf{[1.5.2]}$$

para una respuesta correcta y

$$P\{x_{vi} = 0|\beta_v,\delta_i\} = 1/[1 + \exp(\beta_v - \delta_i)] \qquad \textbf{[1.5.3]}$$

para una respuesta incorrecta. Estas dos expresiones se suman en una sola porque juntas cubren todo lo que puede sucederle a $x_{v\iota}$.

Cuando se administra un conjunto de L ítems a una muestra de N personas, el resultado es un conjunto de N×L respuestas ($(x_{v\iota})$) para n = 1, N; ι = 1, L y el paréntesis doble (( )) nos recuerda que se está implicando una tabla completa de valores de $x_{v\iota}$ N por L. Estos datos pueden ser organizados en forma de matriz como la que se muestra en la figura 1.5.1. Las sumas marginales de los renglones y columnas de esta matriz de datos son los puntajes de las personas.

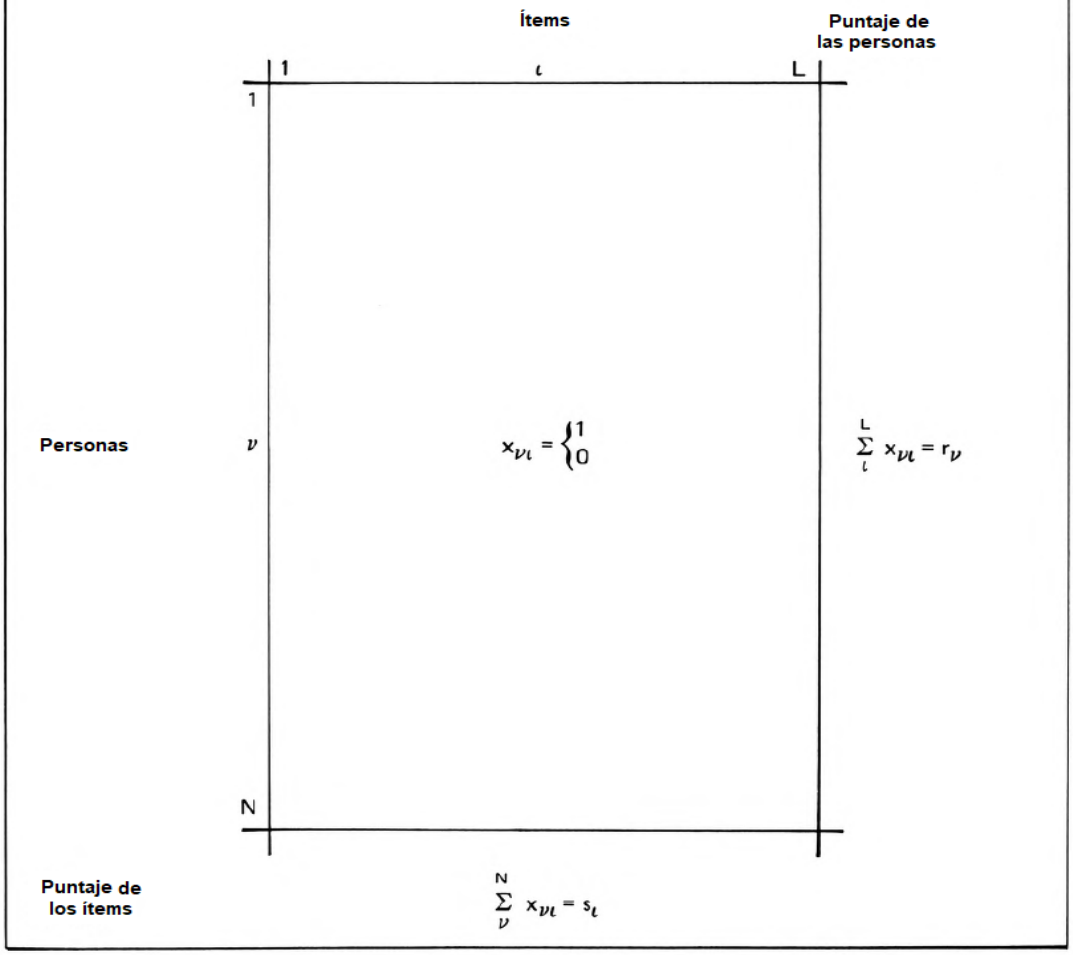
$$r_v = \sum_{\iota}^{L} x_{v\iota} \qquad v = 1, N$$

y los puntajes de los ítems.

$$s_\iota = \sum_{v}^{N} x_{v\iota} \qquad \iota = 1, L$$

**FIGURA 1.5.1**

**MATRIZ DE DATOS DE RESPUESTAS OBSERVADAS**



¿Qué sucede cuando analizamos estos datos como si estuvieran gobernados por el modelo de la Ecuación 1.5.1? De acuerdo con ese modelo, las únicas influencias sistemáticas en la producción de $x_{vi}$ son las habilidades de N personas ($\beta_v$) y las dificultades de L ítems ($\delta_i$). De todo ellos resulta, además de estos parámetros, que los $x_{vi}$ son modelados para ser independientes los unos de los otros. Esto significa que la probabilidad de toda la matriz de datos ($(x_{vi})$), dado el modelo y sus parámetros ($\beta_v$) y ($\delta_i$) se pueden expresar como el producto de la probabilidad de cada $x_{vi}$ por separado, dados por la Ecuación 1.5.1 aplicados en su totalidad sobre $v = 1, N$ y todos los ítems $i = 1, L$.

Este producto aplicado en su totalidad es

$$P\{((x_{vi}))|(\beta_v),(\delta_i)\} = \prod_{v}^{N}\prod_{i}^{L} = \left\{\frac{\exp[x_{vi}(\beta_v - \delta_i)]}{1+\exp(\beta_v - \delta_i)}\right\} \qquad \textbf{[1.5.4]}$$

Cuando movemos los operadores de producto continuados $\prod_{v}^{N}$ y $\prod_{i}^{L}$ en el numerador de la Ecuación 1.5.4 en la expresión exponencial

$$\exp[x_{vi}(\beta_v - \delta_i)]$$

se convierten en los operadores de sumatoria

$$\sum_{v}^{N} \text{y} \sum_{i}^{L} \text{ para que}$$

$$\prod_{v}^{N}\prod_{i}^{L} = \exp[x_{vi}(\beta_v - \delta_i)] = \exp\left[\sum_{v}^{N}\sum_{i}^{L} x_{vi}(\beta_v - \delta_i)\right]$$

entonces, desde

$$\sum_{v}^{N}\sum_{i}^{L} x_{vi}\beta_v = \sum_{v}^{N} r_v\,\beta_v$$

y

$$\sum_{v}^{N}\sum_{i}^{L} x_{vi}\delta_i = \sum_{i}^{L} s_i\,\delta_i,$$

La ecuación 1.5.4 se convierte

$$P\{((x_{vi}))|(\beta_v),(\delta_i)\} = \left\{\frac{\exp\left[\sum_v^N r_v\,\beta_v - \sum_i^L s_i\delta_i\right]}{\prod_v^N\prod_i^L[1+(\exp(\beta_v - \delta_i)]}\right\} \qquad \textbf{[1.5.5]}$$

La Ecuación 1.5.5 es importante porque muestra que para estimar los parámetros ($\beta_v$) y ($\delta_i$), solo necesitamos las sumas marginales de la matriz de datos, ($r_v$) y ($s_i$). Debido a que es la única forma en que los datos ($(x_{vi})$) aparecen en la Ecuación 1.5.5. Por lo tanto, los puntajes de la persona ($r_v$) y los puntajes de los ítems ($s_i$) contienen toda la información modelada sobre las medidas de las personas y las calibraciones de los ítems.

Finalmente, el numerador de la Ecuación 1.5.5 puede factorizarse en dos factores con lo que el modelo de probabilidad de la matriz de datos queda:

$$P\{((x_{vi}))|(\beta_v),(\delta_i)\} = \left\{\frac{[\exp(\sum_v^N r_v \beta_v)][\exp(-\sum_i^L s_i\delta_i)]}{\prod_v^N \prod_i^L [1+(\exp(\beta_v - \delta_i)]}\right\} \quad \textbf{[1.5.6]}$$

La Ecuación 1.5.6 es importante porque muestra que los parámetros de las personas y de los ítems pueden estimarse independientemente uno del otro. La separación de

$$\left(\sum_v^N r_v \beta_v\right)$$

y

$$\left(\sum_i^L s_i\delta_i\right)$$

en la Ecuación 1.5.6 hace posible condicionar cualquiera de los conjuntos de parámetros al estimar el otro conjunto. Esto significa, en el lenguaje de la estadística, que los puntajes ($r_v$) y ($s_i$) son suficientes para estimar las medidas de las personas y las calibraciones de los ítems.

Con esta base podemos usar los puntajes de las personas ($r_v$) para eliminar los parámetros de la persona ($\beta_v$) de la Ecuación 1.5.6 al calibrar los ítems. Esto libera las calibraciones de los ítems de las características modeladas de las personas y a su vez se obtienen calibraciones de los ítems independientes de la muestra. En cuanto a la medición de personas, podríamos usar los puntajes de los ítems ($s_i$) para eliminar los parámetros ($\delta_i$) para los ítems de la Ecuación 1.5.6. Cuando llegamos a la medición de la persona, sin embargo, encontraremos más conveniente trabajar directamente con las calibraciones de los ítems estimados ($d_i$).

Hay varias formas en que la Ecuación 1.5.6 puede usarse para estimar valores para ($\beta_v$) y ($\delta_i$). La forma ideal es usar las estadísticas suficientes para personas ($r_v$) para condicionar los parámetros de las personas ($\beta_v$) de la ecuación. Esto conduce a una verosimilitud condicional que involucra solo los parámetros del ítem ($\delta_i$) y que se pueden estimar a partir de dicha verosimilitud condicional (Fischer y Scheiblechner, 1970; Andersen, 1972a, 1972b; Wright y Douglas, 1975b, 1977b; Allerup y Sorber, 1977; Gustafsson, 1977).

Pero este método ideal es poco práctico e innecesario. Los tiempos de computación son excesivos. Los errores de redondeo limitan su aplicación a pruebas de cincuenta ítems como máximo. En cualquier caso, los resultados son numéricamente equivalentes a los de métodos más rápidos y más robustos. Una alternativa conveniente y práctica consiste es usar la Ecuación 1.5.6 directamente. Para aprender más sobre la estimación incondicional de los parámetros de los ítems, véase Wright y Panchapakesan (1969), Wright y Douglas (1975b, 1977a, 1977b) y el Capítulo 3, Sección 3.4. de este libro.

Sin embargo este método incondicional, a menudo es innecesariamente detallado y costoso para su aplicación práctica. Si las personas que utilizamos para calibrar los ítems no se distribuyen demasiado asimétricamente en el eje de habilidad y no están muy lejos del objetivo, pueden ser más o menos representados por su media y su varianza, con lo que podemos usar un método muy simple y viable para estimar las dificultades de los ítems. Este método, llamado PROX, fue sugerido por primera vez por Leslie Cohen en 1973 (ver Wright y Douglas, 1977a; Wright, 1977).

## 1.6 UN PROCEDIMIENTO DE ESTIMACIÓN SIMPLE Y PRÁCTICO

En este libro se usarán tres métodos de estimación de los parámetros. El método general incondicional llamado UCON requiere una computadora y un programa de computadora como BICAL (Wright y Mead, 1976). UCON se discute e ilustra en el Capítulo 3. Un segundo método llamado UFORM se puede hacer a mano con la ayuda de las tablas simples que figuran en el Apéndice C, que se discuten y aplican en el Capítulo 7. El tercer método, denominado PROX, es completamente manejable a mano. Además, la simplicidad de PROX nos ayuda a ver cómo funciona el modelo de Rasch para resolver problemas de medición. Las derivaciones de las ecuaciones UFORM y PROX se dan en Wright y Douglas (1975a, 1975b).

PROX asume que las habilidades de la persona ($\beta_v$) se distribuyen aproximadamente en forma normal con media M y desviación estándar σ, las dificultades de los ítems ($\delta_i$) también se distribuyen aproximadamente en forma normal con dificultad promedio H y desviación estándar de dificultades ω.

Si

$$\beta_v \sim N(M, \sigma^2)$$

y

$$(\delta_i \sim N(H, \omega^2),$$

entonces para cualquier persona $v$ con puntaje $r_v$ en una prueba de L ítems se sigue que

$$b_v = H + X\ell n\,[r_v/(L - r_v)] \qquad \textbf{[1.6.1]}$$

y para cualquier ítem ι con puntajes de ítems $s_i$ en una muestra de N personas se deduce que

$$d_i = M + Y\ell n[(N - s_i)/s_i] \qquad \textbf{[1.6.2]}$$

Los coeficientes X e Y son factores de expansión que responden en el caso de X a la dispersión de las dificultades de los ítems y en el caso de Y a la dispersión de las habilidades de las personas.

En particular:

$$X = (1 + \omega^2/2.89)^{1/2} \qquad \textbf{[1.6.3]}$$

y

$$Y = (1 + \sigma^2/2.89)^{1/2} \qquad \textbf{[1.6.4]}$$

El valor $2.89 = 1.7^2$ proviene del factor escalar 1.7 que hace que la ojiva logística coincida aproximadamente con la ojiva normal. Esto es porque los valores de 1.7z de la ojiva logística nunca están a más de uno por ciento de diferencia de la ojiva normal para valores de z.

Las estimaciones $b_v$ y $d_i$ tienen errores estándar

$$ES(b_v) = X\,[L/r_v(L - r_v)]^{1/2} \qquad \textbf{[1.6.5]}$$

$$ES(d_i) = Y\,[N/s_i(N - s_i)]^{1/2}\,. \qquad \textbf{[1.6.6]}$$

Este método de estimación se puede aplicar directamente a los puntajes de ítems observados ($s_i$) calculando el puntaje simple en lógitos del ítem $i$ como:

$$x_i = \ell n[(N - s_i)/s_i] \quad \textbf{[1.6.7]}$$

y el lógito de puntaje de ítems de la persona $v$ como

$$y_v = \ell n[r_v/(L - r_v)] \quad \textbf{[1.6.8]}$$

Los factores de expansión X e Y son estimados por las expresiones

$$X = [(1 + U/2.89)/(1 - UV/8.35)]^{1/2}\ , \quad \textbf{[1.6.9]}$$

para el factor de expansión en lógitos de la persona y

$$Y = [(1 + V/2.89)/(1 - UV/8.35)]^{1/2}\ , \quad \textbf{[1.6.10]}$$

para el factor de expansión en lógitos del ítem.

En estas expresiones $2.89 = 1.7^2$ y $8.35 = 2.89^2 = 1.7^4$, con lo que

$$U = \left(\sum_i^L x_i^2 - Lx.^2\right)/(L - 1), \quad \textbf{[1.6.11]}$$

es la varianza del ítem en lógitos y

$$V = \left(\sum_v^N y_v^2 - Ny.^2\right)/(N - 1), \quad \textbf{[1.6.12]}$$

la varianza de la persona en lógitos.

Para completar esta estimación, establecemos el centro de la prueba en cero para que H = 0. Entonces:

$$d_i = M + Yx_i = Y(x_i - x.)\ , \quad \textbf{[1.6.13]}$$

para cada dificultad del ítem, y

$$b_v = H + Xy_v = Xy_v \quad \textbf{[1.6.14]}$$

por cada habilidad de persona.

Los errores estándar son

$$ES(d_i) = Y[N/s_i(N - s_i)]^{1/2} \simeq 2.5/\ N^{1/2} \quad \textbf{[1.6.15]}$$

y

$$ES(b_v) = X[L/r_v(L - r_v)]^{1/2} \cong 2.5/\ L^{1/2}\ . \quad \textbf{[1.6.16]}$$

Finalmente, la media y la desviación estándar estimadas de la muestra de personas se escriben

$$M \approx -Yx. \quad \textbf{[1.6.17]}$$

$$\sigma \approx 1.7\left(Y^2 - 1\right)^{1/2} \quad \textbf{[1.6.18]}$$

Una vez que hemos estimado $b_v$ y $d_i$, podemos usarlos para obtener la diferencia entre lo que predice el modelo y los datos que realmente hemos observado. Estos residuos del modelo se calculan estimando las expectativas del modelo en cada $x_{vi}$ a partir de $b_v$ y $d_i$, restando esta expectativa del $x_{vi}$ que se observó. La expectativa del modelo para $x_{vi}$ es:

$$E\{x_{vi}\} = \pi_{vi}$$

con la varianza del modelo

$$V\{x_{vi}\} = \pi_{vi}(1 - \pi_{vi})$$

donde

$$\pi_{vi} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)].$$

Un residuo estandarizado sería

$$z_{vi} = (X_{vi} - \pi_{vi})/[\pi_{vi}(1 - \pi_{vi})]^{1/2} . \quad \textbf{[1.6.19]}$$

Si los datos se ajustan al modelo, este residuo estandarizado debería distribuirse más o menos normalmente con media cero y varianza uno.

Si estimamos $\pi_{vi}$ de $p_{vi}$ donde

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)] \quad \textbf{[1.6.20]}$$

entonces podemos usar las distribuciones de error

$z_{vi} \sim N(0,1)$ y

$z_{vi}^2 \sim x_1^2$

que nos servirán como pautas para evaluar en qué medida nuestro modelo de medición puede utilizarse con cualquier otro conjunto particular de datos.

Podemos calcular para cada persona la suma de sus residuos cuadráticos $z_{vi}^2$. De acuerdo con el modelo, esta suma de desviaciones normales al cuadrado debe aproximarse a una distribución ji-cuadrado con aproximadamente (L – 1) grados de libertad. Esto nos da este estadístico ji-cuadrado

$$\sum_{i}^{L} z_{vi}^2 = C_v^2 \sim x_{f_v}^2 \quad \textbf{[1.6.21]}$$

con grados de libertad

$$f_v = (L - 1)(N - 1)/N \quad \textbf{[1.6.22]}$$

y una media cuadrática

$$v_v = C_v^2/f_v \sim F_{f_v,\infty} \quad \textbf{[1.6.23]}$$

que se aproxima a una distribución F cuando las respuestas de las personas se ajustan al modelo.

La suma de residuos cuadráticos para cada ítem se puede usar de la misma manera para evaluar el ajuste del ítem. Dados los ítems

$$\sum_{v}^{N} z_{vi}^2 = C_i^{\,2} \sim x_{f_i}^2 \qquad \textbf{[1.6.24]}$$

con

$$f_i = (N-1)(L-1)/L \qquad \textbf{[1.6.25]}$$

y

$$v_i = C_i^2 / f_i \sim F_{f_i,\infty} \qquad \textbf{[1.6.26]}$$

Finalmente, dado que $x_{vi}$ solo puede adquirir valores de uno o cero, podemos usar la definición de $p_{vi}$ dada en la Ecuación 1.6.20 para calcular $z_{vi}$ y $z_{vi}^2$ directamente como

$$z_{vi} = (2x_{vi} - 1)\exp[(2x_{vi} - 1)(d_i - b_v)/2] \qquad \textbf{[1.6.27]}$$

y

$$z_{vi}^{\,2} = \exp[(2x_{v,} - 1)(d_i - b_v)]. \qquad \textbf{[1.6.28]}$$

Esta relación también se puede trabajar de forma inversa. Si ya tenemos un $z_{vi}^2$ y deseamos calcular la probabilidad de la respuesta observada $x_{vi}$ a la que se refiere para decidir si esa respuesta es o no demasiado improbable para creer, entonces podemos usar

$$P\{x_{vi}|b_{v,}d_i\} = 1/\left(1 + z_{vi}^{\,2}\right). \qquad \textbf{[1.6.29]}$$

En contraste con el $p_{vi}$ de la Ecuación 1.6.20 que es la probabilidad estimada de una respuesta correcta, la probabilidad de la Ecuación 1.6.29 se aplica a $x_{vi}$ independientemente del valor que tome, si $x_{vi}$=1 para una respuesta correcta o si $x_{vi}$=0 para una incorrecta.

## 1.7 CÓMO APARECEN LOS VALORES ESTADÍSTICOS TRADICIONALES EN LA MEDICIÓN DE RASCH

Las secciones 1.1, 1.2 y 1.3 discuten el propósito de las pruebas, el uso de los puntajes y los problemas de generalidad y linealidad al construir medidas. Las secciones 1.4, 1.5 y 1.6 describen una solución simple y práctica para estos problemas de medición. Debido a que la formulación matemática puede ser nueva, parecería que usar el modelo de Rasch para analizar los ítems nos alejará de la estadística tradicional con la que estamos familiarizados. Esto no es así.

La aplicación del modelo de Rasch en el desarrollo de la prueba nos da nuevas versiones de las antiguas estadísticas. Estas nuevas estadísticas contienen toda la información familiar que ya conocemos, pero en una forma que resuelve la mayoría de los problemas de medición que siempre han acosado a los constructores de pruebas. Para mostrar esto, examinaremos los tres ítems tradicionales más comunes y las estadísticas de personas, para ver qué tan próximos se encuentran de las correspondientes medidas de Rasch.

**El valor p del ítem**

La estadística tradicional más familiar es el valor p (o "p-valor") del ítem, que es la proporción de personas que responden correctamente al ítem en una muestra específica. La Ecuación 1.6.2 para estimar PROX proporciona una forma conveniente de formular la relación entre el p-valor tradicional y la dificultad de Rasch del ítem. Si el valor de p para el ítem ι se expresa como

$$p_\iota = s_\iota / N$$

en el que $s_\iota$ es el número de personas que respondieron correctamente al ítem L en la muestra de N personas. Entonces la dificultad de Rasch para el ítem estimada por PROX es

$$d_\iota = M + \left(1 + \sigma^2 / 2.89\right)^{1/2} \ell n\left[\left(1 - p_\iota\right)/p_\iota\right] \qquad \textbf{[1.7.1]}$$

La Ecuación 1.7.1 muestra que la dificultad de Rasch $d_\iota$ del ítem está en una relación de uno a uno con el p-valor del ítem representado por $p_\iota$. También muestra que la relación uno a uno es curvilínea con base en la habilidad de la media M y la varianza $\sigma^2$ de la muestra de calibración.

Lo que hace el modelo de Rasch es usar la función logística (o función lógito)

$$\ell n\left[\left(1 - p_\iota\right)/p_\iota\right]$$

que conduce a transformar el p-valor del ítem – que no es lineal en la variable implicada – en un nuevo valor. Este nuevo valor en lógitos expresa la dificultad del ítem en una escala de intervalos iguales que permite realizar de manera fácil la corrección posterior del p-valor del ítem para la habilidad M y varianza $\sigma^2$ de la muestra que se está calibrando.

Esta corrección se realiza escalando el lógito para eliminar los efectos de la varianza $\sigma^2$ de la muestra y traducir el valor escalado en lógitos eliminando a su vez los efectos de la media M de la muestra. Las dificultades de Rasch no solo se encuentran en una escala de intervalos iguales, sino también se liberan de la media de la habilidad observada y la varianza de la muestra de calibración. El p-valor del ítem $p_\iota$ tiene un error estándar (ES) binomial dado por:

$$ES(p_\iota) = \left[p_\iota\left(1 - p_\iota\right)/N\right]^{1/2} \qquad \textbf{[1.7.2]}$$

por lo que la dificultad del ítem PROX $d_\iota$ se asocia estrechamente con su propio error estándar:

$$ES(d_\iota) = \left(1 + \sigma^2/2.89\right)^{1/2} \left[1/Np_\iota\left(1 - p_\iota\right)\right]^{1/2} \qquad \textbf{[1.7.3]}$$

Pero hay dos diferencias importantes entre las ecuaciones 1.7.2 y 1.7.3. A diferencia del error estándar del p-valor en la Ecuación 1.7.2, el error estándar de Rasch en la Ecuación 1.7.3 es corregido para la varianza de habilidad $\sigma^2$ de la muestra de calibración. La segunda diferencia entre estas dos formulaciones es más sutil, pero aún más importante.

Los errores estándar del p-valor tradicional en la Ecuación 1.7.2 son máximos a la mitad de los p-valores cercanos a un medio y es cero en los extremos cuyo p-valor es cero o uno. Esto hace que parezca que sabemos más del ítem que tiene el error estándar más pequeño del p-valor cuando, de hecho, sabemos menos. Esto se debe a que el p-valor del ítem se centra en la muestra calibrada al mismo tiempo que en el ítem. Si la muestra se aleja del sitio donde apunta el ítem, entonces su p-valor se acerca a los extremos 0 y 1, haciendo que el error estándar se aproxime a cero. Esto nos garantiza que el p-valor de la muestra particular está en un extremo, pero no nos

dice nada más sobre el ítem. Por lo tanto, aunque nuestro conocimiento del valor estuviera aumentando, en cambio nuestra información con respecto a la dificultad real del ítem está decreciendo. Cuando los p-valores de los ítems son cero o uno, la muestra de calibración que esperábamos nos informaría cómo funciona el ítem se revela como demasiado capaz o demasiado incapaz de interactuar con el ítem. De modo que sabemos en qué dirección se encuentra el ítem pero no sabemos la posición en la que se encuentra.

En contraste, el error estándar de Rasch para $d_i$ varía de una manera más razonable. La expresión $p_i(1 - p_i)$ que es cero cuando $p_i$ es cero o uno, aparece en el denominador de la Ecuación 1.7.3 en lugar del numerador, como ocurre en la Ecuación 1.7.2. En consecuencia, el error estándar de Rasch es menor en $p_i = 0.5$, donde la muestra está centrada en el ítem y así nos da la mayor información sobre cómo funciona ese ítem. En los extremos, sin embargo, donde tenemos la menor información, el error estándar de Rasch tiende al infinito recordándonos que no hemos aprendido casi nada sobre ese ítem con ayuda de la muestra.

**La correlación punto biserial del ítem.**

El segundo parámetro tradicional de un ítem, ampliamente utilizado es la correlación punto biserial entre las respuestas dicotómicas de las personas de la muestra ante un ítem y su puntaje total en la prueba. La correlación punto-biserial tiene dos características que interfieren con su utilidad como índice de qué tan bien correlaciona el ítem con el conjunto de ítems en el que aparece. Primero, no hay una base clara para determinar un valor de referencia para establecer si es aceptable el ítem. Rechazar la hipótesis estadística de que la correlación punto-biserial es cero no produce un criterio estadístico satisfactorio para validar un ítem. La segunda característica es que la magnitud de la correlación punto-biserial está sustancialmente influenciada por la distribución de los puntajes de la muestra usada en la calibración. La correlación punto-biserial de un ítem dado es más grande cuando la dispersión de puntajes de las personas es amplia y a su vez se centran en el ítem. Por el contrario, si la varianza de los puntajes de las personas disminuye o el nivel de la muestra se aleja del nivel en el que se encuentra el ítem, haciendo que el p-valor se acerque a cero o uno, con ello la correlación tiende a cero independientemente de la calidad del ítem.

El estadístico de Rasch que corresponde en significado a la correlación punto-biserial es el residuo cuadrático medio del ítem, de la Ecuación 1.6.26. El residuo cuadrático medio no solo es sensible a los ítems que no se correlacionan con el puntaje del examen, sino también con valores inesperadamente grandes de la correlación punto-biserial. Esto sucede, por ejemplo, cuando una variable no modelada produce una interacción local entre una característica única de un ítem y alguna característica idiosincrática de algunos miembros de la muestra para la calibración.

En contraste con la correlación punto-biserial, el residuo cuadrático medio de Rasch para el ítem se asocia a una útil distribución estadística de referencia. El valor de referencia para probar la hipótesis estadística de que un ítem pertenece a la prueba es una media cuadrática de uno, con un error estándar de $(2 / f)^{1/2}$ para $f$ grados de libertad. De este modo, el grado en que la media cuadrática excede el valor esperado de uno, puede ser probado por su significancia estadística ante cualquier nivel de significancia considerada.

La media cuadrática de Rasch para el ítem es casi indiferente de la distribución de la habilidad que tiene la muestra de calibración. Esto proporciona una prueba de ajuste del ítem que se centra en las características de la muestra y de los ítems que permanecen una vez que se hayan removido los valores modelados para la dificultad del ítem y la habilidad de la persona.

**El puntaje de la prueba de la persona**

El parámetro estadístico tradicional y más familiar para es el puntaje en la prueba, es decir, el número de aciertos que la persona obtiene en la prueba. Una vez más podemos usar la estimación de PROX por ser un procedimiento que muestra la conexión entre el puntaje tradicional que obtiene la persona y la habilidad de Rasch. Usando la Ecuación de estimación PROX (1.6.1) tenemos:

$$b_v = H + \left(1 + \omega^2/2.89\right)^{1/2} \ell n[r_v/(L - r_v)] \qquad \textbf{[1.7.4]}$$

con un error estándar de

$$ES(b_v) = \left(1 + \omega^2/2.89\right)^{1/2} [L/r_v(1 - r_v)]^{1/2} \qquad \textbf{[1.7.5]}$$

en el cual

$r_v$ = puntaje del examen de la persona $v$,
L = cantidad de ítems en la prueba,
H = nivel de dificultad promedio de la prueba y
$\omega^2$ = varianza en las dificultades de los ítems de prueba.

Al igual que con el p-valor del ítem, la función de transformación en lógitos hace que los puntajes no lineales de las personas cambien a una métrica aproximadamente lineal. También vemos que la escala del valor en lógitos ocupa el ancho de la prueba, representado en la Ecuación 1.7.4 por la varianza de las dificultades de los ítems $\omega^2$ y, además, se traslada para ajustar el nivel de dificultad de la prueba H, de modo que la habilidad estimada para la persona se libere de los efectos locales de la prueba y se convierta en una medida independiente de la prueba.

El error estándar de esta medida es mínimo en puntajes cercanos al 50 por ciento de aciertos, donde tenemos la mayor información sobre la persona y va al infinito en los puntajes de cero y 100 por ciento, donde tenemos la menor información sobre la persona.

Mientras que las prácticas de prueba tradicionales casi siempre enfatizan el análisis de la validez de los ítems, casi nunca se presta atención a la validez del patrón de respuestas que conduce al puntaje de la persona. Hasta donde sabemos, nadie calcula el coeficiente punto biserial de una persona con el propósito de determinar la relación entre las respuestas que la persona da a cada ítem y los valores p supuestamente relevantes del ítem. Esta sería una forma razonable de aplicar el coeficiente de correlación punto-biserial tradicional a la supervisión de la validez del puntaje de la persona.

El enfoque de Rasch para la validez del puntaje de la persona se describe en las Ecuaciones 1.6.19 a 1.6.23 y se discute e ilustra detalladamente en los Capítulos 4 y 7.

Hay otras conexiones que se pueden hacer entre las estadísticas de prueba tradicionales y las estadísticas de Rasch. Podríamos revisar aquí las diversas formas en que el modelo de Rasch aborda la validez y la confiabilidad (o fiabilidad) tradicional, el manejo de pruebas referidas a norma o de pruebas referidas a criterio, la equiparación (o igualación) de formas y las pruebas de desempeño. Pero cada uno de estos temas merece una profunda discusión que, de hecho, es el propósito de los capítulos que siguen. El paso siguiente que vamos a ver es el funcionamiento del procedimiento de estimación de PROX para resolver un problema simple en la construcción de una prueba.

# 2 CALIBRACIÓN DE ÍTEMS A MANO

## 2.1 INTRODUCCIÓN

Este capítulo describe e ilustra en detalle un procedimiento extremadamente simple para calibrar ítems con el modelo de Rasch. El procedimiento, llamado PROX, aproxima bastante bien los resultados obtenidos por procedimientos más elaborados y, por lo tanto, más precisos. Con este procedimiento se alcanzan los objetivos básicos del análisis de Rasch, a saber, la linealización de la escala latente y el ajuste de los efectos locales de la distribución de las habilidades de la muestra de personas. La suposición que hace que PROX sea tan simple es que los efectos en la calibración del ítem sobre la distribución de las habilidades de la muestra puede ser adecuadamente explicado con base solamente en la media y la desviación estándar. Esta suposición hace que PROX sea tan simple que se puede aplicar fácilmente a mano.

En la práctica, lo más conveniente es hacer la calibración de los ítems con ayuda de la computadora. Sin embargo, PROX brinda la oportunidad de ilustrar el análisis de ítems con el modelo de Rasch en forma detallada, favoreciendo una mejor comprensión del proceso involucrado, en caso de que la infraestructura informática no esté disponible o sea urgente verificar la plausibilidad de los resultados de salida del programa, en este caso PROX proporciona un método para calibrar ítems que no requiere nada más que la distribuciones observadas en los puntajes de los ítems y de las personas, una calculadora portátil, papel y lápiz.

Los datos para ilustrar PROX provienen de la administración de los 18 ítems de la Prueba de los Cubos de Knox, una subprueba de la Escala de Puntajes de Arthur (Arthur, 1947) aplicada a 35 estudiantes en Grados 2 a 7. Nuestro análisis de estos datos muestra cómo el análisis de ítems de Rasch puede ser útil no sólo para gestionar la construcción de bancos de ítems nacionales, sino también para atender el problema de medición más pequeño imaginable, es decir, una prueba corta aplicada a un grupo de examinados en un salón de clase.

Al utilizar las respuestas correctas/incorrectas de los estudiantes para cada ítem de la prueba, trabajamos detalladamente cada etapa del procedimiento para el análisis de ítems PROX. El Capítulo 3 revisa la forma de analizar los mismos datos por computadora, con el procedimiento PROX y otro más preciso, denominado UCON, utilizado en la mayoría de los programas informáticos para el análisis de ítems de Rasch. Estos pasos detallados ofrecen una ilustración sistemática del procedimiento de análisis de ítems con los cuales comparar resultados y comprender las salidas de los programas de computadora. También demuestran la facilidad de los cálculos a mano usando PROX (PROX se deriva y describe detalladamente en Wright y Douglas, 1975b, 1976, 1977a; Cohen, 1976, y Wright, 1977). Finalmente, ilustran el desarrollo empírico de un rasgo o variable latente. Cada paso se mueve desde los datos observados hasta la variable inferida, desde los límites de lo observado a través de los puntajes de prueba hasta el otro extremo de las mediciones inferidas independientemente de la prueba.

## 2.2 LA PRUEBA DE LOS CUBOS DE KNOX

La Escala de Puntajes de Arthur cubre una variedad de tareas mentales, entre las cuales la Prueba de los Cubos de Knox implica un único rasgo latente. Para tener éxito en esta subprueba se requiere aplicar la atención visual y la memoria a corto plazo ante una tarea simple de secuencias que parece ser independiente del aprendizaje escolar y, por lo tanto, se toma como indicador de la capacidad o habilidad intelectual no verbal.

La Prueba de los Cubos de Knox utiliza cinco cubos de una pulgada de lado. Cuatro de los cubos están fijos a una tabla, separados a cada dos pulgadas, y el quinto cubo se usa para realizar una serie de golpes sobre los otros cuatro. Se hará referencia a los cuatro cubos fijos, de izquierda a derecha, como "1", "2", "3" y "4", para evitar confusiones al especificar cualquier serie de golpes particular que se vaya a efectuar. En la versión original de la prueba utilizada para este ejemplo, hay 18 series de golpes que van desde secuencias de dos pasos (1-4) y (2-3) a una secuencia de siete golpes (4-1-3-4-2-1-4). Por lo general, se administra la prueba a cada sujeto en dos ocasiones, utilizando otra subprueba de la batería. Sin embargo, para nuestro análisis solo necesitamos usar la primera administración.

Las 18 series se presentan en la Figura 2.2.1. Estos son los 18 "ítems" de la prueba. Obsérvese que los ítems 1 y 2 requieren una secuencia de dos golpes; los ítems 3 a 6, definen una secuencia de tres golpes; los ítems 7 a 10 una secuencia de cuatro golpes; los ítems 11 a 13 una secuencia de cinco golpes, los ítems 14 a 17 una secuencia de seis golpes y, finalmente, el Ítem 18, una secuencia de siete golpes.

**FIGURA 2.2.1**

**NOMBRE DEL ÍTEM Y ORDEN DE LA SECUENCIA DE GOLPES PARA LA PRUEBA DE LOS CUBOS DE KNOX**

| NOMBRE DEL ITEM | SECUENCIA DE GOLPES | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | | | | | |
| 2 | 2 | 3 | | | | | |
| 3 | 1 | 2 | 4 | | | | |
| 4 | 1 | 3 | 4 | | | | |
| 5 | 2 | 1 | 4 | | | | |
| 6 | 3 | 4 | 1 | | | | |
| 7 | 1 | 4 | 3 | 2 | | | |
| 8 | 1 | 4 | 2 | 3 | | | |
| 9 | 1 | 3 | 2 | 4 | | | |
| 10 | 2 | 4 | 3 | 1 | | | |
| 11 | 1 | 3 | 1 | 2 | 4 | | |
| 12 | 1 | 3 | 2 | 4 | 3 | | |
| 13 | 1 | 4 | 3 | 2 | 4 | | |
| 14 | 1 | 4 | 2 | 3 | 4 | 1 | |
| 15 | 1 | 3 | 2 | 4 | 1 | 3 | |
| 16 | 1 | 4 | 2 | 3 | 1 | 4 | |
| 17 | 1 | 4 | 3 | 1 | 2 | 4 | |
| 18 | 4 | 1 | 3 | 4 | 2 | 1 | 4 |

## 2.3 DATOS PARA EL ANÁLISIS DE ÍTEMS

Las respuestas de 35 estudiantes a una sola administración de los 18 ítems de la Prueba de los Cubos de Knox aparecen en la Tabla 2.3.1. Estas respuestas están organizadas en una matriz de datos de personas contra ítems. La respuesta correcta de un estudiante a cada ítem se registra como un 1 y un respuesta incorrecta como 0. Los ítems se han enumerado en la parte superior en el orden de administración.

Los puntajes de los estudiantes, es decir, el número de aciertos logrados por cada estudiante, aparecen en la última columna a la derecha de cada renglón. Los puntajes de los ítems, es decir, el número total de respuestas correctas a cada ítem, se dan en la parte inferior de cada columna.

La inspección de la Tabla 2.3.1 muestra que el orden de administración es muy similar al orden de dificultad. Los ítems 1, 2 y 3 son respondidos correctamente por todos los estudiantes. Un segundo nivel de dificultad, ligeramente más grande, se observa en los ítems 4 a 9. Después siguen los ítems 10 y 11 que muestran un fuerte aumento en la dificultad. Los ítems 12 a 17 son respondidos correctamente por unos cuantos estudiantes y ningún estudiante logra responder correctamente al ítem 18. Solo 12 estudiantes obtienen por lo menos un acierto en los ítems 12 a 17 y solo cinco de dichos estudiantes logran uno o más de la secuencia de seis golpes exitosamente.

## 2.4 CALIBRANDO ÍTEMS Y MIDIENDO PERSONAS

El plan general para realizar el análisis de ítems comienza con la edición de los datos en la Tabla 2.3.1 para eliminar las personas e ítems para los cuales no hay estimaciones definitivas de su habilidad o de su dificultad, es decir, son los casos en que todas son respuestas correctas o todas son respuestas incorrectas. Esto significa que la persona 35 y los ítems 1, 2, 3 y 18 deben ponerse aparte, dejando para el análisis solo a 34 personas y 14 ítems. Una vez hecho esto, la información restante sobre las personas y los ítems queda resumida en una distribución de puntajes de las personas y una distribución de puntajes de los ítems.

A continuación, estas distribuciones de puntajes se representan como proporciones de su máximo valor posible y se registra su frecuencia de ocurrencia. Las proporciones se convierten luego en logaritmo del momio, o lógitos, calculando para cada ítem el logaritmo natural de la proporción de respuestas incorrectas dividida por la proporción de respuestas correctas y para las personas se calcula el logaritmo natural de la proporción de aciertos dividida por la proporción de fallas. Esto convierte a las proporciones, que estaban acotadas por 0 y 1, a una nueva escala que se extiende desde $-\infty$ hasta $+\infty$ y es lineal en la variable subyacente.

La "dificultad del ítem" en la variable medida aumenta con la proporción de respuestas incorrectas. La "habilidad de la persona" aumenta con la proporción de respuestas correctas. Ahora se calcula la media y la varianza para cada distribución de lógitos y la media en lógitos de los ítems se usa para centrar los lógitos en el cero de la escala. Esta elección de origen para la nueva escala es forzosamente arbitraria, pero es algo que debe hacerse. Basar la escala en los ítems en lugar de las personas y colocarla en el centro de los ítems actuales es una acción natural y conveniente.

Las varianzas en lógitos de los ítems y de las personas se usan para calcular dos factores de expansión, uno para los ítems y el otro para las personas. Estos factores se usan para calcular las dificultades finales, independientes de la prueba y las habilidades independientes de las personas. Los factores se necesitan porque las dificultades aparentes relativas de los ítems dependen de cuán dispersas son las habilidades en la muestra de personas. Conforme más dispersas estén las personas, más similar será la dificultad aparente. Esto también es cierto para la habilidad aparente. Cuanto más dispersa sean las dificultades de los ítems en la prueba, más similar aparentarán ser las habilidades de las personas. Estos efectos de dispersión de la muestra y del rango de la prueba deben ser eliminados de las estimaciones de las dificultades de los ítems y de las habilidades de las personas, estas estimaciones deben hacerse independientes de la muestra y de la prueba.

Finalmente, se calculan los errores estándar de las estimaciones. Los errores estándar son necesarios para evaluar la precisión de las estimaciones. Dependen de los mismos factores de expansión junto con qué tanto la dificultad de los ítems se centra en las habilidades de las personas y, a su vez, si las habilidades de las personas se centran en las dificultades de los ítems. Conforme más centrado en el objetivo se tenga a los ítems y a las personas, más precisas serán sus estimaciones y, por lo tanto, más pequeños sus errores estándar.

## TABLA 2.3.1

## RESPUESTAS ORIGINALES DE 35 PERSONAS A 18 ITEMS EN LA PRUEBA KNOX CUBE

| NOMBRE DE LA PERSONA | NOMBRE DEL ITEM | | | | | | | | | | | | | | | | | | PUNTAJE DE LA PERSONA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 14 |
| 8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 11 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 12 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 13 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 14 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 15 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 13 |
| 16 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 17 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 18 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 11 |
| 19 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 20 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 11 |
| 21 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 12 |
| 22 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 12 |
| 24 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 14 |
| 25 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 26 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 27 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| 28 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 29 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 10 |
| 30 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 31 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 32 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 33 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 34 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 12 |
| 35 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| PUNTAJE DEL ITEM | 35 | 35 | 35 | 32 | 31 | 30 | 31 | 27 | 30 | 24 | 12 | 6 | 7 | 3 | 1 | 1 | 1 | 0 | |

### Paso 1. Organización de la matriz de datos

La matriz de datos en la Tabla 2.3.1 se ha reorganizado en la Tabla 2.4.1 para que los puntajes de las personas estén ordenados de menor a mayor, con sus respectivas proporciones dadas por la columna del extremo derecho y los puntajes de los ítems se ordenan de mayor a menor con sus proporciones dadas en la fila inferior.

### Paso 2. Edición de la matriz de datos

La matriz de datos de las respuestas de las persona contra los ítems en la Tabla 2.4.1 también ha sido editada eliminando todos los ítems que fueron contestados correctamente por todos o por nadie, así como eliminando todas las personas que obtuvieron puntajes perfectos o que no respondieron correctamente ningún ítem.

Las líneas limítrofes dibujadas en la Tabla 2.3.1 muestran los ítems y personas eliminadas por el proceso de edición. Los ítems 1, 2 y 3 se eliminaron porque todas las personas los respondieron correctamente. La eliminación de estos tres ítems provocó la eliminación de la Persona 35, porque solo contestó estos tres ítems correctamente y por lo tanto ya no tuvo ningún acierto después de que los tres ítems fueron eliminados. El ítem 18 fue eliminado porque ninguna persona lo respondió correctamente.

La edición de una matriz de datos puede requerir de varios ciclos porque la eliminación de ítems puede implicar que se eliminen personas y viceversa. Por ejemplo, si hubiera habido una persona que contestó acertadamente todos los ítems excepto el 18, entonces la eliminación de este ítem hubiera dejado a esta persona con un puntaje perfecto en los ítems restantes, en cuyo caso esta persona también tendría que haber sido eliminada.

¿Por qué se eliminan algunos ítems y algunas personas? Si en una muestra de personas nadie acertara en un cierto ítem, entonces podría decirse que ese ítem es demasiado difícil para esta muestra de personas. Sin embargo, no se tendría información adicional disponible sobre qué tan difícil es en realidad ese ítem. Si todos respondieran correctamente a un ítem entonces este ítem sería demasiado fácil para estas personas, pero de nuevo, no hay más información disponible sobre exactamente qué tan fácil es realmente. Para poder hacer una estimación definitiva de un ítem muy fácil debemos encontrar por lo menos una persona medible que lo responda incorrectamente, y para un ítem muy difícil, al menos una persona medible debe contestarlo correctamente. Es decir, debemos "encerrar" al ítem entre personas de modo que al menos una ellas sea más capaz y al menos una sea menos capaz que lo que mide la dificultad del ítem. Por supuesto, solo una persona por debajo de un ítem muy fácil o uno muy difícil no dará una estimación muy precisa de la dificultad de ese ítem en particular.

Por lo tanto, tenemos datos insuficientes en nuestro ejemplo como para evaluar los ítems extremos 1, 2, 3 y 18. Sabemos que los ítems 1, 2 y 3 parecen ser muy fáciles y que el ítem 18 parece ser muy difícil para estas personas, pero no tenemos suficiente información para especificar estimaciones definitivas de las dificultades de estos cuatro ítems.

En cuanto a las personas extremas, ¿Puede decirse que las personas con puntaje cero no saben nada? ¿Un puntaje de 100% es indicativo de personas que "lo saben todo" o tal vez solo han respondido preguntas fáciles? Para hacer una estimación definitiva de una persona, debemos "encerrarla" entre ítems que sean más fáciles y más difíciles que la habilidad de la persona.

Los puntajes límite de 0% y 100%, tanto para los ítems como para las personas, representan información incompleta. Nos dicen en qué dirección buscar una estimación de la habilidad de la persona o de la dificultad del ítem, pero no nos dicen qué tan lejos llegar en esa dirección. Para obtener suficiente información para hacer una estimación definitiva de la ubicación de la persona

**TABLA 2.4.1**

**RESPUESTAS ORDENADAS Y EDITADAS DE 34 PERSONAS A LOS 14 ÍTEMS**

| NOMBRE DE LA PERSONA | NOMBRE DEL ITEM | | | | | | | | | | | | | | PUNTAJE EDITADO DE LA PERSONA | PROPORCIÓN DE 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 5 | 7 | 6 | 9 | 8 | 10 | 11 | 13 | 12 | 14 | 15 | 16 | 17 | | |
| 25 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | .14 |
| 4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | .21 |
| 33 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | .21 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | .29 |
| 27 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | .29 |
| 11 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | .36 |
| 12 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | .36 |
| 17 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | .43 |
| 19 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | .43 |
| 30 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | .43 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 7 | .50 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 13 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 7 | .50 |
| 16 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 26 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 28 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 29 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 7 | .50 |
| 31 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | .57 |
| 18 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 8 | .57 |
| 14 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | .57 |
| 32 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | .57 |
| 20 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 8 | .57 |
| 21 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 9 | .64 |
| 22 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 9 | .64 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 9 | .64 |
| 34 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 9 | .64 |
| 15 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 10 | .71 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 11 | .79 |
| 24 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 11 | .79 |
| PUNTAJE EDITADO DEL ITEM | 32 | 31 | 31 | 30 | 30 | 27 | 24 | 12 | 7 | 6 | 3 | 1 | 1 | 1 | | |
| PROPORCIÓN DE 34 | .94 | .91 | .91 | .88 | .88 | .79 | .71 | .35 | .21 | .18 | .09 | .03 | .03 | .03 | | |

o del ítem en la variable latente, debemos encontrar algunos ítems muy fáciles y algunos muy difíciles para esas personas, así como personas más listas y otras más deficientes para esos ítems, de forma que cada ítem y cada persona quede encerrada entre observaciones. De esta forma podremos estimar dónde se encuentran en la variable medida.

### Paso 3. Obteniendo calibraciones iniciales de los ítems

A partir de la matriz de datos de la Tabla 2.4.1, construimos una distribución agrupada de los 10 puntajes incorrectos diferentes en los ítems y sus lógitos. También calculamos la media y la varianza de la distribución de los lógitos de estos ítems respecto de una prueba de 14 ítems, como se hace en la tabla 2.4.2.

| | EXPLICACIÓN DE LA TABLA 2.4.2 | NOTACIÓN Y FÓRMULAS |
|---|---|---|
| **Columna 1** | de la Tabla 2.4.2. Proporciona los nombres de los ítems de cada grupo de puntajes de los ítems. | |
| **Columna 2** | da el puntaje del ítem que caracteriza a cada grupo de puntajes. Debido a que hay 10 diferentes grupos de puntajes en este ejemplo, G=10 y el índice los grupos de puntajes corre de 1 a 10. | $s_i \qquad i = 1,G$ |
| **Columna 3** | presenta la frecuencia de los ítems en cada puntaje. La suma de esta frecuencias a lo largo de los G=10 grupos de puntajes de los ítems, se aplica a los L=14 ítems que se están calibrando | $f_i$<br>$L = \sum_i^G f_i$ |
| **Columna 4** | convierte los puntajes de los ítems en proporciones de aciertos en la muestra de N=34 personas. | $p_i = s_i/N$ |
| **Columna 5** | es la conversión de la proporción de aciertos $p_i$ en la proporción de fallas $1\text{-}p_i$. | $1 - p_i = (N - s_i)/N$ |
| **Columna 6** | convierte la proporción de respuestas incorrectas en lógitos. Cada grupo de puntajes de los ítems en lógitos es el logaritmo natural de la proporción de respuestas incorrectas dividida por su proporción de aciertos.<br><br>Puede facilitarse esta conversión con los valores de lógitos: $\ell n[p/(1-p)]$ dada en la tabla 2.4.3. | $x_i = \ell n\left[(1 - p_i)/p_i\right]$ |
| **Columna 7** | es el producto de la frecuencia del ítem y el lógito de respuestas incorrectas. | $f_i\, x_i$ |
| **Columna 8** | es el producto de las frecuencias de respuestas incorrectas en el ítem y el lógito, elevado al cuadrado. | $f_i\, x_i^2$ |

**TABLA 2.4.2**

**DISTRIBUCIÓN AGRUPADA DE LOS 10 PUNTAJES DIFERENTES DE LOS ÍTEMS POR 34 PERSONAS**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| INDICE DE GRUPO DE PUNTAJE DEL ITEM | NOMBRE DEL ITEM | PUNTAJE DEL ITEM | FRECUENCIA DEL ITEM | PROPORCIÓN CORRECTA | PROPORCIÓN INCORRECTA | LÓGITO INCORRECTO | FRECUENCIA POR LÓGITO | PUNTAJE DE FRECUENCIA POR LÓGITO | CALIBRACIÓN INICIAL DEL ITEM |
| $i$ | | $s_i$ | $f_i$ | $p_i = s_i/N$ | $1 - p_i$ | $x_i = \ell n \left[\frac{1-p_i}{p_i}\right]$ | $f_i x_i$ | $f_i x_i^2$ | $d_i^o = x_i - \bar{x}.$ |
| 1 | 4 | 32 | 1 | .94 | .06 | - 2.75* | - 2.75 | 7.56 | - 2.94 |
| 2 | 5, 7 | 31 | 2 | .91 | .09 | - 2.31 | - 4.62 | 10.67 | - 2.50 |
| 3 | 6, 9 | 30 | 2 | .88 | .12 | - 1.99 | - 3.98 | 7.92 | - 2.18 |
| 4 | 8 | 27 | 1 | .79 | .21 | - 1.32 | - 1.32 | 1.74 | - 1.51 |
| 5 | 10 | 24 | 1 | .71 | .29 | - 0.90 | - 0.90 | 0.81 | - 1.09 |
| 6 | 11 | 12 | 1 | .35 | .65 | + 0.62 | + 0.62 | 0.38 | + 0.43 |
| 7 | 13 | 7 | 1 | .21 | .79 | + 1.32 | + 1.32 | 1.74 | + 1.13 |
| 8 | 12 | 6 | 1 | .18 | .82 | + 1.52 | + 1.52 | 2.31 | + 1.33 |
| 9 | 14 | 3 | 1 | .09 | .91 | + 2.31 | + 2.31 | 5.34 | + 2.12 |
| 10 | 15, 16, 17 | 1 | 3 | .03 | .97 | + 3.48 | +10.44 | 36.33 | + 3.29 |

$\sum_i^{10} f_i = 14$ $\quad N = 34$ $\quad \sum_i^{10} f_i x_i = 2.64$ $\quad \sum_i^{10} f_i x_i^2 = 74.81$

$\bar{x}. = \sum_i^{10} f_i x_i / \sum_i^{10} f_i = 2.64/14 = 0.19$

$U = (\sum_i^{10} f_i x_i^2 - \sum_i^{10} f_i x_i^2) / (\sum_i^{10} f_i - 1) = (74.81 - 0.51)/13 = 5.72$

Valor del atajo U'

14/6 = 2.33 $\quad$ Superior = 3.29 x 2.33 = 7.67 $\quad$ Inferior = ( - 2.94) + ( - 2.50 x 1.33) = − 6.27 $\quad$ $U' = \{2\,[7.67 - (-6.27)]/13\}^2 = 4.6$

*Estos valores se obtienen de la Tabla 2.4.3 donde $\ell n = [.06/.94] = -2.75$. Si se hacen estos cálculos con $s_i$ y N como $\ell n[(N - s_i)/s_i]$ entonces $\ell n[2/32] = -2.77$. La diferencia entre - 2.75 y -2.77 se debe al redondeo en $s_i/N = 32/34 = 0.941176 \ldots \simeq 0.94$.

**TABLA 2.4.3**

## LÓGITOS A PARTIR DE PROPORCIONES

Lógito= ℓn [Proporción / (1 - Proporción)]

| PROPORCIÓN[1] | LÓGITO | PROPORCIÓN | LÓGITO | PROPORCIÓN | LÓGITO | PROPORCIÓN | LÓGITO |
|---|---|---|---|---|---|---|---|
| .01 | – 4.60 | .26 | – 1.05 | 51 | 0.04 | .76 | 1.15 |
| .02 | – 3.89 | .27 | – 0.99 | .52 | 0.08 | .77 | 1.21 |
| .03 | – 3.48 | .28 | – 0.94 | .53 | 0.12 | .78 | 1.27 |
| .04 | – 3.18 | .29 | – 0.90 | .54 | 0.16 | .79 | 1.32 |
| .05 | – 2.94 | .30 | – 0.85 | .55 | 0.20 | .80 | 1.39 |
| .06 | – 2.75 | .31 | – 0.80 | .56 | 0.24 | .81 | 1.45 |
| .07 | – 2.59 | .32 | – 0.75 | .57 | 0.28 | .82 | 1.52 |
| .08 | – 2.44 | .33 | – 0.71 | .58 | 0.32 | .83 | 1.59 |
| .09 | – 2.31 | .34 | – 0.66 | .59 | 0.36 | .84 | 1.66 |
| .10 | – 2.20 | .35 | – 0.62 | .60 | 0.41 | .85 | 1.73 |
| .11 | – 2.09 | .36 | – 0.58 | .61 | 0.45 | .86 | 1.82 |
| .12 | – 1.99 | .37 | – 0.53 | .62 | 0.49 | .87 | 1.90 |
| .13 | – 1.90 | .38 | – 0.49 | .63 | 0.53 | .88 | 1.99 |
| .14 | – 1.82 | .39 | – 0.45 | .64 | 0.58 | .89 | 2.09 |
| .15 | – 1.73 | .40 | – 0.41 | .65 | 0.62 | .90 | 2.20 |
| .16 | – 1.66 | .41 | – 0.36 | .66 | 0.66 | .91 | 2.31 |
| .17 | – 1.59 | .42 | – 0.32 | .67 | 0.71 | .92 | 2.44 |
| .18 | – 1.52 | .43 | – 0.28 | .68 | 0.75 | .93 | 2.59 |
| .19 | – 1.45 | .44 | – 0.24 | .69 | 0.80 | .94 | 2.75 |
| .20 | – 1.39 | .45 | – 0.20 | .70 | 0.85 | .95 | 2.94 |
| .21 | – 1.32 | .46 | – 0.16 | .71 | 0.90 | .96 | 3.18 |
| .22 | – 1.27 | .47 | – 0.12 | .72 | 0.94 | .97 | 3.48 |
| .23 | – 1.21 | .48 | – 0.08 | .73 | 0.99 | .98 | 3.89 |
| .24 | – 1.15 | .49 | – 0.04 | .74 | 1.05 | .99 | 4.60 |
| .25 | – 1.10 | .50 | – 0.00 | .75 | 1.10 | | |

[1] Para los puntajes de las personas esta "proporción" se obtiene con el número de respuestas correctas r divididas por el número de ítems de la prueba L. Entonces la medida en lógitos de la habilidad de las personas es $\ell n[(r/L)/(1-r/L)] = \ell n[r/(L-r)]$.

Para los puntajes de los ítems esta "proporción" se obtiene con el número de respuestas incorrectas (N-s) dividida por el tamaño de la muestra N. Entonces la dificultad del ítem en lógitos es $\ell n\{[(N-s)/N]/[1-(N-s)/N]\} = \ell n[(N-s)/s]$.

**TABLA 2.4.4**

**DISTRIBUCIÓN AGRUPADA DE PUNTAJES OBSERVADOS DE LAS PERSONAS EN 14 ÍTEMS.**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| POSIBLE PUNTAJE | FRECUENCIA PERSONAL | PROPORCIÓN CORRECTA | LÓGITO CORRECTO | FRECUENCIA POR LÓGITO | PUNTAJE DE FRECUENCIA POR LÓGITO | MEDIDA INICIAL DE LA PERSONA |
| $r$ | $n_r$ | $p_r = r/L$ | $y_r = \ell n\left[\frac{p_r}{1-p_r}\right]$ | $n_r y_r$ | $n_r y_r^2$ | $b_r^o = y_r$ |
| 1 | 0 | .07 | - 2.59* | 0.00 | 0.00 | - 2.59 |
| 2 | 1 | .14 | - 1.82 | - 1.82 | 3.31 | - 1.82 |
| 3 | 2 | .21 | - 1.32 | - 2.64 | 3.48 | - 1.32 |
| 4 | 2 | .29 | - 0.90 | - 1.80 | 1.62 | - 0.90 |
| 5 | 2 | .36 | - 0.58 | - 1.16 | 0.67 | - 0.58 |
| 6 | 3 | .43 | - 0.28 | - 0.84 | 0.24 | - 0.28 |
| 7 | 12 | .50 | 0.00 | 0.00 | 0.00 | 0.00 |
| 8 | 5 | .57 | 0.28 | 1.40 | 0.39 | 0.28 |
| 9 | 4 | .64 | 0.58 | 2.32 | 1.35 | 0.58 |
| 10 | 1 | .71 | 0.90 | 0.90 | 0.81 | 0.90 |
| 11 | 2 | .79 | 1.32 | 2.64 | 3.48 | 1.32 |
| 12 | 0 | .86 | 1.82 | 0.00 | 0.00 | 1.82 |
| 13 | 0 | .93 | 2.59 | 0.00 | 0.00 | 2.59 |

$\sum_r^{13} n_r = 34$ $\quad L = 14$ $\quad \sum_r^{13} n_r y_r = -1.00$ $\quad \sum_r^{13} n_r y_r^2 = 15.35$

$y_\cdot = \sum_r^{13} n_r y_r / \sum_r^{13} n_r = -1.00/34 = -0.03$

$V = \left(\sum_r^{13} n_r y_r^2 - \sum_r^{13} n_r y_\cdot^2\right) / \left(\sum_r^{13} n_r - 1\right) = (15.35 - 0.03)/33 = 0.46$

Valor del atajo V'

34/6 = 5.67 $\quad$ Superior = (1.32 x 2) + 0.90 + (2.67 x 0.58) = 5.09 $\quad$ Inferior = - 1.82 + (-1.32 x 2) + (-0.90 x 2) + (-0.58 x 0.67) = - 6.65

$V' = \{2\,[5.09 - (-6.65)]/33\}^2 = 0.5$

* Estos valores proceden de la Tabla 2.4.3 donde $\ell n[.07/.93]=-2.59$. Si se usan r y L como en $\ell n[r/(L-r)]=-2.59$ entonces $\ell n[1/13]=-2.56$.
La diferencia entre -2.59 y -2.56 se debe al redondeo r/L=1/14=0.071428... ≃0.07.

| | | |
|---|---|---|
| | Se calcula en la Columna 6 la media y la varianza en lógitos para el ítem a partir de los valores en las Columnas 7 y 8 y se presenta debajo de ambas. | $x. = \sum_{i}^{G} f_i \; x_i / L$<br>$U = \left( \sum_{i}^{G} f_i \; x_i^2 - L x.^2 \right) / (L - 1)$ |
| **Columna 9** | Da los valores de la Columna 6 centrada restando su media. Estas son las calibraciones iniciales de los ítems, listas para ser corregidas por el efecto de dispersión de la muestra. | $d_i^{\,o} = x_i - x.$ |

En este ejemplo, la media y la varianza han sido calculadas a partir de los valores de las Columnas 7 y 8. La calibración a mano se puede facilitar todavía más con una expresión que sirve de atajo para la desviación estándar propuesta por Mason y Odeh (1968). Para hacer esto, se deben sumar los lógitos de los ítems de la Columna 9 (o de la Columna 6) para los sextos primeros y últimos de los ítems ordenados por dificultad, y se toma el cuadrado del doble de la diferencia de estas sumas divididas por el número de ítems menos uno.

Para los datos en la Tabla 2.4.2:

a. Una sexta parte de los ítems es 14/6 = 7/3, o 2.33 ítems en cada extremo.

b. Los lógitos de respuestas incorrectas de los ítems para los tres ítems superiores en la Columna 9 son 3.29, que multiplicado por 2.33 es $3.29 \times 2.33 = 7.67$.

c. El lógito de repuestas incorrectas del ítem inferior es -2.94 y para los siguientes dos ítems es -2.50. Entonces debe tomarse el ítem más bajo, -2.94, más 7/3 - 3/3 = 4/3 de los siguientes dos ítems, que proporciona $(-2.94) + (-2.50 \times 1.33) = -6.27$.

d. La diferencia entre 7.67 y –6.27 es 13.94.

e. El doble de esta cantidad dividida por la cantidad de ítems menos uno y elevado al cuadrado se convierte en una estimación de la varianza $[2(13.94)/13]^2 = 4.6$.

Este valor obtenido con el atajo de 4.6 es algo menor que 5.7 pero el número de ítems es pequeño y la distribución es más plana que la distribución normal asumida por el atajo.

La finalización de los pasos en la Tabla 2.4.2 proporciona los valores iniciales para las dificultades de los ítems en preparación para el ajuste que compensará el efecto de la dispersión de la muestra.

### Paso 4. Obtención de medidas iniciales para la persona

En la Tabla 2.4.4, realizamos pasos idénticos a los de la distribución agrupada de puntajes de las personas para obtener la distribución de puntajes en lógitos de las personas y por lo tanto los valores iniciales para las habilidades que se asocian con cada puntaje posible en la prueba.

| EXPLICACIÓN DE LA TABLA 2.4.4 | | NOTACIÓN Y FÓRMULAS |
|---|---|---|
| **Columna 1** | de la Tabla 2.4.4. Proporciona el puntaje posible de las personas 1 a 13. | r = 1, L-1 |

| | EXPLICACIÓN DE LA TABLA 2.4.4 | NOTACIÓN Y FÓRMULAS |
|---|---|---|
| **Columna 2** | da la frecuencia de personas observadas en cada puntaje. El total de personas N = 34 es igual a la suma de estas frecuencias de r = 1 a r = 13. | $n_r$<br>$N = \sum_{r}^{L-1} n_r$ |
| **Columna 3** | es la proporción de cada puntaje en la prueba de L = 14 ítems. | $p_r = r / L$ |
| **Columna 4** | es el lógito de respuestas correctas de esa proporción usando la Tabla 2.4.3. | $y_r = \ell n[p_r/(1-p_r)]$ |
| **Columna 5** | es el producto de la frecuencia de las personas por el lógito de respuestas correctas. | $n_r\, y_r$ |
| **Columna 6** | es el producto de la frecuencia de las personas por el lógito de respuestas correctas elevado al cuadrado. | $n_r\, y_r^2$ |
| **Columna 7** | repite los valores de la Columna 4 porque en lo que respecta a esta prueba, los puntajes en lógitos ya están centrados respecto de los lógitos de los ítems. Estas son las medidas iniciales de la persona antes de la corrección para el ámbito de dificultades de la prueba. | $b_r^{\,o} = y_r$ |

La media y la varianza para la distribución de lógitos de puntajes de las personas se dan en la parte inferior de la Tabla 2.4.4, con ayuda del atajo para el cálculo aproximado de la varianza.

Obsérvese que estamos interesados no solo en los puntajes observados en esta muestra sino también en las mediciones implicadas por cualquier puntaje posible que pueda observarse en esta prueba de 14 ítems, los puntajes no observadas de 1, 12 y 13 se han agregado a la Tabla 2.4.4, junto con las medidas iniciales para estos puntajes. El modelo de medición especifica que las medidas son equivalentes a estos puntajes incluso cuando ninguna persona en la muestra realmente lo obtenga.

Para resumir el procedimiento que hemos realizado hasta ahora (dejaremos que cada ítem defina sus puntajes en cada grupo de ítems, con el objeto de simplificar la notación, de modo que el índice del ítem corra de 1 a 14 en lugar de los grupos de puntaje que corren de 1 a 10 ítems):

Para una prueba de L' ítems que se aplican a N' personas, eliminamos todos los ítems que nadie contesta correctamente y que nadie contesta incorrectamente, así como todas las personas que no tienen aciertos o que no tienen fallas, hasta que no queden ítems o personas con estas características.

Sea $s_i$ el número de personas que aciertan el ítem i, donde i = 1 a L y sea $n_r$ el número de personas que tienen r ítems correctos, para r = 1 a L – 1, calculamos la media y la varianza sobre los ítems de los logaritmos del momio de respuestas incorrectas (o lógitos de los ítems respondidos incorrectamente) en la muestra para cada uno de los L ítems y también calculamos la media y la varianza sobre las personas de los logaritmos de momio de aciertos (o los lógitos de puntajes correctos) en la prueba para cada una de las N personas.

Obtenemos, para cada ítem i, su lógito de respuestas incorrectas en la muestra de N personas

$$x_i = \ell n[(N - s_i)/s_i]$$

y la media y la varianza sobre los lógitos de los L ítems.

$$x. = \sum_i^L x_i / L$$

$$U = \sum_i^L (x_i - x.)^2 / (L - 1)$$

Y obtenemos para cada puntaje r el lógito de respuestas correctas en la prueba de L ítems,

$$y_r = \ell n[r/(L - r)]$$

y la media y la varianza sobre los lógitos de los puntajes de las N personas.

$$y. = \sum_r^{L-1} n_r y_r / N$$

$$V = \sum_r^{L-1} n_r (y_r - y.)^2 / (N - 1)$$

Ahora estamos listos para ajustar las calibraciones y medidas iniciales en las Tablas 2.4.2 y 2.4.4 para los efectos locales de la distribución de las habilidades de las personas de la muestra y de la distribución de dificultades de los ítems de la prueba.

### Paso 5. Cálculo de los factores de expansión[2]

Calculamos los factores de expansión para las estimaciones iniciales de las calibraciones de los ítems y de las medidas de las personas, que corrigen las calibraciones de los ítems en función de la dispersión de la muestra y las medidas de las personas para el rango de los ítems de la prueba. De las Tablas 2.4.2 y 2.4.4 tenemos U = 5.72 y V = 0.46 (obtenidos con los valores del atajo U'= 4.6 y V'= 0.5).

a. El factor de expansión de la habilidad de las personas debido al rango de la prueba es:

$$x = \left[\frac{1 + U/2.89}{1 - UV/8.35}\right]^{1/2}$$

$$= \left[\frac{1 + 5.72/2.89}{1 - (5.72)(0.46)/8.35}\right]^{1/2}$$

$$= \left[\frac{2.98}{0.68}\right]^{1/2} = 2.09$$

o con el valor obtenido por el atajo:

$$x' = \left[\frac{1 + U'/2.9}{1 - U'V'/8.4}\right]^{1/2}$$

$$= \left[\frac{1 + 4.6/2.9}{1 - (4.6)(0.5)/8.4}\right]^{1/2}$$

$$= \left[\frac{2.59}{0.73}\right]^{1/2} = 1.9$$

[2] Para una explicación, ver el Capítulo Uno, Sección 1.6.

**TABLA 2.4.5**

**ESTIMACIONES FINALES DE LAS DIFICULTADES DE LOS ÍTEMS PARA 34 PERSONAS.**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| GRUPO DE PUNTAJE DEL ITEM | NOMBRE DEL ITEM | CALIBRACIÓN INICIAL DEL ITEM | FACTOR DE EXPANSIÓN DE DISPERSIÓN | CALIBRACION CORREGIDA DEL ITEM | PUNTAJE DEL ITEM | ERROR ESTANDAR DE CALIBRACIÓN |
| i | | $d_i^\circ$ | Y | $d_i = Yd_i^\circ$ | $s_i$ | SE ($d_i$) |
| 1 | 4 | − 2.94 | 1.31 | − 3.85 | 32 | .95 |
| 2 | 5, 7 | − 2.50 | 1.31 | − 3.28 | 31 | .79 |
| 3 | 6, 9 | − 2.18 | 1.31 | − 2.86 | 30 | .70 |
| 4 | 8 | − 1.51 | 1.31 | − 1.98 | 27 | .56 |
| 5 | 10 | − 1.09 | 1.31 | − 1.43 | 24 | .49 |
| 6 | 11 | + 0.43 | 1.31 | 0.56 | 12 | .47 |
| 7 | 13 | + 1.13 | 1.31 | 1.48 | 7 | .56 |
| 8 | 12 | + 1.33 | 1.31 | 1.74 | 6 | .59 |
| 9 | 14 | + 2.12 | 1.31 | 2.78 | 3 | .79 |
| 10 | 15, 16, 17 | + 3.29 | 1.31 | 4.31 | 1 | 1.33 |

N = 34

$$ES(d_i) = Y[N/s_i(N - s_i)]^{1/2}$$

b. El factor de expansión para las dificultades de los ítems debido a la dispersión de la muestra es:

$$Y = \left[\frac{1 + V/2.89}{1 - UV/8.35}\right]^{1/2}$$

$$= \left[\frac{1 + 0.46/2.89}{1 - (5.72)(0.46)/8.35}\right]^{1/2}$$

$$= \left[\frac{1.16}{0.68}\right]^{1/2} = 1.31$$

o con el valor obtenido por el atajo:

$$Y' = \left[\frac{1 + V'/2.9}{1 - U'V'/8.4}\right]^{1/2}$$

$$= \left[\frac{1 + 0.5/2.9}{1 - (4.6)(0.5)/8.4}\right]^{1/2}$$

$$= \left[\frac{1.17}{0.73}\right]^{1/2} = 1.3$$

### Paso 6. Corrección de las Calibraciones de los Ítems por Efecto de la Dispersión de la Muestra

De la Tabla 2.4.5 obtenemos las calibraciones finales corregidas de los ítems y su error estándar con base en el factor Y de expansión por dispersión de la muestra.

| | EXPLICACIÓN DE LA TABLA 2.4.5 | NOTACIÓN Y FÓRMULAS |
|---|---|---|
| **Columna 1** | contiene el nombre del ítem. | |
| **Columna 2** | repite las calibraciones iniciales de los ítems de la Columna 9 de la Tabla. 2.4.2. (Recuérdese que cuando los ítems se agrupan por puntajes el índice i corre de 1 a G, total de grupos de puntaje de ítems, en lugar de 1 a L que es el índice de los ítems individuales). | $d_i^o = x_i - x.$ $\quad i = 1, G$ |
| **Columna 3** | es el factor de expansión de la dificultad de los ítems Y = 1.31 por la dispersión de la muestra. | $Y$ |
| **Columna 4** | son las calibraciones corregidas de los ítems, obtenidas al multiplicar cada valor inicial de la Columna 2 por el factor de expansión 1.31. | $d_i = Y\, d_i^o$ $= Y(x_i - x.)$ |
| **Columna 5** | nos recuerda el número de personas que acertaron los ítems en cada grupo de puntajes de respuestas correctas. | $s_i$ |

**TABLA 2.4.6**

**ESTIMADOS FINALES DE LAS MEDIDAS DE LAS PERSONAS PARA TODOS LOS PUNTAJES POSIBLES DE LOS 14 ÍTEMS.**

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| POSIBLE PUNTAJE DE LA PRUEBA | MEDIDA INICIAL | FACTOR DE EXPANSIÓN DEL RANGO DE LA RUEBA | MEDIDAS CORREGIDAS | ERROR ESTANDAR DE LAS MEDIDAS CORREGIDAS |
| r | $b_r^o$ | X | $b_r = Xb_r^o$ | ES ($b_r$) |
| 1 | - 2.59 | 2.09 | – 5.41 | 2.17 |
| 2 | - 1.82 | 2.09 | – 3.80 | 1.60 |
| 3 | - 1.32 | 2.09 | – 2.76 | 1.36 |
| 4 | - 0.90 | 2.09 | – 1.88 | 1.24 |
| 5 | - 0.58 | 2.09 | – 1.21 | 1.17 |
| 6 | - 0.28 | 2.09 | – 0.59 | 1.13 |
| 7 | 0.00 | 2.09 | 0.00 | 1.12 |
| 8 | 0.28 | 2.09 | 0.59 | 1.13 |
| 9 | 0.58 | 2.09 | 1.21 | 1.17 |
| 10 | 0.90 | 2.09 | 1.88 | 1.24 |
| 11 | 1.32 | 2.09 | 2.76 | 1.36 |
| 12 | 1.82 | 2.09 | 3.80 | 1.60 |
| 13 | 2.59 | 2.09 | 5.41 | 2.17 |

L = 14

$$ES\ (b_r) = X\,[L/r(L-r)]^{1/2}$$

| | EXPLICACIÓN DE LA TABLA 2.4.5 | NOTACIÓN Y FÓRMULAS |
|---|---|---|
| **Columna 6** | es el error estándar de las calibraciones de los ítems. | $ES(d_i) = Y[N/s_i(N - s_i)]^{1/2}$ |

### Paso 7. Corrección de las medidas de las persona para el efecto del rango de la prueba.

En la Tabla 2.4.6 obtenemos las medidas corregidas finales de las personas y sus errores estándar con base en el factor de expansión X del rango de la prueba.

| | EXPLICACIÓN DE LA TABLA 2.4.6 | NOTACIÓN Y FÓRMULAS |
|---|---|---|
| **Columna 1** | Contiene todos los puntajes posibles, porque queremos obtener medidas para cada posible puntaje en la prueba de 1 a L – 1 en todos los resultados realmente observados. | $r = 1, L - 1$ |
| **Columna 2** | Repite las medidas iniciales de las personas de la Columna 7 de la Tabla 2.4.4. | $b_r^o = y_r$ |
| **Columna 3** | Es el factor de expansión de las habilidades de las personas X = 2.09 debido al rango de dificultades de la prueba. | X |
| **Columna 4** | Son las medidas corregidas para las personas obtenidas al multiplicar cada valor inicial de la Columna 2 por el factor de expansión 2.09. | $b_r = Xb_r^o = Xy_r$ |
| **Columna 5** | Es el error estándar de las medidas corregidas de la persona. | $ES(b_r) = X[L/r(L - r)]^{1/2}$ |

## 2.5 DISCUSIÓN

El procedimiento de análisis de ítems del algoritmo PROX se describió detalladamente no solo porque permite obtener las calibraciones de los ítems y, por lo tanto, las medidas de las personas, sino también porque es el método de análisis más simple que conjuga de manera lógica y directa la interacción entre ítems y personas. La idea primordial en que se basa este análisis es que la probabilidad de éxito está determinada por la habilidad de la persona y la dificultad del ítem. Se supone que una persona más capaz siempre tiene mayores posibilidades de éxito en cualquier ítem que otra persona menos capaz. También se parte del supuesto de que cualquier persona siempre debe tener una mayor posibilidad de éxito en un ítem fácil que en uno difícil. En la medida en que éste sea el caso, la probabilidad del éxito de cualquier persona en cualquier ítem se puede especificar como una consecuencia de la diferencia entre la posición de la persona en una sola variable y la posición en esa misma variable. Éste es el modelo de Rasch para el análisis de los ítems y la construcción de pruebas y, de hecho, es el modelo fundamental que está implícito en el análisis de ítems de todos quienes trabajan con puntajes no ponderados (Andersen, 1977).

Toda la información observada sobre la posición de una persona en la variable, por ejemplo, su habilidad, se supone que está expresada en sus respuestas al conjunto de ítems que se le presentan en la prueba, haciendo un conteo no ponderado del número de ítems que responde correctamente. Para la dificultad de los ítems, se supone que la información observada está completamente contenida en el conteo no ponderado de las personas de la muestra que responden correctamente a ese ítem.

Por supuesto, el modelo de interacción entre la persona y el ítem es una idealización y solo pretende obtener una aproximación a lo que sucede en la realidad. Todas las ideas, sin embargo, son a fin de cuentas, aproximaciones o abstracciones de la experiencia. Su valor puede juzgarse solamente en términos de su utilidad, es decir, de su relevancia demostrable respecto de la situación en estudio y de su simplicidad. Este capítulo ha ilustrado la simplicidad y la conveniencia potencial del análisis de ítems con el modelo de Rasch. Su utilidad puede testimoniarse por miles de aplicaciones. Nuestra siguiente tarea es mostrar cómo se obtienen resultados similares a los que acabamos de obtener a mano con el procedimiento PROX, cuando se trabaja utilizando la computadora con UCON que es un procedimiento más elaborado.

# 3. CALIBRACIÓN DE ÍTEMS POR COMPUTADORA

## 3.1 INTRODUCCIÓN

En este capítulo, mostramos y describimos la salida de un programa de computadora para la calibración de ítems con el modelo de Rasch utilizando los procedimientos de estimación PROX y UCON (Wright y Panchapakesan, 1969; Wright y Douglas, 1975b, 1977a, 1977b). Los datos de la Prueba de los Cubos de Knox que analizamos a mano en el Capítulo 2 se usan aquí con fines ilustrativos. Los procedimientos de estimación son realizados por el programa informático BICAL (Wright y Mead, 1976[3]). La precisión y la utilidad de la calibración manual descrita en el Capítulo 2 se evalúan comparando las estimaciones a "mano" con los que produce al análisis por computadora de los mismos datos. La idea de dar a conocer los pasos que podrían realizarse manualmente tiene el propósito de facilitar la comprensión y el uso de la salida del programa de computadora.

Nos moveremos poco a poco por la salida del programa para destacar su organización y uso. BICAL produce los resultados de las Tablas 3.2.1 a 3.2.9. Los valores estadísticos para los ítems y las personas obtenidos con el algoritmo PROX a mano y por computadora se comparan en las Tablas 3.3.1 a 3.3.5.

## 3.2 SALIDA DE BICAL DEL ANÁLISIS CON PROX DE LOS DATOS DE LA PRUEBA DE LOS CUBOS DE KNOX.

La primera página de la salida, en la Tabla 3.2.1, recapitula las especificaciones de control necesarias para aplicar la versión 1976 de BICAL a este proyecto de calibración (para más detalles consulte el manual que acompaña al programa BICAL). En la parte superior tenemos el título del proyecto, los diferentes parámetros de control, el registro de columnas leídas, la clave de respuestas correctas y una copia del registro de la primera persona leída. Finalmente, esta salida reporta que hay 18 ítems y 34 personas que entraron en este análisis.

**TABLA 3.2.1**

**ESPECIFICACIONES DE CONTROL DEL PROGRAMA**

**PRUEBA DE LOS CUBOS DE KNOX**

**CONTROL DE PARÁMETROS**

| NITEM | NGROP | MINSC | MAXSC | LREC | KCAB | PUNTAJE |
|---|---|---|---|---|---|---|
| 18 | 10 | 1 | 17 | 21 | 1 | 0 |

**COLUMNAS SELECCIONADAS**

```
          1         2         3         4
1********0*********0*********0*********0*****

   11111111111111111

   11111111111111111
```

**CLAVE**

**PRIMERA PERSONA**

```
00111111100000000000
```

**NÚMERO DE ÍTEMS 18**

**NÚMERO DE PERSONAS 34**

[3] N.T. BICAL no está disponible actualmente, pero este capítulo es útil porque describe el algoritmo básico para el análisis por computadora.

Los parámetros de control para este proyecto fueron:

| | | |
|---|---|---|
| 1. Número de ítems (NITEM): | 18 | Originalmente había 18 ítems en la PCK. |
| 2. Tamaño de subgrupo más pequeño (NGROP): | 10 | Se deben formar subgrupos de al menos 10 personas para analizar el ajuste del ítem. |
| 3. Puntaje mínimo (MINSC): | 1 | Puntaje mínimo que se utilizará es 1. |
| 4. Puntaje máximo (MAXSC): | 17 | Puntaje máximo que se utilizará es 17. |
| 5. Longitud del registro a leer (LREC): | 21 | Los datos provienen de las primeras 21 posiciones del registro de persona. [La columna selecciona la tarjeta (enumerada en la Tabla 3.2.1 debajo de "COLUMNA SELECCIONADA") especifica las 18 columnas de las respuestas de la prueba.] |
| 6. Procedimiento de calibración (KCAB): | 1 | El procedimiento de calibración que se utilizará es PROX. [El código de selección es: 1 = PROX, 2 = UCON] |
| 7. Opción de calificación (SCORE): | 0 | Los datos ya están calificados.<br>[El código de control es: 0 = datos a calificar en forma dicótoma de acuerdo con la clave de respuestas correctas; 1 = los datos son enteros sucesivos; 2 = calificar como "correcto", si el valor de respuesta es igual o menor que la clave suministrada, en caso contrario se califica como "Incorrecto"; 3 = calificar como "correcto", si el valor de respuesta es igual o mayor que la clave suministrada, en caso contrario se califica como "incorrecto".] |

La Tabla 3.2.2 proporciona las frecuencias de respuesta de cada ítem para cada valor de respuesta. Esta tabla puede acomodar hasta cinco valores de respuesta que especifique el usuario. Un valor de columna “desconocido" registra la cuenta de todos los demás valores encontrados. La columna final es para la clave. La clave especifica el valor declarado como correcto cuando los datos aún no se han calificado. Como muestra la Tabla 3.2.2, los datos de la PCK se ingresaron como previamente calificados, por lo tanto la clave correcta está formada por un vector de “1” como se tiene en las Tablas 3.2.1 y 3.2.2. Cada ítem se identifica a la izquierda por su número secuencial en el orden original de los ítems de la prueba, como son leídos por BICAL. El nombre del ítem sirve como identificación y está formado por cuatro caracteres. Para la PCK se asignó el nombre del ítem de acuerdo con la cantidad de golpes requeridos.

La Tabla 3.2.2 permite examinar las respuestas observadas para detectar las discrepancias obvias respecto de nuestro plan de prueba y con ello sugerir posibles explicaciones para los desajustes graves. La distribución de las respuestas sobre los distractores en ítems de opción múltiple, puede revelar la influencia indebida de algunos distractores en particular. Los efectos que genera un tiempo insuficiente aparecen en la acumulación de respuestas en la columna DESCONOCIDO hacia el final de la prueba. Los efectos relativos a la inexperiencia en responder un examen aparece como acumulación de respuestas DESCONOCIDO en los ítems iniciales (ítems uno o dos), de la prueba.

Vemos nuevamente en la Tabla 3.2.2 lo que ya aprendimos en la Tabla 2.3.1, a saber, que los primeros tres ítems son contestados correctamente por las 34 personas, que el ítem 18 no fue respondido correctamente por ninguna persona y que hay un cambio rápido de respuestas que pasa de una mayoría correcta a una mayoría incorrecta entre los ítems 9 y 11. Dado que el nombre del ítem proporciona el número de golpes en la serie, vemos que este cambio ocurre cuando la acción pasa de una serie de cuatro golpes a una de cinco.

**TABLA 3.2.2**

**FRECUENCIA DE RESPUESTAS EN CADA ALTERNATIVA DE RESPUESTA**

| NÚM SECUEN-CIAL | NOMBRE DEL ÍTEM | FRECUENCIAS ALTERNATIVAS DE RESPUESTA 0 | 1 | | | | DESCO-NOCIDO | CLAVE |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 34 | 0 | 0 | 0 | 0 | 1 |
| 2 | 2 | 0 | 34 | 0 | 0 | 0 | 0 | 1 |
| 3 | 3 | 0 | 34 | 0 | 0 | 0 | 0 | 1 |
| 4 | 3 | 2 | 32 | 0 | 0 | 0 | 0 | 1 |
| 5 | 3 | 3 | 31 | 0 | 0 | 0 | 0 | 1 |
| 6 | 3 | 4 | 30 | 0 | 0 | 0 | 0 | 1 |
| 7 | 4 | 3 | 31 | 0 | 0 | 0 | 0 | 1 |
| 8 | 4 | 7 | 27 | 0 | 0 | 0 | 0 | 1 |
| 9 | 4 | 4 | 30 | 0 | 0 | 0 | 0 | 1 |
| 10 | 4 | 10 | 24 | 0 | 0 | 0 | 0 | 1 |
| 11 | 5 | 22 | 12 | 0 | 0 | 0 | 0 | 1 |
| 12 | 5 | 28 | 6 | 0 | 0 | 0 | 0 | 1 |
| 13 | 5 | 27 | 7 | 0 | 0 | 0 | 0 | 1 |
| 14 | 6 | 31 | 3 | 0 | 0 | 0 | 0 | 1 |
| 15 | 6 | 33 | 1 | 0 | 0 | 0 | 0 | 1 |
| 16 | 6 | 33 | 1 | 0 | 0 | 0 | 0 | 1 |
| 17 | 6 | 33 | 1 | 0 | 0 | 0 | 0 | 1 |
| 18 | 6 | 34 | 0 | 0 | 0 | 0 | 0 | 1 |

La Tabla 3.2.3 reporta el proceso de edición. Es un resumen del trabajo de la rutina de edición que elimina sucesivamente los registros de personas con puntaje cero o perfecto, así como los ítems contestados correctamente por todas las personas o que ninguna persona acertó. Este proceso continúa hasta que todas las personas o todos los ítems sean detectados y removidos. El proceso de edición conduce a la matriz final de respuestas ítem-persona que pasará a ser analizada.

La Tabla 3.2.3 muestra que inicialmente no había personas con puntajes perfectos o nulos, y que 18 ítems entraron en la corrida del programa, sin que ninguna persona obtuviera un puntaje inferior a 1 o superior a 17, dejando 34 personas para la calibración (La persona número 35 que aparece en la Tabla 2.3.1 ya había sido eliminada manualmente del conjunto de datos). Los ítems 1, 2 y 3 se eliminaron después en el proceso de edición porque todas las personas los respondieron correctamente y también se eliminó el ítem 18 porque nadie lo respondió correctamente. Después de esta edición, la muestra utilizada para esta calibración sigue siendo de 34 personas, pero ahora solo quedan los 14 ítems que se pueden calibrar, con el mismo puntaje mínimo de 1 pero con un nuevo puntaje máximo de 13.

La Tabla 3.2.4 muestra la distribución de las personas sobre los puntajes de la PCK. La escala del histograma se ajusta de acuerdo con el factor de escala impreso debajo del gráfico. La distribución de puntajes de las personas brinda una imagen de cómo respondió la muestra de personas al conjunto de ítems. También muestra qué tan bien "apuntan" estos ítems a las personas y cuán relevantes fueron las personas seleccionadas para este calibración. Para obtener una mejor calibración, las personas deberían estar distribuidas más o menos uniformemente en un rango de puntajes, alrededor y por encima del centro de la prueba. En nuestra muestra, vemos una distribución simétrica alrededor de una moda en el puntaje de 7.

**TABLA 3.2.3**

**PROCESO DE EDICIÓN**

**PRUEBA DE LOS CUBOS DE KNOX**

| | | |
|---|---|---|
| NÚMERO DE PUNTAJES NULOS | | 0 |
| NÚMERO DE PUNTAJES PERFECTOS | | 0 |
| NÚMERO DE ÍTEMS SELECCIONADOS | | 18 |
| NÚMERO DE ÍTEMS NOMINALES | | 18 |
| PERSONAS POR DEBAJO DE | 1 | 0 |
| PERSONAS POR ARRIBA DE | 17 | 0 |
| PERSONAS RESTANTES | | 34 |
| TOTAL DE PERSONAS | | 34 |

ÍTEMS RECHAZADOS

| NÚMERO DEL ÍTEM | NOMBRE DEL ÍTEM | RESPUESTAS CORRECTAS | |
|---|---|---|---|
| 1 | 2 | 34 | PUNTAJE MÁS ALTO |
| 2 | 2 | 34 | PUNTAJE MÁS ALTO |
| 3 | 3 | 34 | PUNTAJE MÁS ALTO |
| 18 | 6 | 0 | PUNTAJE MÁS BAJO |

| | | |
|---|---|---|
| PERSONAS ELIMINADAS | = | 0 |
| PERSONAS RESTANTES | = | 34 |
| ÍTEMS ELIMINADOS | = | 4 |
| ÍTEMS RESTANTES | = | 14 |
| PUNTAJE MÍNIMO | = | 1 |
| PUNTAJE MÁXIMO | = | 13 |

**TABLA 3.2.4**

**DISTRIBUCIÓN DE HABILIDADES DE LA MUESTRA DE PERSONAS**

| PUNTAJE | CONTEO | PROPORCIÓN | DISTRIBUCIÓN DE HABILIDADES<br>2 4 6 8 10<br>*********0*********0*********0*********0*********0 |
|---|---|---|---|
| 1 | 0 | 0.0 | |
| 2 | 1 | 0.03 | x |
| 3 | 2 | 0.06 | xx |
| 4 | 2 | 0.06 | xx |
| 5 | 2 | 0.06 | xx |
| 6 | 3 | 0.09 | xxx |
| 7 | 12 | 0.36 | xxxxxxxxxxxx |
| 8 | 5 | 0.15 | xxxxx |
| 9 | 4 | 0.12 | xxxx |
| 10 | 1 | 0.03 | x |
| 11 | 2 | 0.06 | xx |
| 12 | 0 | 0.0 | |
| 13 | 0 | 0.0 | |
| 14 | 0 | 0.0 | |

CADA X = 2.94 POR CIENTO

La Tabla 3.2.5 presenta como se distribuyen los ítems por su facilidad. La escala aparece en la parte baja de la gráfica. Los ítems 4 a 9 parecen bastante fáciles, siendo el ítem 8 el más difícil de ellos. El ítem 10 es ligeramente más difícil. El ítem 11 es mucho más difícil seguido de los ítems 12 y 13 todavía más difíciles. Por último, los ítems 15, 16 y 17 son tan difíciles que solamente una persona pudo contestarlos correctamente.

**TABLA 3.2.5**

**DISTRIBUCIÓN DE LA FACILIDAD DE LOS ÍTEMS DE LA PRUEBA**

| ÍTEM | CONTEO | PROPORCIÓN | DISTRIBUCIÓN DE FACILIDAD DE LOS ÍTEMS<br>2 4 6 8 10<br>*********0*********0*********0*********0*********0 |
|---|---|---|---|
| 4 | 32 | 0.94 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 5 | 31 | 0.91 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 6 | 30 | 0.88 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 7 | 31 | 0.91 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 8 | 27 | 0.79 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 9 | 30 | 0.88 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 10 | 24 | 0.71 | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx |
| 11 | 12 | 0.35 | xxxxxxxxxxxxxxxxx |
| 12 | 6 | 0.18 | xxxxxxxxx |
| 13 | 7 | 0.21 | xxxxxxxxxx |
| 14 | 3 | 0.09 | xxxx |
| 15 | 1 | 0.03 | x |
| 16 | 1 | 0.03 | x |
| 17 | 1 | 0.03 | x |

CADA X = 2.00 PORCIENTO

La Tabla 3.2.6 presenta la información estimada. En la parte superior están la dificultad y los factores de expansión de habilidad obtenidos con PROX. Observe que estos valores son idénticos a los que se obtuvieron manualmente en el Capítulo 2. En el cuerpo de la tabla, las primeras cuatro columnas tienen el número secuencial del ítem, el nombre del ítem, la dificultad del ítem y su error estándar.

**TABLA 3.2.6**

**CALIBRACIÓN CON PROX**

FACTOR DE EXPANSIÓN PARA LA DIFICULTAD 1.31
FACTOR DE EXPANSIÓN PARA LA HABILIDAD 2.10

| NÚMERO SECUENCIAL | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | ERROR ESTÁNDAR |
|---|---|---|---|
| 4 | 3 | −3.865 | 0.833 |
| 5 | 3 | −3.294 | 0.691 |
| 6 | 3 | −2.876 | 0.608 |
| 7 | 4 | −3.294 | 0.691 |
| 8 | 4 | −2.007 | 0.485 |
| 9 | 4 | −2.876 | 0.608 |
| 10 | 4 | −1.388 | 0.430 |
| 11 | 5 | 0.547 | 0.410 |
| 12 | 5 | 1.767 | 0.514 |
| 13 | 5 | 1.518 | 0.485 |
| 14 | 6 | 2.805 | 0.691 |
| 15 | 6 | 4.321 | 1.160 |
| 16 | 6 | 4.321 | 1.160 |
| 17 | 6 | 4.321 | 1.160 |

La tabla 3.2.7 muestra la medida en lógitos de la habilidad y su error estándar para cada puntaje en la PCK, con el número de personas de la muestra que obtienen cada puntaje. Podemos ver la frecuencia de respuestas de las personas de la muestra para cada puntaje bruto, su habilidad y el error estándar implicado por dicho puntaje. La media de habilidad de la muestra y la desviación estándar aparecen en el parte inferior de la tabla.

**TABLA 3.2.7**

**MEDICIÓN POR PROX**

**TABLA DE EQUIVALENCIA DEL PUNTAJE COMPLETO**

| PUNTAJE BRUTO | CONTEO | HABILIDAD DE LA PERSONA | ERROR ESTÁNDAR |
|---|---|---|---|
| 13 | 0 | 5.40 | 1.51 |
| 12 | 0 | 3.77 | 1.11 |
| 11 | 2 | 2.73 | 0.94 |
| 10 | 1 | 1.93 | 0.86 |
| 9 | 4 | 1.24 | 0.81 |
| 8 | 5 | 0.61 | 0.78 |
| 7 | 12 | 0.00 | 0.78 |
| 6 | 3 | −0.61 | 0.78 |
| 5 | 2 | −1.24 | 0.81 |
| 4 | 2 | −1.93 | 0.86 |
| 3 | 2 | −2.73 | 0.94 |
| 2 | 1 | −3.77 | 1.11 |
| 1 | 0 | −5.40 | 1.51 |

MEDIA DE HABILIDAD = -0.06

ERROR ESTÁNDAR DE HABILIDAD = 1.14

La Tabla 3.2.8 proporciona las curvas características de los ítems y estadísticos de ajuste. Las pruebas de ajuste se hacen dividiendo la muestra que participó en la calibración en subgrupos de habilidades por nivel de puntaje. Se hicieron tres grupos a partir de la muestra de la PCK, las 10 personas con puntajes de 1 a 6, las 12 personas en el puntaje 7 y las 12 personas con puntajes de 8 a 13. El control sobre el tamaño del grupo y, por lo tanto, el número de grupos utilizados, se determina con el parámetro de control NGROP. Una evaluación de la invariancia de la dificultad de los ítems sobre estos grupos de habilidades se hace comparando las estimaciones de dificultad de cada ítem sobre los diferentes grupos. Por lo tanto, las pruebas de ajuste son dependientes de la muestra. No obstante, una vez que las estimaciones de dificultad pasan estas pruebas, dichas estimaciones resultan independientes de la muestra, específicamente de la muestra que intervino para hacer los cálculos. Por supuesto, el hecho de que se obtenga un ajuste exitoso del ítem en una muestra no garantiza que se tendrá buen ajuste en otra. Sin embargo, como los grupos de habilidades dentro de una muestra determinada están ordenados por puntajes, obtenemos información sobre la estabilidad de las dificultades de los ítems a lo largo de diversas habilidades y, por lo tanto, podemos ver si nuestros ítems son invariantes sobre estos grupos de habilidades particulares, lo cual nos permite afirmar que nuestros ítems pueden ser utilizados como instrumentos objetivos de medición.

**TABLA 3.2.8**

**CURVAS CARACTERÍSTICAS DE LOS ÍTEMS Y ANÁLISIS DE AJUSTE CON PROX**

| | | CURVA CARACTERÍSTICA DEL ÍTEM | | | DIFERENCIA RESPECTO A LO ESPERADO | | | ANÁLISIS DE AJUSTE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NÚM SECUEN-CIAL | NOMBRE DEL ÍTEM | 1ER GRUPO | 2O GRUPO | 3ER GRUPO | 1ER GRUPO | 2O GRUPO | 3ER GRUPO | EN EL GRUPO | ENTRE GRUPOS | TOTAL | ÍNDICE DE DISCRIMI-NACIÓN | PUNTO BISERIAL |
| 4 | 3 | 0.80 | 1.00 | 1.00 | −0.05 | 0.02 | 0.01 | 0.51 | 0.33 | 0.49 | 1.30 | 0.40 |
| 5 | 3 | 0.70 | 1.00 | 1.00 | −0.08 | 0.04 | 0.01 | 0.65 | 0.58 | 0.64 | 1.35 | 0.42 |
| 6 | 3 | 0.70 | 0.92 | 1.00 | −0.02 | −0.03 | 0.02 | 0.96 | 0.26 | 0.90 | 1.07 | 0.40 |
| 7 | 4 | 0.90 | 0.83 | 1.00 | 0.12 | −0.13 | 0.01 | 1.52 | 3.89 | 1.73 | 0.48 | 0.23 |
| 8 | 4 | 0.40 | 0.92 | 1.00 | −0.16 | 0.04 | 0.04 | 0.43 | 1.06 | 0.48 | 1.49 | 0.69 |
| 9 | 4 | 0.60 | 1.00 | 1.00 | −0.12 | 0.05 | 0.02 | 0.21 | 0.95 | 0.27 | 1.44 | 0.61 |
| 10 | 4 | 0.30 | 0.75 | 1.00 | −0.14 | −0.05 | 0.08 | 0.80 | 1.15 | 0.83 | 1.40 | 0.54 |
| 11 | 5 | 0.0 | 0.33 | 0.67 | −0.12 | −0.03 | 0.01 | 0.70 | 0.82 | 0.71 | 1.04 | 0.54 |
| 12 | 5 | 0.0 | 0.17 | 0.33 | −0.04 | 0.02 | −0.06 | 0.76 | 0.35 | 0.72 | 0.80 | 0.42 |
| 13 | 5 | 0.0 | 0.0 | 0.58 | −0.05 | −0.18 | 0.14 | 0.26 | 2.38 | 0.44 | 1.49 | 0.58 |
| 14 | 6 | 0.0 | 0.08 | 0.17 | −0.02 | 0.03 | −0.04 | 0.98 | 0.23 | 0.91 | 0.81 | 0.20 |
| 15 | 6 | 0.0 | 0.0 | 0.08 | 0.00 | −0.01 | 0.02 | 0.17 | 0.18 | 0.17 | 1.33 | 0.33 |
| 16 | 6 | 0.0 | 0.0 | 0.08 | 0.00 | −0.01 | 0.02 | 0.17 | 0.18 | 0.17 | 1.33 | 0.33 |
| 17 | 6 | 0.0 | 0.0 | 0.08 | 0.00 | −0.01 | 0.02 | 0.17 | 0.18 | 0.17 | 1.33 | 0.33 |
| RANGO DE PUNTAJES | | 1–6 | 7–7 | 8–13 | N = 10 | 12 | 12 | 31 | 3 | 34 | GRADO DE LIBERTAD | |
| MEDIA DE HABILIDAD | | −1.74 | 0.00 | 1.28 | | | | 0.25 | 0.82 | 0.24 | ERROR ESTÁNDAR | |

En la sección "Curva característica del Ítem CCI" (o ICC por sus siglas en inglés) de la Tabla 3.2.8 tenemos la proporción de respuestas correctas dadas por cada grupo de habilidades para cada ítem. El rango de puntajes y la media de habilidad para cada grupo se tienen en la parte inferior de cada columna. Esperamos que estas CCI crezcan a medida que avanzamos de izquierda a derecha, de los grupos de puntajes de menor habilidad a los de mayor habilidad, como podemos ver que lo hace la mayoría de los casos de la Tabla 3.2.8. Sin embargo, el Ítem 7 muestra un patrón inverosímil. ¡La proporción de personas que contesta correctamente en el grupo de puntaje bajo es mayor que la proporción del grupo intermedio!

En el panel del medio de la Tabla 3.2.8 tenemos las diferencias en las proporciones de la CCI para cada grupo de habilidad entre los observados y los predichos por el modelo de Rasch. Aquí podemos ver dónde ocurren las mayores desviaciones proporcionales y qué dirección tienen. Nuevamente, el ítem 7 está desalineado de los otros ítems, especialmente para el grupo de habilidad más baja.

En la sección "Análisis de Ajuste" de la Tabla 3.2.8, tenemos una serie de ajustes cuadráticos medios. Estos estadísticos de ajuste son residuos medios cuadráticos estandarizados para las respuestas ítem-persona promediados sobre las personas y dividido en dos componentes, uno entre los grupos de habilidades y el otro dentro de cada grupo de habilidades. Estos promedios cuadráticos aumentan en magnitud por arriba del valor de referencia de 1 cuando la CCI observada se aleja de la CCI esperada, es decir, cuando muchas personas de alta habilidad fallan al responder un ítem fácil o demasiadas personas de baja habilidad responden correctamente un ítem difícil. La significancia estadística de los valores grandes se puede juzgar comparando la media cuadrática observada contra su valor esperado de 1 en términos de los errores estándar esperados que se presentan en la parte inferior de la tabla.

El cuadrado medio "total" evalúa el acuerdo general entre la variable definida por el ítem y la variable definida por todos los demás ítems en toda la muestra. Solo el ítem 7 está significativamente desalineado, con una media cuadrática observada de 1.73, más que tres veces su error estándar esperado de 0.24 por arriba de su valor esperado de 1.

La media cuadrática “entre grupos” (o intergrupal) evalúa el acuerdo entre lo observado por la curva característica del ítem y la curva del modelo de Rasch que mejor se ajusta sobre la habilidad de los grupos. Nuevamente, el ítem 7 está desalineado con una media cuadrática 3.89, más de tres veces su error estándar de 0.82 por arriba de 1.

La media cuadrática "dentro del grupo" (o intragrupal) resume el grado de desajuste que se tiene dentro de un grupo de habilidad una vez que el desajuste “entre grupos" ha sido eliminado del "total". El ítem 7 muestra un desajuste de 1.52 contra un valor esperado de 1 y un error estándar de 0.25.

El índice de discriminación que se muestra en la penúltima columna de la Tabla 3.2.8 describe la tendencia lineal de las desviaciones del modelo a través de grupos de habilidades expresadas alrededor del valor modelado de 1. Cuando este índice está cerca de 1, entonces las CCI observadas y esperadas están muy cercanas entre sí respecto de los puntos de referencia definidos por los grupos de habilidades.

Si el índice es sustancialmente menor que 1, entonces la CCI observada es más plana que la esperada y el ítem particular no logra diferenciar entre las habilidades, en tanto que otros ítems sí lo hacen. Esta condición, por supuesto, corre al parejo de una baja correlación punto-biserial, o correlación entre la respuesta del ítem y el puntaje total de la prueba. Sin embargo, el índice de discriminación está menos influenciado en su magnitud que la correlación punto-biserial dependiendo de qué tan centrada o dispersa es la habilidad de las personas en la muestra.

Si el índice es sustancialmente mayor que 1, entonces el ítem aparenta distinguir la habilidad más claramente que los ítems promedio de la prueba y debería investigarse la causa de esta inusual "discriminación". Casi siempre se descubre que es causada por la interacción local entre

una característica secundaria del ítem y una característica secundaria de la muestra, una condición dependiente de la muestra que, al identificarse, se considera como demasiado idiosincrática para ser útil en un sistema de medición en general.

Las estadísticas de ajuste en la Tabla 3.2.8 muestran que el Ítem 7 desajusta tanto "entre" como "dentro" de los grupos de habilidades y que su curva característica es muy plana. Ningún otro ítem muestra un desajuste significativo. El ítem 14 muestra una baja correlación punto-biserial pero sus estadísticos de ajuste no están desalineados y, al igual que los ítems 15, 16 y 17, su baja correlación punto-biserial se debe principalmente a su dificultad para estas personas. Observe cómo la magnitud de la correlación punto-biserial fluctúa ampliamente con el nivel de ICC, ¡independientemente de qué tan bueno sea el ajuste de los ítems!

Esto nos deja con el desajuste que observamos para el Ítem 7. ¿Qué debemos concluir acerca de este desajuste? ¿Se debe a una deficiencia general del Ítem 7 o se debe a una interacción entre el Ítem 7 y algunos patrones aberrantes de respuesta de las personas? Podría ocurrir que el Ítem 7 funcione satisfactoriamente con la mayoría de las personas y que el desajuste observado aquí se deba a respuestas irregulares de unas cuantas personas. Si ese fuera el caso, podríamos decidir conservar en la prueba el Ítem 7 y plantearnos, en cambio, la plausibilidad de los patrones de respuesta de estas pocas personas inusuales. Discutiremos más a detalle sobre el ajuste del ítem y de la persona en el Capítulo 4.

**TABLA 3.2.9**

**RESUMEN DE CALIBRACIÓN DE LOS ÍTEMS POR PROX**

**EN ORDEN CONSECUTIVO**

| NÚM SECUEN-CIAL | NOMBRE DEL ÍTEM | DIFICUL-TAD DEL ÍTEM | ÍNDICE DE DISCRIMI-NACIÓN | AJUSTE MEDIA CUA-DRÁTICA |
|---|---|---|---|---|
| 4 | 3 | −3.87 | 1.30 | 0.49 |
| 5 | 3 | −3.29 | 1.35 | 0.64 |
| 6 | 3 | −2.88 | 1.07 | 0.90 |
| 7 | 4 | −3.29 | 0.48 | 1.73 |
| 8 | 4 | −2.01 | 1.49 | 0.48 |
| 9 | 4 | −2.88 | 1.44 | 0.27 |
| 10 | 4 | −1.39 | 1.40 | 0.83 |
| 11 | 5 | 0.55 | 1.04 | 0.71 |
| 12 | 5 | 1.77 | 0.80 | 0.72 |
| 13 | 5 | 1.52 | 1.49 | 0.44 |
| 14 | 6 | 2.80 | 0.81 | 0.91 |
| 15 | 6 | 4.32 | 1.33 | 0.17 |
| 16 | 6 | 4.32 | 1.33 | 0.17 |
| 17 | 6 | 4.32 | 1.33 | 0.17 |
| MEDIA | | 0.00 | 1.19 | 0.62 |
| D.E. | | 3.15 | 0.31 | 0.42 |

**EN ORDEN DE DIFICULTAD**

| NÚM SECUEN-CIAL | NOMBRE DEL ÍTEM | DIFICUL-TAD DEL ÍTEM | ÍNDICE DE DISCRIMI-NACIÓN | AJUSTE MEDIA CUA-DRÁTICA |
|---|---|---|---|---|
| 4 | 3 | −3.87 | 1.30 | 0.49 |
| 5 | 3 | −3.29 | 1.35 | 0.64 |
| 7 | 4 | −3.29 | 0.48 | 1.73 |
| 6 | 3 | −2.88 | 1.07 | 0.90 |
| 9 | 4 | −2.88 | 1.44 | 0.27 |
| 8 | 4 | −2.01 | 1.49 | 0.48 |
| 10 | 4 | −1.39 | 1.40 | 0.83 |
| 11 | 5 | 0.55 | 1.04 | 0.71 |
| 13 | 5 | 1.52 | 1.49 | 0.44 |
| 12 | 5 | 1.77 | 0.80 | 0.72 |
| 14 | 6 | 2.80 | 0.81 | 0.91 |
| 16 | 6 | 4.32 | 1.33 | 0.17 |
| 17 | 6 | 4.32 | 1.33 | 0.17 |
| 15 | 6 | 4.32 | 1.33 | 0.17 |

**EN ORDEN DE AJUSTE**

| NÚM SECUEN-CIAL | NOMBRE DEL ÍTEM | DIFICUL-TAD DEL ÍTEM | ÍNDICE DE DISCRIMI-NACIÓN | AJUSTE MEDIA CUA-DRÁTICA | CORRELACIÓN PUNTO-BISERIAL |
|---|---|---|---|---|---|
| 16 | 6 | 4.32 | 1.33 | 0.17 | 0.33 |
| 17 | 6 | 4.32 | 1.33 | 0.17 | 0.33 |
| 15 | 6 | 4.32 | 1.33 | 0.17 | 0.33 |
| 9 | 4 | −2.88 | 1.44 | 0.27 | 0.61 |
| 13 | 5 | 1.52 | 1.49 | 0.44 | 0.58 |
| 8 | 4 | −2.01 | 1.49 | 0.48 | 0.69 |
| 4 | 3 | −3.87 | 1.30 | 0.49 | 0.40 |
| 5 | 3 | −3.29 | 1.35 | 0.64 | 0.42 |
| 11 | 5 | 0.55 | 1.04 | 0.71 | 0.54 |
| 12 | 5 | 1.77 | 0.80 | 0.72 | 0.42 |
| 10 | 4 | −1.39 | 1.40 | 0.83 | 0.54 |
| 6 | 3 | −2.88 | 1.07 | 0.90 | 0.40 |
| 14 | 6 | 2.80 | 0.81 | 0.91 | 0.20 |
| 7 | 4 | −3.29 | 0.48 | 1.73 | 0.23 |

La Tabla 3.2.9 resume la información de la calibración de ítems en tres arreglos útiles. Tenemos para cada ítem su nombre, dificultad, índice de discriminación y media cuadrática de ajuste total. Aparece primero por orden de serie, después por orden de dificultad y en tercer lugar por orden de ajuste. Mientras que en el ejemplo de la PCK tenemos solo unos pocos ítems, en pruebas más largas es muy conveniente reordenar los ítems por dificultad y por ajuste porque nos ayuda a encontrar ítems que desajustan al modelo y captar el patrón de desajuste si es que lo hubiera. En nuestro ejemplo vemos nuevamente que el ítem con el mayor desajuste es el Ítem 7, que se nos presenta en la parte inferior del tercer ordenamiento de la Tabla 3.2.9.

## 3.3 COMPARACIÓN DE PROX A MANO CONTRA PROX POR COMPUTADORA

Ahora podemos comparar los resultados de la estimación de PROX para las dificultades del ítem y las medidas de las personas obtenidas a mano contra las producidas por computadora. Los datos obtenidos a mano con PROX para las dificultades de los ítems y las medidas de las personas provienen de las Tablas 2.4.5 y 2.4.6 del Capítulo 2. Los datos en PROX por computadora provienen de los cuadros 3.2.6 y 3.2.7. Todos los datos se han recogido en las Tablas 3.3.1 y 3.3.2.

En la Tabla 3.3.1, se tiene la calibración de cada ítem obtenida a mano y por computadora. También se proporciona el error estándar para cada ítem calculado a mano y por computadora. Los resultados de PROX a mano y PROX por computadora son prácticamente los mismos.

**TABLA 3.3.1**

**COMPARACIÓN DE CALIBRACIONES DE LOS ÍTEMS Y SUS ERRORES ESTÁNDAR PARA PROX A MANO Y POR COMPUTADORA**

| | CALIBRACIÓN | | | ERROR ESTÁNDAR | | |
|---|---|---|---|---|---|---|
| Ítem | Manual | Computadora | Diferencia | Manual | Computadora | Diferencia |
| 4 | −3.9 | −3.9 | 0.0 | 1.0 | 0.8 | 0.2 |
| 5 | −3.3 | −3.3 | 0.0 | 0.8 | 0.7 | 0.1 |
| 6 | −2.9 | −2.9 | 0.0 | 0.7 | 0.6 | 0.1 |
| 7 | −3.3 | −3.3 | 0.0 | 0.8 | 0.7 | 0.1 |
| 8 | −2.0 | −2.0 | 0.0 | 0.6 | 0.5 | 0.1 |
| 9 | −2.9 | −2.9 | 0.0 | 0.7 | 0.6 | 0.1 |
| 10 | −1.4 | −1.4 | 0.0 | 0.5 | 0.4 | 0.1 |
| 11 | +0.6 | +0.5 | 0.1 | 0.5 | 0.4 | 0.1 |
| 12 | +1.7 | +1.8 | −0.1 | 0.6 | 0.5 | 0.1 |
| 13 | +1.5 | +1.5 | 0.0 | 0.6 | 0.5 | 0.1 |
| 14 | +2.8 | +2.8 | 0.0 | 0.8 | 0.7 | 0.1 |
| 15 | +4.3 | +4.3 | 0.0 | 1.3 | 1.2 | 0.1 |
| 16 | +4.3 | +4.3 | 0.0 | 1.3 | 1.2 | 0.1 |
| 17 | +4.3 | +4.3 | 0.0 | 1.3 | 1.2 | 0.1 |
| MEDIA | 0.00 | 0.00 | | 0.82 | 0.71 | |
| D.E. | 3.13 | 3.15 | | 0.29 | 0.29 | |

Las medidas de habilidad y sus errores estándar se dan en la Tabla 3.3.2. Nuevamente, las diferencias entre los dos métodos son mínimas. Solo los errores estándar de medida muestran una diferencia de magnitud muy variada. Esta diferencia se debe al uso de una fórmula más precisa pero también más laboriosa utilizada en la computadora para PROX. Por lo tanto, excepción hecha de los errores estándar de medida, el muy simple procedimiento PROX realizado a mano y el más preciso PROX por computadora producen virtualmente los mismos resultados.

**TABLA 3.3.2**

**COMPARACIÓN DE MEDIDAS DE PERSONA Y SUS ERRORES ESTÁNDAR PARA PROX A MANO Y POR COMPUTADORA**

| | MEDIDA | | | ERROR ESTÁNDAR | | |
|---|---|---|---|---|---|---|
| Puntaje | Manual | Computadora | Diferencia | Manual | Computadora | Diferencia |
| 1 | −5.4 | −5.4 | 0.0 | 2.2 | 1.5 | 0.7 |
| 2 | −3.8 | −3.8 | 0.0 | 1.6 | 1.1 | 0.5 |
| 3 | −2.8 | −2.7 | −0.1 | 1.4 | 0.9 | 0.5 |
| 4 | −1.9 | −1.9 | 0.0 | 1.2 | 0.9 | 0.3 |
| 5 | −1.2 | −1.2 | 0.0 | 1.2 | 0.8 | 0.4 |
| 6 | −0.6 | −0.6 | 0.0 | 1.1 | 0.8 | 0.3 |
| 7 | 0.0 | 0.0 | 0.0 | 1.1 | 0.8 | 0.3 |
| 8 | +0.6 | +0.6 | 0.0 | 1.1 | 0.8 | 0.3 |
| 9 | +1.2 | +1.2 | 0.0 | 1.2 | 0.8 | 0.4 |
| 10 | +1.9 | +1.9 | 0.0 | 1.2 | 0.9 | 0.3 |
| 11 | +2.8 | +2.7 | 0.1 | 1.4 | 0.9 | 0.5 |
| 12 | +3.8 | +3.8 | 0.0 | 1.6 | 1.1 | 0.5 |
| 13 | +5.4 | +5.4 | 0.0 | 2.2 | 1.5 | 0.7 |
| MEDIA | 0.00 | 0.00 | | 1.42 | 0.98 | |
| D.E. | 3.08 | 3.06 | | 0.39 | 0.25 | |

## 3.4 ANÁLISIS DE LA PCK CON EL PROCEDIMIENTO UCON

Ahora que hemos visto cómo se compara PROX a mano con PROX por computadora, podemos recurrir a un procedimiento un poco más elaborado pero también más preciso que no es adecuado para el trabajo manual, pero sí es conveniente y económico de aplicar por computadora. Es el procedimiento UCON, desarrollado por Wright y Panchapakesan en 1966 (1969), revisado y probado posteriormente por Wright y Douglas en 1974 (1975b, 1977a) y Wright y Mead en 1975 (1975, 1976). La salida de la computadora de UCON es similar en forma a la de PROX. El análisis de UCON de los datos de la PCK se muestra en las Tablas 3.4.1 a 3.4.4. Solo se presentan las tablas que contienen resultados diferentes del análisis PROX.

La Tabla 3.4.1 proporciona los ítems de la prueba con sus calibraciones UCON y sus errores estándar. UCON usa las dificultades de los ítems obtenidas con PROX como punto de partida, que se incluyen en la columna del extremo derecho de la tabla. La Tabla 3.4.2 muestra la medida de la habilidad obtenida con UCON, asociada con cada puntaje y el error estándar para cada medida. Los errores estándar más grandes en los puntajes 7 y 8, para habilidades entre ± 2 lógitos son causados por la distribución bimodal de las dificultades de los ítems que se muestran en la Tabla 3.4.1. Seis de los 14 ítems tienen dificultades por debajo de -3.2 lógitos, mientras que otros seis tienen dificultades superiores a +1.8 lógitos. Esto deja solo dos ítems que funcionan en el rango de 5 lógitos entre -3.2 y +1.8 y los errores estándar de medida en esa región son por consiguiente más altos.

**TABLA 3.4.1**

**CALIBRACIÓN CON UCON**

FACTOR DE EXPASIÓN DE DIFICULTAD 1.31

FACTOR DE EXPASIÓN DE HABILIDAD 2.10

NÚMERO DE ITERACIONES = 7

| NÚMERO SECUENCIAL | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | ERROR ESTÁNDAR | ÚLTIMO CAMBIO EN DIFICULTAD | DIFICULTAD PROX |
|---|---|---|---|---|---|
| 4 | 3 | −4.186 | 0.816 | −0.025 | −3.865 |
| 5 | 3 | −3.648 | 0.709 | −0.023 | −3.294 |
| 6 | 3 | −3.220 | 0.647 | −0.021 | −2.876 |
| 7 | 4 | −3.648 | 0.709 | −0.023 | −3.294 |
| 8 | 4 | −2.241 | 0.547 | −0.015 | −2.007 |
| 9 | 4 | −3.220 | 0.647 | −0.021 | −2.876 |
| 10 | 4 | −1.498 | 0.489 | −0.009 | −1.388 |
| 11 | 5 | 0.760 | 0.456 | 0.006 | 0.547 |
| 12 | 5 | 2.135 | 0.556 | 0.015 | 1.767 |
| 13 | 5 | 1.861 | 0.529 | 0.014 | 1.518 |
| 14 | 6 | 3.214 | 0.705 | 0.022 | 2.805 |
| 15 | 6 | 4.564 | 1.076 | 0.027 | 4.321 |
| 16 | 6 | 4.564 | 1.076 | 0.027 | 4.321 |
| 17 | 6 | 4.564 | 1.076 | 0.027 | 4.321 |

RAÍZ MEDIA CUADRÁTICA = 0.022

**TABLA 3.4.2**

**MEDICIÓN CON UCON**

TABLA COMPLETA DE EQUIVALENCIA DE PUNTAJES

| PUNTAJE BRUTO | CONTEO | HABILIDAD DE LA PERSONA | ERROR ESTÁNDAR |
|---|---|---|---|
| 13 | 0 | 5.09 | 1.14 |
| 12 | 0 | 4.11 | 0.95 |
| 11 | 2 | 3.31 | 0.92 |
| 10 | 1 | 2.53 | 0.93 |
| 9 | 4 | 1.71 | 0.96 |
| 8 | 5 | 0.81 | 1.03 |
| 7 | 12 | −0.22 | 1.07 |
| 6 | 3 | −1.19 | 0.97 |
| 5 | 2 | −1.96 | 0.86 |
| 4 | 2 | −2.61 | 0.81 |
| 3 | 2 | −3.21 | 0.81 |
| 2 | 1 | −3.86 | 0.88 |
| 1 | 0 | −4.73 | 1.10 |

HABILIDAD MEDIA = - 0.16

D.E. DE HABILIDAD = 1.45

TABLA 3.4.3

## CURVAS CARACTERÍSTICAS Y ANÁLISIS DE AJUSTE CON UCON

| | | CURVA CARACTERÍSTICA DEL ÍTEM | | | DIFERENCIA RESPECTO A LO ESPERADO | | | ANÁLISIS DE AJUSTE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NÚM SECUE-NCIAL | NOMBRE DEL ÍTEM | 1ER GRUPO | 2O GRUPO | 3ER GRUPO | 1ER GRUPO | 2O GRUPO | 3ER GRUPO | EN EL GRUPO | ENTRE GRUPOS | TOTAL | ÍNDICE DE DISCRI-MINACIÓN | PUNTO BISE-RIAL |
| 4 | 3 | 0.80 | 1.00 | 1.00 | −0.04 | 0.02 | 0.00 | 0.39 | 0.20 | 0.37 | 1.13 | 0.40 |
| 5 | 3 | 0.70 | 1.00 | 1.00 | −0.06 | 0.03 | 0.01 | 0.55 | 0.37 | 0.54 | 1.17 | 0.42 |
| 6 | 3 | 0.70 | 0.92 | 1.00 | 0.01 | −0.04 | 0.01 | 0.97 | 0.26 | 0.91 | 0.97 | 0.40 |
| 7 | 4 | 0.90 | 0.83 | 1.00 | 0.14 | −0.14 | 0.01 | 1.72 | 4.68 | 1.98 | 0.60 | 0.23 |
| 8 | 4 | 0.40 | 0.92 | 1.00 | −0.09 | 0.03 | 0.03 | 0.45 | 0.45 | 0.45 | 1.20 | 0.69 |
| 9 | 4 | 0.60 | 1.00 | 1.00 | −0.09 | 0.05 | 0.01 | 0.20 | 0.60 | 0.24 | 1.22 | 0.61 |
| 10 | 4 | 0.30 | 0.75 | 1.00 | −0.04 | −0.03 | 0.05 | 0.82 | 0.46 | 0.79 | 1.10 | 0.54 |
| 11 | 5 | 0.0 | 0.33 | 0.67 | −0.06 | 0.06 | −0.01 | 0.80 | 0.47 | 0.78 | 0.96 | 0.54 |
| 12 | 5 | 0.0 | 0.17 | 0.33 | −0.02 | 0.08 | −0.06 | 0.99 | 0.76 | 0.97 | 0.82 | 0.42 |
| 13 | 5 | 0.0 | 0.0 | 0.58 | −0.02 | −0.11 | 0.13 | 0.30 | 1.49 | 0.41 | 1.34 | 0.58 |
| 14 | 6 | 0.0 | 0.08 | 0.17 | −0.01 | 0.05 | −0.04 | 1.39 | 0.72 | 1.33 | 0.82 | 0.20 |
| 15 | 6 | 0.0 | 0.0 | 0.88 | 0.00 | −0.01 | 0.01 | 0.14 | 0.07 | 0.13 | 1.08 | 0.33 |
| 16 | 6 | 0.0 | 0.0 | 0.08 | 0.00 | −0.01 | 0.01 | 0.14 | 0.07 | 0.13 | 1.08 | 0.33 |
| 17 | 6 | 0.0 | 0.0 | 0.08 | 0.00 | −0.01 | 0.01 | 0.14 | 0.07 | 0.13 | 1.08 | 0.33 |
| RANGO DE PUNTAJES | | 1–6 | 7–7 | 8–13 | N = 10 | 12 | 12 | 31 | 3 | 34 | GRADO DE LIBERTAD | |
| MEDIA DE HABILIDAD | | −2.30 | −0.22 | 1.67 | | | | 0.25 | 0.82 | 0.24 | ERROR ESTÁNDAR | |

La tabla 3.4.3 proporciona las Curvas Características de los Ítems que se tienen en la tabla 3.2.8, las discrepancias de estas CCI respecto de las CCI esperadas por el modelo con base en las estimaciones de UCON y las medias cuadráticas de ajuste que resultan del análisis del mismo UCON. La tabla 3.4.4 resume la calibración de UCON en la misma forma que la Tabla 3.2.9 presentó las calibraciones con PROX.

**TABLA 3.4.4**

**RESUMEN DE CALIBRACIÓN DE LOS ÍTEMS POR UCON**

**EN ORDEN CONSECUTIVO**

| NÚM SECUENCIAL | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | ÍNDICE DE DISCRIMINACIÓN | AJUSTE MEDIA CUADRÁTICA |
|---|---|---|---|---|
| 4 | 3 | −4.19 | 1.13 | 0.37 |
| 5 | 3 | −3.65 | 1.17 | 0.53 |
| 6 | 3 | −3.22 | 0.97 | 0.90 |
| 7 | 4 | −3.65 | 0.60 | 1.98 |
| 8 | 4 | −2.24 | 1.20 | 0.44 |
| 9 | 4 | −3.22 | 1.22 | 0.23 |
| 10 | 4 | −1.50 | 1.10 | 0.79 |
| 11 | 5 | 0.76 | 0.96 | 0.77 |
| 12 | 5 | 2.14 | 0.82 | 0.97 |
| 13 | 5 | 1.86 | 1.34 | 0.40 |
| 14 | 6 | 3.21 | 0.82 | 1.33 |
| 15 | 6 | 4.56 | 1.08 | 0.13 |
| 16 | 6 | 4.56 | 1.08 | 0.13 |
| 17 | 6 | 4.56 | 1.08 | 0.13 |
| MEDIA | | 0.00 | 1.04 | 0.65 |
| D.E. | | 3.44 | 0.19 | 0.53 |

**EN ORDEN DE DIFICULTAD**

| NÚM SECUENCIAL | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | ÍNDICE DE DISCRIMINACIÓN | AJUSTE MEDIA CUADRÁTICA |
|---|---|---|---|---|
| 4 | 3 | −4.19 | 1.13 | 0.37 |
| 5 | 3 | −3.65 | 1.17 | 0.53 |
| 7 | 4 | −3.65 | 0.60 | 1.98 |
| 6 | 3 | −3.22 | 0.97 | 0.90 |
| 9 | 4 | −3.22 | 1.22 | 0.23 |
| 8 | 4 | −2.24 | 1.20 | 0.44 |
| 10 | 4 | −1.50 | 1.10 | 0.79 |
| 11 | 5 | 0.76 | 0.96 | 0.77 |
| 13 | 5 | 1.86 | 1.34 | 0.40 |
| 12 | 5 | 2.14 | 0.82 | 0.97 |
| 14 | 6 | 3.21 | 0.82 | 1.33 |
| 16 | 6 | 4.56 | 1.08 | 0.13 |
| 17 | 6 | 4.56 | 1.08 | 0.13 |
| 15 | 6 | 4.56 | 1.08 | 0.13 |

**EN ORDEN DE AJUSTE**

| NÚM SECUENCIAL | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | ÍNDICE DE DISCRIMINACIÓN | AJUSTE MEDIA CUADRÁTICA | CORRELACIÓN PUNTO-BISERIAL |
|---|---|---|---|---|---|
| 16 | 6 | 4.56 | 1.08 | 0.13 | 0.33 |
| 17 | 6 | 4.56 | 1.08 | 0.13 | 0.33 |
| 15 | 6 | 4.56 | 1.08 | 0.13 | 0.33 |
| 9 | 4 | −3.22 | 1.22 | 0.23 | 0.61 |
| 4 | 3 | −4.19 | 1.13 | 0.37 | 0.40 |
| 13 | 5 | 1.86 | 1.34 | 0.40 | 0.58 |
| 8 | 4 | −2.24 | 1.20 | 0.44 | 0.69 |
| 5 | 3 | −3.65 | 1.17 | 0.53 | 0.42 |
| 11 | 5 | 0.76 | 0.96 | 0.77 | 0.54 |
| 10 | 4 | -1.50 | 1.10 | 0.79 | 0.54 |
| 6 | 3 | 3.22 | 0.97 | 0.90 | 0.40 |
| 12 | 5 | 2.14 | 0.82 | 0.97 | 0.42 |
| 14 | 6 | 3.21 | 0.82 | 1.33 | 0.20 |
| 7 | 4 | 3.65 | 0.60 | 1.98 | 0.23 |

En el Capítulo 2 demostramos la viabilidad de la calibración manual y mostramos el cálculo en detalle. Esto se hizo para proporcionar una base para entender la forma en que el programa de computadora realiza la misma tarea y la salida de resultados. La comparación de PROX a mano y PROX por computadora demuestra su comparabilidad. En UCON tenemos un algoritmo que proporciona una mayor precisión. Nuestro siguiente paso es comparar UCON contra PROX.

## 3.5 COMPARACIÓN DE UCON CONTRA PROX CON LOS DATOS DE LA PCK.

La Tabla 3.5.1 proporciona las calibraciones y los errores estándar para los datos de la PCK producidos por los métodos UCON y PROX. Las diferencias de calibración entre UCON y PROX corren alrededor de ± 0.3 lógitos. La diferencia entre sus errores estándar es a lo sumo ± 0.1 lógitos.

**TABLA 3.5.1**

**COMPARACIÓN DE CALIBRACIONES DE ÍTEMS Y ERRORES ESTÁNDAR PARA UCON Y PROX POR COMPUTADORA**

| | CALIBRACIÓN | | | ERROR ESTÁNDAR | | |
|---|---|---|---|---|---|---|
| Ítem | UCON | PROX | Diferencia | UCON | PROX | Diferencia |
| 4 | −4.2 | −3.9 | −0.3 | 0.8 | 0.8 | 0.0 |
| 5 | −3.6 | −3.3 | −0.3 | 0.7 | 0.7 | 0.0 |
| 6 | −3.2 | −2.9 | −0.3 | 0.6 | 0.6 | 0.0 |
| 7 | −3.6 | −3.3 | −0.3 | 0.7 | 0.7 | 0.0 |
| 8 | −2.2 | −2.0 | −0.2 | 0.6 | 0.5 | 0.1 |
| 9 | −3.2 | −2.9 | −0.3 | 0.6 | 0.6 | 0.0 |
| 10 | −1.5 | −1.4 | −0.1 | 0.5 | 0.4 | 0.1 |
| 11 | +0.8 | +0.5 | 0.3 | 0.5 | 0.4 | 0.1 |
| 12 | +2.1 | +1.8 | 0.3 | 0.6 | 0.5 | 0.1 |
| 13 | +1.9 | +1.5 | 0.4 | 0.5 | 0.5 | 0.0 |
| 14 | +3.2 | +2.8 | 0.4 | 0.7 | 0.7 | 0.0 |
| 15 | +4.6 | +4.3 | 0.3 | 1.1 | 1.2 | −0.1 |
| 16 | +4.6 | +4.3 | 0.3 | 1.1 | 1.2 | −0.1 |
| 17 | +4.6 | +4.3 | 0.3 | 1.1 | 1.2 | −0.1 |
| MEDIA | 0.00 | 0.00 | | 0.74 | 0.71 | |
| E.S. | 3.44 | 3.15 | | 0.22 | 0.29 | |

La Tabla 3.5.2 proporciona las medidas de las personas y sus errores estándar para UCON y PROX. Las diferencias en las medidas que se alcanzan entre los métodos UCON y PROX son del orden de ± 0.7 lógitos.

Vemos que al usar el procedimiento UCON – que es más preciso, porque toma en cuenta las distribuciones particulares de las dificultades de los ítems y las habilidades de las personas – se produce una diferencia tangible para los datos de la PCK. Como hemos visto, los ítems de la PCK tienen una distribución bimodal muy distintiva pero que no es manejada por el procedimiento PROX. Desde luego, estas diferencias entre PROX y UCON nunca son tan grandes como un error

estándar, y por lo tanto no pueden ser consideradas como estadísticamente significativas, sin embargo, podrían preocupar a algunos evaluadores. El origen de estas discrepancias se puede atribuir a que este ejemplo de la PCK está hecho sobre pocos datos y a la bimodalidad de las dificultades de los ítems. Para conjuntos de datos más grandes y para distribuciones de dificultad de ítems más uniformes, las diferencias en los resultados que se obtienen con PROX y UCON son prácticamente inapreciables.

**TABLA 3.5.2**

**COMPARACIÓN DE MEDIDAS DE LAS PERSONAS Y ERRORES ESTÁNDAR PARA UCON Y PROX POR COMPUTADORA**

| | MEDIDA | | | ERROR ESTÁNDAR | | |
|---|---|---|---|---|---|---|
| Puntaje | UCON | PROX | Diferencia | UCON | PROX | Diferencia |
| 1 | −4.7 | −5.4 | 0.7 | 1.1 | 1.5 | −0.4 |
| 2 | −3.9 | −3.8 | −0.1 | 0.9 | 1.1 | −0.2 |
| 3 | −3.2 | −2.7 | −0.5 | 0.8 | 0.9 | −0.1 |
| 4 | −2.6 | −1.9 | −0.7 | 0.8 | 0.9 | −0.1 |
| 5 | −2.0 | −1.2 | −0.8 | 0.9 | 0.8 | 0.1 |
| 6 | −1.2 | −0.6 | −0.6 | 1.0 | 0.8 | 0.2 |
| 7 | −0.2 | 0.0 | −0.2 | 1.0 | 0.8 | 0.2 |
| 8 | 0.8 | 0.6 | 0.2 | 1.0 | 0.8 | 0.2 |
| 9 | 1.7 | 1.2 | 0.5 | 1.0 | 0.8 | 0.2 |
| 10 | 2.5 | 1.9 | 0.6 | 0.9 | 0.9 | 0.0 |
| 11 | 3.3 | 2.7 | 0.6 | 0.9 | 0.9 | 0.0 |
| 12 | 4.1 | 3.8 | 0.3 | 1.0 | 1.1 | −0.1 |
| 13 | 5.1 | 5.4 | −0.3 | 1.1 | 1.5 | −0.4 |
| MEDIA | 0.0 | 0.0 | | 1.0 | 1.0 | |
| E.S | 3.2 | 3.1 | | 0.1 | 0.3 | |

## 3.6 ALGORITMO INFORMÁTICO PARA PROX

Aquí presentamos una implementación concisa del procedimiento PROX, adecuado para programarse en computadora:

1. Edite la matriz de datos binarios (o dicótomos) de respuestas persona-ítem de manera que ninguna persona tenga un puntaje cero o una calificación perfecta, así como ningún ítem tenga cero respuestas correctas o un puntaje perfecto. Esta edición puede ir más allá de una sola etapa si es que la eliminación de un ítem conduce a eliminar algunas personas, y viceversa. El resultado final es un vector de puntajes de los ítems $s_i$ donde i corre de 1 a L y un vector de frecuencias de puntajes de las personas $n_r$ donde r va de 1 a L- 1.

2. Haga: $$x_i = \ell_n[(N - s_i)/s_i] \quad \textbf{[3.6.1]}$$

$$x_. = \sum_i^L x_i / L \quad \textbf{[3.6.2]}$$

$$y_r = \ell_n[r/(L - r)] \quad \textbf{[3.6.3]}$$

$$y. = \sum_{r}^{L-1} n_r \, y_r / N \qquad \textbf{[3.6.4]}$$

$$D = \sum_{i}^{L} (x_i - x.)^2 / 2.89(L - 1) \qquad \textbf{[3.6.5]}$$

$$B = \sum_{r}^{L-1} n_r \left(y_r - y.\right)^2 / 2.89(N - 1) \qquad \textbf{[3.6.6]}$$

$$G = BD \qquad \textbf{[3.6.7]}$$

3. Calcule los factores de expansión:

$$X = [(1 + D)/(1 - G)]^{1/2} \qquad \textbf{[3.6.8]}$$

$$Y = [(1 + B)/(1 - G)]^{1/2} \qquad \textbf{[3.6.9]}$$

4. Estime las dificultades del ítem como:

$$d_i = Y(x_i - x.), \qquad \text{para} \qquad i = 1, L \qquad \textbf{[3.6.10]}$$

5. Con errores estándar de:

$$ES(d_i) = Y \, [\, N / \, s_i(N - s_i)]^{1/2} \qquad \textbf{[3.6.11]}$$

6. Las estimaciones de habilidad para este conjunto de ítems están dadas por:

$$b_r = Xy_{r'} \qquad \text{para} \qquad r = 1, L - 1 \qquad \textbf{[3.6.12]}$$

7. Con errores estándar de:

$$ES(b_r) = X \, [\, L / \, r \, (L - r)]^{1/2} \qquad \textbf{[3.6.13]}$$

## 3.7 PROCEDIMIENTO INCONDICIONAL UCON

El modelo de Rasch para observaciones binarias (o dicótomas) define la probabilidad de una respuesta $x_{vi}$ al ítem i por la persona $v$ como:

$$P\{x_{vi} | \beta_v, \delta_i\} = \exp[x_{vi}(\beta_v - \delta_i)] \, / \, [1 + \exp(\beta_v - \delta_i)] \qquad \textbf{[3.7.1]}$$

donde $x_{vi} = \begin{cases} 1 \text{ si es correcto} \\ 0 \text{ en caso contrario} \end{cases}$

$\beta_v$ = parámetro de habilidad de la persona $v$,

$\delta_i$ = parámetro de dificultad del ítem i.

La verosimilitud $\Lambda$ de la matriz de datos $((x_{vi}))$ es el productorio (también llamado multiplicatorio u operador producto) de la Ecuación [3.7.1] sobre todos los valores de $v$ y de i, donde L es el número de ítems y N es el número de personas con puntajes en la prueba entre 0 y L, ya que puntajes de 0 y L conducen a valores indefinidos para los estimados de habilidad.

$$\Lambda = \exp\left[\sum_{v}^{N} \sum_{i}^{L} x_{vi} \, (\beta_v - \delta_i)\right] / \prod_{v}^{N} \prod_{i}^{L} [1 + \exp \, (\beta_v - \delta_i)] \qquad \textbf{[3.7.2]}$$

Al tomar logaritmos y haciendo:

$$\sum_{i}^{L} x_{vi} = r_v \qquad \text{puntaje de la persona } v$$

y

$$\sum_{v}^{N} x_{vi} = s_i \qquad \text{el puntaje del ítem i,}$$

el logaritmo de la verosimilitud λ puede escribirse:

$$\lambda = \ell_n \Lambda = \sum_{v}^{N} r_v \beta_v - \sum_{i}^{L} s_i \delta_i - \sum_{v}^{N}\sum_{i}^{L} \ell_n [1 + \exp(\beta_v - \delta_i)] \qquad \textbf{[3.7.3]}$$

La reducción de la matriz de datos $((x_{vi}))$ a sus marginales $(r_v)$ y $(s_i)$ y la separación de $r_v\beta_v$ y $s_i\delta_i$ en la Ecuación 3.7.3 establecen la suficiencia de $r_v$ para estimar $\beta_v$ y de $s_i$ para estimas $\delta_i$, así como la objetividad de tales estimaciones.

Es importante reconocer, por supuesto, que aunque $r_v$ y $s_i$ conducen a estimadores suficientes de $\beta_v$ y de $\delta_i$, por sí mismos no son satisfactorios como estimadores. El puntaje $r_v$ de la persona no es independiente de las dificultades de los ítems particulares que respondió en la prueba. Tampoco el puntaje $s_i$ del ítem es independiente de la distribución de habilidades de las personas que respondieron al ítem. Para lograr la independencia de estos factores locales se requiere ajustar los valores observados de $r_v$ y $s_i$ para la dificultad relativa del ítem y las distribuciones de habilidad de las personas hasta para producir lo que deseamos: medidas de las personas independientes de la prueba y calibraciones de ítems independientes de la muestra.

Con la condición lateral $\sum_{i}^{L} \delta_i = 0$ que restringe la indeterminación del origen en los parámetros de respuesta, la primera y segunda derivadas parciales de λ con respecto a $\beta_v$, y $\delta_i$ se escriben:

$$\frac{\partial \lambda}{\partial \beta_v} = r_v - \sum_{i}^{L} \pi_{vi} \qquad v = 1, N \qquad \textbf{[3.7.4]}$$

$$\frac{\partial^2 \lambda}{\partial \beta_v^2} = -\sum_{i}^{L} \pi_{vi}(1 - \pi_{vi}) \qquad \textbf{[3.7.5]}$$

y

$$\frac{\partial \lambda}{\partial \delta_i} = -s_i + \sum_{v}^{N} \pi_{vi} \qquad i = 1, L \qquad \textbf{[3.7.6]}$$

$$\frac{\partial^2 \lambda}{\partial \delta_i^2} = \sum_{v}^{N} \pi_{vi}(1 - \pi_{vi}) \qquad \textbf{[3.7.7]}$$

donde $\pi_{vi} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)]$

Estas son las ecuaciones necesarias para estimar la máxima verosimilitud incondicional. Las soluciones para las estimaciones de dificultad de los ítems en las ecuaciones 3.7.6 y 3.7.7 dependen de la presencia de valores para las estimaciones de habilidad de las personas. Dado que los puntajes no ponderados son estadísticos suficientes para estimar las habilidades en la prueba, todas las personas con puntajes idénticos obtienen idénticas estimaciones de habilidad. Por lo tanto, podemos agrupar a las personas por su puntaje, haciendo que:

$b_r$ sea la habilidad estimada para cualquier persona con puntaje r,
$d_i$ sea la estimación de dificultad del ítem i,
$n_r$ sea el número de personas con puntaje r,

y escribiendo la probabilidad estimada de que una persona con un puntaje r responda correctamente al ítem i como:

$$p_{ri} = \exp(b_r - d_i)/[1 + \exp(b_r - d_i)] \qquad \textbf{[3.7.8]}$$

Entonces $\sum_{\nu}^{N} \pi_{\nu i} \approx \sum_{r}^{L-1} n_r p_{ri}$ , en lo que se refiere a las estimaciones.

Un algoritmo conveniente para calcular los estimados ($d_i$) es:

1. Defina un conjunto inicial de ($b_r$) como

$$b_r^{(0)} = \ell n\left(\frac{r}{L-r}\right) \qquad r = 1, L-1 \qquad \textbf{[3.7.9]}$$

2. Defina un conjunto inicial de ($d_i$), centrado en d.= 0, como

$$d_i^{(0)} = \ell n\left(\frac{N-s_i}{s_i}\right) - \left[\sum_{i}^{L} \ell n\left(\frac{N-s_i}{s_i}\right)\right]/L \qquad i = 1, L \qquad \textbf{[3.7.10]}$$

   donde $b_r^{(0)}$ es la máxima verosimilitud estimada de $\beta_v$ para una prueba de L ítems equivalentes centrados en cero y $d_i^{(0)}$ es la máxima verosimilitud estimada de $\delta_i$, igualmente centrada para una muestra de N personas de igual habilidad.

3. Aplicar el método de Newton-Raphson a la Ecuación 3.7.6 para mejorar cada $d_i$ de acuerdo con la expresión:

$$d_i^{(j+1)} = d_i^{(j)} - \frac{-s_i + \sum_r^{L-1} n_r p_{ri}^{(j)}}{-\sum_r^{L-1} n_r p_{ri}^{(j)}\left(1 - p_{ri}^{(j)}\right)} \qquad i = 1, L \qquad \textbf{[3.7.11]}$$

   hasta alcanzar la convergencia en $\left|d_i^{(j+1)} - d_i^{(j)}\right| < .01$

   donde $$p_{ri}^{(j)} = \exp\left(b_r - d_i^{(j)}\right) / \left[1 + \exp\left(b_r - d_i^{(j)}\right)\right] \qquad \textbf{[3.7.12]}$$

   y el conjunto actual de ($b_r$) vienen dados por el ciclo anterior.

4. Recentrar el conjunto de ($d_i$) en d.= 0.

5. Usando este conjunto mejorado de ($d_i$), aplique el método de Newton a la Ecuación 3.7.4 para mejorar cada $b_r$ de acuerdo con esta expresión:

$$b_r^{(m+1)} = b_r^{(m)} - \frac{-r - \sum_i^L p_{ri}^{(m)}}{-\sum_i^L p_{ri}^{(m)}\left(1 + p_{ri}^{(m)}\right)} \qquad r = 1, L-1 \qquad \textbf{[3.7.13]}$$

   hasta la convergencia en $\left|b_r^{(m+1)} - b_r^{(m)}\right| < .01$

donde $$p_{ri}^{(m)} = \exp\left(b_r^{(m)} - d_i\right) / \left[1 + \exp\left(b_r^{(m)} - d_i\right)\right] \quad \textbf{[3.7.14]}$$

y el conjunto actual de ($d_i$) se obtuvieron en el ciclo anterior.

6. Repita los pasos (3) a (5) hasta que las estimaciones sucesivas de todo el conjunto de ($d_i$) se estabilice en

$$\sum_{i}^{L} \left(d_i^{(k+1)} - d_i^{(k)}\right)^2 / \ L < .00001 \quad \textbf{[3.7.15]}$$

lo cual generalmente ocurre en tres o cuatro ciclos.

7. Use los recíprocos de las raíces cuadradas negativas definidas en la Ecuación 3.7.7 como estimadores asintóticos de los errores estándar de los estimados de dificultad:

$$ES(d_i) = \left[ \sum_{r}^{L-1} n_r p_{ri} \left(1 - p_{ri}\right) \right]^{-1/2} \qquad i = 1, L \quad \textbf{[3.7.16]}$$

Andersen (1973) demostró que la presencia de los parámetros de habilidad ($\beta_v$) en la Ecuación de verosimilitud de esta aproximación incondicional conduce a estimaciones sesgadas de las dificultades ($\delta_i$). Las simulaciones realizadas para probar UCON en 1966 indicaron que al multiplicar los estimadores centrados de las dificultades de los ítems por el factor [(L – 1) / L] compensa la mayor parte de este sesgo. Ver Wright y Douglas (1975b o 1977a) para una discusión y evaluación del factor de corrección de sesgo [(L – 1) / L].

# 4 EL ANÁLISIS DE AJUSTE

## 4.1 INTRODUCCIÓN

Los procedimientos para la calibración de los ítems a mano se dieron en el Capítulo 2, los ejemplos de salida de las calibraciones hechas por un programa informático se discutieron en el Capítulo 3. Sin embargo, estos procedimientos de calibración no son más que una parte del análisis completo de una muestra de datos. El modelo de Rasch hace ciertos supuestos plausibles sobre lo que sucede cuando una persona responde a un ítem y un análisis completo debe incluir una evaluación de qué tan bien se ajustan los datos a estos supuestos. Cuando, por ejemplo, una persona responde correctamente todos los ítems difíciles de una prueba, pero falla en varios ítems fáciles, estaremos sorprendidos porque nos parecerá un patrón poco plausible de respuestas. Si bien podríamos examinar a ojo los registros individuales para determinar qué tan plausibles son, en la práctica queremos realizar este tipo de apreciaciones sobre una base sistemática y manejable. Queremos ser capaces de ser específicos y objetivos en nuestras reacciones ante las observaciones inverosímiles.

Incluso si el modelo de medición llega a ajustarse en una aplicación en particular, no podríamos predecir de antemano qué tan bien otros nuevos ítems (o incluso los viejos) continuarían funcionando en cada situación en la que podrían aplicarse, ni podríamos saber de antemano cómo responderían todas las personas. Por lo tanto, si somos serios en nuestros intentos de medir, debemos examinar cada aplicación para ver qué tan bien cada conjunto de respuestas corresponde con nuestras expectativas del modelo. Debemos evaluar no solo la plausibilidad de las respuestas de la muestra, sino también la plausibilidad de las respuestas de cada persona al conjunto de ítems en su prueba. Para hacer esto debemos examinar la respuesta de cada persona en cada ítem para determinar si es consistente con el patrón general de respuestas observadas.

## 4.2. LA MATRIZ DE RESPUESTA DE LA PCK

Comenzamos el estudio del análisis de ajuste volviendo a la matriz de datos ítems-personas de la Prueba de los Cubos de Knox (PCK) de la Tabla 2.4.1. En esta tabla tenemos las respuestas editadas y ordenadas de 34 personas a 14 ítems de la PCK. El proceso de edición eliminó los ítems acertado por todos o por nadie, así como las personas que responden correctamente todos o ninguno de los ítems. Las personas e ítems restantes ordenan en forma creciente de aciertos.

Esta matriz de ítems-personas, de 1 y 0, es el registro completo de las respuestas de las personas utilizables para esta prueba. Por inspección vemos que la dificultad creciente de los ítems de la PCK divide a la matriz aproximadamente en dos triángulos: un triángulo inferior izquierdo dominado por respuestas correctas significadas por 1 y un triángulo superior derecho dominado por respuestas incorrectas significadas por 0.

Este es el patrón que esperamos. A medida que los ítems se vuelven más difíciles, recorriendo de izquierda a derecha la Tabla 2.4.1, la cadena de aciertos de cualquier persona en particular debería evolucionar gradualmente hasta terminar en una cadena de fallas en los ítems demasiado difíciles para dicha persona. Del mismo modo, cuando examinamos el patrón de respuestas para cualquier ítem procediendo de la parte inferior de la Tabla 2.4.1 hacia arriba en

la columna del ítem donde las personas se ordenaron por habilidad decreciente, esperamos que la cadena de aciertos evolucione en la parte inferior hasta acabar como fallas, a medida que las personas se vuelven demasiado bajas en su habilidad para acertar este ítem.

De nuestra calibración de los ítems de la PCK tenemos estimaciones de las dificultades del ítem ($d_i$) y de las habilidades ($b_r$) asociados con los posibles puntajes (r) en esta prueba. En la Tabla 4.2.1 mostramos las respuestas de la Tabla 2.4.1 a la que hemos agregado resultados de nuestra calibración, las dificultades de los ítems ($d_i$) en la parte inferior y las habilidades ($b_r$) asociadas con cada puntaje en la columna derecha. Las dificultades de los ítems y los puntajes de habilidad en la Tabla 4.2.1 fueron las estimadas a mano con PROX en el Capítulo 2.

Observe cómo la Tabla 4.2.1 está organizada en seis secciones para resaltar el patrón de respuestas. Los 14 ítems están divididos en los 7 más fáciles y los 7 más difíciles. Las 34 personas se dividen en 10 que obtuvieron un puntaje por debajo de siete aciertos, 12 que tuvieron exactamente siete y otras 12 que tuvieron un puntaje por arriba de siete. En la sección inferior izquierda solo hay respuestas con 1. Todas las personas con habilidad superior respondieron correctamente todos los ítems más fáciles. En la sección superior derecha solo hay valores de 0. Todas las personas de menor habilidad responden incorrectamente los ítems más difíciles. Pero en las otras cuatro secciones hay un patrón de 1 y 0 que debe analizarse.

Cuando examinamos los patrones de respuestas inesperadas "correctas" e "incorrectas" en nuestros datos, encontramos que la Tabla 4.2.1 presenta varias excepciones al patrón de todos los 1 seguido de todos los 0. Por supuesto, no esperamos que todas las personas fallen por primera vez en un punto particular y luego continúen haciéndolo en todos los ítems más difíciles. Lo que esperamos es una serie de aciertos y fallas que finalmente conduzca a una serie de fallas conforme los ítems se van haciendo demasiado difíciles. Sin embargo, algunas de las excepciones en la Tabla 4.2.1 parecen exceder incluso esta expectativa. Para facilitar su revisión hemos encerrado en un círculo las respuestas que parecen ser más inesperadas, dado el patrón general.

Las ubicaciones de estas respuestas aparentemente sorprendentes nos llevan a examinar más de cerca algunos de los registros de personas en la Tabla 4.2.1.

**Para la Persona 2,** el patrón de respuestas es quizá demasiado razonable: todos los 1 seguidos por todos los 0.

**Para la Persona 29,** este patrón, en cambio, es bastante desconcertante; muestra tanto fallas en ítems fáciles como acierto en uno difícil.

El patrón esperado lo vemos en los registros de las Personas 12 o 23. Cada registro muestra una cadena de 1, cerca de los cuales se alternan algunos 1 y 0, seguidos de un cadena de 0.

Si regresamos a los cuatro registros más cuestionables, vemos:

**La persona 11,** falló el ítem 4 pero acertó los ítems 5 a 9 antes de fallar todos los ítems restantes.

**La persona 17,** respondió correctamente el ítem 4, falló en el ítem 5, acertó los ítems 6 a 10 y luego falló los ítems restantes.

**Persona 13,** acertó los ítems 4 y 5, falló en los ítems 6 y 7, acertó los ítems 8 a 12 y luego falló en los restantes.

**TABLA 4.2.1**

**PATRONES DE RESPUESTA DE 34 PERSONAS A 14 ÍTEMS DE LA TABLA 2.4.1**

| | NOMBRE DE LA PERSONA | Ítems FÁCILES → | | | | | | | Ítems DIFÍCILES → | | | | | | | PUNTAJE r | PROPORCIÓN DE 14 | HABILIDAD DE PUNTUACIÓN $b_r$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | (4) | (5) | (7) | (6) | 9 | (8) | 10 | 11 | 13 | (12) | (14) | 15 | 16 | 17 | | | |
| Personas Bajas | 25 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | .14 | −3.8 |
| | 4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | .21 | −2.8 |
| | 33 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | .21 | |
| | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | .29 | −1.9 |
| | 27 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | .29 | |
| | (11) | (0) | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | .36 | −1.2 |
| | (32) | 1 | 1 | 1 | (0) | 1 | (0) | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | .36 | |
| | (17) | 1 | (0) | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | .43 | |
| | 19 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | .43 | −0.6 |
| | 30 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | .43 | |
| | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | (3) | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | (1) | 0 | 0 | 0 | 0 | 7 | .50 | |
| | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | 8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | (13) | 1 | 1 | (0) | (0) | 1 | 1 | 1 | 1 | 0 | (1) | 0 | 0 | 0 | 0 | 7 | .50 | 0.0 |
| | 16 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | 26 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | 28 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| | (29) | 1 | 1 | (0) | 1 | 1 | (0) | 1 | 1 | 0 | 0 | (1) | 0 | 0 | 0 | 7 | .50 | |
| | 31 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | .50 | |
| Personas Altas | 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | .57 | |
| | 18 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 8 | .57 | |
| | 14 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | .57 | 0.6 |
| | 32 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | .57 | |
| | 20 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 8 | .57 | |
| | 21 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 9 | .64 | |
| | 22 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 9 | .64 | 1.2 |
| | 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 9 | .64 | |
| | 34 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 9 | .64 | |
| | 15 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 10 | .71 | 1.9 |
| | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 11 | .79 | |
| | 24 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 11 | .79 | 2.8 |
| | **PROPORCIÓN DE 34** | .94 | .91 | .91 | .88 | .88 | .79 | .71 | .35 | .21 | .19 | .09 | .03 | .03 | .03 | | | |
| | **Dificultad del Ítem** $d_i$ | −3.9 | −3.3 | −3.3 | −2.9 | −2.9 | −2.0 | −1.4 | 0.6 | 1.5 | 1.7 | 2.8 | 4.3 | 4.3 | 4.3 | | | |

Dificultad de los ítems ($d_i$) y habilidades de puntuación ($b_r$) de las Tablas 2.4.5 y 2.4.6

**La persona 29,** acertó los ítems 4 a 6, falló en los ítems 7 y 8, acertó en los ítems 9 a 11, falló en los ítems 12 y 13 y luego pasó el ítem 14 antes de fallar todos los ítems restantes.

Hay algunos otros registros que también podrían examinarse como los casos de las Personas 3 y 12, pero a medida que echamos una ojeada a esta pequeña matriz, podemos ver que los otros registros son menos excepcionales.

Ahora que hemos encontrado algunos casos de respuestas posiblemente irregulares, queremos una forma sistemática de juzgar el grado de excepción que hemos observado en estos patrones de respuesta.

El modelo de Rasch fundamenta la calibración y la medición en dos expectativas: (1) que una persona más capaz siempre debe tener una mayor probabilidad de éxito en cualquier ítem que otra persona menos capaz, y (2) que cualquier persona siempre debería ser más propensa a responder mejor un ítem más fácil que uno más difícil. Cuando un patrón observado de respuestas muestra desviaciones significativas a estas expectativas, podemos usar los detalles del modelo y las estimaciones de personas e ítems para calcular un índice de qué tan inesperado es un registro de respuestas de cualquier persona en particular o de cualquier ítem.

## 4.3 EL ANÁLISIS DE AJUSTE A MANO

El primer paso en nuestro análisis de plausibilidad de las respuestas o ajuste (FIT en inglés) consiste en observar la diferencia ($b_v - d_i$) entre las estimaciones de habilidad $b_v$ y la dificultad $d_i$ para cada persona e ítem. Cuando esta diferencia es positiva, significa que el ítem debe ser fácil para la persona. Cuanto más positiva sea la diferencia, más fácil será el ítem y, por lo tanto, mayor será nuestra expectativa de que la persona tendrá éxito. Del mismo modo, cuando la diferencia entre la habilidad de la persona y la dificultad del ítem se vuelve más y más negativa, el ítem debería ser más y más difícil para esa persona, y con ello aumentará nuestra expectativa de su falla.

Para enfocar nuestra aplicación en estas ideas, hemos tomado de la Tabla 4.2.1 las respuestas de las seis personas con los patrones más inverosímiles a los siete ítems en cuales tienen lugar sus respuestas poco plausibles. Estas respuestas seleccionadas se presentan en la Tabla 4.3.1. Con esta tabla podemos estudiar más fácilmente las respuestas "correctas" o “incorrectas" más inesperadas.

Para empezar, podemos tabular el número de respuestas inesperadas para las personas e ítems en la Tabla 4.3.1 llegando a un conteo simple con el cual describir lo que está ocurriendo. Vemos que las personas 13 y 29 están en la peor condición con tres respuestas inesperadas cada uno. Sin embargo, este conteo simple no nos dice cómo ponderar y, por lo tanto, cómo juzgar el grado en el que estas respuestas son inesperadas.

Una forma en que los estadísticos piensan sobre los resultados de eventos probabilísticos como el lanzamiento de dados o de una moneda o la forma de obtener un ítem correcto en una prueba, es definiendo el valor esperado de la variable, que se presenta en cualquier respuesta $x_{vi}$, para la persona $v$ en el ítem i, como la probabilidad $\pi_{vi}$ de que dicha respuesta ocurra. Esto es útil porque, si tuviéramos que obtener una respuesta $x_{vi}$ muchas veces y su génesis estuviera más o menos gobernada por la probabilidad $\pi_{vi}$, entonces esperaríamos que ocurriera un acierto más o menos $\pi_{vi}$ número de veces, del mismo modo como se espera que el "6" resulte aproximadamente una sexta parte del tiempo cuando tiramos los dados o que “cara” aparezca aproximadamente la mitad de veces que arrojamos una moneda.

**TABLA 4.3.1**

**RESPUESTAS PERSONA-ÍTEM ($x_{vi}$) CON RESPUESTAS INESPERADAS ENCERRADAS EN UN CÍRCULO**

| PERSONA | ÍTEM 4 | 5 | 7 | 6 | 8 | 12 | 14 | NÚMERO DE RESPUESTAS INESPERADAS | HABILIDAD DE LA PERSONA* |
|---|---|---|---|---|---|---|---|---|---|
| 11 | (0) | 1 | 1 | 1 | 1 | 0 | 0 | 1 | −1.2 |
| 12 | 1 | 1 | 1 | (0) | (0) | 0 | 0 | 2 | −1.2 |
| 17 | 1 | (0) | 1 | 1 | 1 | 0 | 0 | 1 | −0.6 |
| 3 | 1 | 1 | 1 | 1 | 1 | (1) | 0 | 1 | 0.0 |
| 13 | 1 | 1 | (0) | (0) | 1 | (1) | 0 | 3 | 0.0 |
| 29 | 1 | 1 | (0) | 1 | (0) | 0 | (1) | 3 | 0.0 |
| Número de Respuestas Inesperadas | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 11 | |
| Dificultad del ítem * | −3.9 | −3.3 | −3.3 | −2.9 | −2.0 | 1.7 | 2.8 | | |

"1" Esperado
"0" Inesperado

"0" Esperado
"1" Inesperado

* De las tablas 2.4.5 y 2.4.6

Nuestro modelo estima la probabilidad $\pi_{vi}$ de las instancias de respuesta $x_{vi}$ como

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)]$$

Donde $b_v$ = la medida de habilidad estimada de la persona $v$

y $d_i$=la dificultad de calibración estimada del ítem i.

Por lo tanto, podemos usar $p_{vi}$ como una estimación del valor esperado de instancias de $x_{vi}$.

La misma teoría nos dice que la varianza esperada de instancias de $x_{vi}$ es $\pi_{vi}(1 - \pi_{vi})$, que podemos estimar con $p_{vi}(1 - p_{vi})$. El resultado es un residuo estandarizado estimado $z_{vi}$ para cada $x_{vi}$ de

$$z_{vi} = (x_{vi} - p_{vi})/[p_{vi}(1 - p_{vi})]^{1/2}. \quad \textbf{[4.3.1]}$$

Para estimar este residuo estandarizado $z_{vi}$, restamos al valor observado $x_{vi}$ su valor estimado esperado $p_{vi}$ y estandarizamos esta diferencia residual dividiendo entre

$$[p_{vi}(1 - p_{vi})]^{1/2}$$

que es la desviación estándar binomial estimada a partir de las observaciones. En la medida en que nuestros datos se aproximan al modelo, esperamos que el residuo estimado $z_{vi}$ se distribuya de forma aproximadamente normal con una media de 0 y una varianza cercana a 1.

Por lo tanto, como un criterio aproximado pero útil para el ajuste de los datos al modelo, podemos examinar hasta qué punto estos residuos estandarizados se aproximan a una distribución normal, es decir:

$$z_{vi} \sim N(0,1)$$

o sus cuadrados se aproximen a una distribución ji-cuadrado con un grado de libertad, es decir:

$$z_{vi}^2 \sim x_1^2$$

Los valores de referencia de 0 para la media y 1 para la desviación estándar y las distribuciones de referencia N (0,1) y $x_1^2$ nos ayudan a decidir cuando los residuos estandarizados estimados se desvían significativamente de las expectativas del modelo. Este análisis de los residuos sugerirá si podemos utilizar los ítems para hacer mediciones, o si debemos trabajar más en ellos y en el diseño de la prueba para alinearlos con las expectativas razonables. También permitirá identificar a personas que no respondieron a la prueba de manera plausible.

Cuando un residuo cuadrático en particular $z_{vi}^2$ se vuelve muy grande, nos preguntamos si algo inesperado sucedió cuando la persona ν respondió el ítem i. Por supuesto, una sola respuesta inesperada es menos indicativa de problemas que una cadena de valores inesperadamente grandes de $z_{vi}^2$. Entonces el impacto acumulado de estos valores tomados sobre los ítems en una persona o sobre las personas para un ítem, generaría preocupación por la verosimilitud de la medida de la persona o de la calibración del ítem y, por lo tanto, pondría en duda el significado de la medición de esa persona o de la calibración de ese ítem.

Tomando en cuenta que $x_{vi}$, adquiere solo los dos valores de "0" y "1", podemos expresar estos residuos estandarizados en términos de las estimaciones $b_v$ y $d_i$.

De la ecuación 4.3.1 tenemos

$$z_x = (x - p)/[p(1-p)]^{1/2}$$

Entonces cuando $x = 0$ entonces $z_0 = (-p)/[p(1-p)]^{1/2} = -[p/(1-p)]^{1/2}$

y $x = 1$ entonces $z_1 = (1-p)/[p(1-p)]^{1/2} = +[(1-p)/p]^{1/2}$.

Si partimos de $p = \exp(b-d)/[1+\exp(b-d)]$

Entonces $p/(1-p) = \exp(b-d)$

y $(1-p)/p = \exp(d-b)$.

Por lo tanto e $z_0 = -\exp[(b-d)/2]$ $\qquad z_0^2 = \exp(b-d)$

y $z_1 = +\exp[(d-b)/2]$ $\qquad z_1^2 = \exp(d-b)$

o en general $z = (2x-1)\exp[(2x-1)(d-b)/2]$ **[4.3.2]**

$$z^2 = \exp[(2x-1)(d-b)] \quad \textbf{[4.3.3]}$$

Por lo tanto, exp (b – d) indica lo inesperado de una respuesta incorrecta a un ítem relativamente fácil, mientras que exp (d – b) indica lo inesperado de una respuesta correcta a un ítem relativamente difícil. Los valores de $z_0^2$= exp (b – d) y $z_1^2$= exp (d – b) determinarse para cada $x_{vi}$ de valores 0 o 1 y luego pueden acumularse sobre los ítems para evaluar la plausibilidad de la medida de cualquier persona, o sobre todas las personas para determinar la verosimilitud de la calibración de cualquier ítem.

Para evaluar las respuestas inesperadas en la Tabla 4.3.1, reemplazamos cada instancia de una respuesta inesperada por la diferencia entre la medida de la habilidad para esa persona y la dificultad de calibración para ese ítem. Para la persona 11 en el ítem 4, la respuesta incorrecta inesperada asociada con la habilidad de persona $b_v$ de –1.2 y la dificultad de ítem $d_i$ de –3.9 produce una diferencia ($b_v$ – $d_i$) de (–1.2) – (–3.9) = +2.7.

Esta diferencia de 2.7 para la Persona 11 en el ítem 4 se escribe en la posición de la respuesta inesperada en la matriz de la Tabla 4.3.2 donde también hemos calculado las diferencias para las otras instancias de respuestas inesperadas en la Tabla 4.3.1.

**TABLA 4.3.2**

**DIFERENCIAS HABILIDAD - DIFICULTAD ($b_v$ - $d_i$) PARA RESPUESTAS INESPERADAS**

| | ÍTEM | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| PERSONA | 4 | 5 | 7 | 6 | 8 | 12 | 14 | HABILIDAD DE LA PERSONA |
| 11 | 2.7 | | | | | | | – 1.2 |
| 12 | | | | 1.7 | 0.8 | | | – 1.2 |
| 17 | | 2.7 | | | | | | – 0.6 |
| 3 | | | | | | 1.7 | | 0.0 |
| 13 | | | 3.3 | 2.9 | | 1.7 | | 0.0 |
| 29 | | | 3.3 | | 2.0 | | 2.8 | 0.0 |
| Dificultad del ítem | – 3.9 | – 3.3 | – 3.3 | – 2.9 | – 2.0 | 1.7 | 2.8 | |

Dado que "1" esperado "0" inesperado El dato se calcula (b - d)

Dado que "0" esperado "1" inesperado el dato se calcula (d - b)

Las respuestas incorrectas inesperadas se han registrado como (b – d), pero las respuestas correctas inesperadas se han registrado como (d – b). Esto se debe a que si la respuesta es incorrecta, es decir, x = 0, entonces el índice de inesperado es exp (b – d), pero cuando la respuesta es correcta, es decir, x = 1, entonces el índice es exp (d – b).

La evidencia de que la respuesta es inesperada en la Tabla 4.3.2 se muestra con una diferencia positiva, ya sea (b – d) o (d – b). Los valores correspondientes para $z^2$ se consultar en la Tabla 4.3.3 que proporciona los valores de $z_0^2$= exp (b – d) para las respuestas incorrectas inesperadas o los valores $z_1^2$ = exp (d – b) para las respuestas correctas inesperadas. El dato $C_x$ en la Columna 1 de la Tabla 4.3.3 es $C_0$ = (b – d) cuando x = 0 y la respuesta es incorrecta, o también $C_1$ = (d – b) cuando x = 1 y la respuesta es correcta.

**TABLA 4.3.3**

**ESTADÍSTICAS DE DESAJUSTE**

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| DIFERENCIA ENTRE LA HABILIDAD DE LA PERSONA Y LA DIFICULTAD DEL ÍTEM C* | RESIDUO CUADRÁTICO ESTANDARIZADO $z^2 = \exp C$ | IMPROBABILIDAD DE LA RESPUESTA $p = 1/(1 + z^2)$ | EFICIENCIA RELATIVA DE LA OBSERVACIÓN $I = 400p(1 - p)$ | NÚMERO DE ÍTEMS REQUERIDOS PARA MANTENER LA MISMA PRECISIÓN $L = 1000/I$ |
| −0.6,0.4 | 1 | .50 | 100 | 10 |
| 0.5,0.9 | 2 | .33 | 90 | 11 |
| 1.0,1.2 | 3 | .25 | 75 | 13 |
| 1.3,1.5 | 4 | .20 | 65 | 15 |
| 1.6,1.7 | 5 | .17 | 55 | 18 |
| 1.8,1.8 | 6 | .14 | 50 | 20 |
| 1.9,2.0 | 7 | .12 | 45 | 22 |
| 2.1 | 8 | .11 | 40 | 25 |
| 2.2 | 9 | .10 | 36 | 28 |
| 2.3 | 10 | .09 | 33 | 30 |
| 2.4 | 11 | .08 | 31 | 32 |
| 2.5 | 12 | .08 | 28 | 36 |
| 2.6 | 13 | .07 | 25 | 40 |
| 2.7 | 15 | .06 | 23 | 43 |
| 2.8 | 16 | .06 | 21 | 48 |
| 2.9 | 18 | .05 | 20 | 50 |
| 3.0 | 20 | .05 | 18 | 55 |
| 3.1 | 22 | .04 | 16 | 61 |
| 3.2 | 25 | .04 | 15 | 66 |
| 3.3 | 27 | .04 | 14 | 73 |
| 3.4 | 30 | .03 | 12 | 83 |
| 3.5 | 33 | .03 | 11 | 91 |
| 3.6 | 37 | .03 | 10 | 100 |
| 3.7 | 40 | .02 | 9 | 106 |
| 3.8 | 45 | .02 | 9 | 117 |
| 3.9 | 49 | .02 | 8 | 129 |
| 4.0 | 55 | .02 | 7 | 142 |
| 4.1 | 60 | .02 | 6 | 156 |
| 4.2 | 67 | .02 | 6 | 172 |
| 4.3 | 74 | .01 | 5 | 189 |
| 4.4 | 81 | .01 | 5 | 209 |
| 4.5 | 90 | .01 | 4 | 230 |
| 4.6 | 99 | .01 | 4 | 254 |

*Para las respuestas incorrectas cuando x=0, se tiene $C_0 = (b - d)$. Para las respuestas correctas cuando x=1, se tiene $C_1 = (d - b)$.

**TABLA 4.3.4**

**AJUSTE CUDRÁTICO MEDIO ($z^2_{vi}$) PARA RESPUESTAS INESPERADAS**

| | ÍTEM | | | | | | | DESAJUSTE TOTAL DE LA PERSONA |
|---|---|---|---|---|---|---|---|---|
| PERSONA | 4 | 5 | 7 | 6 | 8 | 12 | 14 | |
| 11 | 15 | | | | | | | 15 |
| 12 | | | | 6 | 2 | | | 8 |
| 17 | | 15 | | | | | | 15 |
| 3 | | | | | | 6 | | 6 |
| 13 | | | 27 | 18 | | 6 | | 51 |
| 29 | | | 27 | | 7 | | 17 | 51 |
| Desajuste Total del Ítem | 15 | 15 | 54 | 24 | 9 | 12 | 17 | 146 |

"1" esperada "0" inesperada

"0" esperada "1" inesperada

Podemos ubicar la diferencia +2.7 en (b – d) para la Persona 11 en el ítem 4 en la primera columna de la Tabla 4.3.3 y leer el valor correspondiente de $z^2$ en la Columna 2 igual a 15. Este valor y todos los demás valores para las diferencias en la Tabla 4.3.2 han sido registrados en la Tabla 4.3.4, que ahora contiene todos los $z^2$ para cada instancia inesperada que hayamos observado para las seis personas y los siete ítems. En los marginales de la Tabla 4.3.4 están las sumas de estos $z^2$ para cada persona e ítem. Estas sumas indican cuán inesperado es el patrón de respuestas de la persona o del ítem.

En la Columna 3 de la Tabla 4.3.3 tenemos $p = 1/(1 + z^2)$, que indica qué tan improbable es la respuesta observada. Esto proporciona un nivel de significancia para la hipótesis nula del ajuste de cualquier respuesta particular. Con nuestro ejemplo que tiene una diferencia (b – d) de 2.7 tenemos un nivel de significancia de 0.06 contra la hipótesis nula de que la respuesta de la Persona 11 al ítem 4 esté de acuerdo con el modelo. Los valores están distribuidos aproximadamente como $\chi^2$ con 1 grado de libertad cada uno. Cuando se acumulan sobre los ítems para una persona o sobre las personas para un ítem, las sumas resultantes se distribuyen aproximadamente como $\chi^2$ con (L – 1) grados de libertad para una persona y (N – 1) grados de libertad para un ítem.

En la Columna 4 de la Tabla 4.3.3 tenemos $I = 400p(1 – p)$, se trata de un índice de la eficiencia relativa con la que una observación en un valor (b – d) dado puede proporcionar información sobre la interacción de la persona y del ítem. Este índice se escala por el factor 400 para que dé

la cantidad de información disponible para la observación como un porcentaje de la información máxima que podría proporcionar en (b – d) =0, es decir, justo en el objetivo. El porcentaje de información en una observación se puede usar para juzgar el valor de cualquier ítem particular para medir una persona. Esto se puede hacer considerando cuánto de la cantidad de información se perdería si se elimina ese ítem de la prueba. Por lo tanto, un valor I de 23% para la Persona 11 en el Ítem 4 nos da una indicación de cuánto ganamos al incluir el Ítem 4 en la medición de la Persona 11 o de cuánto perderíamos si elimin——ramos ese ítem 4.

La forma en que el concepto de información o eficiencia participa al juzgar el valor de una observación es a través de su relación con la precisión de la medición. La precisión de una medición depende de la cantidad de ítems que tiene un registro y de la relevancia de cada ítem para una persona en particular. Podemos evaluar de forma simple la contribución que hace cada ítem a nuestro conocimiento de la persona si calculamos el porcentaje que representa respecto del mejor ítem posible. Eso es lo que proporcionan los valores de I en la Columna 4.

Si el ítem y la persona están cerca el uno del otro, es decir, en el "blanco" u objetivo, entonces el ítem contribuye más a la medida de la persona que si el ítem y la persona están alejados. Cuanto mayor sea la diferencia entre el ítem y la persona, mayor será el número de ítems necesarios para obtener una medida de precisión comparable y, como resultado, menos eficiente resulta cada ítem.

Por ejemplo, se requieren cinco ítems de 20% para proporcionar tanta información sobre una persona de lo que podría ser proporcionada por un ítem de 100%. Por lo tanto, cuando (b – d) es aproximadamente 3.0, se necesitan unas cuatro o cinco veces más ítems para proporcionar tanta información como podrían proporcionar ítems que estuvieran a menos de un lógito de la persona, es decir, en la región $|b - d| < 1$.

En general, la longitud de prueba necesaria para mantener un nivel específico de precisión en una medición es inversamente proporcional a la eficiencia relativa de los ítems usados. La última columna de la Tabla 4.3.3 presenta el número L de ítems poco eficientes que son necesarios para coincidir con la precisión de 10 ítems que estén justo en el "blanco" u objetivo esperado[4].

Para facilitar el uso de la Tabla 4.3.3, se ha organizado en cuatro secciones:

| | | |
|---|---|---|
| Justo en el "blanco" u objetivo | $\|b - d\| < 2$ | La eficiencia del ítem es 45% o mayor, en la región $\|b–d\| <1$ es 79% o mayor. El desajuste es difícil de detectar. |
| Ligeramente fuera del "blanco" u objetivo | $2 < \|b - d\| < 3$ | Eficiencia pobre, menos de 45%. El desajuste se llega a detectar cuando se acumulan respuestas inesperadas. |
| Fuera del "blanco" u objetivo | $3 < \|b - d\| < 4$ | Eficiencia muy pobre, menos de 18%. Incluso las respuestas individuales inesperadas pueden indicar irregularidades significativas en las respuestas. |
| Extremadamente lejos del "blanco" u objetivo | $4 < \|b - d\|$ | Eficiencia virtualmente nula, menos de 7%. Las respuestas inesperadas son siempre inaceptables. |

[4] La palabra "target" usada por los autores para designar a la persona o grupo focales de la prueba, se podría denominar el "blanco" o la "diana" en el caso de un tiro al blanco, pero esta denominación no pareció suficientemente apropiada para el sentido de la traducción y por ello se optó por usar "objetivo".

## 4.4 REGISTROS DESAJUSTADOS EN LAS PERSONAS

Al examinar los renglones de la Tabla 4.3.4 ante valores altos de $z^2$ en los registros de las personas, hallamos que los valores acumulados más grandes se tienen con las personas 13 y 29. Se trata de las dos personas cuyo patrón de respuestas es más cuestionable y por ello revisaremos sus registros con mayor detalle.

**TABLA 4.4.1**
**ANÁLISIS DETALLADOS DE AJUSTE PARA LAS PERSONAS 29 Y 13**

| | | ITEM | 4 | 5 | 7 | 6 | 9 | 8 | 10 | 1 | 13 | 12 | 14 | 15 | 16 | 17 | PUNTAJE r | SUMA DE CUADRADOS $\Sigma z^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DIFICULTAD $d_i$ | −3.9 | −3.3 | −3.3 | −2.9 | −2.9 | −2.0 | −1.4 | 0.6 | 1.5 | 1.7 | 2.8 | 4.3 | 4.3 | 4.3 | | |
| PERSONA 29 (b=0.0) | | x | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 7 | |
| | x = 0 | (b - d) | | | 3.3 | | | 2.0 | | | −1.5 | −1.7 | | −4.3 | −4.3 | −4.3 | | |
| | x = 1 | (d - b) | −3.9 | −3.3 | | −2.9 | −2.9 | | −1.4 | 0.6 | | | 2.8 | | | | | |
| | | $z^2$ | 0 | 0 | 27 | 0 | 0 | 7 | 0 | 2 | 0 | 0 | 17 | 0 | 0 | 0 | | 53 |
| PERSONA 13 (b=0.0) | | x | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 7 | |
| | x = 0 | (b - d) | | | 3.3 | 2.9 | | | | | −1.5 | | −2.8 | −4.3 | −4.3 | −4.3 | | |
| | x = 1 | (d - b) | −3.9 | −3.3 | | | −2.9 | −2.0 | −1.4 | 0.6 | | 1.7 | | | | | | |
| | | $z^2$ | 0 | 0 | 27 | 18 | 0 | 0 | 0 | 2 | 0 | 6 | 0 | 0 | 0 | 0 | | 53 |

La Tabla 4.4.1 detalla los vectores de respuesta para las personas 13 y 29 sobre los 14 ítems. Para cada persona mostramos sus respuestas de 0 o 1, sus diferencias calculadas (b – d) o (d – b), dependiendo de si la respuesta es 0 para incorrecta o 1 para correcta, y su valor correspondiente de $z^2$. Las sumas de la fila de $z^2$ para la Persona 13 y la Persona 29 son, casualmente, 53. De acuerdo con el modelo, estos valores $z^2$ acumulados deberían seguir una distribución ji-cuadrado con 1 grado de libertad por cada $z^2$ menos el grado de libertad necesario para estimar la medida b de la persona.

Además, cualquier suma de $z^2$, cuando se divide por sus grados de libertad, debe tener una distribución de la media cuadrática con la expresión $v = \Sigma z^2/f$, que puede ser convenientemente evaluada conforme al estadístico t:

$$v = \sum z^2 / f \qquad \text{y} \qquad f = L - 1 \qquad \textbf{[4.4.1]}$$

$$t = [\ell n(v) + v - 1][f/8]^{1/2} \sim N(0,1) \qquad \textbf{[4.4.2]}$$

que se aproxima a una distribución normal unitaria.

Para la Persona 13 tenemos:

$$v_{13} = \sum_{i}^{14} z^2_{13i} / (14 - 1) = 53/13 = 4.1$$

para cual

$$t_{13} = [\ell n(v_{13}) + v_{13} - 1][13/8]^{1/2} = [1.4 + 4.1 - 1][1.3] = 5.8,$$

que es un valor bastante improbable para t, si el rendimiento de esta persona estuviera ajustado al modelo.

Para la Persona 29 observamos los mismos resultados y el mismo estadístico t. Con un desajuste tan importante parece razonable diagnosticar ambos registros como fuentes de datos inaceptables, tanto para la medición de estas dos personas como para la calibración de los ítems.

## 4.5 REGISTROS DE ÍTEMS DESAJUSTADOS

También podemos ver en la Tabla 4.3.4 que los Ítems 7 y 6 muestran el mayor desajuste, especialmente el Ítem 7 que tiene un valor acumulado $z^2$ de 54. En la Tabla 4.5.1 analizamos los vectores de datos completos de estos dos ítems, mostrando para cada persona la respuesta de 0 o 1, los datos asociados (b – d) o (d – b) y sus respectivos $z^2$.

$$v_7 = \sum_{v}^{34} z^2_{v7} / (34 - 1) = 57/33 = 1.7$$

para cual

$$t_7 = [\ell n(v_7) + v_7 - 1][33/8]^{1/2} = [0.5 + 1.7 - 1][2.0] = 2.4$$

que es un valor de t que parece igualmente improbable, si el ítem estuviera ajustado al modelo.

**TABLA 4.5.1**

**ANALISIS COMPLETO DE AJUSTE PARA LOS ÍTEMS 7 Y 6**

| | | | ITEM 7 (d = − 3.3) | | | | ITEM 6 (d = − 2.9) | | |
|---|---|---|---|---|---|---|---|---|---|
| PERSONA | HABILIDAD b | RESPUESTA x | x = 0 (b - d) | x = 1 (d - b) | $z^2$ | RESPUESTA x | x = 0 (b - d) | x = 1 (d - b) | $z^2$ |
| 25 | − 3.8 | 0 | − 0.5 | | 1 | 1 | | 0.9 | 3 |
| 4 | − 2.8 | 1 | | − 0.5 | 1 | 0 | + 0.1 | | 1 |
| 33 | − 2.8 | 1 | | − 0.5 | 1 | 0 | + 0.1 | | 1 |
| 1 | − 1.9 | 1 | | − 1.4 | 0 | 1 | | − 1.0 | 0 |
| 27 | − 1.9 | 1 | | − 1.4 | 0 | 1 | | − 1.0 | 0 |
| 11 | − 1.2 | 1 | | − 2.1 | 0 | 1 | | − 1.7 | 0 |
| 12 | − 1.2 | 1 | | − 2.1 | 0 | 0 | + 1.7 | | 6 |
| 17 | − 0.6 | 1 | | − 2.7 | 0 | 1 | | − 2.3 | 0 |
| 19 | − 0.6 | 1 | | − 2.7 | 0 | 1 | | − 2.3 | 0 |
| 30 | − 0.6 | 1 | | − 2.7 | 0 | 1 | | − 2.3 | 0 |
| 2 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 3 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 5 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 6 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 8 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 9 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| (13) | 0.0 | 0 | + 3.3 | | 27 | 0 | + 2.9 | | 18 |
| 16 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 26 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 28 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| (29) | 0.0 | 0 | + 3.3 | | 27 | 1 | | − 2.9 | 0 |
| 31 | 0.0 | 1 | | − 3.3 | 0 | 1 | | − 2.9 | 0 |
| 10 | + 0.6 | 1 | | − 3.9 | 0 | 1 | | − 3.5 | 0 |
| 18 | + 0.6 | 1 | | − 3.9 | 0 | 1 | | − 3.5 | 0 |
| 14 | + 0.6 | 1 | | − 3.9 | 0 | 1 | | − 3.5 | 0 |
| 32 | + 0.6 | 1 | | − 3.9 | 0 | 1 | | − 3.5 | 0 |
| 20 | + 0.6 | 1 | | − 3.9 | 0 | 1 | | − 3.5 | 0 |
| 21 | + 1.2 | 1 | | − 4.5 | 0 | 1 | | − 4.1 | 0 |
| 22 | + 1.2 | 1 | | − 4.5 | 0 | 1 | | − 4.1 | 0 |
| 23 | + 1.2 | 1 | | − 4.5 | 0 | 1 | | − 4.1 | 0 |
| 34 | + 1.2 | 1 | | − 4.5 | 0 | 1 | | − 4.1 | 0 |
| 15 | + 1.9 | 1 | | − 5.2 | 0 | 1 | | − 4.8 | 0 |
| 7 | + 2.8 | 1 | | − 6.1 | 0 | 1 | | − 5.7 | 0 |
| 24 | + 2.8 | 1 | | − 6.1 | 0 | 1 | | − 5.7 | 0 |
| SUMA DE CUADRADOS | | | | | 57 | | | | 29 |

Para el ítem 6

$$v_6 = \sum_{v}^{34} z^2_{v6}/(34-1) = 29/33 = 0.9$$

para el cual

$$t_6 = [\ell n(v_6) + v_6 - 1][33/8]^{1/2} = [-0.1 + 0.9 - 1][2.0] = 0.4,$$

que obviamente no es un desajuste significativo.

Encontramos que la media cuadrática para el Ítem 7 es significativa pero que la media cuadrática del ítem 6 no lo es. Sin embargo, si examinamos nuevamente la Tabla 4.5.1, vemos que las dos personas 13 y 29, cuyo desajuste es significativo, son quienes más contribuyen al desajuste de estos dos ítems. En este momento tenemos la oportunidad de mejorar el ajuste de los datos al modelo, ya sea eliminando el Ítem 7 y observando lo que sucede o eliminando a las personas 13 y 29.

## 4.6 BREVE RESUMEN DEL ANÁLISIS DE AJUSTE

Para cualquier respuesta de la Persona $v$ al Ítem i

$x_{vi} = 0$ si es "incorrecto" y

$x_{vi} = 1$ si es "correcto"

El residuo cuadrático medio estándar se convierte en

$z^2_{vi} = \exp(b - d)$, para $x_{vi} = 0$, incorrecto, y

$z^2_{vi} = \exp(d - b)$, para $x_{vi} = 1$, correcto.

Para evaluar el ajuste general de la persona $v$, sumados su vector de residuos cuadráticos estandarizados ($z_{vi}^2$) sobre los ítems de la prueba de i = 1, L y calculamos su estadístico de desajuste personal como:

$$v_v = \sum_{i}^{L} z^2_{vi}/(L-1) \sim F_{L-1,\infty} \qquad \textbf{[4.6.1]}$$

con

$$t_v = [\ell n(v_v) + v_v - 1][(L-1)/8]^{1/2} \sim N(0,1) \qquad \textbf{[4.6.2]}$$

Para evaluar el ajuste del Ítem i, sumamos el vector de residuos cuadráticos estandarizados ($z_{vi}^2$) del ítem sobre la muestra de personas de v = 1, N y calculamos el estadístico de desajuste como:

$$v_i = \sum_{v}^{N} z^2_{vi}/(N-1) \sim F_{N-1,\infty}$$

con

$$t_i = [\ell n(v_i) + v_i - 1][(N-1)/8]^{1/2} \sim N(0,1) \qquad \textbf{[4.6.4]}$$

## 4.7 ANÁLISIS DE AJUSTE POR COMPUTADORA

En el análisis de ajuste hecho a mano, vimos que ciertos registros de personas e ítems tenían residuos que hemos calificado como significativos. Después de haber mostrado los procedimientos para el análisis de ajuste a mano pasamos al análisis por computadora y volvemos a nuestra calibración de la PCK con los 18 ítems y las 34 personas. En la calibración de la PCK vemos en la sección izquierda de la Tabla 4.7.1, que la media cuadrática de ajuste del Ítem 7 produce el mayor desajuste con un valor de 1.98 no muy lejos del valor 1.7 encontrado en nuestro cálculo manual. Del análisis de desajuste de las personas sabemos que las personas 13 y 29 contribuyeron grandemente a este desajuste en el Ítem 7. Si careciéramos de esta información al momento de hacer la calibración, podríamos haber considerado desechar al ítem 7 debido a su alto valor cuadrático medio y decidido recalibrar la prueba con el Ítem 7 eliminado. Al hacer esto, se obtienen los resultados que se encuentran en la sección central de la Tabla 4.7.1. Ahora vemos que el ítem 6 alcanza un desajuste de 2.73 a pesar de que previamente tuvo una media cuadrática de 0.90, que ajustaba bien cuando hicimos la calibración de los 14 ítems. Este cambio en el estado del ítem 6 es problemático. No parece que tenemos un buen conjunto de ítems. Sin embargo, demos un paso más y recalibremos una vez más, pero esta vez eliminando los dos Ítems 7 y 6. Los resultados están en la sección derecha de la Tabla 4.7.1. Por desgracia, ahora nos encontramos que el ítem 8 se ha desajustado. Estos intentos de encontrar un conjunto adecuado de ítems parecen estar condenados al fracaso.

**TABLA 4.7.1**
**ANÁLISIS DE AJUSTE CON UCON: DESECHADO DE ÍTEMS**

| ÍTEMS ORDENADOS POR AJUSTE | | | | ÍTEMS ORDENADOS POR AJUSTE | | | | ÍTEMS ORDENADOS POR AJUSTE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NÚMERO CONSECUTIVO | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | AJUSTE MN SQ | NÚMERO CONSECUTIVO | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | AJUSTE MN SQ | NÚMERO CONSECUTIVO | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | AJUSTE MN SQ |
| 16 | 6 | 4.56 | 0.13 | | | | | | | | |
| 17 | 6 | 4.56 | 0.13 | 16 | 6 | 4.45 | 0.13 | | | | |
| 15 | 6 | 4.56 | 0.13 | 17 | 6 | 4.45 | 0.13 | 16 | 6 | 4.29 | 0.13 |
| 9 | 4 | − 3.22 | 0.23 | 15 | 6 | 4.45 | 0.13 | 17 | 6 | 4.29 | 0.13 |
| 4 | 3 | − 4.19 | 0.37 | 9 | 4 | − 3.75 | 0.22 | 15 | 6 | 4.29 | 0.13 |
| 13 | 5 | 1.86 | 0.40 | 4 | 3 | − 4.75 | 0.40 | 9 | 4 | − 4.30 | 0.18 |
| 8 | 4 | − 2.24 | 0.44 | 13 | 5 | 1.77 | 0.42 | 4 | 3 | − 5.38 | 0.35 |
| 5 | 3 | − 3.65 | 0.53 | 5 | 3 | − 4.20 | 0.54 | 13 | 5 | 1.61 | 0.44 |
| 11 | 5 | 0.76 | 0.77 | 11 | 5 | 0.65 | 0.68 | 5 | 3 | − 4.79 | 0.64 |
| 10 | 4 | − 1.50 | 0.79 | 14 | 6 | 3.11 | 0.74 | 11 | 5 | 0.43 | 0.67 |
| 6 | 3 | − 3.22 | 0.90 | 12 | 5 | 2.04 | 0.83 | 14 | 6 | 2.96 | 0.77 |
| 12 | 5 | 2.14 | 0.97 | 10 | 4 | − 1.81 | 0.88 | 10 | 4 | − 2.19 | 0.78 |
| 14 | 6 | 3.21 | 1.33 | 8 | 4 | − 2.66 | 1.03 | 12 | 5 | 1.89 | 0.82 |
| 7 | 4 | − 3.65 | 1.98 | 6 | 3 | − 3.75 | 2.73 | 8 | 4 | − 3.11 | 1.29 |
| TODAS LAS PERSONAS Y TODOS LOS ÍTEMS | | | | ELIMINANDO ÍTEM 7 | | | | ELIMINANDO ÍTEMS 7 Y 6 | | | |
| L = 14 N = 34 | | | | L = 13 N = 34 | | | | L = 12 N = 34 | | | |

Supongamos, en cambio, que una vez terminada la primera calibración de los ítems de la PCK, decidimos evaluar el ajuste de las personas. El análisis por computadora para el ajuste de las personas que se muestra en la Tabla 4.7.2, también identifica que las personas 13 y 29 son las que producen los mayores estadísticos de desajuste. Así que volvamos a recalibrar los 14 ítems completos, pero con estas dos personas eliminadas. Ahora, en la Tabla 4.7.3, vemos que las medias cuadráticas son lo suficientemente pequeñas en todos los ítems como para dejarnos satisfechos. Quitando los dos registros de las personas inadecuadas hemos conseguido que todos los ítems estén en buen acuerdo con el modelo.

**TABLA 4.7.2**

**ANÁLISIS DE AJUSTE DE LAS PERSONAS CON UCON**

| PERSONA | PUNTAJE r | HABILIDAD UCON b | DESAJUSTE UCON v | |
|---|---|---|---|---|
| 25 | 2 | − 4.4 | 0.5 | |
| 4 | 3 | − 3.7 | 0.4 | |
| 33 | 3 | − 3.7 | 0.9 | |
| 1 | 4 | − 3.1 | 0.3 | |
| 27 | 4 | − 3.1 | 0.3 | |
| 11 | 5 | − 2.3 | 0.8 | |
| 12 | 5 | − 2.3 | 0.5 | |
| 17 | 6 | − 1.4 | 1.0 | |
| 19 | 6 | − 1.4 | 0.2 | |
| 30 | 6 | − 1.4 | 0.2 | |
| 2 | 7 | − 0.3 | 0.1 | |
| 3 | 7 | − 0.3 | 1.4 | |
| 5 | 7 | − 0.3 | 0.1 | |
| 6 | 7 | − 0.3 | 0.1 | |
| 8 | 7 | − 0.3 | 0.1 | |
| 9 | 7 | − 0.3 | 0.1 | |
| (13) | 7 | − 0.3 | [5.7] | ( A mano PROX = 4.1 ) |
| 16 | 7 | − 0.3 | 0.6 | |
| 26 | 7 | − 0.3 | 0.1 | |
| 28 | 7 | − 0.3 | 0.6 | |
| (29) | 7 | − 0.3 | [6.6] | ( A mano PROX = 4.1 ) |
| 31 | 7 | − 0.3 | 0.1 | |
| 10 | 8 | + 1.0 | 0.2 | |
| 14 | 8 | + 1.0 | 0.2 | |
| 18 | 8 | + 1.0 | 0.4 | |
| 20 | 8 | + 1.0 | 0.4 | |
| 32 | 8 | + 1.0 | 0.2 | |
| 21 | 9 | + 2.0 | 0.2 | |
| 22 | 9 | + 2.0 | 0.2 | |
| 23 | 9 | + 2.0 | 0.7 | |
| 34 | 9 | + 2.0 | 0.7 | |
| 15 | 10 | + 3.0 | 0.2 | |
| 7 | 11 | + 3.9 | 0.4 | |
| 24 | 11 | + 3.9 | 0.9 | |
| | | Media | 0.7 | |
| | | Desviación Estándar | 1.6 | |

**TABLA 4.7.3**

**ANÁLISIS DE AJUSTE CON UCON: PERSONAS ELIMINADAS**

| NÚMERO CONSECUTIVO | NOMBRE DEL ÍTEM | DIFICULTAD DEL ÍTEM | AJUSTE MN | AJUSTE SQ |
|---|---|---|---|---|
| 7 | 4 | – 5.70 | 0.10 | |
| 16 | 6 | 5.27 | 0.13 | |
| 17 | 6 | 5.27 | 0.13 | |
| 15 | 6 | 5.27 | 0.13 | |
| 8 | 4 | – 2.87 | 0.17 | |
| 9 | 4 | – 3.73 | 0.21 | |
| 6 | 3 | – 4.24 | 0.34 | |
| 13 | 5 | 2.34 | 0.38 | |
| 4 | 3 | – 4.84 | 0.40 | |
| 14 | 6 | 4.43 | 0.55 | |
| 5 | 3 | – 4.24 | 0.64 | |
| 11 | 5 | 1.51 | 0.70 | |
| 12 | 5 | 3.01 | 0.99 | |
| 10 | 4 | – 1.48 | 1.03 | |

Borrando Personas 13 y 29

L = 14 N = 32

Parece claro que fueron los registros de respuestas impredecibles de estas dos personas los que produjeron que nos parecieran desajustados primero el ítem 7 y luego el ítem 6. Por lo tanto, hemos aprendido que si eliminamos ítems de manera sucesiva sin analizar el ajuste de las personas, podemos llegar a creer que los ítems desajustan cuando, de hecho, son los registros de respuesta de unas pocas personas irregulares los que están causando los problemas. El tamaño de muestra utilizado en nuestro ejemplo es muy pequeño, lo cual exagera el impacto de las dos personas irregulares, pero incluso las muestras grandes no borran por completo la posible influencia contaminante de los registros de personas irregulares, con la agravante de que en una muestra grande los registros defectuosos pueden ser muy difíciles de detectar y, por lo tanto, seguir siendo desconocidos a menos que se realicen rutinariamente pruebas explícitas de ajuste de las personas.

# 5 CONSTRUCCIÓN DE UNA VARIABLE

## 5.1 GENERALIZANDO LA DEFINICIÓN DE UNA VARIABLE

En los capítulos 2, 3 y 4 hemos mostrado cómo exponer y evaluar las relaciones entre los instrumentos que intentan hacer una medición, es decir, los ítems de la prueba y los objetos que pretende medir, es decir, las personas. Esto nos prepara para el presente capítulo, que se ocupa de la manera de definir una variable.

Con un procedimiento de calibración viable y un método para la evaluación del ajuste, resulta práctico dirigir nuestra atención a una actividad mucho más importante, a saber, el análisis crítico de los ítems calibrados para ver si implican la posibilidad de que una variable sea generalizable en forma útil. Queremos saber si nuestros ítems calibrados se distribuyen de manera que muestren una dirección coherente y significativa. Si no se distribuyen de ninguna manera entonces todo lo que hemos logrado es definir un punto, tal vez en alguna variable o tal vez no. Pero la variable en sí misma, cualquiera que esta sea, podrá seguir siendo oscura.

Ahora intentaremos mostrar cómo se pueden usar los ítems calibrados para definir una variable y cómo saber si tiene sentido la definición operacional resultante de esa variable. Comenzaremos examinando el grado en que la dispersión de las dificultades de los ítems excede sustancialmente el error estándar de sus estimaciones, es decir, el grado en que los datos le dan una dirección a la variable. Por ejemplo, supongamos que consideramos las estimaciones de dos dificultades de ítems con sus respectivos errores estándar. ¡Para que estos dos ítems puedan definir una línea entre ellos, deberían tener una diferencia entre sus estimaciones sustancialmente mayor que el error estándar de esta diferencia! Solo si las dos estimaciones están bien separados por varios de estos errores estándar comenzaremos a ver una línea entre los dos ítems que sugieren una dirección para la variable que definen.

Sin embargo, si la comparación estas dos estimaciones no difieren más que en uno o dos errores estándar, entonces se superpondrían sustancialmente y, por lo tanto, no podríamos asumir que estos dos valores difieren y por ello no definirían una dirección para la variable. En cambio, los dos ítems pueden definir un punto, pero sin dirección.

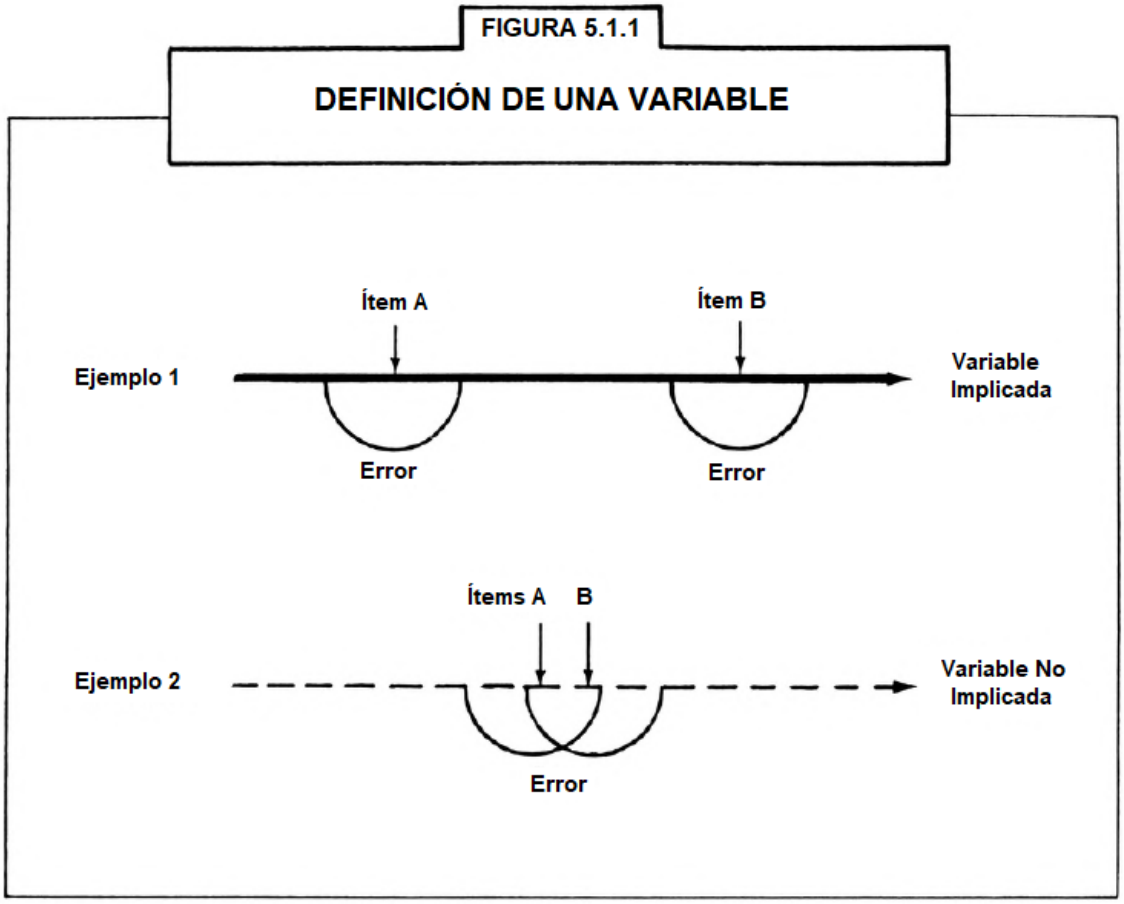
La figura 5.1.1 ilustra esto. En el Ejemplo 1 tenemos los ítems A y B separados entre sí por varios errores estándar. Incluso con dos ítems ya comenzamos a ver una dirección para la variable, al menos como lo permiten estos dos ítems. En el segundo ejemplo, sin embargo, encontramos dos ítems tan cerca uno del otro que, teniendo en cuenta sus errores estándar, no son separables. Hemos encontrado un punto. Pero no se ha establecido ninguna dirección, por lo que no puede implicarse una variable.

Como ejemplo de definición de la variable, continuaremos nuestro estudio de los datos de la PCK a ver qué tan bien sus ítems logran definir una variable y qué variable parece ser.

## 5.2 DEFINICIÓN DE LA VARIABLE DE LA PCK.

Los ítems de la PCK se integran por una serie de golpes que crece en longitud aumentando el número de golpes y crece en complejidad por la distancia entre golpes consecutivos y el número de inversiones en la dirección del movimiento.

FIGURA 5.1.1




En la Figura 5.2.1 se tienen los 18 ítems que componen originalmente a la PCK. Cada ítem se describe por un nombre numérico, correspondiente a la serie de golpes y al patrón de orden de dichos golpes.

La Tabla 5.2.1 se centra en los 14 ítems de la PCK que fueron calibrados en los Capítulos 2 y 3. Los ítems 1, 2 y 3 no están incluidos porque fueron demasiado fáciles para las 34 personas de la muestra y el ítem 18 tampoco está incluido por ser demasiado difícil. La tabla 5.2.1 proporciona los nombres de los ítems, las series de golpes, las dificultades de los ítems y sus errores estándar. El rango de dificultades de estos 14 ítems va desde -4.2 lógitos a +4.6 lógitos.

Las dificultades de los ítems en la Tabla 5.2.1 los hacen cuantitativamente explícitos para nuestra definición de la variable de la PCK ubicando los 14 ítems en sus posiciones calibradas a lo largo de la recta de la variable. Esto se hace en la Figura 5.2.2. Debido a que varios ítems tienen la misma dificultad o están muy cercanos en términos de sus errores estándar, difícilmente pueden distinguirse, por lo que solo estamos mostrando los ocho ítems que mejor se ubican a lo largo de la variable de la PCK. Los semicírculos en la Columna 2 de la Figura 5.2.2 muestran una tolerancia de un error estándar alrededor de cada dificultad estimada. Podemos ver que los ítems 4, 6, 8 y 10 definen el extremo fácil de la variable. Luego hay una brecha indefinida bastante amplia en el medio. Finalmente, los ítems 11, 12, 14 y 16 definen el extremo difícil. Los patrones de golpes en la columna 3 describen el movimiento que se realiza a lo largo de la variable en términos del incremento en el número de golpes y de la complejidad del patrón. La columna 4 da la distribución a lo largo de la variable de la PCK de las 34 personas que participaron en la calibración inicial.

**FIGURA 5.2.1**

**DESCRIPCIÓN DE LA VARIABLE PCK**

| NOMBRE DEL ÍTEM | SERIE DE GOLPES | NÚMERO DE BLOQUES | PATRÓN DE GOLPES |
|---|---|---|---|
| 1 | 1 – 4 | 4 3 2 1 | |
| 2 | 2 – 3 | 4 3 2 1 | |
| 3 | 1 – 2 – 4 | 4 3 2 1 | |
| 4 | 1 – 3 – 4 | 4 3 2 1 | |
| 5 | 2 – 1 – 4 | 4 3 2 1 | |
| 6 | 3 – 4 – 1 | 4 3 2 1 | |
| 7 | 1 – 4 – 3 – 2 | 4 3 2 1 | |
| 8 | 1 – 4 – 2 – 3 | 4 3 2 1 | |
| 9 | 1 – 3 – 2 – 4 | 4 3 2 1 | |

FIGURA 5.2.1

## DESCRIPCIÓN DE LA VARIABLE PCK (Continuación)

| NOMBRE DEL ÍTEM | SERIE DE GOLPES | NÚMERO DE BLOQUES | PATRÓN DE GOLPES |
|---|---|---|---|
| 10 | 2 – 4 – 3 – 1 | 4 3 2 1 | |
| 11 | 1 – 3 – 1 – 2 – 4 | 4 3 2 1 | |
| 12 | 1 – 3 – 2 – 4 – 3 | 4 3 2 1 | |
| 13 | 1 – 4 – 3 – 2 – 4 | 4 3 2 1 | |
| 14 | 1 – 4 – 2 – 3 – 4 – 1 | 4 3 2 1 | |
| 15 | 1 – 3 – 2 – 4 – 1 – 3 | 4 3 2 1 | |
| 16 | 1 – 4 – 2 – 3 – 1 – 4 | 4 3 2 1 | |
| 17 | 1 – 4 – 3 – 1 – 2 – 4 | 4 3 2 1 | |
| 18 | 4 – 1 – 3 – 4 – 2 – 1 – 4 | 4 3 2 1 | |

**TABLA 5.2.1**

**CALIBRACIÓN DE LA VARIABLE DE LA PCK CON ÍTEMS EN ORDEN DE DIFICULTAD**

| NOMBRE DEL ÍTEM | SERIE DE GOLPES | CALIBRACIÓN DEL ÍTEM | ERROR ESTÁNDAR |
|---|---|---|---|
| 4 | 1 – 3 – 4 | -4.2 | 0.8 |
| 5 | 2 – 1 – 4 | -3.6 | 0.7 |
| 7 | 1 – 4 – 3 – 2 | -3.6 | 0.7 |
| 6 | 3 – 4 – 1 | -3.2 | 0.6 |
| 9 | 1 – 3 – 2 – 4 | -3.2 | 0.6 |
| 8 | 1 – 4 – 2 – 3 | -2.2 | 0.6 |
| 10 | 2 – 4 – 3 – 1 | -1.5 | 0.5 |
| 11 | 1 – 3 – 1 – 2 – 4 | 0.8 | 0.5 |
| 13 | 1 – 4 – 3 – 2 – 4 | 1.9 | 0.5 |
| 12 | 1 – 3 – 2 – 4 – 3 | 2.1 | 0.6 |
| 14 | 1 – 4 – 2 – 3 – 4 – 1 | 3.2 | 0.7 |
| 15 | 1 – 3 – 2 – 4 – 1 – 3 | 4.6 | 1.1 |
| 16 | 1 – 4 – 2 – 3 – 1 – 4 | 4.6 | 1.1 |
| 17 | 1 – 4 – 3 – 1 – 2 – 4 | 4.6 | 1.1 |
| | Media | 0.0 | |
| | Desviación Estándar | 3.4 | |

Calibración UCON de la Tabla 3.4.1

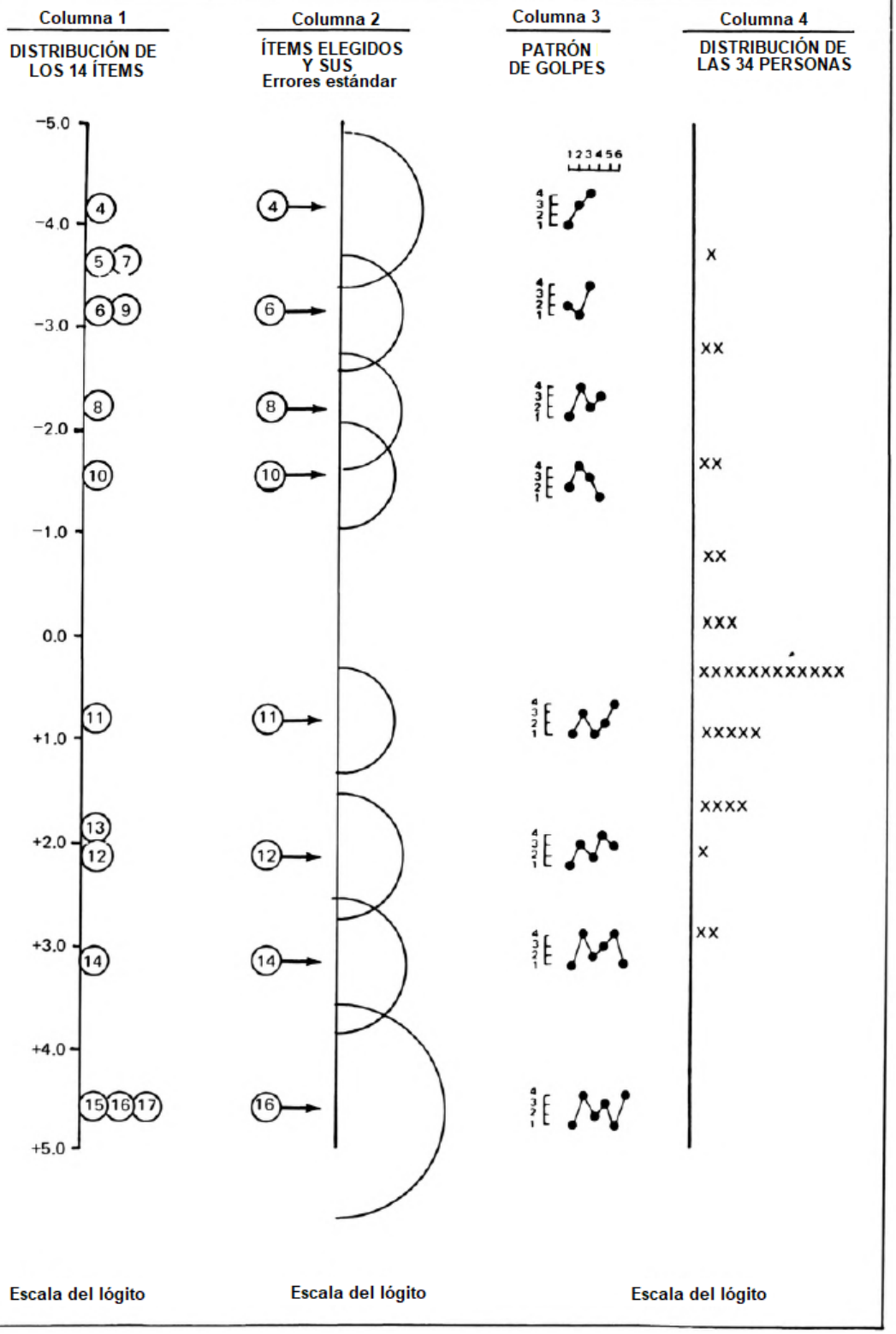
Vemos que la mayoría de las personas de la muestra se encuentran en el centro de la prueba. Pero eso es justo en esa zona donde tenemos una gran brecha en los ítems de prueba. Hemos descubierto algo importante y útil para nosotros, a saber, que nuestro instrumento de prueba es más débil en el punto donde está la moda de nuestra muestra. Resulta claro que, si queremos discriminar entre la mayoría de las personas encontradas en el rango medio de la PCK, entonces debemos construir algunos ítems adicionales cuya dificultad se encuentre dentro del rango medio y que serán más apropiados para medir las habilidades dentro de este rango medio.

## 5.3 INTENSIFICAR Y EXTENDER LA VARIABLE DE LA PCK

Para mejorar la medición a lo largo de la variable de la PCK, especialmente en el rango medio, se requiere un mayor desarrollo de los ítems. Necesitamos ítems que llenen el vacío en la definición original de la variable y necesitamos ítems más fáciles y más difíciles para extender el rango de la variable. Sin embargo, dado que toda nuestra muestra contestó acertadamente los tres ítems más fáciles, extendiendo la variable de la PCK hacia niveles más fáciles puede resultar tarea difícil. Tendríamos que localizar algunas personas mucho menos capaces que las disponibles entre las 34 de la muestra original para poder calibrar ítems más fáciles. Por otro lado, solo hay un ítem de la PCK en el cual fallaron todas las personas de nuestra muestra. Podría ser útil intentar agregar algunos ítems más difíciles que los ítems 15, 16 y 17, bajo el supuesto de que con una muestra de personas más capaces podríamos obtener calibraciones útiles de estos ítems más difíciles y así extender la variable de la PCK hacia valores más altos.

FIGURA 5.2.2

**DEFINICIÓN DE LA VARIABLE DE LA PCK POR LA DISTRIBUCIÓN DE LAS DIFICULTADES DE LOS ÍTEMS, PATRÓN DE GOLPE Y DISTRIBUCIÓN DE LAS HABILIDADES DE LAS PERSONAS**



Con estas consideraciones en mente, se emprendió el desarrollo de la variable de la PCK. Se conservaron los 18 ítems originales de la PCK y se agregaron diez nuevos ítems. La PCK original proviene de la Forma II de la escala de puntajes de Arthur[5]. Examinamos la Forma I y encontramos tres ítems que no se usan en la Forma II (Arthur, 1943). A estos tres ítems se les agregaron siete más. Cinco ítems fueron diseñados para llenar la brecha en el rango medio, cuatro ítems fueron diseñados para extender la variable de la PCK hacia arriba y se esperaba que uno de los ítems de la Forma I pudiera ubicarse cerca de los ítems 5, 6 y 7 anteriores. Las series de golpes para estos ítems adicionales y sus ubicaciones en la variable de la PCK se muestran en las Figuras 5.3.1 y 5.3.2.

La Figura 5.3.1 muestra el ítem de la Forma I y los cinco nuevos ítems diseñados para llenar la brecha entre los viejos ítems 10 y 11 de la PCK, a su vez, la Figura 5.3.2 presenta los cuatro ítems diseñados para extender la PCK hacia la región del Ítem 18. El resultado es una nueva Forma de prueba, la PCKB, que contiene los 18 ítems originales junto con los 10 ítems nuevos. Este nuevo instrumento de 28 ítems se administró a una muestra de 101 personas y se calibraron los ítems 2 a 25. El ítem 1 siguió siendo demasiado fácil y los ítems 26, 27 y 28 fueron demasiado difíciles para ser calibrados.

La columna 6 en la Tabla 5.3.1 proporciona las calibraciones de la nueva PCKB. El resto de la tabla 5.3.1 muestra la relación entre la PCK original y las calibraciones de la nueva PCKB. La Columna 1 contiene los nombres de los 14 ítems originales de la PCK. La columna 2 muestra sus calibraciones originales tomadas de la Tabla 3.4.4. Se observa en la Columna 6 que ahora hemos obtenido calibraciones de los ítems 2, 3 y 18 de la PCK original, tres de los ítems originales que no pudieron calibrarse en nuestro primer estudio con 34 personas.

La columna 3 de la Tabla 5.3.1 aplica el ajuste necesario para que las calibraciones previas de la PCK queden alineadas con sus nuevas calibraciones en la nueva PCKB. Esto se hace recorriendo o trasladando las calibraciones en la columna 2 con el valor constante 0.4 que es la posición media de la PCK original en las nuevas calibraciones de la PCKB. Esto causa que la Columna 3 y la Columna 5 tengan la misma media de 0.4.

En la Tabla 5.3.1 vemos que los nuevos ítems 12 a 16 de la PCKB caen más o menos donde se esperaba, aunque tal vez un poco hacia el lado fácil. El ítem 25 de la PCKB junto con el ítem 18 de la PCK amplía el alcance de la variable de la PCK en 2 lógitos más hacia arriba, pero no hemos encontrado a nadie que responda exitosamente a los ítems 26, 27 y 28 de la PCKB.

La Figura 5.3.3 compara las dificultades de los ítems que aparecieron tanto en las calibraciones hechas con la PCK y con la PCKB. Cada uno de los 14 ítems se encuentra en la Figura 5.3.3 con su par de estimaciones de dificultad. Si los ítems se ajustan al modelo de medición, entonces esperamos que estas estimaciones de sus dificultades sean estadísticamente equivalentes.

Por lo tanto en la medida en que los 14 puntos caigan sobre la recta de identidad, se estaría probando la invariancia de estas 14 dificultades. Como se muestra en la figura 5.3.3, los 14 puntos se encuentran dentro de las líneas de control de calidad del 95%. Este es el patrón al cual el modelo dice que deben aproximarse los ítems para ser útiles como instrumentos de medición.

[5] NT. Se utiliza la palabra "forma" como sinónimo de "batería" de ítems que integran una "versión" o "cuadernillo" de una prueba.

**FIGURA 5.3.1**

**CINCO NUEVOS ÍTEMS DISEÑADOS PARA LLENAR EL TRECHO VACÍO EN LA PCK Y UN ÍTYEM FÁCIL DE LA PCK, FORMA I**

| NOMBRE DEL ÍTEM Original PCK | NOMBRE DEL ÍTEM Nueva PCKB | SERIES DE GOLPES | NÚMERO DE BLOQUES | PATRÓN DE GOLPES |
|---|---|---|---|---|
| | 7 | 1 – 2 – 3 – 4 | 4 3 2 1 | |
| 10 | 11 | 2 – 4 – 3 – 1 | 4 3 2 1 | |
| | 12 | 3 – 1 – 4 – 2 | 4 3 2 1 | |
| | 13 | 2 – 1 – 4 – 3 | 4 3 2 1 | |
| | 14 | 4 – 2 – 1 – 3 | 4 3 2 1 | |
| | 15 | 1 – 2 – 3 – 4 – 3 | 4 3 2 1 | |
| | 16 | 1 – 2 – 3 – 4 – 2 | 4 3 2 1 | |
| 11 | 17 | 1 – 3 – 1 – 2 – 4 | 4 3 2 1 | |

PCK Original Ítems 1-6 / Ítems 7-9 equivalen en la nueva PCKB Ítems 1-6 / Ítems 8-10

**FIGURA 5.3.2**

**CUATRO NUEVOS ÍTEMS DE LA PCKB DISEÑADOS PARA EXTENDER LA PCK PARA MEDIR A LAS PERSONAS MÁS CAPACES**

| NOMBRE DEL ÍTEM Original PCK | Nueva PCKB | SERIES DE GOLPES | NÚMERO DE BLOQUES | PATRÓN DE GOLPES |
|---|---|---|---|---|
| 17 | 23 | 1 – 4 – 3 – 1 – 2 – 4 | 4 3 2 1 | |
| 18 | 24 | 4 – 1 – 3 – 4 – 2 – 1 – 4 | 4 3 2 1 | |
| | 25 | 3 – 2 – 4 – 1 – 3 – 4 – 2 | 4 3 2 1 | |
| | 26 | 2 – 4 – 3 – 1 – 4 – 2 – 3 | 4 3 2 1 | |
| | 27 | 1 – 4 – 2 – 3 – 2 – 4 – 3 | 4 3 2 1 | |
| | 28 | 1 – 3 – 4 – 2 – 4 – 2 – 3 | 4 3 2 1 | |

Ítems 12 -16 originales de la PCK se convierten en ítems 18-22 de la nueva PCKB

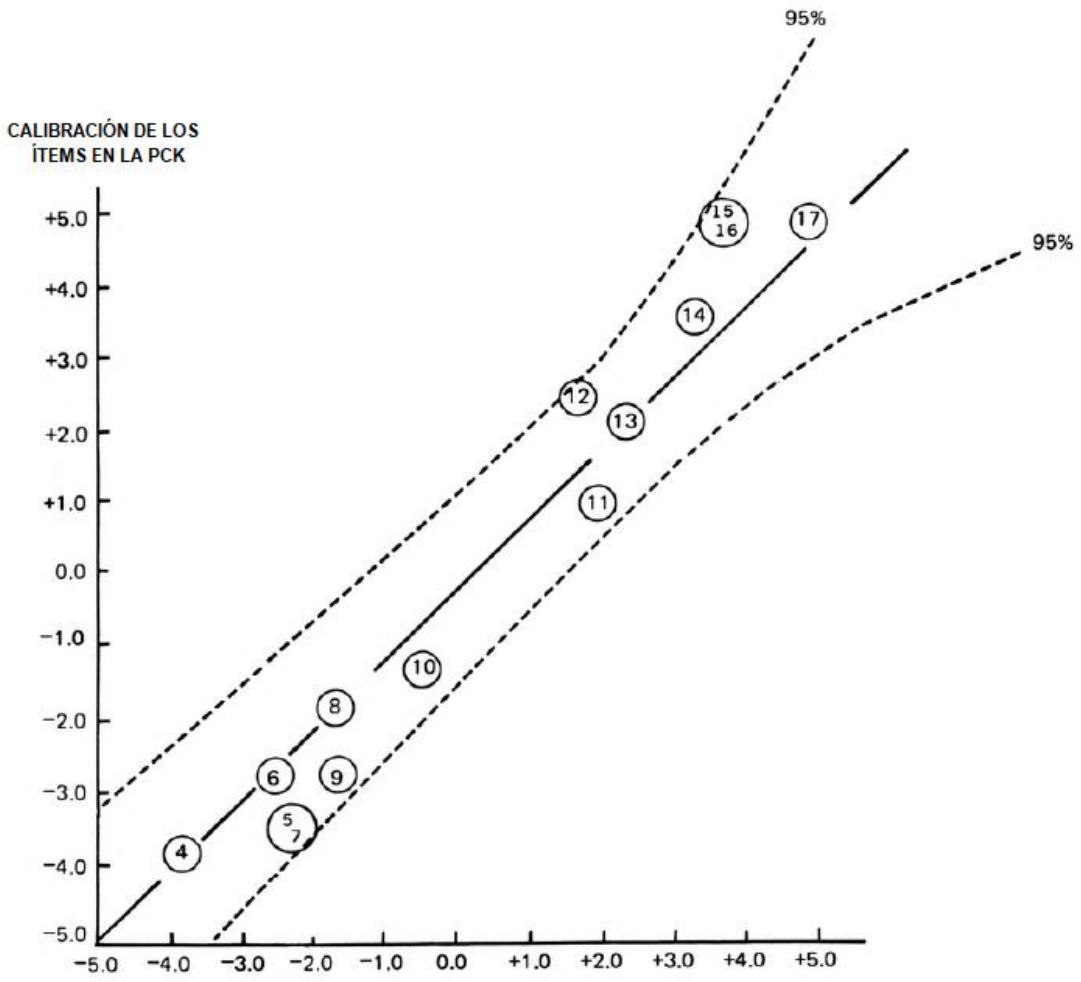
**TABLA 5.3.1**

**CALIBRACIÓN DE LA PCKB**

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| | PCK ORIGINAL | | | PCKB NUEVA | |
| Nombre del ítem | Calibración PCK | | Nombre del ítem | Calibración PCKB | |
| | Sin ajustar | Ajustada* | | Ítem original | Todos los ítems |
| 2 | | | 2 | | −6.0 |
| 3 | | | 3 | | −5.6 |
| 4 | −4.2 | −3.8 | 4 | −3.8 | −3.8 |
| 5 | −3.6 | −3.2 | 5 | −2.3 | −2.3 |
| 6 | −3.2 | −2.8 | 6 | −2.5 | −2.5 |
| | | | 7 | | −4.0 |
| 7 | −3.6 | −3.2 | 8 | −2.3 | −2.3 |
| 8 | −2.2 | −1.8 | 9 | −1.8 | −1.8 |
| 9 | −3.2 | −2.8 | 10 | −1.8 | −1.8 |
| 10 | −1.5 | −1.1 | 11 | −0.8 | −0.8 |
| | | | 12 | | 0.1 |
| | | | 13 | | −0.6 |
| | | | 14 | | −0.3 |
| | | | 15 | | −1.3 |
| | | | 16 | | −0.5 |
| 11 | 0.8 | 1.2 | 17 | 2.2 | 2.2 |
| 12 | 2.1 | 2.5 | 18 | 1.6 | 1.6 |
| 13 | 1.9 | 2.3 | 19 | 2.2 | 2.2 |
| 14 | 3.2 | 3.6 | 20 | 3.1 | 3.1 |
| 15 | 4.6 | 5.0 | 21 | 3.6 | 3.6 |
| 16 | 4.6 | 5.0 | 22 | 3.6 | 3.6 |
| 17 | 4.6 | 5.0 | 23 | 4.7 | 4.7 |
| 18 | | | 24 | | 6.5 |
| | | | 25 | | 6.0 |
| Media | 0.0 | 0.4 | | 0.4 | 0.0 |
| Desviación Estándar | 3.4 | 3.4 | | 2.8 | 3.4 |
| PCK: | L = 14 | N = 34 | PCKB: | L = 24 | N = 101 |

* Las calibraciones del Capítulo 3 para los 14 ítems originales de la PCK de la columna 2 tuvieron un corrimiento de 0.4 lógitos, haciendo que las medias de calibraciones de ese capítulo sean iguales a las medias de las nuevas calibraciones de la PCKB. Esta nueva media se calculó en la Columna 5.

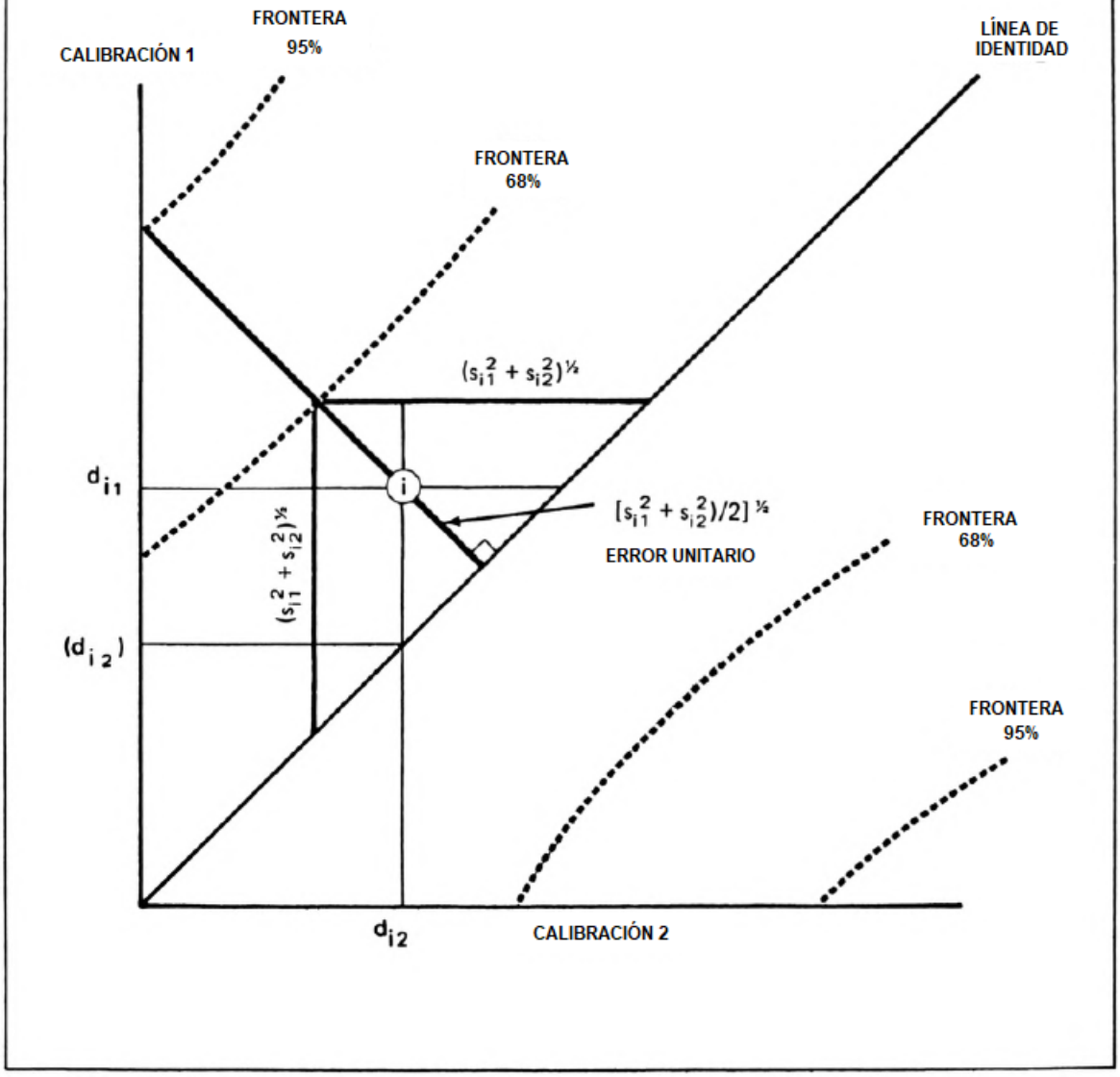
**FIGURA 5.3.3**
**GRÁFICA DE CALIBRACIONES DE LOS ÍTEMS PCK VS PCKB**




Los ítems se identifican con los nombres en la PCK original.

FIGURA 5.4.1

MANERA DE CONSTRUIR LAS LÍNEAS DE CONTROL DE CALIDAD AL 68% Y 95%




## 5.4 LÍNEAS DE CONTROL PARA GRÁFICAS DE IDENTIDAD

La Figura 5.3.3 contiene un par de líneas de control de calidad del 95% que nos ayudan a ver qué tanto los 14 puntos que representan a los ítems se ajustan a la expectativa de nuestro modelo de invariancia en la dificultad del ítem. En la gráfica que se utiliza para evaluar la invariancia de la dificultad de los ítems y, por lo tanto, la calidad de los ítems, estas líneas del 95% facilitan ver de qué manera se alinean los puntos de los ítems de la gráfica con la identidad esperada.

La figura 5.4.1 muestra cómo se dibujan las líneas de control de calidad. Cada gráfica compara una serie de pares de calibraciones de los ítems. Cada ítem tiene una dificultad $d_i$ y un error estándar $s_i$ de dos calibraciones independientes en las que apareció el mismo ítem. Por lo tanto, para cada ítem i contamos con los pares de valores ($d_{i1}$, $s_{i1}$) y ($d_{i2}$, $s_{i2}$). Dado que cada par de calibraciones se aplica a un ítem, esperamos que las dos dificultades $d_{i1}$ y $d_{i2}$ permitan estimar la misma dificultad $\delta_i$ una vez que se haya hecho una sola traslación necesaria para establecer un origen común a ambos conjuntos de ítems. También esperamos que el error de estas estimaciones pueda ser estimado por medio de $s_{i1}$ y $s_{i2}$.

Tenemos un estimador estadístico para probar el grado en que los dos valores $d_i$ estiman la misma dificultad $\delta_i$, es decir:

$$t_{i12} = (d_{i1} - d_{i2})/\left(s_{i1}^2 + s_{i2}^2\right)^{1/2} \qquad \sim N(0,1) \qquad \textbf{[5.4.1]}$$

en el cual $\left(s_{i1}^2 + s_{i2}^2\right)^{1/2}$ estima el error estándar esperado de la diferencia entre las dos estimaciones independientes $d_{i1}$ y $d_{i2}$ del único parámetro $\delta_i$. Podemos introducir esta prueba de la calidad de cada punto en la gráfica dibujando las líneas de frontera o límites de control de calidad aproximadamente a dos errores estándar a cada lado de la línea de identidad.

Dado que la unidad de error estándar de diferencia paralelo a cualquier eje de la trama es

$$\left(s_{i1}^2 + s_{i2}^2\right)^{1/2}$$

la unidad de error perpendicular a la línea de identidad de 45 grados debe ser

$$\left(s_{i1}^2 + s_{i2}^2\right)^{1/2}$$

Dos unidades de error perpendiculares a la línea de identidad en cada dirección producen un par líneas de control de aproximadamente 95% de calidad. La distancia perpendicular $D_{i1\,2}$ entre estas líneas de control de calidad y la línea de identidad se puede escribir:

$$D_{i1\,2} = 2\left[\left(s_{i1}^2 + s_{i2}^2\right)/2\right]^{1/2} \qquad \textbf{[5.4.2]}$$

Cuando $s_{i1}$ y $s_{i2}$ son suficientemente similares de modo que la media de sus cuadrados sea aproximadamente igual al cuadrado de sus medias, es decir.

$$\left(s_{i1}^2 + s_{i2}^2\right)/2 \simeq \left[\left(s_{i_2} + s_{i_2}\right)/2\right]^2$$

entonces la distancia $D_{i1\,2}$ de la línea de identidad a un límite de confianza del 95% puede aproximarse por:

$$D_{i1\,2} = 2\left[\left(s_{i1}^2 + s_{i2}^2\right)/2\right]^{1/2}$$

$$\simeq 2[(s_{i1} + s_{i2})/2] = s_{i1} + s_{i2}.$$

Por lo tanto, para los ítems i = 1, K están disponibles sus pares de calibraciones y pueden dibujarse las distancias ($s_{i1} + s_{i2}$) perpendiculares a la línea de identidad, partiendo de cada punto (que representa a cada ítem), con ello se pueden localizar las líneas de confianza del 95% que permiten evaluar la estabilidad general de las calibraciones de los ítems representados en la gráfica.

## 5.5 CONECTANDO DOS PRUEBAS

El método habitual para equiparar pruebas se basa en la igualación de puntajes equi-percentilares. Este procedimiento requiere una muestra de personas lo suficientemente grande y lo suficientemente distribuida para asegurar una definición adecuada de cada conexión de puntaje a percentil. Con la medición de Rasch es posible contar con un método más económico y controlado para construir un banco de ítems. Se pueden insertar enlaces de 10 a 20 ítems comunes en lo que podrían ser pruebas diferentes. Cada prueba puede administrarse a su propia muestra de personas y ninguna persona necesita responder más de una prueba. Pero todos los ítems en todas las pruebas pueden relacionarse posteriormente conectado a través de la red de enlaces (o eslabones) de ítems comunes.

Comencemos con un ejemplo simple tradicional para equiparar (igualar) dos pruebas A y B de 60 ítems que podrían administrarse simultáneamente a una muestra de al menos 1200 personas como se representa en la parte superior de la Figura 5.5.1. Este es un plan probable para obtener una definición detallada de percentiles de puntajes, necesaria para la equiparación exitosa de percentiles. Cada persona debe responder ambas pruebas, es decir, 120 ítems.

Por el contrario, un enfoque de Rasch podría hacer el mismo trabajo haciendo que cada persona responda una sola prueba de 60 ítems. Para lograr esto, una tercera prueba C de 60 ítems está compuesta por 30 ítems de cada una de las pruebas originales A y B. Luego cada una de estas tres pruebas se da a una muestra de 400 personas como se muestra en la parte inferior de la Figura 5.5.1. En este caso, cada persona responde una sola prueba, pero se calibran los 120 ítems juntos a través de los dos enlaces (o eslabones) de 30 ítems que conectan a las tres pruebas. El trabajo que tiene cada persona al responder la prueba es tan solo la mitad del requerido por el plan de equiparación (o igualación) equi-percentilar.

Al equiparar (o igualar) las calificaciones separadas de cada prueba con el modelo de Rasch, se produce un par de dificultades para cada eslabón o ítem de enlace. De acuerdo con el modelo, las estimaciones en cada par es estadísticamente equivalente excepto por una sola constante de translación común a todos los pares del enlace. Cuando se unen dos pruebas, A y B, mediante un enlace (o eslabón) común de K ítems, cada prueba se administra a su propia muestra de N personas, y $d_{iA}$ y $d_{iB}$ son las dificultades estimadas del ítem i en cada prueba, con errores estándar de aproximadamente $2.5/N^{½}$, entonces la única constante necesaria para trasladar todas las dificultades de los ítems al calibrar la Prueba B respecto de la escala de la Prueba A es

$$G_{A\,B} = \sum_{i}^{K} (d_{i\,A} - d_{i\,B})/K \qquad \textbf{[5.5.1]}$$

con un error estándar de aproximadamente $3.5/(NK)^{½}$ lógitos.

La calidad de este enlace puede evaluarse mediante el estadístico de ajuste

$$\sum_{i}^{K} (d_{i\,A} - d_{i\,B} - G_{A\,B})^2 (N/12)\,[K/(K-1)] \sim \chi^2_K \qquad \textbf{[5.5.2]}$$

el cual, de acuerdo con el modelo debe distribuirse aproximadamente como ji-cuadrado con K grados de libertad.

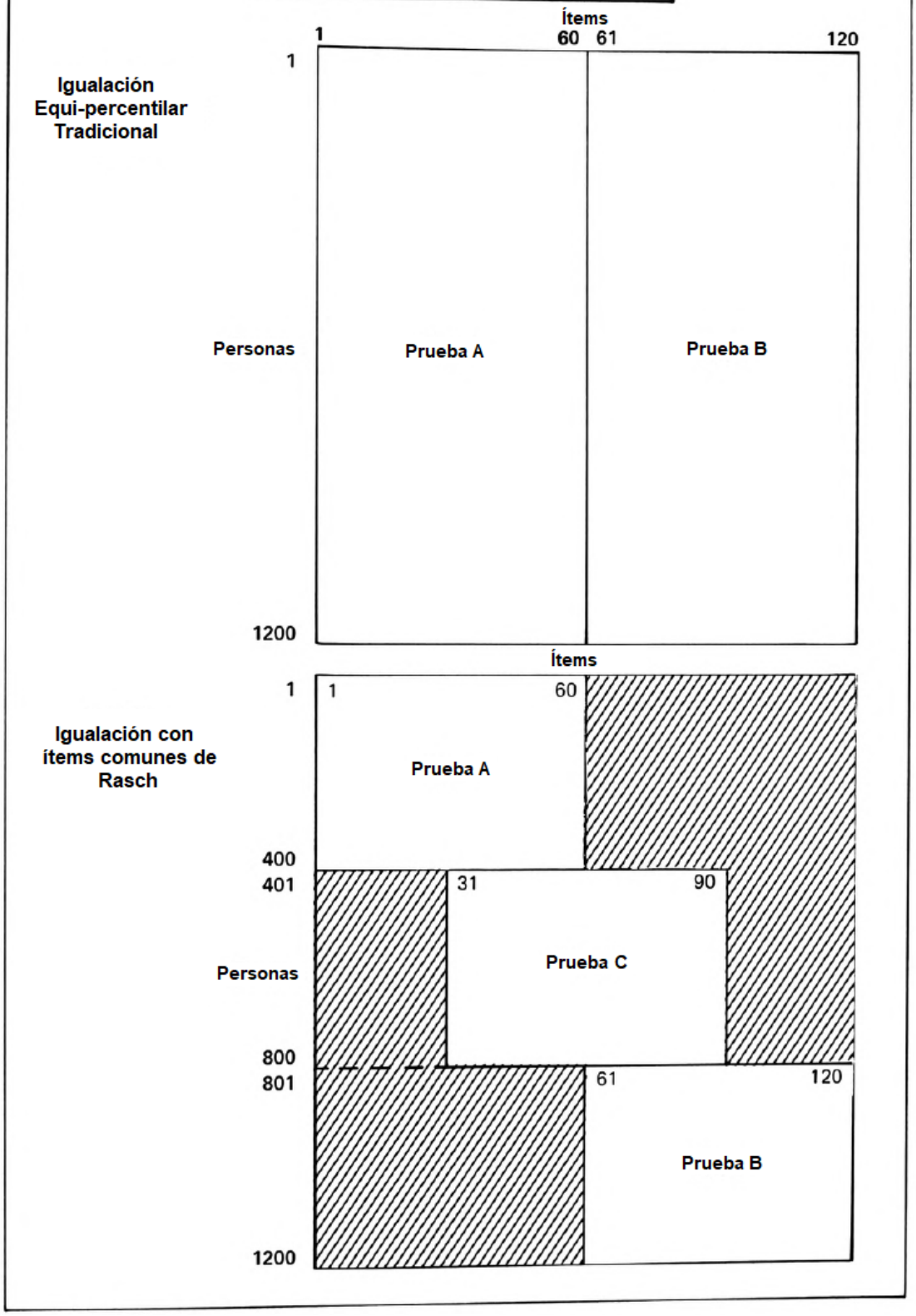
El ajuste individual de cualquier ítem en el enlace puede ser evaluado por esta expresión:

$$(d_{i\,A} - d_{i\,B} - G_{A\,B})^2 (N/12)[K/(K-1)] \sim \chi^2_1 \qquad \textbf{[5.5.3]}$$

que según el modelo debe tener aproximadamente una distribución ji-cuadrado con un grado de libertad.

FIGURA 5.5.1

DISEÑO DE EQUIPARACIÓN TRADICIONAL Y DE RASCH



Al usar estos estadísticos ji-cuadrado para juzgar la calidad del enlace, no debemos olvidar cómo se ven afectados por el tamaño de la muestra. Cuando N excede unos 500 casos, los estadísticos ji-cuadrado pueden detectar defectos del enlace tan pequeños como para hacer que la diferencia sea tangible en $G_{A\,B}$. Cuando las muestras de calibración son grandes, es más útil la raíz cuadrática media para medir el desajuste. Este estadístico puede usarse para estimar el aumento en lógitos que ocasionan los defectos del enlace en el error de calibración.

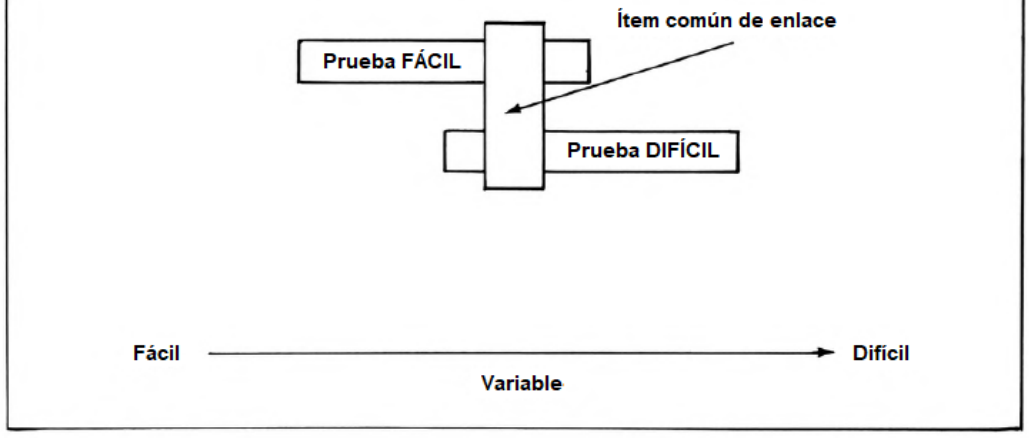
Al decidir cómo evaluar el ajuste del enlace debemos tener en cuenta que una incertidumbre aleatoria de menos de 0.3 lógitos en la dificultad de un ítem no implica ningún efecto práctico en la medición de la persona (Wright y Douglas, 1975a, 35-39). Debido a la manera en que el tamaño de la muestra interviene para el cálculo de la dificultad del ítem y, por lo tanto, en la evaluación de la calidad del enlace, podemos inferir que una muestra de 200 personas y un enlace de 10 buenos ítems, siempre serán más que suficientes para supervisar la validez de enlaces mayores a 0.3 lógitos. De manera práctica hemos descubierto que podemos construir bancos de ítems útiles con unidades muestrales tan pequeñas como 100 personas.

## 5.6 CONSTRUCCIÓN DE BANCOS DE ÍTEMS

A medida que establecemos y ampliamos la definición de una variable al agregar nuevos ítems tenemos los elementos iniciales para construir un banco de ítems. Con una planificación cuidadosa, podemos agregar sistemáticamente nuevos ítems y de esta manera construir un banco de ítems calibrados útiles para una amplia variedad de aplicaciones de medición. A medida que aumenta el número de ítems, se multiplican los problemas para administrar el banco. No es solo la cuestión de qué tan bien seleccionamos y combinamos los ítems y las personas, sino cómo gestionamos eficazmente la colección resultante de ítems calibrados. El modelo de medición de Rasch proporciona una aproximación concreta bien definida para la gestión del banco de ítems.

La estructura básica necesaria para calibrar un conjunto de ítems en una sola variable es por medio de un ítem común de enlace de un ítem común compartido por el conjunto de ítems de enlace de la prueba de enlace, que conecta a dos pruebas que, en otras condiciones, serían completamente diferentes. Una prueba fácil y una prueba difícil podrían estar vinculadas por un conjunto de ítems comunes como se ilustra en la Figura 5.6.1. En este ejemplo, los ítems de enlace son los “difíciles” de una prueba FÁCIL pero al mismo tiempo serían los ítems "fáciles" en la prueba DIFÍCIL.

FIGURA 5.6.1
ENLACE CON UN ÍTEM COMÚN



Con dos o más pruebas de enlaces podemos construir una cadena del tipo mostrado en la Figura 5.6.2. Hay que reconocer que la representación en la Figura 5.6.2 es incómoda. La estructura de enlace puede ser representada de modo más conveniente por medio del esquema más simple en la Figura 5.6.3 que enfatiza los enlaces y facilita hacer diagramas con estructuras más complicadas.

Debe considerarse que el número de ítems de un banco y el rango de sus dificultades puede crecer mucho más allá de la posibilidad de administrarle una prueba a una persona, por ello la cadena de ítems debe ser fraccionada en Formas (baterías de ítems o cuadernillos de prueba) que sean manejables tanto en su longitud o número de ítems y en el rango de sus dificultades. En la figura 5.6.3 cada círculo indica una prueba suficientemente corta en el rango de dificultades que es manejable por una muestra de personas elegida convenientemente. Cada línea que conecta a un círculo representa un enlace de ítems comunes compartidos por las dos pruebas a las que se une. Las pruebas crecen en dificultad en sentido horizontal a lo largo de la variable y son de dificultad comparable en el sentido vertical.

**FIGURA 5.6.2**
**UNA CADENA CON DOS ENLACES**

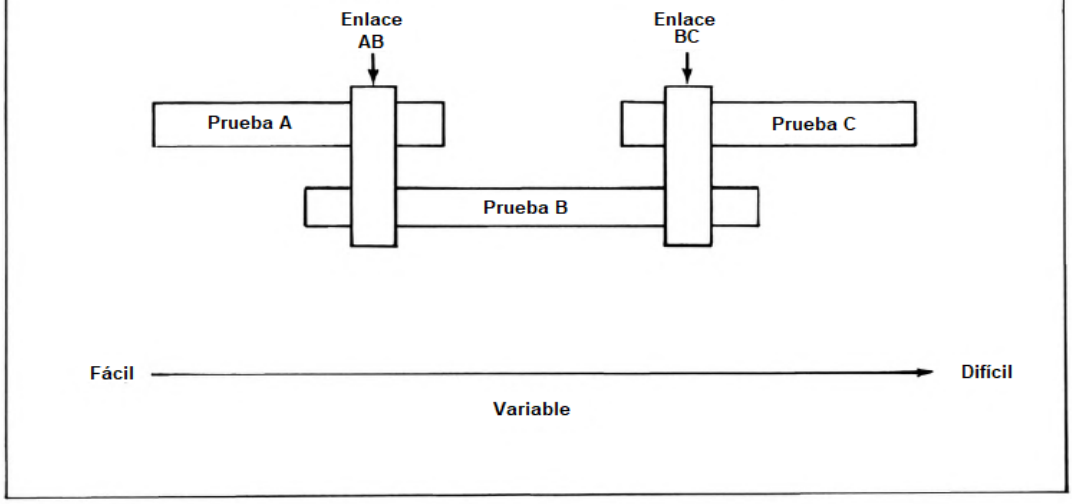
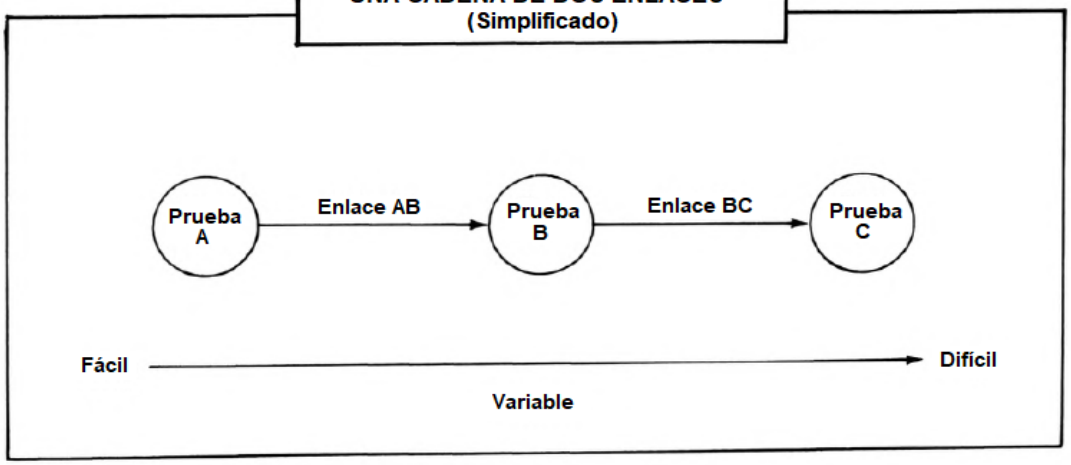



**FIGURA 5.6.3**
**UNA CADENA DE DOS ENLACES (Simplificado)**



Pueden construirse tres enlaces para describir un circuito cerrado (o bucle) como el que se presenta en la Figura 5.6.4. Este circuito (bucle) es una estructura de enlace importante porque agrega una prueba de enlace adicional que sirve para dar coherencia al conjunto. Si los tres enlaces del circuito son consistentes, entonces la suma de las tres traslaciones de enlace debería ser aproximadamente cero.

$$(G_{A\,B} + G_{B\,C} + G_{C\,A}) \simeq 0$$

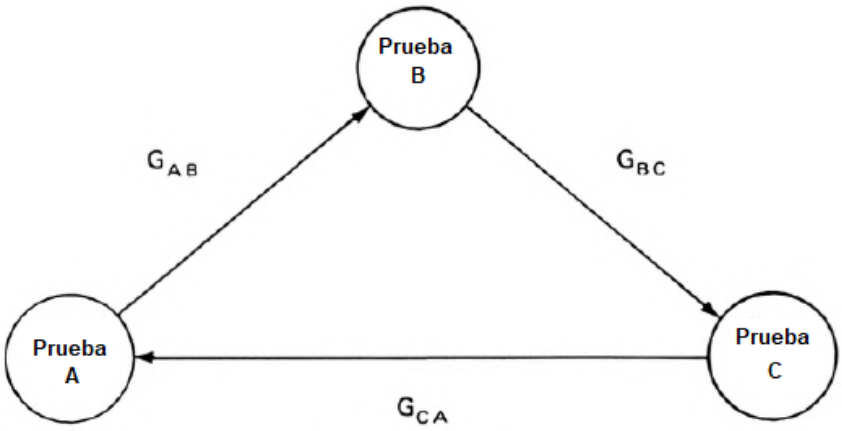
Observe que $G_{A\,B}$ significa el paso desde la Prueba A hacia la Prueba B a medida que recorremos el circuito en sentido de las manecillas del reloj, de modo que $G_{C\,A}$ significa el paso de la prueba C de regreso a la prueba A. La estimación de cero significa "estadísticamente" que la suma de los pasos entre prueba debe ocurrir en cero con menos de uno o dos errores estándar. El error estándar de la suma $G_{A\,B} + G_{B\,C} + G_{C\,A}$ será aproximadamente:

$$3.5(1/N_{A\,B}K_{A\,B} + 1/N_{B\,C}K_{B\,C} + 1/N_{C\,A}K_{C\,A})^{1/2}$$

en el que las N son los tamaños de muestra de calibración y las K son el número de ítems en cada enlace.

**FIGURA 5.6.4**

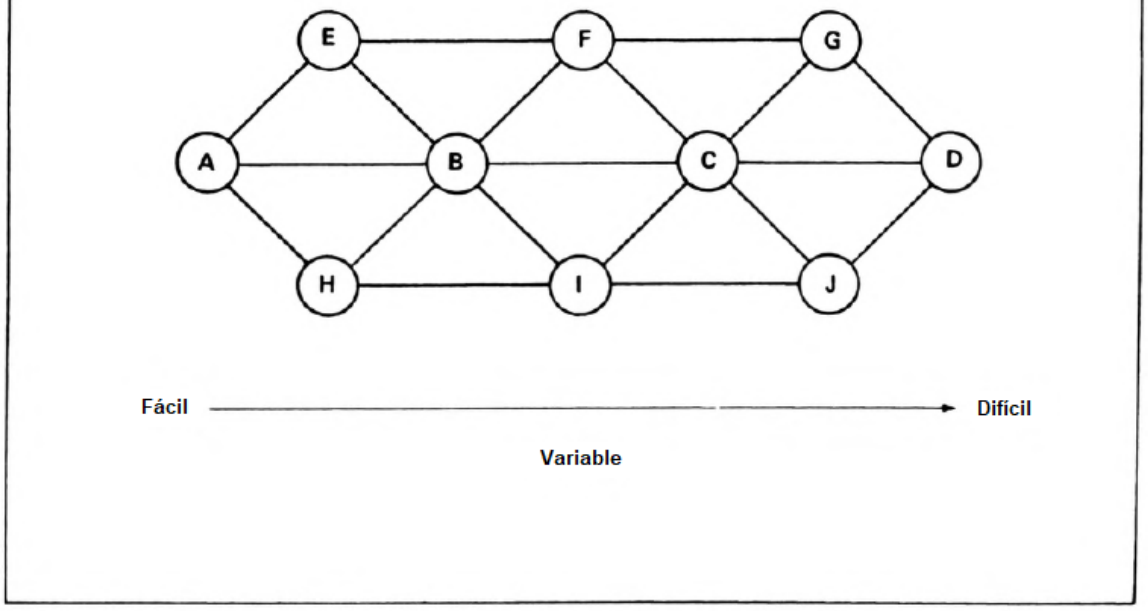
**UN CIRCUITO O BUCLE DE TRES ENLACES**



Con cuatro o más pruebas, podemos construir una red de circuitos. Por ejemplo, una secuencia de pruebas creciente en dificultad podría calibrarse comúnmente mediante una serie de enlaces de conexión como se muestra en la Figura 5.6.5. Estas diez pruebas marcan siete niveles de dificultad entre las pruebas A y la D. Esta red podría conectar diez pruebas de 60 ítems por medio de diecinueve enlaces de 10 ítems para cubrir 600 - 190 = 410 ítems. Si se contara con 200 personas para cada prueba, entonces los 410 ítems podrían evaluarse para una posible calibración conjunta de las respuestas de solo 2,000 personas. Aún en el caso de tan solo 1,000 personas, es decir, 100 por prueba, se dispondría de una estimación sustancial de las posibilidades de construir un banco de ítems para disponer de lo mejor de esos 410 ítems.

FIGURA 5.6.5

**UNA RED CONECTANDO DIEZ PRUEBAS CON DIECINUEVE ENLACES**




Los bloques de construcción de la red de pruebas son circuitos (bucles) de tres pruebas cada uno. Si un circuito se ajusta al modelo de Rasch, entonces la suma de sus tres traslaciones debería ser cero, dentro de uno o dos errores estándar. Por lo tanto, el éxito de la red para el enlace de las calibraciones de los ítems se puede evaluar a partir de las magnitudes y direcciones de las sumas de los circuitos (bucles). Con ello se pueden identificar regiones inestables y se pueden definir acciones para evitarlas o mejorarlas.

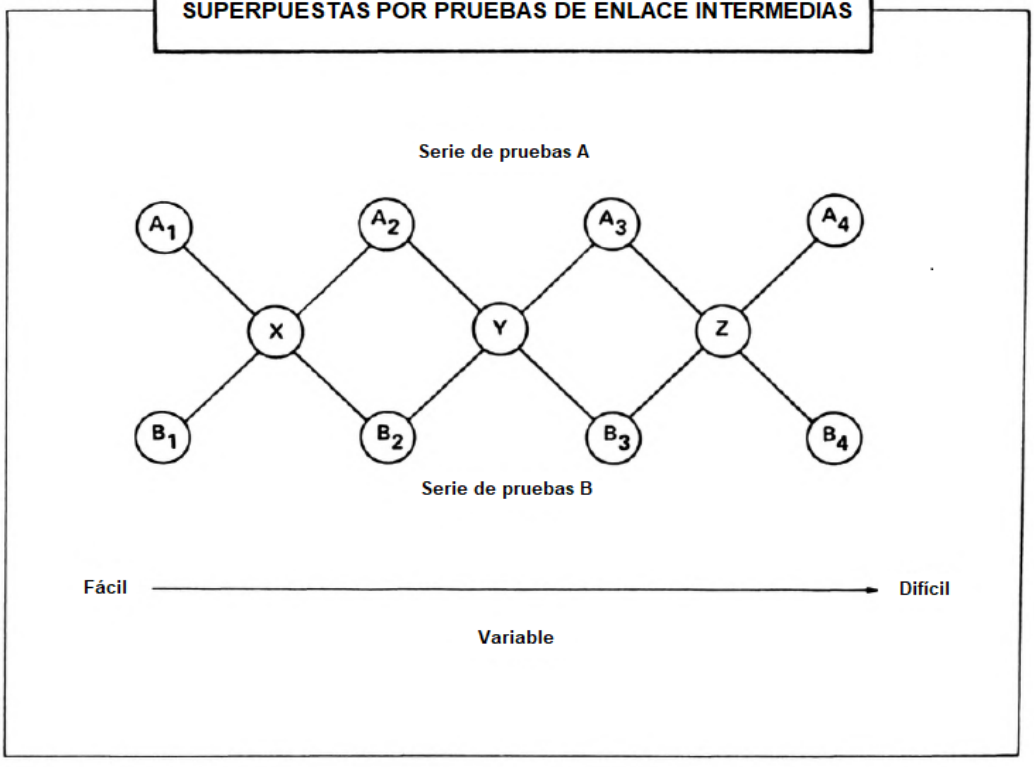
La utilización de redes de pruebas puede conducir a bancos con numerosos ítems comunes calibrados, con una gran dispersión de dificultades, de mayor amplitud a lo que una sola persona pudiera manejar. Los bancos resultantes, debido a la calibración de sus ítems en una variable común, pueden proporcionar los recursos necesarios para tener una prolífica familia de pruebas útiles, largas o cortas, fáciles o difíciles, muy espaciadas en las dificultades de los ítems o enfocadas a un rango corto, todas igualadas automáticamente en sus respectivas medidas.

Estos métodos para producir bancos de ítems pueden aplicarse a pruebas existentes, si fueron construidas cuidadosamente. Supongamos dos series de pruebas consecutivas no superpuestas $A_1$, $A_2$, $A_3$, $A_4$ y $B_1$, $B_2$, $B_3$, $B_4$ que queremos igualar mediante el modelo de Rasch. Las ocho pruebas pueden equipararse conectándolas con nuevas pruebas intermedias X, Y, Z, integradas por ítems comunes a ambas series, como muestra la Figura 5.6.6. Inclusive en una etapa de planificación, las series de pruebas A y B de la figura 5.6.6, podrían vincularse directamente incluyendo los ítems comunes en cada prueba de acuerdo con el patrón propuesto en Figura 5.6.7.

Dado que la coherencia es una preocupación vital en la construcción de un banco de ítems, estamos especialmente interesados en vincular estructuras que maximicen el control estadístico sobre la coherencia conjunta de todas las calibraciones de ítems. Las redes maximizan la cantidad de enlaces entre Formas de pruebas, haciendo que cada una se enlace con el mayor número posible de Formas. En el caso extremo, se puede construir una especie de "armadura" (o red)[6], que es un tipo más complejo de red en la que cada ítem individual de una Forma de prueba la relaciona con todas las demás.

**FIGURA 5.6.6**

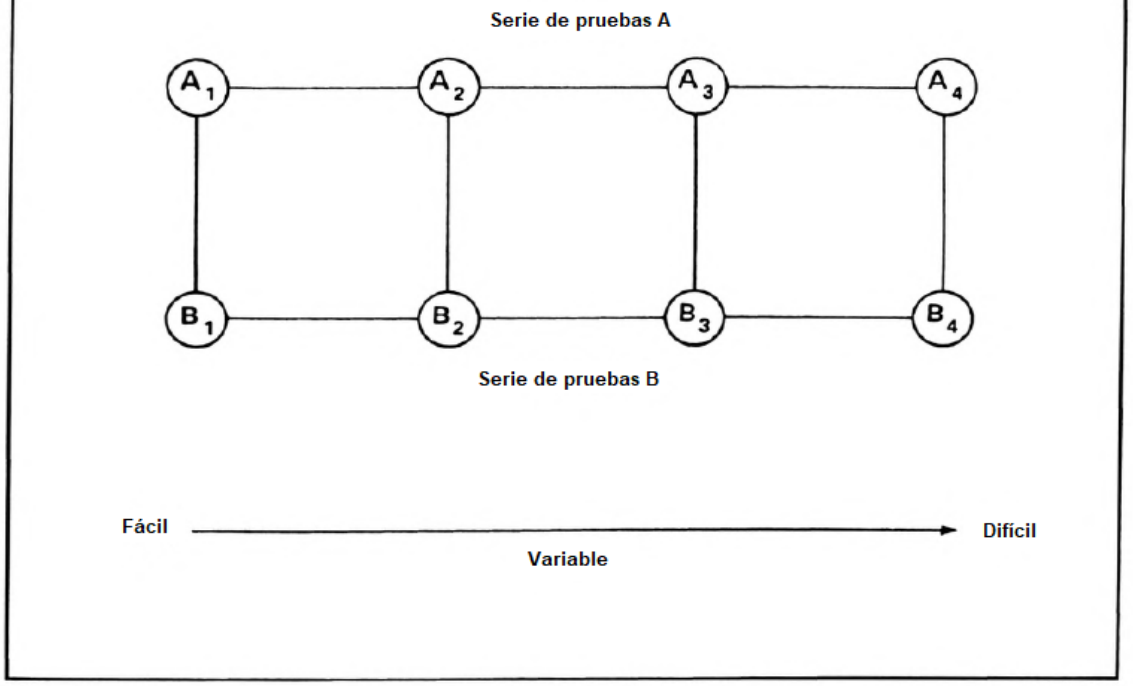
**CONECTANDO DOS SERIES DE PRUEBAS NO SUPERPUESTAS POR PRUEBAS DE ENLACE INTERMEDIAS**




[6] N.T. Se toma como "red" la traducción de "network" y "armadura" para "web", que sería un entramado o red más compleja. Dado que "armadura" es poco usual en el área de evaluación español, se adoptó también la traducción de "red".

FIGURA 5.6.7

CONECTANDO DOS SERIES DE PRUEBAS POR ENTRETEJIDO DE ÍTEMS COMUNES

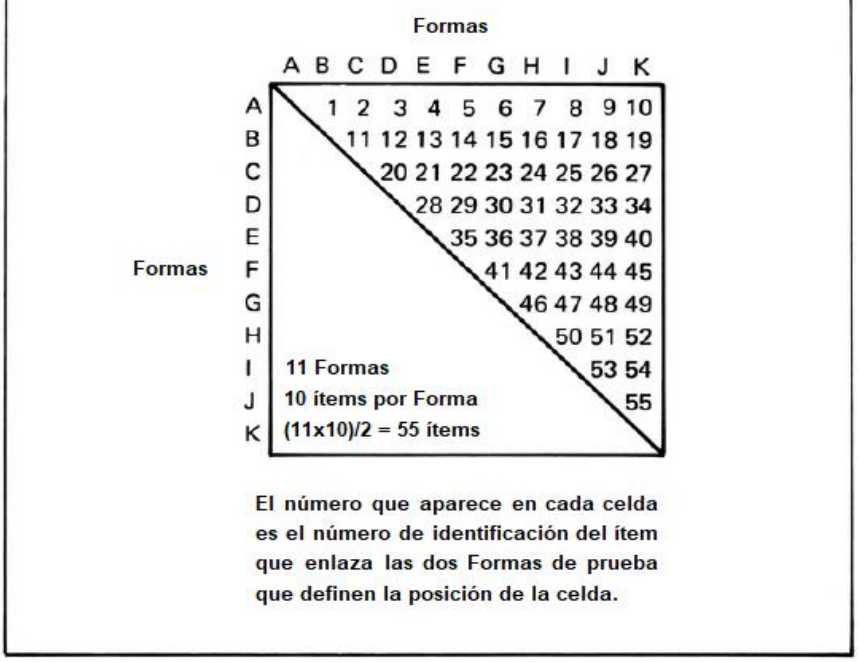



Para ilustrar tomemos un pequeño banco donde usamos 10 ítems por Forma de prueba en una armadura (o red) en la cual cada uno los 10 ítems aparece en una de las otras 10 Formas de prueba diferentes. El conjunto completo de 10 + 1 = 11 Formas constituye una armadura (o red) tejida con 11 × 10/2 = 55 ítems individuales de enlace. Cada una de las 11 Formas está entretejida con todas las demás Formas. El patrón se parece a la imagen de la Figura 5.6.8.

Este diseño de banco puede denominarse de armadura (o red) "completa" porque cada Forma está tejida con todas las demás. Sin embargo en el diseño de redes útiles hay tres restricciones que afectan su construcción: la cantidad total de ítems que queremos calibrar en el banco, el número máximo de ítems que podemos combinar en una sola Forma y la medida en que el banco que tenemos en mente se extiende en dificultades más allá de la habilidad de cualquiera de las personas.

Las condiciones de examen y la habilidad de las personas que responden a cada una de las Formas de prueba limitarán la cantidad de ítems que podemos incluir en una sola Forma. Por lo general, sucede que queremos calibrar muchos más ítems de los que podemos utilizar en una armadura (o red) completa como la de la Figura 5.6.8. Hay dos esquemas posibles para incluir ítems adicionales. El esquema más simple, aunque no el mejor estadísticamente, consiste en diseñar una armadura (o red) completa "nuclear" que utilice una parte de los ítems que podemos incluir en una sola Forma de prueba. Luego completamos el resto de la Forma hasta la longitud o número de ítems requeridos con ítems adicionales "etiquetados". Estos ítems etiquetados fueron calibrados en el banco junto con los ítems de enlace en sus respectivas Formas de prueba.

**FIGURA 5.6.8**
**UNA RED COMPLETA PARA FORMAS PARALELAS**




A diferencia de los ítems de enlace que siempre aparecen en dos Formas, los ítems etiquetados aparecen solo en una de las Formas y, por lo tanto, no brindan ayuda para vincular otras Formas en un solo banco calibrado con ítems comunes.

Otra posibilidad, que es mejor estadísticamente, consiste en incrementar el número de Formas utilizadas mientras se mantiene fijo el número de ítems de cada Forma de prueba hasta el límite requerido. Esto abre la armadura (o red) de manera sistemática, pero utiliza dos veces cada ítem lo cual permite que el par de datos de cada ítem pueda ser utilizado para evaluar la coherencia de las calibraciones del banco. La figura 5.6.9 muestra una red "incompleta" para el diseño de 21 Formas de prueba con 10 ítems cada una, similar a la de la Figura 5.6.8, pero con casi el doble de ítems utilizados en la armadura (o red) incompleta.

La armadura (o red) incompleta en la Figura 5.6.9 es adecuada para vincular un conjunto de Formas de pruebas paralelas. Sin embargo, cuando el banco de ítems tenga un alcance superior a la habilidad de cualquiera de las personas, no bastará ninguna de las armaduras (o redes mostradas) en las Figuras 5.6.8 y 5.6.9 porque no podremos combinar ítems de los extremos fáciles y difíciles en las mismas Formas. Los triángulos de enlace de ítems en las esquinas superiores derechas de las Figuras 5.6.8 y 5.6.9 no serán funcionales y tendrán que ser eliminados. Para mantener el equilibrio del enlace a lo largo de la variable, tendremos que hacer algo en cada extremo de la armadura (o red) para completar las Formas más fáciles y difíciles de modo que los extremos puedan enlazarse tan fuertemente como ocurre en la parte central. La figura 5.6.10 muestra cómo hacerlo sistemáticamente para 21 Formas secuenciales. Todavía tenemos 10 ítems por Forma de prueba pero ahora solo las Formas adyacentes están enlazadas entre sí. No hay ítems comunes que enlacen las Formas más fáciles directamente con las Formas más difíciles. Pero en el rango de la variable, las Formas de prueba que tienen dificultades cercanas entre sí quedan entretejidas con el número máximo de ítems de enlace.

Cada ítem de enlace en las redes que se muestran en las Figuras 5.6.8, 5.6.9 y 5.6.10 podrían referirse a un conglomerado de dos o más ítems que aparecen juntos en cada una de las dos Formas que enlazan. A veces, el diseño o el formato de impresión de los ítems los obliga a agruparse en conglomerados. Esto sucede generalmente en las pruebas de comprensión lectora donde los conglomerados se refieren a un texto o pasaje a leer. También ocurre de manera natural en las pruebas de matemáticas o en las que piden la recuperación de información, donde los conglomerados están referidos a gráficas comunes. Resulta evidente que la presencia de conglomerados aumenta la longitud (o número de ítems) de cada Forma de prueba por un factor igual al tamaño del conglomerado.

El análisis estadístico de una red construida para un banco se simplifica si dicha red es tan completa como la de la Figura 5.6.8. Las medias por renglón de la matriz de enlaces entre las Formas son estimadas por mínimos cuadrados de las dificultades de las Formas de prueba. Solo tenemos que ser cuidadosos con los signos. Si la celda $G_{j\,k}$ estima la diferencia en dificultad $(\delta_j - \delta_k)$ entre las formas j y k, con las dificultades de las Formas centradas en cero, de modo que $\delta_\cdot = 0$, entonces:

$$G_{j\cdot} = \sum_{k}^{M} G_{j\,k} / M \approx \delta_j$$

Las medias por renglón de la matriz de enlace calibran las Formas de prueba en una variable común. Una vez que se obtiene la dificultad de cada Formas, solo se deben agregar a las dificultades de los ítems dentro de ellas para llevar todos los ítems a la variable común que está compartida por dichas Formas.

**FIGURA 5.6.9**

**UNA RED INCOMPLETA PARA FORMAS PARALELAS**

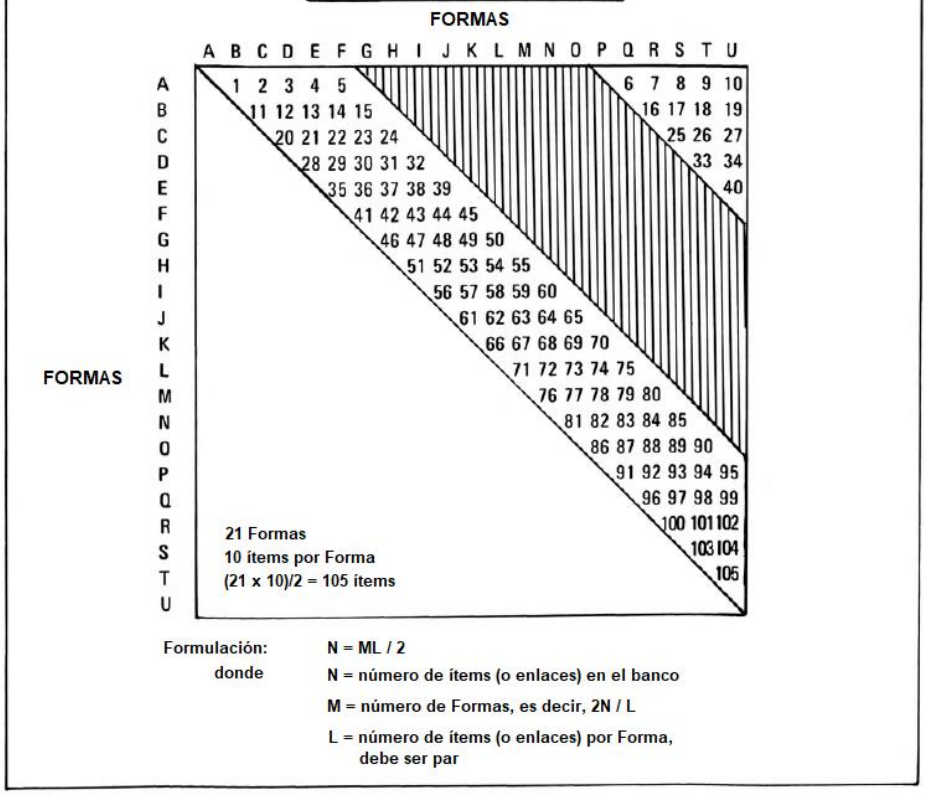
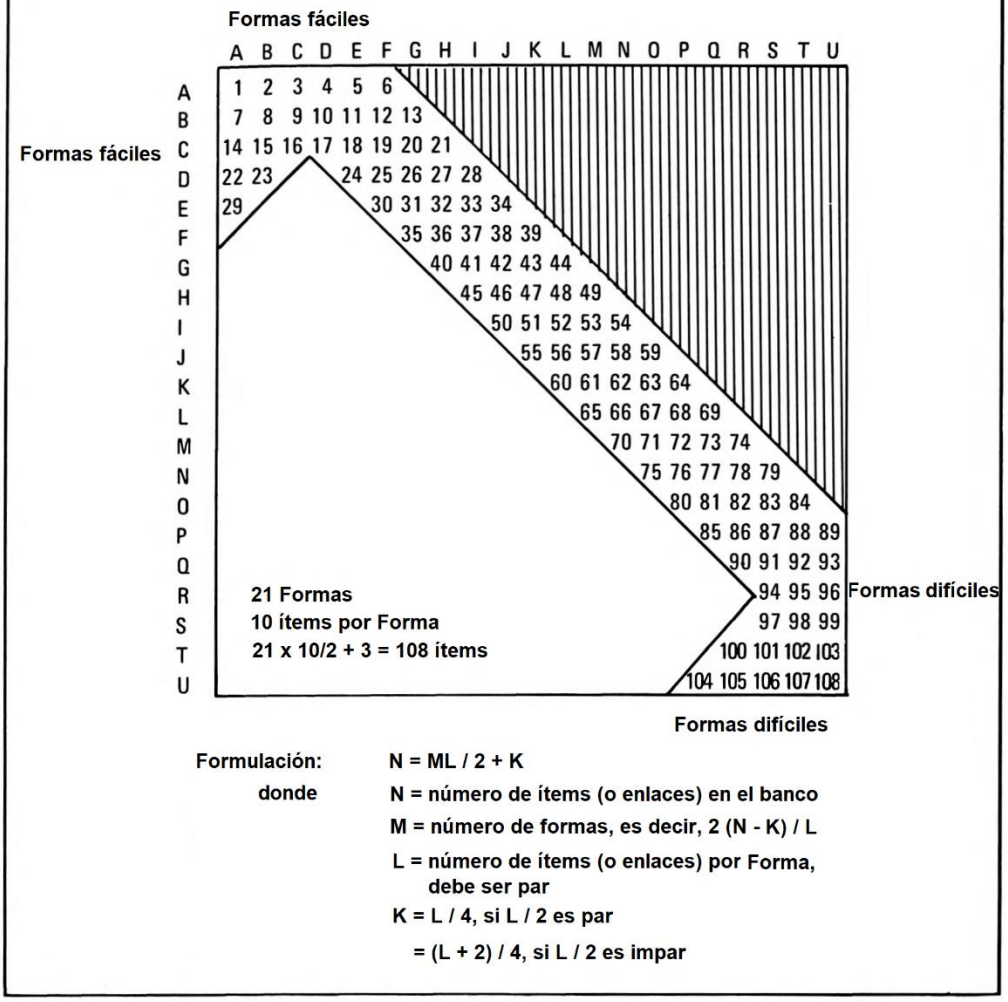


FIGURA 5.6.10
UNA RED INCOMPLETA
PARA FORMAS SECUENCIALES




Las armaduras (o redes) incompletas en las Figuras 5.6.9 y 5.6.10 requieren que estimemos las medias de los valores de los renglones de una matriz con datos faltantes. La simetría oblicua de las matrices de enlace ayuda a la solución de este problema que puede hacerse de manera satisfactoria por iteración o regresión.

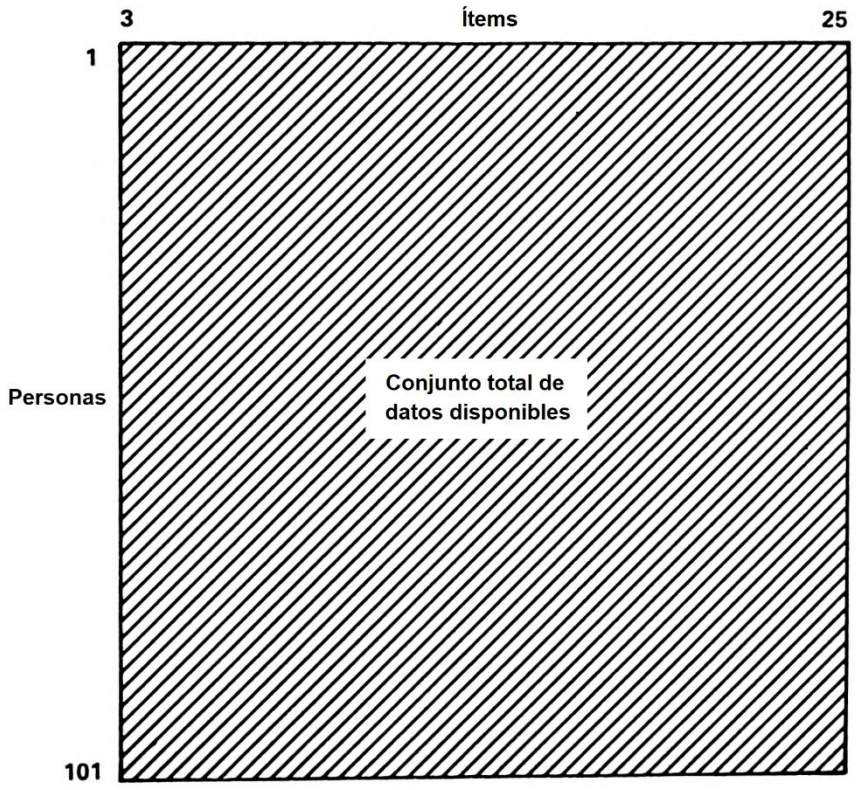
## 5.7 TRANSFERENCIA AL BANCO DE DATOS DE LA PCKB

La PCKB es una prueba corta que puede ser aplicada de manera práctica a 101 personas para que contesten los 23 ítems, lo cual nos da la matriz de respuesta ilustrada en la Figura 5.7.1. Sin embargo, la mayoría de los proyectos de bancos de ítems involucran la calibración de cientos de ítems contestados por miles de examinados. Resulta imposible pedir a cada persona que responda a cada uno de los ítems. Afortunadamente la construcción de un banco de ítems no requiere una tarea tan agobiante. Como vimos en la Sección 5.6, los ítems pueden unirse por medio de una red de enlaces. En general, hay dos tipos posibles de equiparación de Formas de prueba: por personas comunes y por ítems comunes.

Una manera de enlazar Formas separadas es administrarlas a la misma muestra de personas. Podemos ilustrar la equiparación con "personas comunes" con nuestros datos de la PCKB, para ello definimos dos pruebas secuenciales no superpuestas, FÁCIL y DIFÍCIL, y localizamos a todas las personas que produjeron respuestas medibles simultáneamente en ambas pruebas. Este es un intento de equiparación vertical de una prueba fácil y una prueba difícil y podemos esperar que las personas con puntajes utilizables en ambas pruebas sean escasas. Con nuestro ejemplo de la PCKB, solo tenemos 29 de las 101 personas. La imagen de esta equiparación por personas comunes en la Figura 5.7.2 muestra el núcleo de 29 personas de la muestra total que une dos partes no superpuestas de la PCKB, con una prueba FÁCIL DE 9 ítems y una prueba DIFÍCIL de 8 ítems.

FIGURA 5.7.1

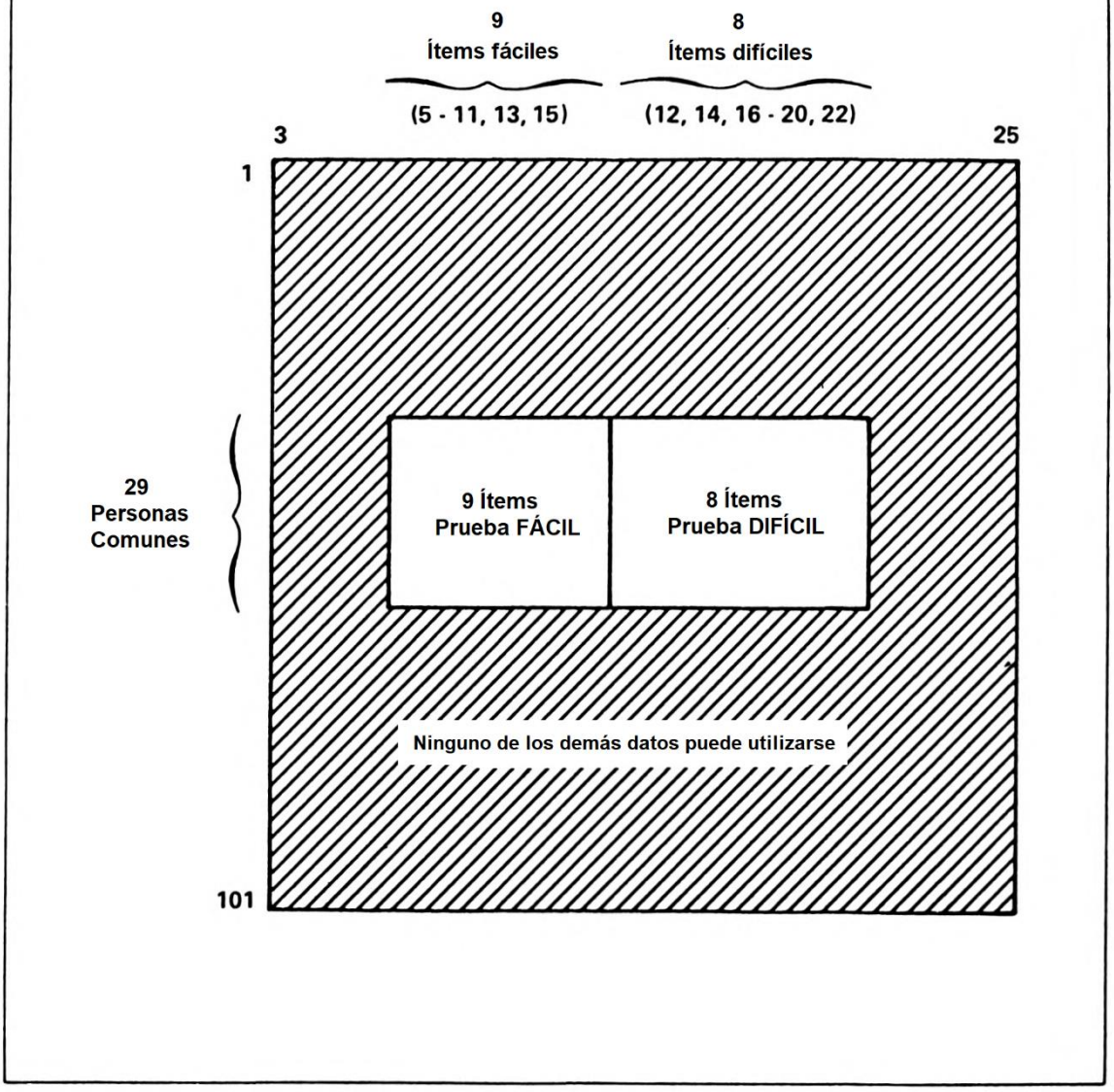
**PERSONAS COMUNES E ÍTEMS COMUNES CON LA PCKB**




El ítem 2 fue descartado por ser demasiado fácil ser útil.

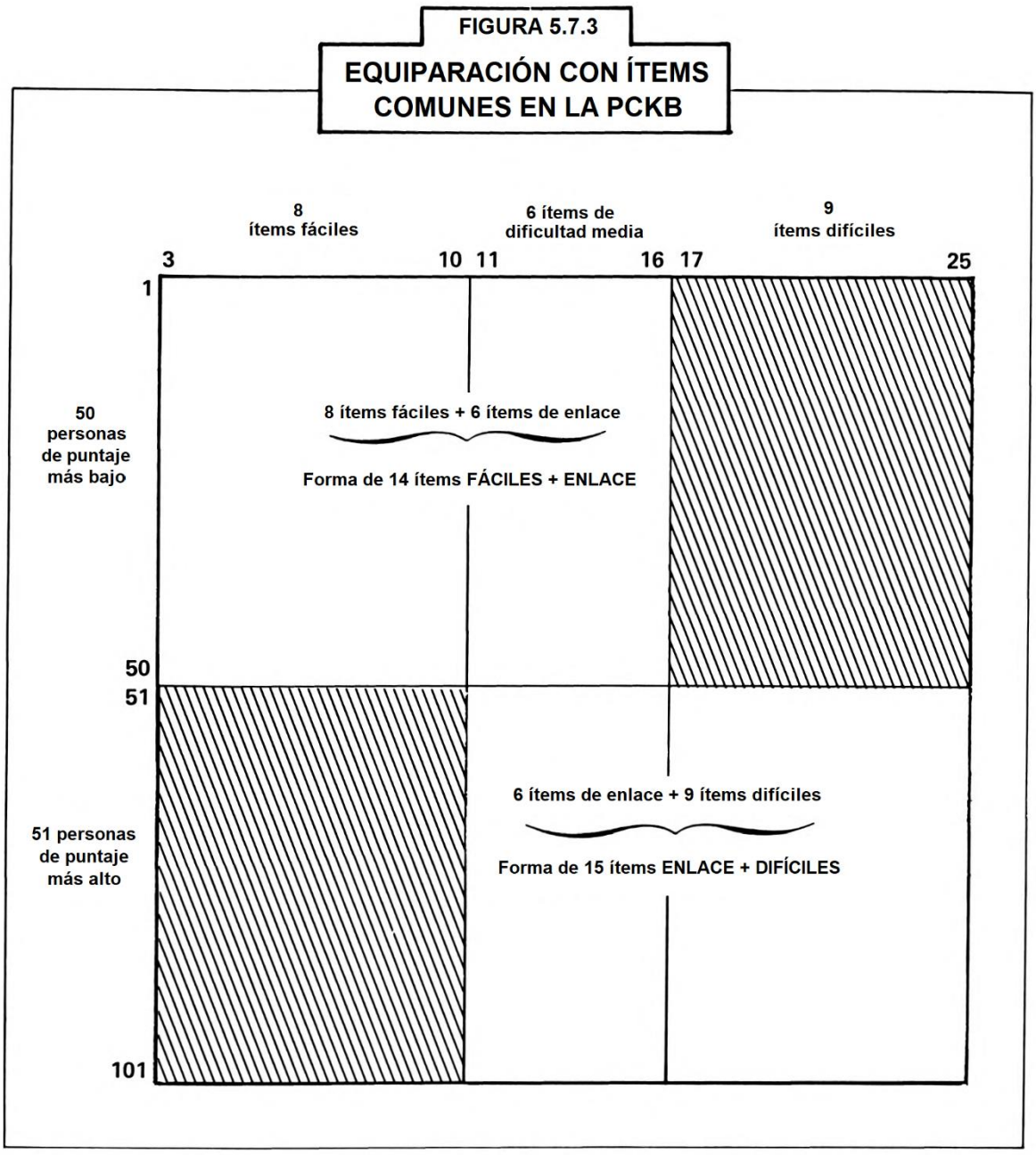
FIGURA 5.7.2

**EQUIPARACIÓN CON PERSONAS COMUNES EN LA PCKB**




Una mejor manera de equiparar Formas de prueba es mediante el uso de ítems comunes. Este enfoque para la PCKB se muestra en la Figura 5.7.3. Allí mostramos ocho ítems fáciles conectados a nueve ítems difíciles por medio de un enlace de seis ítems que produce una Forma FÁCIL + ENLACE de 14 ítems que responden las 50 personas de puntaje más bajo y una Forma de 15 ítems ENLACE + DIFÍCIL respondido por las 51 personas con el puntaje más alta.

FIGURA 5.7.3




## 5.8 EQUIPARACIÓN CON PERSONAS COMUNES EN LA PCKB

Entre las 101 personas que respondieron los ítems 3 a 25 de la PCKB encontramos dos Formas secuenciales que no se intersectan, llamadas FÁCIL y DIFÍCIL, en las cuales se tuvieron 29 personas con puntajes utilizables en ambas Formas de prueba. La Forma FÁCIL se hizo a partir de los ítems 5 a 11, 13 y 15. La Forma DIFÍCIL se hizo con los Ítems 12, 14, 16 a 20 y 22. Las 29 personas fueron aquéllas que permanecieron después de eliminar a las personas de puntajes perfectos en la prueba FÁCIL y las personas de bajo puntaje eliminadas por haber tenido puntajes nulos en la prueba DIFÍCIL.

Las medidas de estas 29 personas en cada Forma de prueba constituyen los datos de personas comunes para enlazar las Formas FÁCIL y DIFÍCIL en una sola. La diferencia entre las dos medias de habilidad permite estimar el corrimiento requerido para llevar las Formas FÁCIL y DIFÍCIL a una escala común. Los parámetros estadísticos de habilidad para las 29 personas en cada Forma de prueba son:

| | Forma FÁCIL | Forma DIFÍCIL | Diferencia |
|---|---|---|---|
| Media de habilidad | 1.49 | -0.57 | 2.06 |
| Desviación estándar | 0.80 | 0.43 | |

El procedimiento de equiparación es el siguiente:

1. Use la diferencia observada en la media de habilidades de la muestra 1.49 – (–0.57) = 2.06 como la diferencia de dificultad estimada entre las dos Formas de prueba.

2. Distribuya proporcionalmente esta diferencia sobre los nueve ítems de la Forma FÁCIL y los ocho ítems de la Forma DIFÍCIL de modo que la dificultad promedio de los 17 ítems se reduzca a cero.

Para los nueve ítems FÁCILES use

$$[(17 – 9) / 17] (2.06) = 0.97$$

Para los ocho ítems DIFÍCILES use

$$[(17 – 8) / 17] (2.06) = 1.09$$

3. Lleve las dos Formas de prueba a una escala común restando 0.97 a la dificultad de cada ítem de la Forma FÁCIL y sumando 1.09 a la dificultas de cada ítem de la Forma DIFÍCIL.

Estos cálculos se muestran en la Tabla 5.8.1. La columna 1 contiene el nombre de los 17 ítems de la PCKB que fueron utilizados en las formas FÁCIL y DIFÍCIL. La columna 2 presenta las calibraciones separadas de los ítems para la Forma FÁCIL. La columna 3 da las calibraciones separadas para la Forma DIFÍCIL. Debido a que estas calibraciones separadas están centradas en su propia Forma de prueba, la suma de las columnas 2 y 3 es igual a cero.

Para referir las calibraciones de las Columnas 2 y 3 a una escala centrada de personas comunes se requiere restar 0.97 de las dificultades de ítems de la Forma FÁCIL en la Columna 2 y agregar 1.09 a las dificultades de los ítems de la Forma DIFÍCIL en la Columna 3. Esto se hace en las Columnas 4 y 5, dando como resultado una escala de personas comunes para los 17 ítems centrados en 0.0.

Para evaluar la eficacia de la equiparación con personas comunes, obtuvimos una calibración combinada de los 17 ítems para las mismas 29 personas. La Columna 6 presenta las calibraciones de referencia y la Columna 7 muestra las diferencias entre la escala de personas comunes y la escala de referencia.

La figura 5.8.1 compara la escala de personas comunes y la escala de referencia. Las pequeñas diferencias entre las dos escalas permiten ver que la técnica de las personas comunes puede producir resultados equivalentes a la calibración combinada de ambas pruebas.

**TABLA 5.8.1**

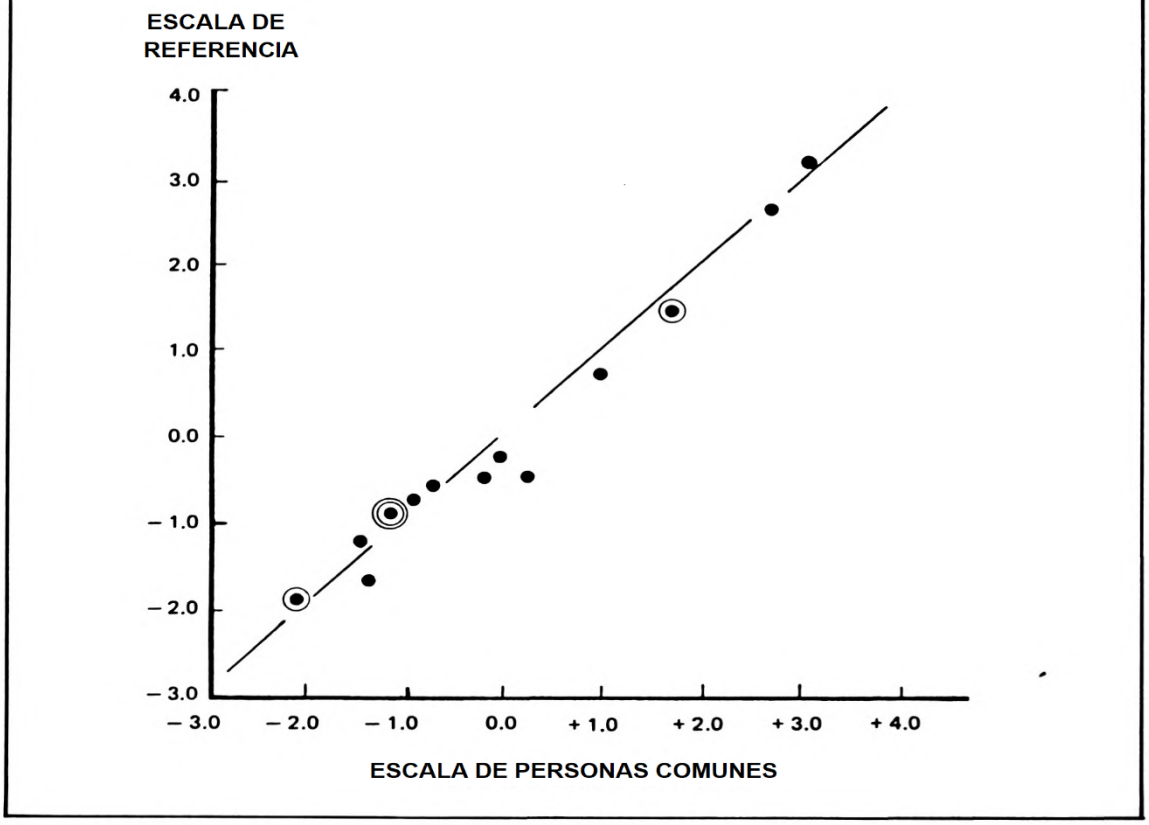
**EQUIPARACIÓN DE LA FORMAS FÁCIL Y DIFÍCIL USANDO PERSONAS COMUNES**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| | Calibraciones separadas | | Escala de personas comunes | | | |
| Nombre del Ítem | FÁCIL $d_E$ | DIFÍCIL $d_H$ | FÁCIL $d_C = d_E - 0.97$ | DIFÍCIL $d_C = d_H + 1.09$ | Calibración de Referencia* $d_R$ | Diferencia $d_C - d_R$ |
| 5 | 0.03 | | −0.94 | | −1.04 | −0.10 |
| 6 | 0.03 | | −0.94 | | −1.04 | −0.10 |
| 7 | −0.94 | | −1.91 | | −2.05 | −0.14 |
| 8 | 0.03 | | −0.94 | | −1.04 | −0.10 |
| 9 | 0.24 | | −0.73 | | −0.82 | −0.09 |
| 10 | 0.43 | | −0.54 | | −0.62 | −0.08 |
| 11 | 1.36 | | 0.39 | | 0.35 | −0.04 |
| 12 | | −1.44 | | −0.32 | −0.10 | 0.22 |
| 13 | −0.22 | | −1.19 | | −1.30 | −0.11 |
| 14 | | −1.25 | | −0.16 | 0.05 | 0.21 |
| 15 | −0.94 | | −1.91 | | −2.05 | −0.14 |
| 16 | | −2.66 | | −1.57 | −1.30 | 0.27 |
| 17 | | −0.12 | | 0.97 | 1.10 | 0.13 |
| 18 | | 0.65 | | 1.74 | 1.81 | 0.07 |
| 19 | | 0.65 | | 1.74 | 1.81 | 0.07 |
| 20 | | 1.83 | | 2.92 | 2.90 | −0.02 |
| 22 | | 2.32 | | 3.41 | 3.36 | −0.05 |
| Media | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 |
| Desviación estándar | 0.70 | 1.70 | 1.62 | | 1.65 | 0.14 |

*Con base en 29 personas que responden los 17 ítems

FIGURA 5.8.1

COMPARACIÓN DE LA ESCALA DE PERSONAS COMUNES CON LA ESCALA DE REFERENCIA




## 5.9 EQUIPARACIÓN CON ÍTEMS COMUNES EN LA PCKB

Para ilustrar la equiparación de ítems comunes hemos dividido los 23 ítems de la PCKB en tres partes: FÁCIL, ENLACE y DIFÍCIL. La Forma FÁCIL + ENLACE contiene ocho ítems FÁCILES y seis ítems de ENLACE para hacer una prueba fácil de 14 ítems. La Forma ENLACE + DIFÍCIL contiene los seis ítems ENLACE comunes más nueve ítems DIFÍCILES que hacen una prueba difícil de 15 ítems.

Cada una de estas Formas de prueba fue calibrada en muestras separadas. La Forma FÁCIL + ENLACE fue calibrada con las 50 personas con puntaje más bajo y la Forma ENLACE + DIFÍCIL fue calibrada con las 51 personas con mayor puntaje. Estas calibraciones están contenidas en la Tabla 5.9.1.

Los pares de calibraciones de los seis ítems de enlace, del 11 al 16, se presentan de nuevo en las Columnas 2 y 3 de la Tabla 5.9.2. Sus diferencias $D = d_E - d_H$ se tienen en la Columna 4. La media de estas diferencias es 4.11, que es la diferencia de dificultad entre la Forma FÁCIL + ENLACE y la Forma ENLACE + DIFÍCIL. Si sustraemos al valor D esta diferencia de 4.11 tendremos los residuos del enlace dado en la Columna 5.

**TABLA 5.9.1**

**CALIBRACIONES DE LOS ÍTEMS EN LAS FORMAS FÁCIL + ENLACE Y ENLACE + DIFÍCIL**

| Nombre del Ítem | FÁCIL + ENLACE | | ENLACE + DIFÍCIL | |
|---|---|---|---|---|
| | Dificultad | Error | Dificultad | Error |
| 3 | − 3.80 | | | |
| 4 | − 2.00 | | | |
| 5 | − 0.37 | | | |
| 6 | − 0.37 | | | |
| 7 | − 2.00 | | | |
| 8 | − 0.37 | | | |
| 9 | 0.06 | | | |
| 10 | 0.20 | | | |
| 11 | 0.97 | .36 | − 2.24 | .49 |
| 12 | 2.08 | .38 | − 1.83 | .44 |
| 13 | 1.58 | .36 | − 3.22 | .73 |
| 14 | 1.95 | .37 | − 2.80 | .61 |
| 15 | 0.84 | .36 | − 3.90 | 1.01 |
| 16 | 1.21 | .36 | − 2.02 | .46 |
| 17 | | | 0.60 | |
| 18 | | | − 0.50 | |
| 19 | | | 0.26 | |
| 20 | | | 1.18 | |
| 21 | | | 1.56 | |
| 22 | | | 1.56 | |
| 23 | | | 2.78 | |
| 24 | | | 4.51 | |
| 25 | | | 4.06 | |
| Media | 0.00 | | 0.00 | |
| Desviación Estándar | 1.68 | | 2.64 | |

Si estos ítems proporcionan un enlace utilizable, entonces sus residuos deberían distribuirse alrededor de cero con el error estándar que predice el modelo. Los errores estándar de estos residuos se calculan por medio de:

$$S_D = \left(S_E{}^2 + S_H{}^2\right)^{1/2}$$

los cuales están dados en la Columna 6. Los residuos estandarizados

$$z = (D - 4.11)/S_D$$

se dan en la Columna 7.

La Figura 5.9.1 es un gráfico de las calibraciones FÁCILES de estos ítems de ENLACE contra sus calibraciones en la Forma DIFÍCIL. Los puntos representan a los ítems que están dentro de las líneas de control de calidad del 95% lo cual demuestra que el corrimiento estimado con base en este enlace puede ser utilizado usar para conectar las dos formas de prueba.

**TABLA 5.9.2**

**ANÁLISIS DE ENLACE**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| | Cálculo del CORRIMIENTO PARA EL ENLACE | | | Verificación del AJUSTE DEL ENLACE | | |
| Nombre del ítem | FÁCIL $d_E$ | DIFÍCIL $d_H$ | Diferencia $D = d_E - d_H$ | Diferencia de residuo $D - 4.11$ | Error estándar del residuo $S_D$ | Residuo estandarizado $z = (D - 4.11)/S_D$ |
| 11 | 0.97 | − 2.24 | 3.21 | − 0.90 | 0.61 | − 1.48 |
| 12 | 2.08 | − 1.83 | 3.91 | − 0.20 | 0.58 | − 0.34 |
| 13 | 1.58 | − 3.22 | 4.80 | 0.69 | 0.81 | 0.85 |
| 14 | 1.95 | − 2.80 | 4.75 | 0.64 | 0.71 | 0.90 |
| 15 | 0.84 | − 3.90 | 4.74 | 0.63 | 1.07 | 0.59 |
| 16 | 1.21 | − 2.02 | 3.23 | − 0.88 | 0.58 | − 1.52 |
| Media | 1.44 | − 2.67 | 4.11 | 0.00 | | − 0.17 ≃ 0 |
| Desviación estándar | 0.52 | 0.79 | 0.76 | 0.76 | | 1.13 ≃ 1 |

Corrimiento para el enlace = $\sum_{i}^{6} D_i/6 = 4.11$

Error estándar del residuo: $S_D = (S_E^2 + S_H^2)^{1/2}$

Media esperada de z es 0

Desviación estándar esperada de z es 1

Nuestro siguiente paso es conectar la Forma FÁCIL + ENLACE a la Forma ENLACE + DIFÍCIL. Para hacer esto conectaremos los dos ENLACES de las Formas FÁCIL y DIFÍCIL. La Tabla 5.9.3 muestra el método utilizado. En la Columna 1 tenemos el nombre del ítem para cada uno de los 23 ítems de la PCKB. Las dificultades de los ítems desde 3 hasta 10 aparecen en la Columna 2. Debido a que haremos referencia a todos los ítems a la Forma FÁCIL, registramos las dificultades para los puntos 3 a 10 directamente en la Columna 6. Para los ítems de ENLACE 11 a 16 tenemos dos conjuntos de dificultades. En la Columna 2 tenemos los estimados de dificultad para ítems 11 a 16 calibrados con los ítems de la Forma FÁCIL. En la columna 3 tenemos los estimados de dificultad para estos mismos ítems obtenidos de su calibración con los ítems de la Forma DIFÍCIL.

A las dificultades de ENLACE $d_H$ agregamos la diferencia de dificultad de enlace de 4.11. A continuación hacemos el promedio de las dificultades de ENLACE $d_E$ con las dificultades de ENLACE $d_H$ que fueron ajustadas por medio del corrimiento de ENLACE de 4.11. El promedio de las dos estimaciones de ENLACE $(d_E + d_H + 4.11)/2$ para los ítems 11 a 16 se da en la Columna 5. Ingresamos estos valores en la Columna 6.

**FIGURA 5.9.1**

**EQUIPARACIÓN POR ENLACE DE ÍTEMS COMUNES**

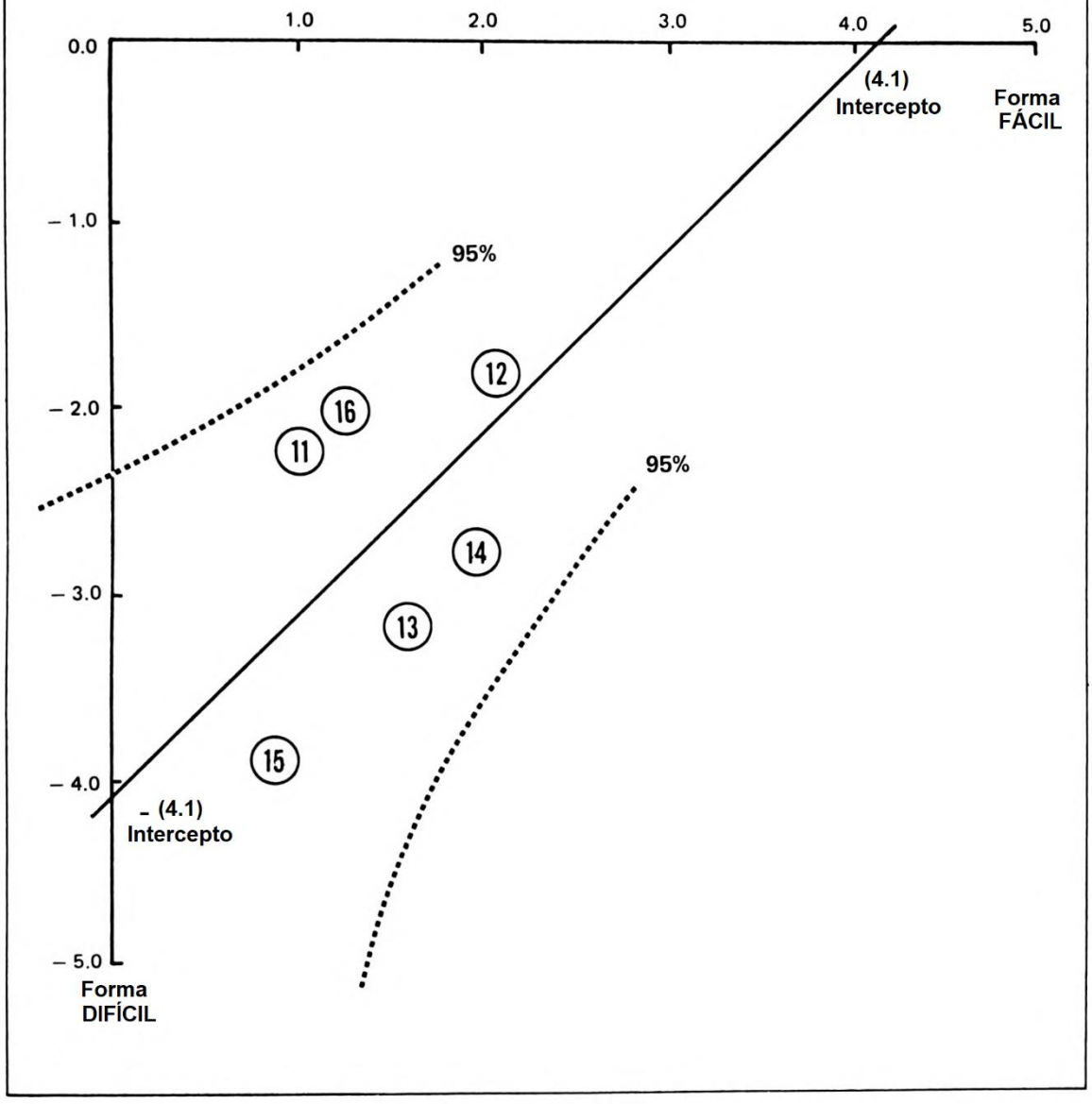


**TABLA 5.9.3**
**EQUIPARACIÓN DE FORMAS FÁCIL Y DIFÍCIL POR ENLACE DE ÍTEMS COMUNES**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| | Calibración de cada Forma | | Corrimiento de FÁCIL+ENLACE | | Escala de ítems comunes | |
| Nombre del ítem | FÁCIL+ENLACE $d_E$ | ENLACE+DIFÍCIL $d_H$ | $d_H + 4.11$ | $(d_E + d_H + 4.11)/2$ | $d_C$ | Centrada $d_C - 2.30$ |
| 3 | −3.80 | | | | − 3.80 | − 6.10 |
| 4 | −2.00 | | | | − 2.00 | − 4.30 |
| 5 | −0.37 | | | | − 0.37 | − 2.67 |
| 6 | −0.37 | | | | − 0.37 | − 2.67 |
| 7 | −2.00 | | | | − 2.00 | − 4.30 |
| 8 | −0.37 | | | | − 0.37 | − 2.67 |
| 9 | 0.06 | | | | 0.06 | − 2.20 |
| 10 | 0.20 | | | | 0.20 | − 2.10 |
| 11 | 0.97 | −2.24 | 1.87 | 1.42 | 1.42 | − 0.92 |
| 12 | 2.08 | −1.83 | 2.28 | 2.18 | 2.18 | − 0.12 |
| 13 | 1.58 | −3.22 | 0.89 | 1.24 | 1.24 | − 1.06 |
| 14 | 1.95 | −2.80 | 1.31 | 1.63 | 1.63 | − 0.67 |
| 15 | 0.84 | −3.90 | 0.21 | 0.53 | 0.53 | − 1.77 |
| 16 | 1.21 | −2.02 | 2.09 | 1.65 | 1.65 | − 0.65 |
| 17 | | 0.60 | 4.71 | | 4.71 | 2.41 |
| 18 | | −0.50 | 3.61 | | 3.61 | 1.31 |
| 19 | | 0.26 | 4.37 | | 4.37 | 2.07 |
| 20 | | 1.18 | 5.29 | | 5.29 | 2.99 |
| 21 | | 1.56 | 5.67 | | 5.67 | 3.37 |
| 22 | | 1.56 | 5.67 | | 5.67 | 3.37 |
| 23 | | 2.78 | 6.89 | | 6.89 | 4.59 |
| 24 | | 4.51 | 8.62 | | 8.62 | 6.32 |
| 25 | | 4.06 | 8.17 | | 8.17 | 5.87 |
| Media | 0.00 | 0.00 | 4.11 | | 2.30 | 0.00 |
| Desviación Estándar | 1.68 | 2.64 | 2.64 | | 3.37 | 3.37 |

Finalmente, para hacer que los ítems de la Forma DIFÍCIL queden en la escala común, agregamos 4.11 a los ítems 17 a 25 de esta Forma DIFÍCIL y los nuevos estimadores nos permiten completar la Columna 6. De esta Forma tenemos en esta columna una nueva escala de los ítems con el promedio de los valores estimados para las dificultades de dos ENLACES y los de los ítems de las Formas DIFÍCIL y FACIL.

La media de esta escala de ítems comunes en la columna 6 es 2.30, así que restamos 2.30 a la dificultad de cada ítem de la Columna 6 para centrar la nueva escala en 0.00 como se muestra en la Columna 7.

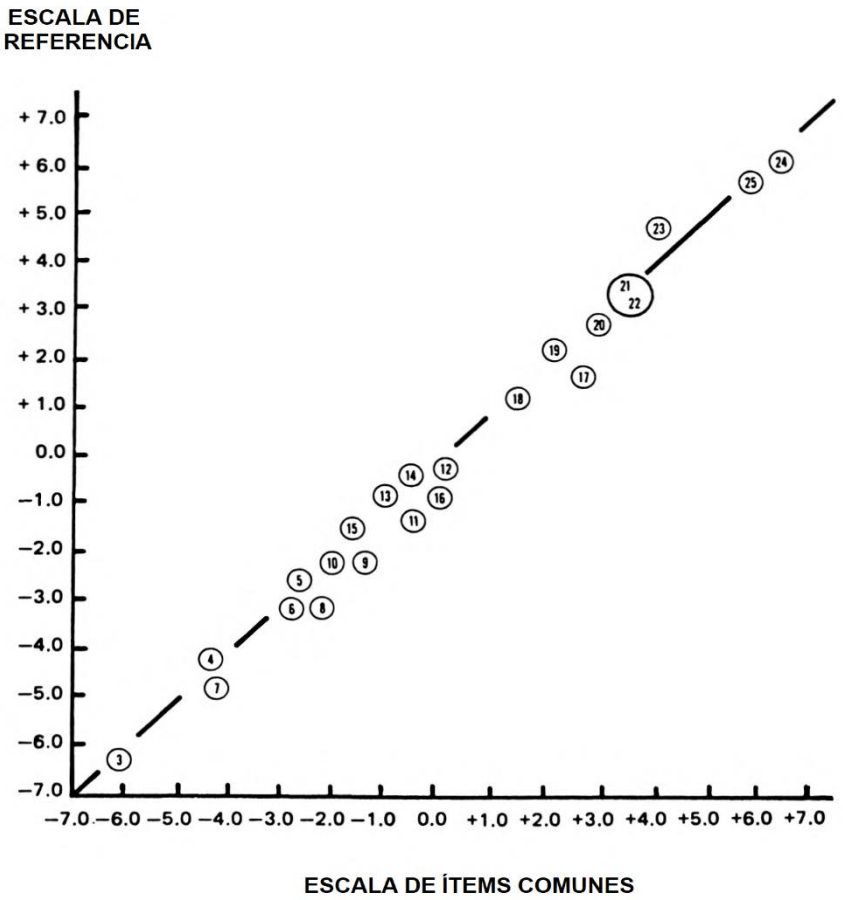
Para comprobar la idoneidad de esta equiparación de ítems comunes, compararemos con las dificultades de los ítems que habríamos tenido si no hubiéramos intentado hacer el enlace sino utilizado las 101 respuestas de las personas a los 23 ítems. Las dificultades de los ítems comunes de la Tabla 5.9.3 se dan en Columna 2 de la Tabla 5.9.4. La Columna 3 proporciona las calibraciones de referencia de los 23 ítems con las 101 personas y la Columna 4 muestra las diferencias entre las dificultades de los ítems comunes $d_C$ y las dificultades de los ítems en la escala de referencia $d_R$. La gráfica de estos valores se presenta en La Figura 5.9.2 que muestra cómo todos los ítems están cerca de lo esperado en la recta de identidad.

**TABLA 5.9.4**
**COMPARACIÓN DE LA EQUIPARACIÓN ON ÍTEMS COMUNES Y LA ESCALA DE REFERENCIA**

| 1<br>Nombre del ítem | 2<br>Escala de ítems comunes<br>$d_C - 2.30$ | 3<br>Escala de referencia<br>$d_R$ | 4<br>Diferencia<br>$d_C - d_R$ |
|---|---|---|---|
| 3 | - 6.10 | -6.20 | .10 |
| 4 | - 4.30 | -4.11 | - .19 |
| 5 | - 2.67 | -2.58 | - .09 |
| 6 | - 2.67 | -2.72 | .05 |
| 7 | - 4.30 | -4.34 | .04 |
| 8 | - 2.67 | -2.58 | - .09 |
| 9 | - 2.24 | -2.06 | - .18 |
| 10 | - 2.10 | -2.06 | - .04 |
| 11 | - 0.92 | -1.03 | .11 |
| 12 | - 0.12 | -0.12 | .00 |
| 13 | - 1.06 | -0.85 | - .21 |
| 14 | - 0.67 | -0.52 | - .15 |
| 15 | - 1.77 | -1.52 | - .25 |
| 16 | - 0.65 | -0.77 | .12 |
| 17 | 2.41 | 1.93 | .48 |
| 18 | 1.31 | 1.36 | - .05 |
| 19 | 2.07 | 2.01 | .06 |
| 20 | 2.99 | 2.88 | .11 |
| 21 | 3.37 | 3.33 | .04 |
| 22 | 3.37 | 3.33 | .04 |
| 23 | 4.59 | 4.52 | .07 |
| 24 | 6.32 | 6.27 | .05 |
| 25 | 5.87 | 5.81 | .06 |
| Media | 0.00 | 0.00 | 0.00 |
| Desviación Estándar | 3.37 | 3.32 | 0.15 |

FIGURA 5.9.2

COMPARACIÓN DE LA EQUIPARACIÓN ON ÍTEMS COMUNES Y LA ESCALA DE REFERENCIA

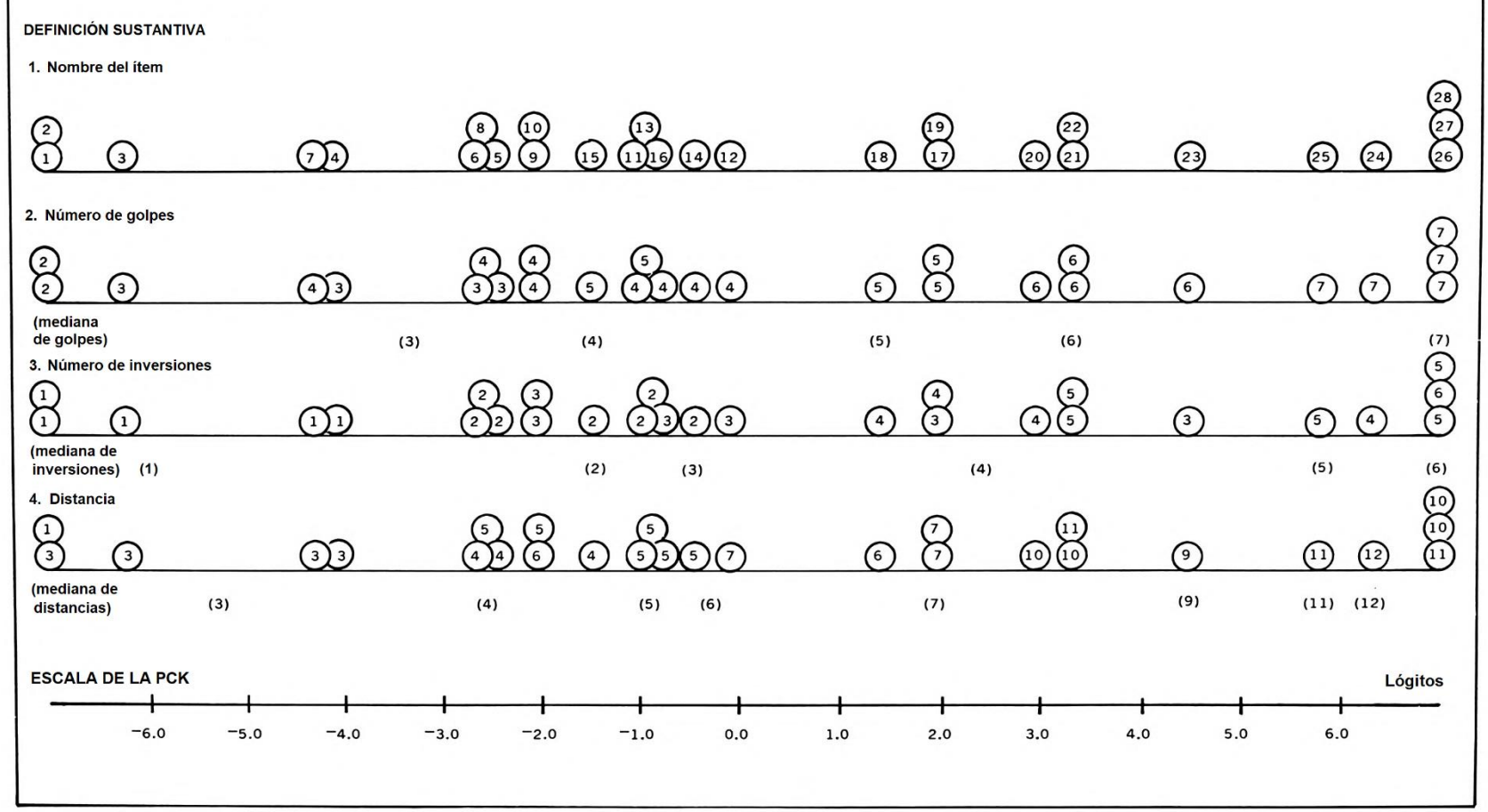



## 5.10 LA VARIABLE DE LA PCK COMO PRUEBA CON REFERENCIA A CRITERIO

Al ubicar los 23 ítems de la PCKB en una sola escala, podemos hacer más explícita la definición de la variable de la PCK. Estos ítems que ahora definen a la variable están construidos por unos cuantos componentes básicos: número de golpes, número de inversiones y distancia total entre los cubos. Es la Forma en que estos componentes subyacentes evolucionan a lo largo de la variable lo que documenta el posible significado de las medidas en la variable de la PCK. La figura 5.10.1 muestra el nivel de dificultad de los ítems de la PCKB junto con su número de golpes, inversiones y distancias.

FIGURA 5.10.1

**DOCUMENTACIÓN DE LA VARIABLE DE LA PCK**
**(101 personas contra 28 ítems)**



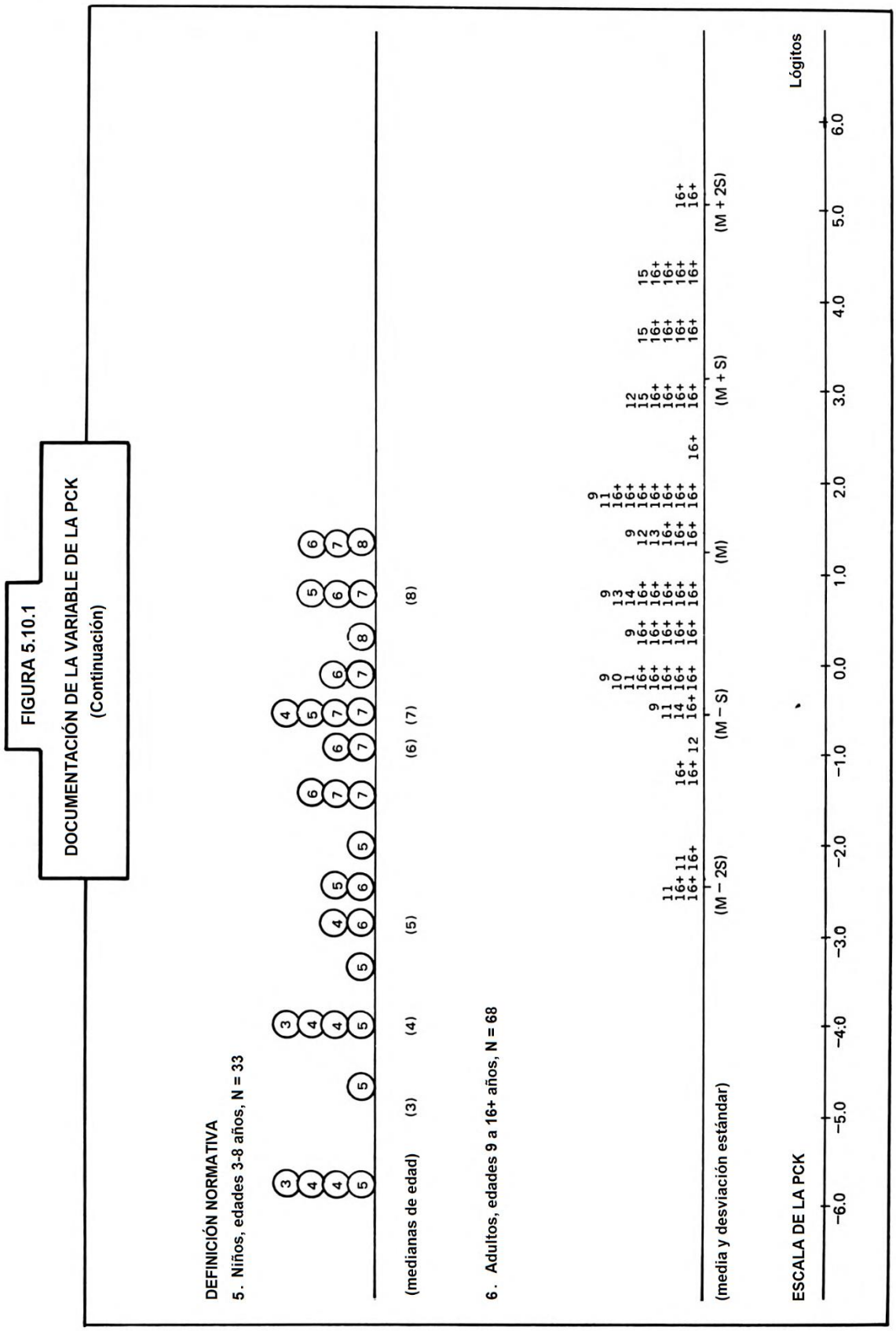


FIGURA 5.10.1
DOCUMENTACIÓN DE LA VARIABLE DE LA PCK
(Continuación)
DEFINICIÓN NORMATIVA
5. Niños, edades 3-8 años, N = 33
(medianas de edad) (3) (4) (5) (6) (7) (8)
6. Adultos, edades 9 a 16+ años, N = 68
(media y desviación estándar) (M − 2S) (M − S) (M) (M + S) (M + 2S)
ESCALA DE LA PCK
−6.0 −5.0 −4.0 −3.0 −2.0 −1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0
Lógitos

El renglón 1 de la figura 5.10.1 contiene los ítems de la PCKB ordenados por sus calibraciones en la variable de acuerdo con la escala de lógitos dada en la parte inferior de la figura. El número de golpes para cada ítem se da en el renglón 2. Los ítems 1 y 2 son ítems de dos golpes respondidos correctamente por la totalidad de 101 personas en la muestra. A medida que avanzamos en la variable, el número de golpes pasa de dos a siete. Debajo del renglón 2, hemos marcado la mediana de dificultad para cada número de golpes de dos a siete. El renglón 3 muestra el número de inversiones en cada ítem y la mediana de niveles de dificultad. El renglón 4 muestra las distancias entre los cubos que intervienen en cada serie de golpes y sus medianas.

El patrón de golpes, inversiones y distancias de la Figura 5.10.1 muestra cómo la variable de la PCK se construye a partir de estas operaciones básicas. Esto proporciona una referencia sustantiva, o criterio, para la variable de la PCK. La imagen resultante nos da una idea de la naturaleza de la variable que subyace a los ítems individuales. En particular, nos muestra cómo generar más ítems para cualquier nivel de dificultad que se desee.

También podemos aprender sobre la variable de la PCK viendo cómo se distribuyen en ella las 101 personas en nuestra muestra. En los renglones 5 y 6 mostramos la posición de cada persona en la variable de acuerdo con su edad en años. Esto nos permite referir la variable a una norma relativa a la mediana de edad de tres a ocho años y nos da una distribución por edades de personas más "maduras" de 9 años de edad o mayores, con una media de 1.3 lógitos y una desviación estándar de 1.9 lógitos. Así, la figura 5.10.1 se convierte en un mapa de la variable que está referido al mismo tiempo a criterio y a norma.

## 5.11 CONTROL DE CALIDAD DE LA CALIBRACIÓN DE LOS ÍTEMS

No podemos esperar que los ítems de un banco conserven sus calibraciones indefinidamente o que funcionen igualmente bien para cualquier persona en la cual se están utilizando. La calidad de la calibración de los ítems debe ser supervisada continuamente. Esto se puede hacer de manera conveniente mediante un examen rutinario de las diferencias entre la manera cómo responden en un momento dado las personas a ítems particulares y cómo esperamos que respondan con base en nuestras calibraciones de los ítems y nuestras medidas de las personas. Estas diferencias son residuos de las expectativas. Un ítem que tiene un residuo sorprendente puede sugerir que aconteció una situación anómala en la prueba o que hay alguna persona peculiar. Sin embargo, las tendencias en los residuos de los ítems pueden ser indicativos de la falla del ítem. Las tendencias a que se presenten problemas en los ítems, a que cambien su dificultad o a que muestran un sesgo respecto de algunos grupos de personas, puede ser revelado por medio de un análisis acumulativo de los residuos del ítem a lo largo del tiempo, de lugares y de grupos de personas. Los ítems problemáticos pueden eliminarse del banco o actualizarse en su dificultad.

El objetivo del control de calidad de los ítems es supervisar de manera permanente la estabilidad de sus calibraciones frente a las posibles influencias ocasionadas por la edad, el sexo, la educación o cualquier otro factor relacionado con las personas y que podría perturbar el funcionamiento de los ítems. Un procedimiento de control de calidad requiere que el uso de cada ítem se vea acompañado de información sobre aspectos educativos y demográficos que proporcionen una base para analizar cuáles otras variables pudieran amenazar la estabilidad de la calibración y, por lo tanto, alterar la interpretación de las respuestas en la prueba. La discusión que sigue se basa en el análisis de ajuste desarrollado en el Capítulo 4.

Para implementar el control de calidad de los ítems, cada vez que se utiliza un ítem registramos estos elementos:

- $x_{vi}$ la respuesta 0 o 1 de la persona v al ítem i
- $b_v$ la habilidad estimada para la persona v derivada de su puntaje en cualquier “prueba" integrada por los ítems calibrados que respondió y
- $(y_v)$ el vector de información demográfica que caracteriza a la persona v.

Cuando las dos piezas de información $x_{vi}$ y $b_v$ se combinan con la dificultad $d_i$ procedente del banco de ítems, podemos determinar un residuo estandarizado $z_{vi}$ que dispondrá de toda la información de esta aplicación del ítem i, que se refiere a la posibilidad de una perturbación en su funcionamiento.

En general, ese residuo estimado $z_{vi}$ es

$$z_{vi} = (x_{vi} - p_{vi})/[p_{vi}(1 - p_{vi})]^{1/2}$$

donde

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)]$$

es la probabilidad de acertar estimada para la persona v en el ítem i y, por lo tanto, la estimación del valor esperado de $x_{vi}$ con base en el modelo.

Dado que $x_{vi}$ solo puede adquirir uno de dos valores, 0 para una respuesta incorrecta y 1 para una respuesta correcta, las posibilidades para $z_{vi}$ y su cuadrado $z_{vi}^2$ se limita a las opciones dadas en Tabla 5.11.1. La improbabilidad de cualquier respuesta particular $x_{vi}$, como una función de su $z_{vi}^2$, es $1/(1+ z_{vi}^2)$. En el ejemplo de la PCKB, hay 101 personas que responden 23 ítems. Estas 23 × 101 respuestas de ítems por personas implican 2323 posibilidades de tener un desajuste. Sin embargo, el desajuste solo se comprueba cuando la diferencia entre la habilidad de la persona $b_v$ y la dificultad del ítem $d_i$ es lo suficientemente grande como para que uno de los valores posibles para la respuesta $x_{vi}$ se vuelva significativamente improbable. Para que esto suceda la diferencia $(b_v - d_i)$ debe ser por lo menos de tres lógitos. En consecuencia, solo hay unas 500 ocasiones de respuestas ítem por persona en las que podría ocurrir un desajuste.

**TABLA 5.11.1**

**RESIDUOS ESTANDARIZADOS DE RESPUESTAS**

| Valor de la respuesta $x_{vi}$ | Residuos Estandarizados: Como desviación en la distribución normal $z_{vi} \sim N(0,1)$ | Residuos Estandarizados: Como ji cuadrado $z_{vi}^2 \sim \chi_1^2$ |
|---|---|---|
| "Incorrecta" 0 | $z_{vi} = -p_{vi}/[p_{vi}(1 - p_{vi})]^{1/2}$<br>$= -[p_{vi}/(1 - p_{vi})]^{1/2}$<br>$= -\exp[(b_v - d_i)/2]$ | $z_{vi}^2 = \exp(b_v - d_i)$ |
| "Correcta" 1 | $z_{vi} = (1 - p_{vi})/[p_{vi}(1 - p_{vi})]^{1/2}$<br>$= [(1 - p_{vi})/p_{vi}]^{1/2}$<br>$= \exp[(d_i - b_v)/2]$ | $z_{vi}^2 = \exp(d_i - b_v)$ |

La tabla 5.11.2 brinda un resumen de las respuestas inesperadas observadas en los datos de la PCKB. La Columna 1 proporciona el rango de la diferencia absoluta entre la habilidad de la persona y la dificultad del ítem. La Columna 2 expresa esta diferencia como $z^2 = \exp(|b - d|)$ y la Columna 3 convierte $z^2$ a la improbabilidad de respuesta $1/[1+z^2]$ que ella implica.

**TABLA 5.11.2**
**RESUMEN DE RESPUESTAS INESPERADAS EN LA PCKB**
**101 PERSONAS POR 23 ÍTEMS**

| Diferencia Habilidad-Dificultad | $z^2$ | Improbabilidad $1/(1+z^2)$ | Conteo posible | Conteo esperado | Conteo observado | Nombre de los ítems | Nombre de las personas |
|---|---|---|---|---|---|---|---|
| Superior a 4.6 | Superior a 99 | Inferior a 0.01 | 226 | 2 | 3 | 4, (7), 9 | (49M), 68F, 93F |
| 3.9 - 4.6 | 49 - 99 | .02 - .01 | 133 | 5 | 3 | 8, 10, 12 | 79M 83F, [95M] |
| 2.9 - 3.8 | 19 - 49 | .05 - .02 | 184 | 20 | 8 | [3], [3], 5 | 12M, 13M, 27F |
| | | | | | | 6, (7), 11 | 47M, (49M), 82F |
| | | | | | | 18, 19 | [95M] 10F |

Hemos contado la cantidad de interacciones ítem-por-persona que podrían caer dentro de cada renglón de la Tabla 5.11.2 y multiplicado este conteo "posible" por su improbabilidad para estimar el conteo que podríamos esperar si estos datos se ajustaran al modelo. Esto fue hecho al multiplicar $(226) \times 0.01 \cong 2$, $(226 + 133) \times 0.02 \cong (2 + 5)$ y $(226 + 133 + 184) \times 0.05 \cong (2 + 5 + 20)$. Los conteos realmente observados en los datos se dan en la Columna 6. Por lo tanto, cuando $(b - d)$ es mayor que 4.6 lógitos, esperamos dos respuestas improbables y observamos tres. Cuando $(b - d)$ está entre 3.9 y 4.6, esperamos alrededor de cinco respuestas improbables y nuevamente observamos tres. Finalmente cuando $(b - d)$ está entre 2.9 y 3.8 esperamos cerca de veinte respuestas improbables pero solo observamos ocho. Estos datos parecen ajustarse bastante bien al modelo.

Cuando revisamos las 14 respuestas más inesperadas de los ítems y las personas que figuran en la Tabla 5.11.2, vemos que están bien distribuidos sobre los ítems y las personas. Solo los ítems 3 y 7 y las personas 49M y 95M aparecen dos veces y los sexos están igualmente representados. Debemos concluir que no se ha detectado ningún signo claro de desajuste sistemático en estos datos.

Sin embargo, al usar el ejemplo de la PCKB para mostrar la aplicación del control de calidad de los ítems, procederemos con un análisis adicional de las seis respuestas más inesperadas. Estas respuestas de las personas 49M, 68F, 79M, 83F, 93F y 95M a los puntos 4, 7, 8, 9, 10 y 12 se dan en la Tabla 5.11.3. Para cada una de tales respuestas incorrectas inesperadas cometidas por las personas capaces al fallar en los ítems fáciles, hemos incluido el valor correspondiente a $(b_v - d_i)$. Además incluimos las características de cada ítem dentro de la variable de la PCK, es decir, su número de golpes, las inversiones y la distancia, así como las características demográficas de sexo, edad y grado para cada persona.

**TABLA 5.11.3**
**LAS SEIS RESPUESTAS MÁS INESPERADAS EN LA PCKB**
**(101 personas por 23 ítems)**

| Habilidad de la persona | Dificultad del ítem $d_i$ | | | | | | Características de la persona | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $b_v$ | −4.3 | −4.1 | −2.6 | −2.1 | −2.1 | −0.1 | Nombre | Sexo | Edad | Grado |
| 0.4 | 0(4.7)* | 1 | 1 | 1 | 1 | 1 | 49 | M | 16+ | 12+ |
| 1.4 | 1 | 0(5.5) | 1 | 1 | 1 | 1 | 68 | F | 16+ | 12+ |
| 1.9 | 1 | 1 | 0(4.5) | 1 | 1 | 1 | 79 | M | 16+ | 12+ |
| 2.4 | 1 | 1 | 1 | 1 | 0(4.5) | 1 | 83 | F | 16+ | 12+ |
| 3.6 | 1 | 1 | 1 | 0(5.7) | 1 | 1 | 93 | F | 16+ | 12+ |
| 4.3 | 1 | 1 | 1 | 1 | 1 | 0(4.4) | 95 | M | 16+ | 12+ |

| Características del ítem ; | | | | | | |
|---|---|---|---|---|---|---|
| Nombre | #7 | #4 | #8 | #9 | #10 | #12 |
| Golpes | 4 | 3 | 4 | 4 | 4 | 4 |
| Inversiones | 0 | 0 | 1 | 2 | 2 | 2 |
| Distancia | 3 | 3 | 5 | 6 | 5 | 7 |

* (b − d) = (0.4) − (−4.3) = 4.7

Podemos esperar que las características asociadas con la dificultad de los ítems – en inversiones y distancia – se incrementen conforme los ítems se van haciendo más difíciles. Por un lado, los seis ítems están en el extremo fácil de la variable. Por otro lado, las seis personas son adultos relativamente capaces. Esto sugiere que, si se llega a detectar una fuente sistemática de desajuste, solo sería una ligera tendencia ocasionada por el descuido, o por falta de atención, entre algunas de las personas mayores que respondieron ítems demasiado fáciles para ellos.

Las matrices de análisis de ajuste, como la de la Tabla 5.11.3, que reúnen las características de la persona y el ítem en las respuestas más inesperadas, son convenientes para supervisar la calidad del funcionamiento del ítem. Estas matrices identifican y sugieren correcciones para las fuentes de falla que ocurren sistemáticamente en los datos.

Los cálculos necesarios para evaluar las respuestas inesperadas se pueden realizar de tres maneras. Las dos primeras son por computadora con el modelo UCON y el método manual explicado en Capítulo 4. La tercera manera es con un método de ajuste burdo (también llamado ajuste crudo o aproximado), que es rápido y a menudo resulta suficiente para efectos prácticos.

Este método burdo de análisis de ajuste consiste en identificar y calcular solo los pocos valores más grandes de $z^2$ observados en un ítem y luego agregarles un 1 por cada persona que responde a ese ítem. Esto supone que toda la perturbación observada en el ítem se debe a sus residuos excepcionales, mientras que el resto del patrón se presenta más o menos como se esperaba.

La Tabla 5.11.4 ilustra este método. Ahí hemos tomado de la Tabla 5.11.3 solo el valor de $z_{vi}^2$ más grande observado en nuestros datos de la PCKB y se le agrega un 1 por cada persona que responde ese ítem, en este caso 100. Esto da $z_{vi}^2 + 100 = \chi^2$ como el valor de ji-cuadrado para ese ítem y $v_i = \chi^2/100$ como su media cuadrática.

Para ver si este método burdo puede ser útil, lo compararemos con UCON y el método manual descrito en el Capítulo 4, pero ahora aplicado a estos datos de la PCKB. En esos procedimientos sumamos los 101 valores de $z^2$ que realmente ocurrieron para hacer nuestro análisis de ajuste de ítem y luego dividimos esta suma de cuadrados por sus 100 grados de libertad para obtener las medias cuadráticas que se muestran en la Tabla 5.11.5.

**TABLA 5.11.4**
**ANÁLISIS DE AJUSTE BURDO PARA SEIS ÍTEMS DE LA PCKB**

| Nombre del ítem | Diferencia Hablidad menos Dificultad | Valor de $z^2$ para el ítem | Estadístico burdo de ajuste | |
|---|---|---|---|---|
| | | | Ji cuadrado $\chi^2 = (z^2 + 100)$ | Media cuadrática $v = \chi^2 / 100$ |
| 9 | 5.7 | 299 | 399 | 4.0 |
| 4 | 5.5 | 245 | 345 | 3.5 |
| 7 | 4.7 | 110 | 210 | 2.1 |
| 10 | 4.5 | 90 | 190 | 1.9 |
| 8 | 4.5 | 90 | 190 | 1.9 |
| 12 | 4.4 | 81 | 181 | 1.8 |

**TABLA 5.11.5**
**COMPARACIÓN DE MÉTODOS DE CONTROL DE CALIDAD DE LOS ÍTEMS APLICADOS A LA PCKB**

| Nombre del ítem | Media cuadrática UCON | Media cuadrática Ajuste manual | Media cuadrática Ajuste burdo |
|---|---|---|---|
| 9 | 3.42 | 3.16 | 4.0 |
| 4 | 2.64 | 2.56 | 3.5 |
| 7 | 1.48 | 1.40 | 2.1 |
| 10 | 1.46 | 1.20 | 1.9 |
| 8 | 1.43 | 1.12 | 1.9 |
| 12 | 1.36 | 0.98 | 1.8 |

Los métodos UCON y de ajuste manual se aproximan bastante entre sí. A pesar de que los cuadrados medios del ajuste burdo son algo más grandes en magnitud, son del mismo orden a los de los otros métodos y sus valores son suficientemente cercanos para tener una idea clara sobre el ajuste relativo de estos seis ítems. La Tabla 5.11.5 sugiere que el método burdo puede ser útil para hacer un análisis rápido del funcionamiento de los ítems.

## 5.12 LA VARIABLE DE LA PCK COMO PRUEBA CON REFERENCIA A NORMA

Las normas no son tan fundamentales para calibrar los bancos de ítems como pudieran ser las distribuciones de las estaturas de las personas al construir una escala, pero es útil disponer de algunas características demográficas de la variable definida por el banco de ítems, porque pueden tener implicaciones normativas en circunstancias particulares. Debido al cambio de énfasis, normar una variable con el modelo de Rasch requiere mucha menos información que normar una prueba. Solo requerimos suficientes ítems para estimar los estadísticos de "normalización" deseados. Una vez normada la variable, cualquier puntaje de cualquier prueba extraída del banco calibrado queda automáticamente normado respecto de la variable.

A menudo estamos satisfechos con una media y una desviación estándar para cada celda de nuestro plan de muestreo normativo. Ambos estadísticos podrían estimarse a partir de una muestra aleatoria de unas 100 personas que responden a una prueba de tan solo dos ítems. Por supuesto, una prueba de 10 a 15 ítems hará un mejor trabajo. No solo las estimaciones serán mejores sino que los ítems adicionales arrojarán errores estándar en torno a los estadísticos de normalización con los cuales hacer una prueba de ajuste para la plausibilidad de los datos. Sin embargo, rara vez serán necesarios más de 15 ítems en una prueba referida a norma. Esto significa que podríamos referir a la norma a seis variables diferentes dedicando 15 ítems en cada una de las seis sub-pruebas, como si se tratara de una prueba compuesta de 90 ítems.

Podemos estimar normas por un método corto a partir de datos de frecuencia en los ítems calibrados del banco, sin tener que medir a las personas individualmente. Esto puede ser útil cuando no resulta pertinente recortar datos de la muestra. Por ejemplo, si buscamos una muestra de la probabilidad de una población, sería preferible no distorsionar las cualidades de la muestra eliminando algunas personas solo porque obtuvieron puntajes cero o perfectos.

El procedimiento de normalización por el método corto se realiza trabajando con el modelo dado un número observado de respuestas correctas para cada ítem calibrado, de este modo:

1. Para cada celda de muestreo en el estudio normativo, seleccione del banco de ítems un conjunto de K ítems calibrados, suficientemente espaciados en dificultad $d_i$ para cubrir la dispersión esperada de las habilidades de la celda particular muestreada. Debe observarse que, en principio, cada celda de muestreo tiene su propia prueba normativa.
2. Administre la prueba de K ítems a una muestra aleatoria de N personas de la celda especificada.
3. Observe el número de personas $s_i$ que responden correctamente en cada ítem.
4. Calcule el logaritmo natural del momio $h_i$ de las respuestas correctas $s_i$ para cada ítem

$$h_i = \ell n\,[s_i/(N - s_i)] \qquad i = 1,K \qquad \textbf{[5.12.1]}$$

5. Realice la regresión de los logaritmos de momio $h_i$ de las dificultades de los ítems $d_i$ sobre los K ítems para obtener el intercepto (u ordenada al origen) A y la pendiente C de la recta de mínimos cuadrados.
6. Estime la media poblacional M y la desviación estándar SD de las habilidades de la celda:

$$M = -A/C \qquad \textbf{[5.12.2]}$$

$$SD = 1.7\left[(1 - C^2)/C^2\right]^{1/2} \qquad \textbf{[5.12.3]}$$

Aplicaremos estos procedimientos a la muestra de 101 personas de la PCKB para ver qué tan bien reproducen la media y la desviación estándar de la muestra que ya hemos estimado de las medidas de cada una de las 101 personas, es decir, M' = 0.19 y SD' = 2.44.

Tenemos las dificultades $d_i$ de los ítems 3 a 25 y, por lo tanto, solo necesitamos calcular los logaritmos naturales de los momios $h_i$ de las respuestas correctas observadas para cada uno de estos ítems. Los valores de $h_i$ se dan en la Columna 3 de la Tabla 5.12.1 con las dificultades correspondientes en la Columna 4.

Al hacer la regresión de los logaritmos de los momios de las respuestas correctas contra las dificultades de los ítems sobre los 23 ítems, obtenemos un intercepto de A = 0.07 y una pendiente de C = – 0.56.

La Ecuación 5.12.2 estima la media de la muestra como

$$M = -A/C$$

$$= -0.07/-0.56$$

$$= 0.13$$

**TABLA 5.12.1**
**LOGARITMOS DE MOMIOS DE LAS RESPUESTAS CORRECTAS Y DIFICULTADES DE LOS ÍTEMS**

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Nombre del ítem | Personas que aciertan $s_i$ | Logaritmo de momio de respuestas correctas $h_i = \ell n[s_i/(N - s_i)]$ | Dificultad del ítem $d_i$ |
| 3 | 98 | 3.49 | −6.20 |
| 4 | 91 | 2.21 | −4.11 |
| 5 | 82 | 1.46 | −2.58 |
| 6 | 83 | 1.53 | −2.76 |
| 7 | 92 | 2.32 | −4.34 |
| 8 | 82 | 1.46 | −2.58 |
| 9 | 78 | 1.22 | −2.06 |
| 10 | 78 | 1.22 | −2.06 |
| 11 | 68 | 0.72 | −1.03 |
| 12 | 57 | 0.26 | −0.12 |
| 13 | 66 | 0.63 | −0.85 |
| 14 | 62 | 0.46 | −0.52 |
| 15 | 73 | 0.96 | −1.51 |
| 16 | 65 | 0.59 | −0.77 |
| 17 | 30 | −0.86 | 1.93 |
| 18 | 37 | −0.55 | 1.36 |
| 19 | 29 | −0.91 | 2.01 |
| 20 | 20 | −1.40 | 2.88 |
| 21 | 16 | −1.67 | 3.33 |
| 22 | 16 | −1.67 | 3.33 |
| 23 | 8 | −2.45 | 4.52 |
| 24 | 2 | −3.90 | 6.27 |
| 25 | 3 | −3.49 | 5.81 |
| | N = 101 | Media | 0.00 |
| | | Desviación estándar | 3.32 |

La Ecuación 5.12.3 estima la desviación estándar de la muestra como

$$SD = 1.7\left[(1 - C^2)/C^2\right]^{1/2}$$

$$= 1.7[(1 - 0.31)/0.31]^{1/2}$$

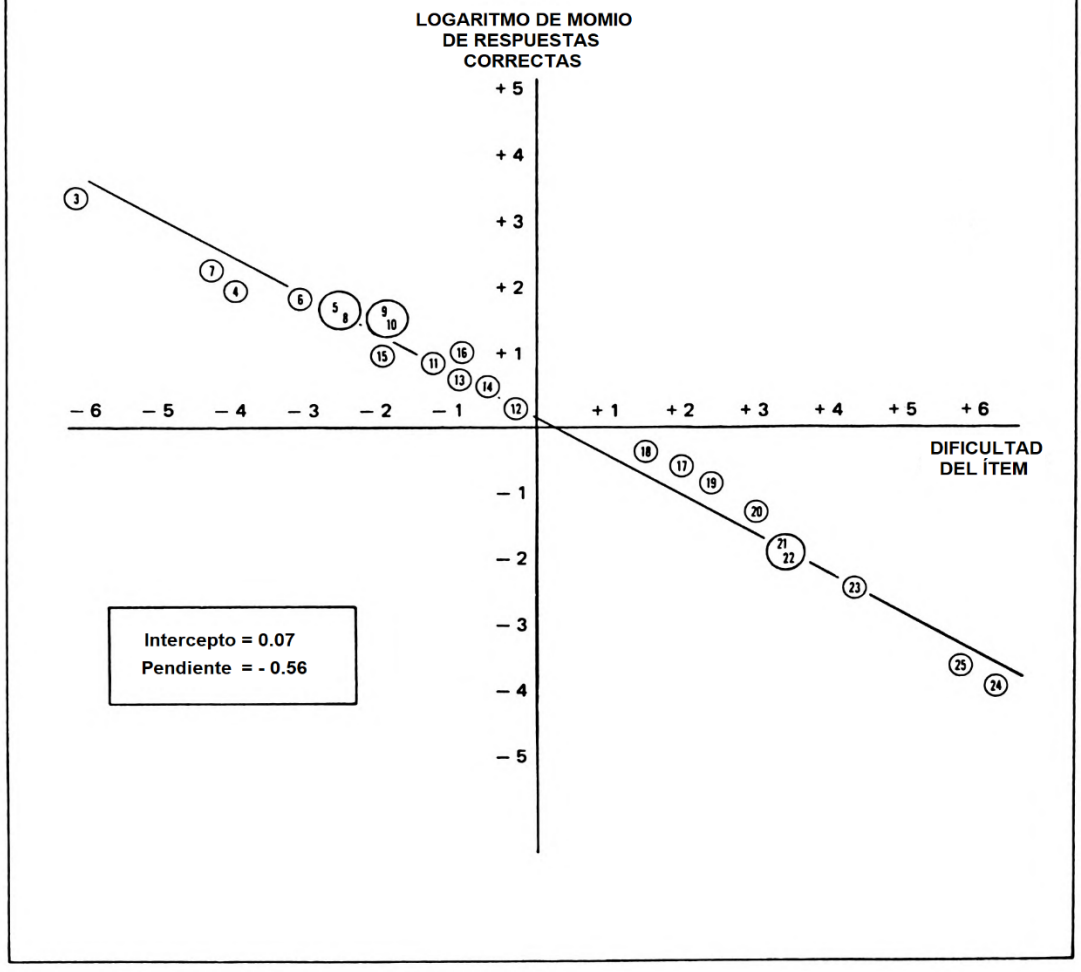
$$= 2.54.$$

Estas estimaciones por el método corto de regresión normativa que dieron 0.13 para la media y 2.54 para la desviación estándar son satisfactoriamente comparables con los valores de 0.19 y 2.44 calculados al medir cada una de las 101 personas y luego calculando su media y desviación estándar de la manera habitual.

La gráfica de los logaritmos naturales de los momios de las respuestas correctas contra las dificultades $d_i$ de los ítem de la Figura 5.12.1 muestra qué tan bien se ajustan estos datos normativos recta esperada por el modelo.

FIGURA 5.12.1

**EL MÉTODO DE NORMALIZADO RÁPIDO APLICADO A LA PCKB PARA ESTIMAR LA MEDIA Y AL DESVIACIÓN ESTÁNDAR DE LA MUESTRA**



# 6 DISEÑO DE PRUEBAS[7]

## 6.1 INTRODUCCIÓN

En el Capítulo 5 vimos cómo establecer la definición operativa de una variable por medio de un banco de ítems calibrados. El siguiente paso es descubrir cómo usar estos ítems calibrados para construir medidas. Para ello debemos considerar dos preguntas relacionadas. En primer lugar debemos encontrar la manera de hacer la mejor selección posible de ítems calibrados de nuestro banco para hacer que cualquier medida que tengamos en mente sea lo más efectiva posible. En segundo lugar, a partir de dicha selección de ítems y de un patrón de respuestas observado en ellos, necesitamos encontrar la manera de evaluar la calidad de esta observación y, si es válida, cómo extraer de ella la medida que buscamos, junto con su error estándar. La primera de estas preguntas, la del diseño óptimo de una prueba, es el tema principal de este capítulo. El Capítulo 7 se encarga de presentar la forma de construir medidas.

## 6.2 EL OBJETIVO DE LA MEDICIÓN

Cuando planificamos una medición, debe haber una persona o un grupo de personas que forman el "blanco" u objetivo[8] acerca de quienes queremos saber más de lo que ya sabemos. Si nos preocupamos por la calidad de las mediciones que pretendemos realizar, entonces deberíamos construir nuestro instrumento de medición con los detalles de este objetivo en mente. Para hacer esto sistemáticamente debemos comenzar por definir tan claramente como podamos lo que esperamos de nuestro "blanco" u objetivo. ¿Dónde suponemos que está ubicado en la variable? ¿Qué tan inciertos estamos de esa ubicación aproximada? ¿Cuál es la habilidad más baja que pensamos puede tener nuestro "blanco" u objetivo? ¿Cuál es la más alta? ¿Cómo se distribuyen entre medio de estos extremos los otros valores posibles?

A veces tenemos conocimiento previo explícito sobre nuestro "blanco" u objetivo. Tal vez nosotros, u otros evaluadores, lo hemos podido medir antes y por ello podemos sugerir su ubicación probable y la dispersión directamente en términos de estas medidas previas sobre la variable y sus errores estándar. Algunas veces podemos usar ítems calibrados a lo largo de la variable, algunos de los cuales creemos que caen probablemente justo en el objetivo, así como otros son demasiado difíciles y otros más son demasiado fáciles. Entonces a partir de las dificultades de estos ítems de referencia podemos tomar indicaciones aproximadas de la ubicación probable del centro y los límites de nuestro objetivo.

De una forma u otra, reunimos y clarificamos nuestras suposiciones sobre nuestro "blanco" u objetivo tan bien como podamos para derivar de ellos el diseño de prueba que tiene la mayor posibilidad de incrementar nuestro conocimiento sobre dicho objetivo.

Obviamente, si ya sabemos todo lo que queremos saber sobre nuestro "blanco" objetivo, entonces no tendríamos que medirlo en primer lugar. Sin embargo, no importa lo poco que sepamos, siempre debemos tener alguna idea acerca de donde está nuestro objetivo. Es esencial tener la mayor claridad posible acerca de ese conocimiento previo, para diseñar la mejor prueba posible.

---

[7] Graham A. Douglas colaboró en la preparación de partes de este capítulo. Ver Wright & Douglas, 1975a.

[8] Ver nota en la página 75.

La especificación del "blanco" u objetivo es una declaración sobre dónde suponemos que está dicho objetivo. Expresamos mejor nuestra conjetura si especificamos dónde se encuentra el centro del objetivo, su supuesta dispersión y tal vez su supuesta forma o distribución. Hagamos:

M = nuestra mejor estimación en cuanto a la ubicación del objetivo
S = nuestra mejor estimación en cuanto a la dispersión del objetivo,
D = nuestra mejor estimación en cuanto a la distribución del objetivo,

entonces podemos describir un "blanco" u objetivo G por la expresión G (M, S, D) y así podemos resumir nuestro conocimiento previo, y por lo tanto, nuestros requisitos de medición para cualquier objetivo que deseamos medir, asumiendo de la mejor manera posible, los valores de los tres parámetros objetivo M, S y D.

Una imagen de un objetivo se da en la Figura 6.2.1. Suponer la ubicación M del objetivo es sumamente fácil. Sin embargo, adivinar la dispersión S y la distribución D nos obliga a pensar en la diferencia que se tiene entre un "blanco" u objetivo cuando se trata de una sola persona, de otro que es el objetivo en un grupo. Para una persona sola, S puede describir la extensión de nuestra incertidumbre respecto a la posición donde se encuentra esa persona. Cuanto mayor es nuestra incertidumbre, mayor es el valor de S.

Si podemos especificar los límites dentro de los cuales nos sentimos bastante seguros de encontrar a la persona, podemos establecer el valor S de modo que $M \pm kS$ defina estos límites. Entonces, incluso si no tenemos una idea clara sobre la distribución D de nuestra incertidumbre entre estos límites, podemos sin embargo esperar que al menos $1 - 1/k^2$ de las posibles medidas caiga dentro $M \pm kS$.

Si vamos más allá y esperamos que las medidas que creemos posibles para la persona se apilan cerca de M, entonces incluso podemos estar dispuestos a asumir que una distribución normal es útil para describir la forma de nuestra incertidumbre. En ese caso, podemos esperar que 0.95 de las posibles medidas caigan dentro de $M \pm 2S$ y prácticamente todas ellas caen dentro de $M \pm 3S$.

Nos referiremos a estas dos distribuciones objetivo como el intervalo de Tchebycheff y como la distribución normal. Podríamos considerar otras distribuciones objetivo, pero estas dos parecen cubrir bastante bien todas las formas de "blanco" u objetivos razonables. Por ejemplo, si llegáramos a sentirnos infelices tratando de imaginar nuestro objetivo como aproximadamente normal, entonces puede ser poco probable que tengamos en mente otra alternativa. Por lo tanto, la alternativa más probable a un "blanco" u objetivo normal es una distribución desconocida, capturada mejor por un intervalo de Tchebycheff. Es importante comprender que todas las posibles formas del objetivo se pueden representar satisfactoriamente con solo dos alternativas razonables, haciendo que una solución única al problema del diseño óptimo de una prueba no solo sea posible sino incluso práctica.

Si el objetivo es un grupo en lugar de un individuo, entonces podemos utilizar S y D para nuestra mejor estimación como la desviación estándar y la distribución de ese grupo. Si pensamos que el grupo tiene una distribución más o menos normal, entonces la asumiremos como nuestra mejor estimación para D. De lo contrario, siempre podemos recurrir al intervalo de Tchebycheff.

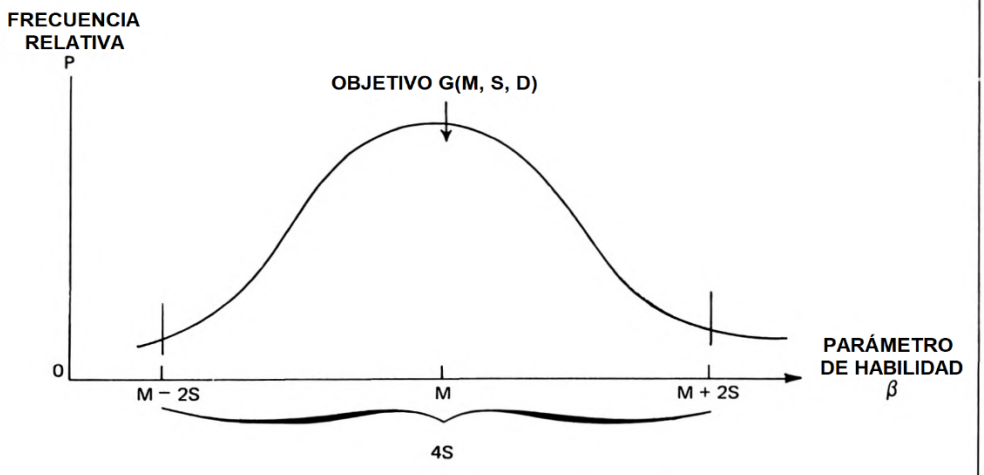
Finalmente, debemos ser explícitos sobre cuán precisa queremos que sea nuestra medición. Después de todo, este es nuestro motivo para hacer mediciones. Es solo porque nuestro conocimiento actual sobre nuestro objetivo es demasiado aproximado para satisfacernos que queremos saber de forma más precisa dónde está dicho objetivo y, si es un grupo en lugar de un solo individuo, conocer su dispersión de manera más precisa. Sin embargo, si el objetivo es un individuo o un grupo, nuestra decisión sobre el valor esperado del error estándar de medida ESM (o SEM, por sus siglas en inglés Standard Error of Measurement) se hará en términos de individuos, para eso a fin de cuentas, es para lo que medimos realmente.

En el caso de que el objetivo sea una persona, queremos que el SEM sea lo suficientemente más pequeño que S como para recompensar nuestros esfuerzos de medición con un incremento útil en la precisión de nuestro conocimiento acerca de la posición donde se encuentra esa persona objetivo. En

el caso de un objetivo grupal, queremos disponer de una estimación mejorada no solo de M, el centro del grupo, sino también de S, su dispersión. La varianza observable de las medidas sobre los valores estimados del grupo, permiten estimar tanto la varianza subyacente de la habilidad $S^2$ como la varianza del error de medida $SEM^2$. Nuestra capacidad de ver la dispersión de nuestro "blanco" u objetivo en el contexto del error de medición depende de que podamos distinguir entre estos dos componentes de la varianza. Puesto que entran en igualmente en la varianza observable de las medidas estimadas, mientras más pequeño sea el valor de $SEM^2$ en comparación con $S^2$, mejor será nuestra identificación y estimación de $S^2$, que es el componente asociado con la dispersión del objetivo. Por lo tanto, para todos los objetivos buscamos un SEM considerablemente más pequeño que S.

FIGURA 6.2.1

LA IMAGEN DE UN OBJETIVO



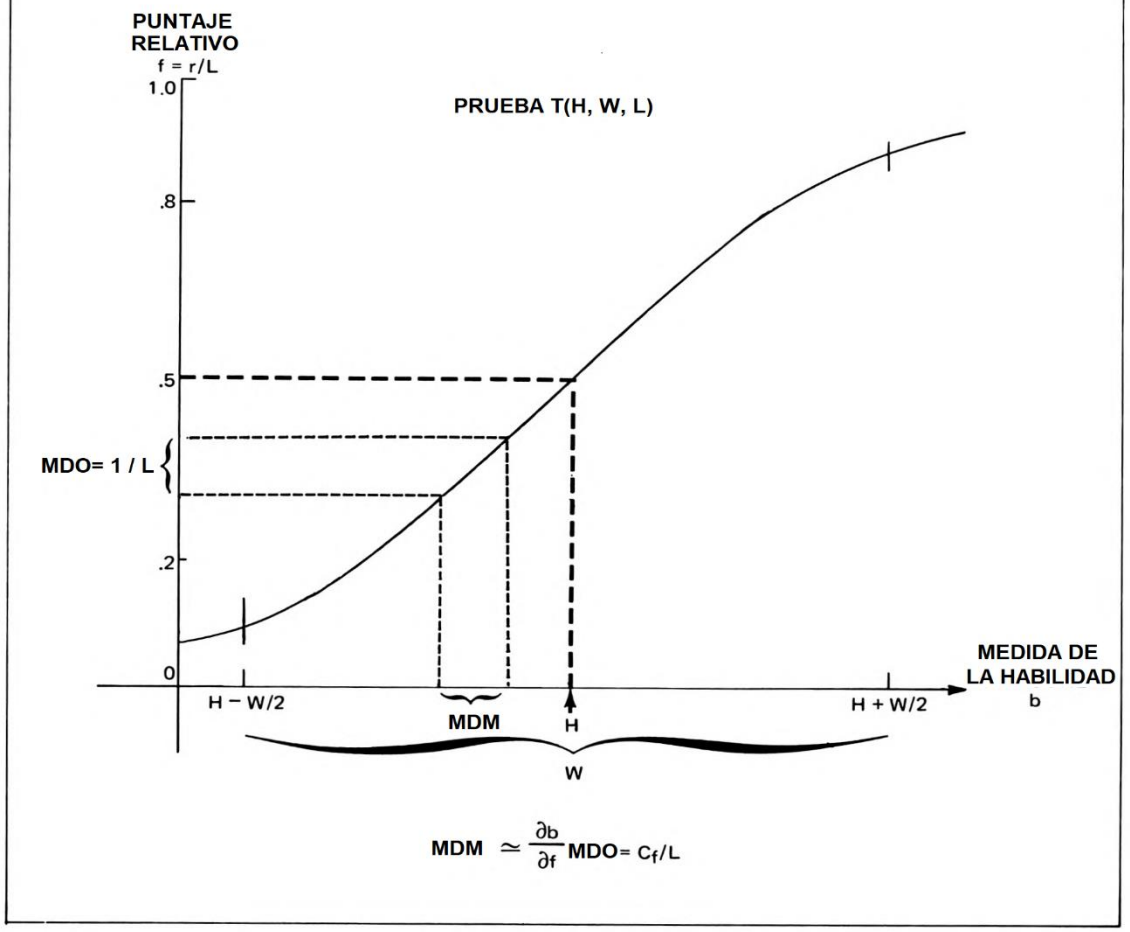

| FORMA D | M ±2S | M ±3S |
|---|---|---|
| Intervalo Tchebycheff | .75+ | .89+ |
| Normal | .95 | .99 |

## 6.3 CARACTERISTICAS MÉTRICAS DE LA PRUEBA

Una prueba es un conjunto de ítems convenientemente calibrados que se eligen para formar de manera conjunta un instrumento de medida. La especificación completa de una prueba es el conjunto de todos los parámetros que caracterizan a estos ítems. Pero cuando examinamos una imagen de cómo funciona una prueba para transformar los puntajes observados en medidas estimadas, vemos que la curva de operación es bastante simple y puede especificarse por medio de unos cuantos parámetros de la prueba. Sabemos que cuando la forma en que funcionan nuestros ítems se ajusta al modelo de Rasch, entonces los únicos parámetros que debemos considerar para determinar las características operativas de una prueba son las dificultades de los ítems. Cuando imponemos una distribución fija razonable sobre estas dificultades, entonces no importa cuántos ítems usemos, podemos reducir el número de parámetros de la prueba a tan solo tres.

**FIGURA 6.3.1**

**LA CURVA OPERATIVA DE UNA PRUEBA**




En la Figura 6.3.1 podemos ver de la forma de la curva operativa de la prueba que sus dos cualidades notables son su posición a lo largo de la variable, que podemos denominar *altura* de la prueba y el rango de habilidades sobre las cuales la prueba puede medir de manera más o menos precisa una característica producida en principio por la dispersión de las dificultades de los ítems, lo que denominaremos su *ancho* o rango

Pero la altura y su ancho o rango no caracterizan a una prueba de manera completa. Si miramos más de cerca a la forma en que la curva de la prueba transforma los puntajes observados en medidas inferidas, veremos que hay una discontinuidad en los puntajes observados que es la que determina el incremento más pequeño en habilidad que podemos medir con una prueba en particular. Esta MDM que es la diferencia más pequeña o Mínima Diferencia Medible (LMD por sus siglas en inglés Least Measurement Difference) depende de la Mínima Diferencia Observable MDO (o LOD por sus siglas en inglés Least Observable Difference). Atendiendo a que el menor cambio posible en el puntaje de una prueba es 1, la MDO en puntaje relativo f = r/L será 1/L. En la sección 6.5 encontraremos que el error estándar de medida, o la menor diferencia aceptable, SEM también depende del número de ítems de la prueba. De hecho SEM = $MDM^{½}$. Por lo tanto, para caracterizar finalmente a la prueba, también debemos especificar su *longitud*.

Por lo anterior, vemos que cualquier diseño de prueba puede definirse más o menos completamente especificando tres características de la prueba; altura, ancho y largo. Si hacemos:

H = altura de la prueba en la variable, es decir es, la dificultad promedio de los ítems seleccionados,
W = ancho de la prueba en dificultades de los ítems, es decir, el rango de sus dificultades
y
L = longitud de la prueba en número de ítems,

entonces podemos describir el diseño de una prueba T por la expresión conveniente T (H, W, L).

En la aplicación práctica del diseño óptimo de una prueba, sin embargo, tendremos que aproximarnos a nuestro diseño óptimo de T para un objetivo G a partir de un conjunto finito de ítems disponibles. Para distinguir en nuestra mente lo que es el diseño óptimo de prueba T (H, W, L) y su realización aproximada, describiremos una prueba real como t (h, w, L) donde

h = dificultad promedio de los ítems utilizados en la prueba,
y
w = estimación del rango de dificultades de los ítems utilizados en la prueba.

## 6.4 LA FORMA DE UNA MEJOR PRUEBA

La mejor prueba es la que proporciona mejores medidas en la región en de la cual se espera realizar la medición[9]. Medir mejor significa medir de manera más precisa. Un diseño óptimo de prueba T (H, W, L) es el que tiene el error mínimo de medición SEM sobre el "blanco" u objetivo G (M, S, D) para la longitud dada L (o dicho en forma equivalente, con la L más pequeña para un valor dado de SEM). 'Sobre el "blanco" u objetivo' implica minimizar la distribución de posibles SEM y, por lo tanto, debe tomarse una postura con respecto a la distribución más probable del "blanco" u objetivo antes de proceder a hacer la minimización de SEM.

Hemos reducido la enorme diversidad de posibles formas centrándonos en dos extremos – el intervalo y distribución normal. ¿Cómo se debe especificar la minimización en cada caso? Para un objetivo normal parece razonable maximizar la precisión promedio, es decir, minimizar el SEM promedio, sobre todo el objetivo.

Para decidir qué hacer cuando el objetivo está definido por medio de un intervalo, necesitamos saber cómo varía el SEM sobre los posibles puntajes de la prueba. Cuando derivamos una forma exacta para la precisión de la medición, encontramos que para pruebas ordinarias con menos de tres lógitos entre elementos adyacentes, se tiene un máximo en la precisión de las mediciones realizadas en el centro de la prueba y disminuye a medida que la prueba y el objetivo están descentrados, es decir, cada vez más alejados entre sí. Para pruebas centradas en sus objetivos esto significa que maximizar la precisión en los límites de un "blanco" dado por un intervalo es una buena forma de maximizar la precisión sobre el intervalo objetivo. Entonces para "blancos" u objetivos de intervalo maximizaremos la precisión en los límites del objetivo.

Cuando derivamos el $SEM^2$ de nuestro modelo de respuesta, descubriremos que es el recíproco de la información sobre la habilidad que suministra la respuesta a cada ítem promediada en la prueba. Dado que los ítems más informativos son los más cercanos a la habilidad que se mide y

[9] Birnbaum (1968, pp. 465-471) intentó cumplir este requisito. Nuestras ideas son consistente con sus esfuerzos, pero las hemos llevado a su conclusión lógica y práctica.

las menos informativas son las más alejadas, la precisión sobre el objetivo no solamente depende de la distribución del "blanco" u objetivo, sino también de la forma de la prueba. Por lo tanto, la cuestión de cuál es la mejor prueba también depende de que tomemos una posición con respecto a la mejor distribución de las dificultades de los ítems en la prueba.

¿Cuáles son las posibilidades razonables? Si queremos medir un objetivo normal, entonces la prueba compuesta por ítems distribuidos en forma normal debería conducir al óptimo de precisión sobre el objetivo. Esta es la conclusión implícita en el análisis de Birnbaum sobre maximizar la información (Birnbaum, 1968, p.467).

Sin embargo, las pruebas con distribución normal en los ítems son inapropiadas para ser construidas. Usar estadísticas de la distribución normal para definir un conjunto de dificultades de los ítems, es una labor tediosa. Más problemático es el extraño concepto de medición a partir de un instrumento compuesto por ítems distribuidos de forma normal. Una prueba normal sería equivalente a una vara de medir (o cinta métrica) cuyas marcas están agrupadas en el medio y dispersas en los extremos. Medir con una vara de medir con esta distribución irregular sería incómodo. A la larga, incluso para "blancos" u objetivos normales, nuestro interés se extiende de manera uniforme sobre todas las habilidades que pueden medirse con una prueba. Una prueba con los ítems igualmente espaciados es la que mejor sirve a nuestros intereses. Esa es la forma en que construimos una vara de medir. El diseño de una prueba correspondiente a una vara de medir con marcas uniformemente distribuidas es en la cual los ítems están espaciados de manera uniforme de lo más fácil a lo más difícil (Birnbaum, 1968, p.466).

Dos distribuciones del "blanco" u objetivo, normal y de intervalo, y dos formas de prueba, normal y uniforme, producen cuatro combinaciones posibles de objetivo y prueba. Wright y Douglas (1975a) investigaron de manera muy extensa estas cuatro combinaciones y encontraron que la prueba normal funciona mejor en un objetivo normal y la prueba uniforme lo hace mejor en un objetivo de intervalo. Sin embargo, cuando compararon las pruebas normales y uniformes en objetivos normales, ambas formas de prueba difieren tan poco en la precisión de su medición que pueden considerarse equivalentes para fines prácticos. En consecuencia, la mejor forma de prueba de propósito general es la prueba uniforme.

## 6.5 LA PRECISIÓN DE UNA MEJOR PRUEBA

Ahora pasemos a la formulación del modelo de respuesta del error estándar de medida SEM de modo que podamos establecer explícitamente qué diseños de prueba maximizan la precisión al minimizar el SEM. Debemos averiguar cómo el diseño de prueba T (H, W, L) influye en el SEM y cómo podemos variar las características H, W y L de la prueba, como respuesta a una especificación del G (M, S, D) para minimizar el SEM sobre ese objetivo.

El modelo de respuesta especifica que:

$$p_{fi} = \exp(b_f - d_i)/[1 + \exp\ (b_f - d_i)] \qquad \textbf{[6.5.1]}$$

donde

$p_{fi}$ = probabilidad de una respuesta correcta en f e i, $d_i$ =
$b_f$ = habilidad estimada en puntaje relativo f = r / L,
$d_i$ = dificultad calibrada del ítem i.

La medida $b_f$ se estima a partir de una prueba de longitud L con ítems $\{d_i\}$ para i = 1, L a través de la siguiente ecuación (para detalles ver las Secciones 1.5 y 3.7):

$$f = \sum_{i}^{L} p_{fi}/L, \qquad \text{for } f = 1/L,\ (L-1)/L \qquad \textbf{[6.5.2]}$$

con varianza asintótica

$$1/\sum_{i}^{L} p_{fi}(1-p_{fi}) = SEM_f^2 \qquad \textbf{[6.5.3]}$$

Este es el cuadrado del error estándar de medida en el puntaje relativo f.

Vemos que $SEM_f$ depende de la suma de $p_{fi}(1-p_{fi})$ sobre i. Por lo tanto, es una función de $b_f$ y todos los $d_i$. Sin embargo, las fluctuaciones en p (1 – p) son bastante suaves para p entre 0.2 y 0.8. Para agilizar la comprensión de la composición de $SEM_f$ podemos reformularlo de modo que el promedio de $p_{fi}(1-p_{fi})$ sobre i sea un componente y la longitud de la prueba L sea el otro.

$$SEM_f = \left\{L/\left[\sum_{i}^{L} p_{fi}(1-p_{fi})\right]\right\}^{1/2} (1/L)^{1/2} = (C_f/L)^{1/2} \qquad \textbf{[6.5.4]}$$

en el cual

$$C_f = \left[\sum_{i}^{L} p_{fi}(1-p_{fi})/L\right]^{-1}$$

En esta expresión, factorizamos en SEM la longitud de prueba L para encontrar un coeficiente de error $c_f$, independiente de la longitud.

Reanudando nuestro estudio de la curva operativa de una prueba dada en la Figura 6.3.1 vemos que la diferencia menos medible en la habilidad MDM es $(\frac{\partial b}{\partial f})$ MDO. Dado que el menor incremento observable en el puntaje relativo es 1 / L, todo lo que necesitamos para completar la formulación de la MDM es la derivada de b con respecto a f, que con base en las ecuaciones 6.5.1 y 6.5.2 es

$$\frac{\partial b}{\partial f} = \left[\sum_{i}^{L} p_{fi}(1-p_{fi})/L\right]^{-1} \qquad \textbf{[6.5.5]}$$

Este es el coeficiente de error $C_f$, por lo tanto, la diferencia menos medible en el puntaje relativo f es

$$MDM_f \simeq C_f/L \qquad \textbf{[6.5.6]}$$

y

$$SEM_f = MDM_f^{1/2} \simeq (C_f/L)^{1/2}$$

Con $SEM_f$ definida de esta forma, observamos que en lo que respecta a la forma de la prueba, $C_f$ es la que requiere minimización. Esto será cierto ya sea que usemos $C_{min}$ para minimizar $SEM_f$ dada L o para minimizar L dado $SEM_f$.

## 6.6 EL COEFICIENTE DE ERROR

Ahora necesitamos saber más sobre este coeficiente de error $C_f$. El ingrediente esencial de $C_f$ es la expresión $p_{fi}(1-p_{fi})$. Esta es la información $I_{fi}$ sobre $b_f$ contenida en una respuesta al ítem i de dificultad $d_i$ (Birnbaum, 1968, p.460-68). Su valor promedio es

$$I_{f.} = \sum_{i}^{L} I_{fi}/L = C_f^{-1}$$

que se calcula sobre los ítems de una prueba y es la información promedio sobre $b_f$ por cada ítem incluido en dicha prueba. Por lo tanto, $C_f$ es el recíproco del promedio de la información de la prueba. Cuanto mayor sea la información obtenida por una prueba menor será $C_f$ y por lo tanto, el $SEM_f$ será también más pequeño y, por consiguiente, la precisión será mayor.

¿Qué valores podemos esperar que tome $C_f$? Abordemos esta pregunta de dos maneras: en términos de la influencia de los valores razonables de $(b_f - d_i)$ en $p_{fi}$ y, para pruebas uniformes, en términos del ancho W de la prueba y las probabilidades límite $p_{fi}$ para i = 1, el ítem más fácil, y $p_{fL}$ para i = L, el ítem más difícil. La probabilidad $p_{fi}$ se definió en la Ecuación 6.5.1.

Comenzando con valores razonables de $(b_f - d_i)$, vemos que cuando $b_f = d_i$ donde la diferencia es cero, se tiene que $p_{fi}=1/2$, $p_{fi}(1-p_{fi})=1/4$ y $C_f=4$, pero cuando $(b_f - d_i)= -2$ se tiene que $p_{fi}=1/8$, $p_{fi}(1-p_{fi})=1/9$ y $C_f=9$. Pero también se tiene que $C_f=9$ cuando $(b_f - d_i)= +2$ y $p_{fi}=7/8$. Dado que un promedio nunca puede ser mayor que su elemento máximo ni menor de su mínimo, podemos usar estas cifras como límites para $C_f$.

Cuando $$-2 < (b_f - d_i) < +2$$

entonces $$1/8 < p_{fi} < 7/8$$

y $$4 < C_f < 9. \qquad \textbf{[6.6.1]}$$

Volviendo a los límites que podemos derivar para $C_f$ con base en el ancho W de la prueba y las probabilidades límite $p_{fi}$ y $p_{fL}$ de una prueba uniforme, podemos usar una expresión para $C_f$ en función de W obtenida por Wright y Douglas, 1975a (ver también Birnbaum, 1968, p.466).

$$C_{fw} = W/(p_{f1} - p_{fL})$$

donde

W = ancho de dificultades de los ítems en una prueba uniforme
$p_{fi}$ = probabilidad de respuesta correcta por $b_f$ en el ítem más fácil en la prueba, y
$p_{fL}$ = probabilidad de respuesta correcta por $b_f$ al ítem más difícil en la prueba.

Cuando $b_f$ está contenido dentro de los límites de dificultad de la prueba y W es mayor que 4, entonces $1/2 < (p_{fi} - p_{fL}) < 1$ y $C_{fw}$ debe estar entre W y 2W, por lo que

Cuando $$d_1 < b_f < d_L$$

y $$W > 4$$

entonces $$W < C_{fw} < 2W. \qquad \textbf{[6.6.2]}$$

De estas consideraciones se deduce que $SEM=(C/L)^{1/2}$ está acotado por

$$2/L^{1/2} < SEM < 3/L^{1/2}$$

para cualquier prueba en la que

$$-2 < (b_f - d_i) < +2,$$

y por

$$(W/L)^{1/2} < SEM < (2W/L)^{1/2}$$

para pruebas uniformes cuando

$$W > 4 \text{ and } d_1 < b_f < d_L.$$

## 6.7 EL DISEÑO DE UNA MEJOR PRUEBA

El diseño óptimo de una prueba depende de relacionar las características del diseño de la prueba T (H, W, L) con las características del objetivo G (M, S, D) para que el SEM se minimice en la región de la variable donde se espera que se realicen las mediciones. La relación entre la prueba y el "blanco" u objetivo que se plasman en la Figura 6.7.1 hacen evidentes los principios generales del diseño óptimo de una prueba. Para hacer coincidir la prueba con el "blanco" u objetivo, apuntemos la altura de la prueba al centro del objetivo, ampliemos la prueba lo suficiente para cubrir la dispersión objetivo y alarguemos la prueba hasta que proporcione la precisión que requerimos.

Para el diseño óptimo de prueba, tanto en un objetivo de intervalo o de distribución normal, seleccionamos un conjunto de ítems equivalentes (donde W = 0) o un conjunto de ítems uniformes con la W indicada en la Tabla 6.7.1. Esta Tabla 6.7.1 proporciona anchos óptimos para pruebas uniformes con objetivos de distribución normal o de intervalo. Por ejemplo, si creemos que el objetivo es aproximadamente normal y suponemos que tiene una desviación estándar S = 1.5, entonces el ancho W óptimo para la prueba es 4. Si, por otro lado, el objetivo se distribuye de manera más uniforme entonces el ancho óptimo puede ser tan grande como 8. Debemos tener en cuenta que para cualquier valor de S, un ancho W pequeño siempre está indicado cuando se espera una forma normal "agrupada" alrededor del "blanco" u objetivo.

La Tabla 6.7.1 también muestra la eficiencia de una regla simple para relacionar el ancho de prueba W a la dispersión S del objetivo. La regla W = 4S se acerca al óptimo W para "blancos" u objetivos que tienen un intervalo estrecho y para objetivos normales amplios. Cuando estamos indefinidos acerca de dónde está nuestro objetivo, estaremos igualmente vagos sobre sus límites. Esa es propiamente la situación en la que estaríamos dispuestos a usar una distribución normal como la forma de nuestro objetivo incierto. En cambio, cuando nuestro objetivo sea estrecho será la ocasión en que estaremos bastante seguros de sus límites objetivo, pero no tan dispuestos a especificar nuestras expectativas en cuanto a su distribución precisa dentro de tales límites estrechos. Así como las formas de intervalo son esperadas naturalmente para objetivos estrechos, del mismo modo las formas normales son inevitables para objetivos amplios, W = 4S es una regla simple de mucha utilidad.

La eficiencia de esta regla simple para "blancos" u objetivos normales y de intervalo se da en las columnas finales de la Tabla 6.7.1. Ahí vemos que su eficiencia casi nunca es inferior a 90 por ciento. Si pasamos de un objetivo de intervalo a otro distribuido en forma normal con una dispersión esperada superior a 1.4, entonces la eficiencia no será inferior a 95 por ciento. Esto significa, por ejemplo, que una prueba de 20 ítems cualesquiera nunca es menos precisa que una prueba óptima de 19 ítems.

Nuestras investigaciones han demostrado que dado el objetivo M, S y D siempre existe un diseño óptimo H y W que permitirá generar un conjunto único de L ítems cuyos parámetros $\{\delta_i\}$ estén uniformemente distribuidos. Sin embargo, este diseño no es más que una idealización y no puede ser perfecto en la práctica. Los bancos de ítems reales son finitos y cada dificultad del ítem no es más que una estimación de su correspondiente parámetro y, por lo tanto, está inevitablemente sujeto a un error de calibración. Nunca podremos seleccionar los ítems exactos estipulados por el diseño óptimo de prueba $\{\delta_i\}$. En cambio, debemos tratar de seleccionar entre los ítems disponibles, un conjunto real de $\{d_i\}$ que esté tan cerca como sea posible a nuestro diseño ideal $\{\delta_i\}$.

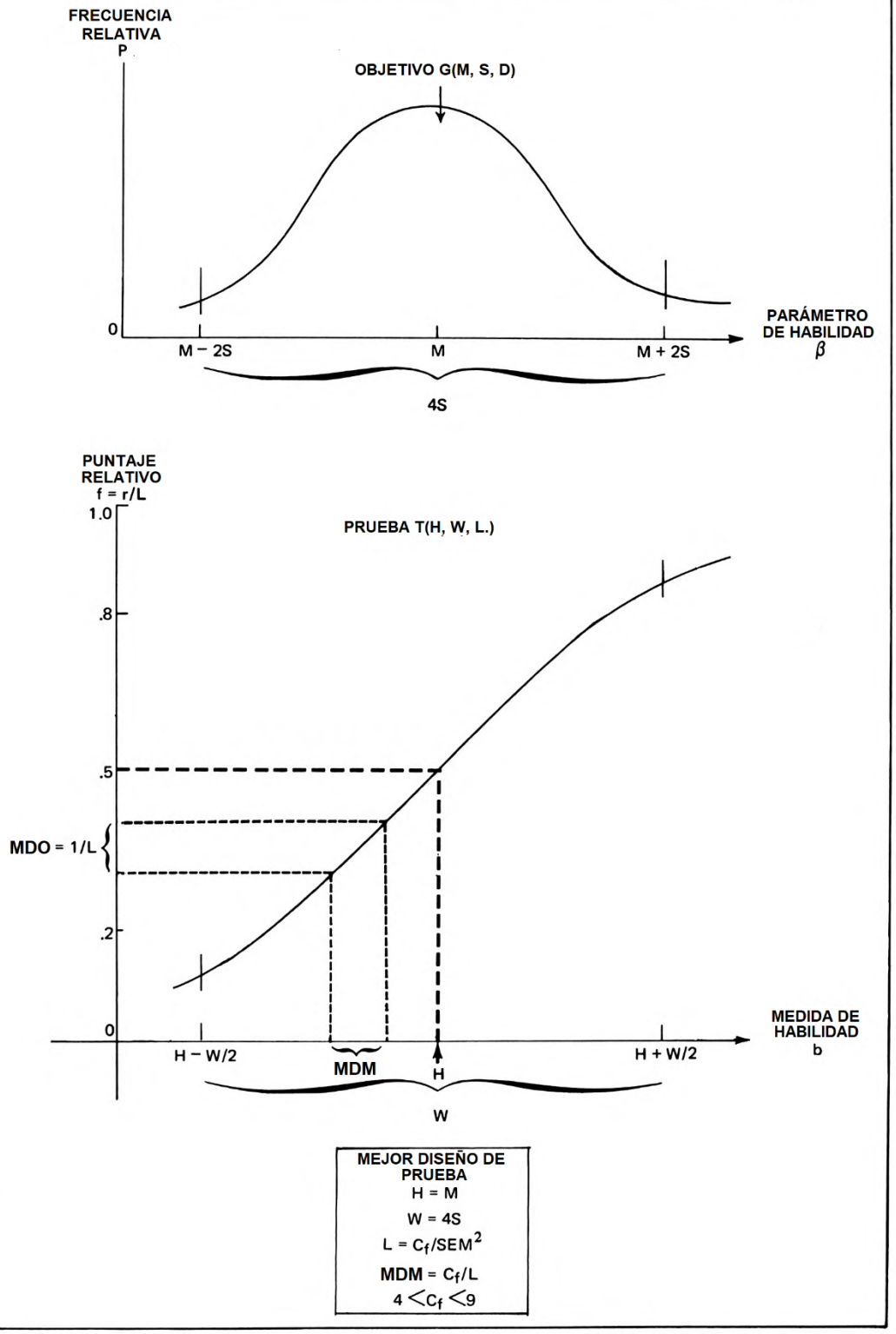
Por lo tanto, paralelamente a la especificación de diseño T (H, W, L) debemos escribir la descripción de la prueba t (h, w, L) que caracteriza la prueba real $\{d_i\}$ que es la que podemos construir en la práctica. Esto nos conduce al problema de estimar h y w.

La altura estimada para la prueba puede determinarse por el promedio de las dificultades estimadas de los ítems de prueba

$$h = \sum_{i}^{L} d_i / L = d. \qquad \textbf{[6.7.1]}$$

FIGURA 6.7.1

**DISTRIBUCIÓN DE UN OBJETIVO Y DE LA CURVA OPERATIVA DE UNA PRUEBA**




**MEJOR DISEÑO DE PRUEBA**

$H = M$

$W = 4S$

$L = C_f/SEM^2$

$MDM = C_f/L$

$4 < C_f < 9$

El ancho estimado para la prueba w se puede determinar a partir del rango de las dificultades estimadas, o tal vez de forma un poco más precisa a partir de una estimación del rango basado en los dos ítems más fáciles $d_1$ y $d_2$ y los dos más difíciles, $d_{L-1}$ y $d_L$.

$$w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)] \quad \textbf{[6.7.2]}$$

**TABLA 6.7.1**

**VALORES ÓPTIMOS DE W PARA MEJORES PRUEBAS UNIFORMES SOBRE OBJETIVOS NORMALES Y DE INTERVALO**

| DESVIACIÓN ESTÁNDAR OBJETIVO S | OBJETIVO NORMAL Error minimizado sobre $N(M,S^2)$ | OBJETIVO DE INTERVALO Error minimizado en $(M \pm 2S)$ | REGLA SIMPLE* W = 4S | EFICIENCIA** | |
|---|---|---|---|---|---|
| | | | | Normal | Intervalo |
| .5 | 0 | 0 | 2 | 94 | 97 |
| .6 | 0 | 0 | 2 | | |
| .7 | 0 | 2 | 3 | 90 | 100 |
| .8 | 0 | 3 | 3 | | |
| .9 | 0 | 4 | 4 | | |
| 1.0 | 0 | 5 | 4 | 89 | 98 |
| 1.1 | 0 | 6 | 4 | | |
| 1.2 | 1 | 6 | 5 | 92 | 96 |
| 1.3 | 2 | 7 | 5 | | |
| 1.4 | 3 | 7 | 6 | | |
| 1.5 | 4 | 8 | 6 | 96 | 91 |
| 1.6 | 5 | 9 | 6 | | |
| 1.8 | 6 | 10 | 7 | 98 | 87 |
| 2.0 | 8 | 11 | 8 | 99 | 84 |

* Esta regla simple es conservadora para objetivos reducidos y más práctica ya que los ítems disponibles están acotados en su posible dispersión. También está cerca del objetivo normal óptimo para objetivos amplios, que es razonable ante cualquier incertidumbre sustancial del objetivo.

** Eficiencia = $C_W / C_{4S} = L_W / L_{4S}$

donde $C_W$ = coeficiente de error mínimo para W óptima

$C_{4S}$ = coeficiente de error para W = 4S.

$L_W$ = longitud de la prueba óptima de ancho W.

$L_{4S}$ = longitud de prueba igualmente precisa de ancho 4S.

## 6.8 LAS REGLAS COMPLETAS PARA EL DISEÑO ÓPTIMO DE PRUEBAS

Ahora estamos en condiciones de dar reglas explícitas, objetivas y sistemáticas para el diseño y uso de la mejor prueba posible. Para diseñar la prueba T (H, W, L) para el objetivo G (M, S, D):

1. A partir de nuestra hipótesis sobre M deducimos H = M.
2. A partir de nuestra hipótesis sobre S derivamos un W óptimo ya sea mediante consulta en la Tabla 6.7.1 o usando la regla simple W = 4S.

3. A partir de nuestros requisitos para la precisión de medida relativa a la SEM que buscamos, derivamos $L=C/SEM^2$. En la Tabla 6.8.1 podemos encontrar un valor bastante preciso para C que, para varios puntajes relativos esperados, conduce a f el valor de C minimizado para W elegido en el Paso 2. Alternativamente, el valor de C puede ser aproximado por una de los reglas simples: C = 6S o C = 6 de las ecuaciones 6.6.1 o 6.6.2.
4. A partir de estos H, W y L, generamos el conjunto de ítems de diseño $\{\delta_i\}$ según la fórmula

$$\delta_i = H - (W/2)[(L - 2i + 1)/L] \qquad \text{for } i = 1,L$$

Luego, para la prueba t (h, w, L) del diseño T (H, W, L):

5. Seleccionamos ítems $d_1$ de nuestro banco de ítems para que se aproximen mejor al conjunto $\{\delta_i\}$ minimizando la discrepancia $(d_i - \delta_i)$.
6. Calculamos

$$h = \sum_i^L d_i / L = d.$$

y

$$w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L - 2)]$$

7. Administramos el conjunto de $\{d_i\}$ como la prueba t (h, w, L).

**TABLA 6.8.1**

**COEFICIENTE DE ERROR $C_{fw}$ PARA ANCHOS DE PRUEBA W Y PUNTAJES RELATIVOS ESPERADOS f PARA PRUEBAS UNIFORMES**

| Puntaje relativo esperado f | Ancho W de la prueba 0 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|---|
| .10 | 10.9 | 11.6 | 13.0 | 13.7 | 15.2 | 16.0 |
| .20 | 6.3 | 6.8 | 7.3 | 8.4 | 10.2 | 11.6 |
| .30 | 4.8 | 5.3 | 5.8 | 7.3 | 9.0 | 10.2 |
| .40 | 4.0 | 4.4 | 5.3 | 6.8 | 8.4 | 10.2 |
| .50 | 4.0 | 4.4 | 5.3 | 6.8 | 8.4 | 10.2 |
| .60 | 4.0 | 4.4 | 5.3 | 6.8 | 8.4 | 10.2 |
| .70 | 4.8 | 5.3 | 5.8 | 7.3 | 9.0 | 10.2 |
| .80 | 6.3 | 6.8 | 7.3 | 8.4 | 10.2 | 11.6 |
| .90 | 10.9 | 11.6 | 13.0 | 13.7 | 15.2 | 16.0 |

$$C_{fW} = W\,[1 - \exp(-W)] \,/\, \{[1 - \exp(-fW)]\,[1 - \exp(-(1 - f)\,W)]\}$$

# 7 CONSTRUYENDO MEDIDAS

## 7.1 EL USO DE UNA VARIABLE PARA CONSTRUIR MEDIDAS

En este capítulo veremos cómo transformar puntajes brutos en medidas. Pero antes de que mostremos esto en las secciones 7.2 y 7.3, revisaremos cómo construir medidas utilizando los ítems que definen una variable en una prueba.

Para construir una medida recolectamos y combinamos una serie de respuestas observadas de modo que permitan sustentar una inferencia acerca de la posición que tiene la habilidad de una persona en una variable dada. Podemos concretar en forma resumida estas observaciones en un puntaje basado en ellas que usamos para inferir la medida de la persona en dicha variable. Sin embargo, la variable por sí misma es más bien una idea más que una experiencia directa. Su naturaleza puede ser inferida solamente de muestras relevantes de observaciones cuidadosamente seleccionadas.

El propósito de la variable es proporcionar una base para comparar a las personas y generalizar su posición relativa. Este propósito requiere que se satisfaga la objetividad en la definición de la variable y la forma en que se hacen las medidas. La idea de la variable trasciende cualquier conjunto particular de observaciones y del mismo modo la medida de la variable debe trascender a las respuestas observadas sobre las cuales está fundamentada. Construir medidas con pruebas requiere ítems objetivamente calibrados con base en las respuestas observadas en cada uno de ellos y con ello las calibraciones traducen estas respuestas en la escala de la variable. Son estos ítems los que definen operacionalmente a la variable y le dan sentido a la medición de las habilidades de la persona.

Las distintas formas de obtener un puntaje particular en una prueba nos conducen a diferentes opiniones de las habilidades de las personas que responden a una prueba. Cuando dos personas obtienen el mismo puntaje, dudosamente pensaríamos ubicar a una persona adelante de otra porque haya contestado correctamente algún ítem en particular. Esto es porque pensamos que cada puntaje es el resultado de la misma exposición a los mismos ítems, dando a cada persona la misma oportunidad de demostrar cierta habilidad. Pero cuando deseamos que idénticos puntajes tengan el mismo significado sin tomar en cuenta cuáles ítems fueron efectivamente contestados, entonces estamos practicando lo que conocemos como medición "independiente del ítem". La práctica generalizada de la medición independiente del ítem en una prueba implica, sin asumir nada más, que la medida es independiente de la prueba dentro de un banco de ítems calibrados.

Un banco de ítems calibrados proporciona un recurso con el cual varios subconjuntos de ítems pueden seleccionarse para formar pruebas específicamente diseñadas con características óptimas. Los puntajes obtenidos con estas pruebas, pueden producir diferentes combinaciones de respuestas "correctas" a diversas elecciones de ítems, sin embargo todas pueden convertirse en medidas comparables a través de las calibraciones del banco. Los procedimientos para obtener medidas comparables para pruebas específicas se presentan en las secciones 7.4 a 7.7.

Para validar estas medidas, sin embargo, debemos evaluar el grado en que las personas en cuestión, han respondido los ítems en la forma en que pretendíamos que fueran respondidos. Las calibraciones de los ítems del banco provienen de ocasiones en las que ocurrió que muchas personas respondieron estos ítems de una manera particular consistente. Este es el contexto en el cual las calibraciones de los ítems adquieren significado. El significado que nos transmiten ahora estas calibraciones depende de cómo responden a estos ítems las nuevas personas que se están midiendo. La validez de sus medidas depende de la presencia de relaciones aceptables

entre lo que realmente observamos y lo que esperamos observar de acuerdo con nuestro modelo de medición y nuestras calibraciones de los ítems. Por lo tanto, antes de poder aceptar cualquier medida como válida, debemos examinar la verosimilitud del patrón de respuestas en que se basa esa medida. El procedimiento para realizar el análisis del ajuste de la persona, necesario para establecer la validez de la medida, se describe en las Secciones 7.8 y 7.9. En estas secciones, mostramos cómo detectar personas desajustadas al modelo y cómo se ven los tipos de desajuste que pueden ocurrir.

Cuando se identifica el desajuste, el siguiente paso es abordar el problema del control de la calidad de la medición que ocasiona este desajuste. Si podemos identificar las circunstancias que conducen al desajuste, entonces podremos extraer de cada registro de respuesta defectuoso una medida que pueda fundamentar al patrón observado de respuestas. Mostramos cómo hacer esto en la Sección 7.10.

## 7.2 CONVERSIÓN DE PUNTAJES CON MEDIDAS POR UCON, PROX Y UFORM

Cuando una persona responde una prueba, la observación resultante de la persona es su puntaje en dicha prueba. Para ver cómo obtener de este puntaje r la medida estimada b que implica, nos referimos al modelo de medición,

$$P\{x_i = 1\} = \pi_i = \exp(\beta - \delta_i)/[1 + \exp(\beta - \delta_i)] \qquad \textbf{[7.2.1]}$$

que especifica cómo la calibración del ítem $\delta_i$ y la medida $\beta$ de la persona están involucradas en la respuesta observada $x_i$. El modelo implica que por cada respuesta de una persona a un ítem podemos "esperar" un valor "probable" intermedio que no sea ni $x_i = 1$ para una respuesta correcta ni tampoco $x_i = 0$ para una respuesta incorrecta, sino más bien un valor entre ellos. Este valor "esperado" es la probabilidad $\pi_i$ para $x_i = 1$ dada en la Ecuación 7.2.1, que funciona igual que nuestra expectativa de que una moneda no cargada cae la mitad del tiempo en “cara”. Ya que "esperamos" un valor en cada lanzamiento de moneda, de modo que la mitad del tiempo caiga “cara” y la otra mitad de tiempo caiga “cruz”, a pesar de que lo que suceda solo puede ser uno o el otro, nuestro valor esperado para un lanzamiento en particular no es 0 ni 1, sino a medio camino entre ellos, en π=½.

Por lo tanto, el valor esperado de la respuesta $x_i$ es

$$E\{x_i\} = \pi_i$$

la probabilidad modelada para una respuesta correcta al ítem i.

Dado que el puntaje de la prueba $r = \sum_i^L x_i$ es la suma de las respuestas a los ítems, el valor esperado de r es la suma de sus expectativas,

$$E\{r\} = E\left\{\sum_i^L x_i\right\} = \sum_i^L E\{x_i\} = \sum_i^L \pi_i$$

Si ahora sustituimos en $\pi_i$, la medida $b_r$ que se estimará para $\beta$ sobre la base del puntaje r y de las calibraciones estimadas $\{d_i\}$ para $\{\delta_i\}$, tenemos una ecuación de estimación que relaciona a r y $b_r$ de la siguiente manera

$$r = \sum_i^L \exp(b_r - d_i)/[1 + \exp(b_r - d_i)]. \qquad \textbf{[7.2.2]}$$

A partir de esta ecuación, el puntaje r de una persona y las calibraciones $\{d_i\}$ de los ítems que respondió, podemos determinar la medida implicada $b_r$.

Una forma de resolver la Ecuación 7.2.2 es usar el algoritmo o procedimiento UCON descrito en el Capítulo 3. La medida estimada por UCON se obtiene realizando j = 1, m iteraciones de

$$b_r^{j+1} = b_r^j + \left(r - \sum_i^L p_{ri}^j\right) / \left[\sum_i^L p_{ri}^j (1 - p_{ri}^j)\right] \qquad \textbf{[7.2.3]}$$

en el cual

$$p_{ri}^{j} = \exp(b_r^{j} - d_i)/[1 + \exp(b_r^{j} - d_i)] \quad \textbf{[7.2.4]}$$

donde el primer valor de $b_r{}^j$ es

$$b_r^{o} = \ell n[r/(L - r)].$$

El algoritmo UCON requiere unas 3 o 4 iteraciones y un criterio de convergencia entre valores sucesivos $b_r{}^j$ como este

$$|b_r^{j+1} - b_r^{j}| < .01 \text{ lógitos.}$$

Cuando se alcanza el criterio de convergencia, entonces la medida estimada es el último valor de $b_r$ a saber

$$b_r = b_r^{j+1} \quad \textbf{[7.2.5]}$$

con error estándar

$$s_r = \left[\sum_{i}^{L} p_{ri}^{j}\ (1 - p_{ri}^{j})\right]^{-1/2} \quad \textbf{[7.2.6]}$$

El algoritmo UCON responde en detalle a la distribución de dificultades de ítems $\{d_i\}$ y así se estima una medida $b_r$ que está completamente libre de cualquier distribución característica que pudiera tener la prueba en función de las dificultades de sus ítems. Si ocurre que las dificultades de los ítems se distribuyen aproximadamente en forma normal con $d_i \sim N(H, \sigma_d^2)$, entonces el algoritmo o procedimiento PROX descrito en el Capítulo 2 es una excelente aproximación al algoritmo UCON.

La medida estimada PROX $b_r$ se puede encontrar sin hacer iteraciones como

$$b_r = h + [1 + (s_d{}^2/2.89)]^{1/2} \ell n[r/(L - r)] \quad \textbf{[7.2.7]}$$

en el cual

$$h = \sum_{i}^{L} d_i /L = d.$$

que proporciona una estimación de la altura de prueba H y

$$s_d{}^2 = \left(\sum_{i}^{L} d_i{}^2 - L d.^2\right)/(L - 1)$$

estima la varianza $\sigma_d{}^2$ de las dificultades de los ítems de la prueba. El error estándar para este $b_r$ es

$$s_r = (1 + s_d{}^2/2.89)^{1/2}\ [L/r(L - r)]^{1/2}. \quad \textbf{[7.2.8]}$$

Dado que es frecuente que las $d_i$ de una muestra de nuevos ítems se aproximen a una distribución normal y dado que es típico que las muestras de personas se distribuyan en forma normal, PROX suele ser útil para calibrar nuevos ítems. Al construir medidas, sin embargo, podemos aprovechar los ítems previamente calibrados y distribuirlos uniformemente $d_i \sim U(H,W)$ sobre el rango de habilidades a medir. Una prueba uniforme como ésta puede describirse completamente por su altura H, su ancho W y su longitud L. Sus medidas pueden calcularse de manera eficiente mediante el algoritmo o procedimiento UFORM descrito en la Sección 7.3.

La medida $b_f$ estimada por UFORM es:

$$b_f = h + w(f - 0.5) + \ell n(A/B) \qquad \textbf{[7.2.9]}$$

donde $A = 1 - \exp(-wf)$

$B = 1 - \exp[-w(1 - f)]$

y $h = \sum_i^L d_i / L = d.$

que estima la altura H de la prueba,

$$w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L - 2)]$$

estima el ancho W de la prueba y

$$f = r/L$$

es el puntaje relativa en la prueba de L ítems (Wright y Douglas, 1975a, 21-23).

El error estándar para este valor $b_f$ es

$$s_f = [(w/L)(C/AB)]^{1/2} \qquad \textbf{[7.2.10]}$$

donde $A = 1 - \exp(-wf)$

$B = 1 - \exp[-w(1 - f)]$

$C = 1 - \exp(-w).$

Para ejemplificar el uso de estos procedimientos hemos escogido nueve personas de nuestra muestra de 101 que contestaron la PCKB. Tres de estas personas están en nivel preescolar, tres más están en nivel primario y los últimos tres son adultos.

En las columnas 2 a 4 de la Tabla 7.2.1 tenemos el sexo, la edad y el grado de estas nueve personas. La Columna 5 contiene sus puntajes en la PCKB. Sus habilidades UCON correspondientes se tienen en la Columna 6.

**TABLA 7.2.1**
**NUEVE PERSONAS SELECCIONADAS DE LA MUESTRA DE LA PCKB**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Grupo de habilidad | Nombre de la persona | Sexo | Edad en años | Grado escolar | Puntaje en la PCKB | Habilidad UCON |
| Preescolar | 3M | M | 3 | Preescolar | 1 | −5.8 |
| | 6F | F | 5 | Preescolar | 3 | −3.9 |
| | 12M | M | 4 | Preescolar | 5 | −2.8 |
| Primaria | 29M | M | 6 | 1 | 10 | −0.9 |
| | 35F | F | 9 | 4 | 11 | −0.5 |
| | 69M | M | 8 | 4 | 15 | 1.4 |
| Adutlo | 88M | M | 17+ | 12+ | 18 | 3.0 |
| | 98F | F | 16 | 11 | 20 | 4.3 |
| | 101F | F | 17+ | 12+ | 21 | 5.2 |

## 7.3 MEDIDAS DE MEJORES PRUEBAS CON UFORM

Previamente a la calibración de los ítems, nuestro conocimiento de sus dificultades proviene de nuestro concepto general que tenemos de la variable que supuestamente definen los ítems. No conocemos la distribución verdadera de esos ítems a lo largo de su variable. En cuanto calibramos los ítems, como hicimos con la PCKB, tenemos una imagen detallada de dónde se ubican. Como resultado podemos utilizar subconjuntos de estos ítems calibrados, seleccionados especialmente para hacer una medición más ágil.

Estas pruebas especialmente diseñadas o "a la medida" pueden variar en longitud, en nivel de dificultad y en el rango de habilidades cubiertas, dependiendo del objetivo de la medición. Las medidas estimadas de estos subconjuntos de ítems pueden hacerse eficientemente porque podemos construir la distribución de dificultades de los ítems de manera que satisfagan nuestro propósito. En particular, si queremos optimizar la eficiencia de nuestros diseños de pruebas, podemos construirlas asegurándonos de que los ítems se distribuyan con dificultades uniformemente espaciadas a lo largo de las medidas de nuestro objetivo. Esto hace que la estimación de medidas con base en puntajes resultantes de estas pruebas, sea enteramente manejable por el algoritmo simple UFORM.

Todo lo que necesitamos para aplicar UFORM es contar con estimaciones de la altura H, el ancho W y a longitud L de la prueba. Con ello, una sola tabla de conversión que organiza los puntajes relativos y el ancho de la prueba, nos proporcionará las medidas que necesitamos para las personas. Una segunda tabla, organizada de manera similar nos dará los coeficientes necesarios para obtener los errores estándar de estas medidas. La Tabla 7.3.1 es una versión reducida de estas medidas relativas y la Tabla 7.3.2 es una tabla reducida de los coeficientes de error. Las tablas completas de medidas relativas y sus coeficientes de error se presentan en el Apéndice Tablas A y B.

Para usar las Tablas 7.3.1 y 7.3.2 (o las tablas A y B), necesitamos el ancho aproximado w de la prueba y los puntajes relativos de las personas f = r/L. Ambos datos determinan la habilidad relativa de la persona $x_{fw}$ y su correspondiente coeficiente de erros $C_{fw}$. Cuando combinamos esta información con la altura h y la longitud L de la prueba, obtenemos la medida $b_{fw} = h + x_{fw}$ y su error estándar $s_{fw} = C_{fw}^{½} / L^{½}$.

Para aplicar las tablas 7.3.1 y 7.3.2 en una prueba en particular, necesitamos valores estimados de las características básicas de la prueba H, W y L. La longitud L es evidente por sí misma. La altura H se estima del promedio de dificultades de los ítems, sea $h = \sum_i^L d_i / L = d$. La estimación del ancho W, sin embargo, puede ser problemático cuando no puede evitarse una distribución irregular de los ítems en los extremos del intervalo.

El ancho de la prueba se puede estimar de varias maneras. Por ejemplo:

$$w_1 = (d_L - d_1)\,[L/(L-1)]$$

$$w_2 = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$$

$$w_3 = [(d_L + d_{L-1} + d_{L-2} - d_3 - d_2 - d_1)/3]\,[L/(L-3)]$$

o también:

$$w_s = 3.5 s_d \qquad \text{donde} \qquad s_d^2 = \left(\sum_i^L d_i^2 - Ld.^2\right)/(L-1)$$

El método que mejor funciona en la práctica es $w_2$, que está basado en la diferencia promedio entre los dos ítems más fáciles y los dos ítems más difíciles. Este procedimiento para estimar el ancho de la prueba se ilustra en la Tabla 7.3.3 donde calculamos w para cinco formas de la PCKB.

**TABLA 7.3.1**

**MEDIDA RELATIVA $x_{fw}$ PARA PRUEBAS UNIFORMES**

| Puntaje relativo | Ancho w de la prueba | | | | |
|---|---|---|---|---|---|
| f | 2 | 4 | 6 | 8 | 10 |
| .1 | − 2.3 | − 2.7 | − 3.2 | − 3.8 | − 4.5 |
| .2 | − 1.5 | − 1.8 | − 2.2 | − 2.6 | − 3.1 |
| .3 | − 0.9 | − 1.1 | − 1.4 | − 1.7 | − 2.1 |
| .4 | − 0.4 | − 0.5 | − 0.7 | − 0.8 | − 1.0 |
| .5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| .6 | 0.4 | 0.5 | 0.7 | 0.8 | 1.0 |
| .7 | 0.9 | 1.1 | 1.4 | 1.7 | 2.1 |
| .8 | 1.5 | 1.8 | 2.2 | 2.6 | 3.1 |
| .9 | 2.3 | 2.7 | 3.2 | 3.8 | 4.5 |

Para mayores detalles ver la Tabla A en el Apéndice

Longitud de la prueba: $L$

Puntaje relativo: $f = r/L$

Altura de la prueba: $h = \sum_{i}^{L} d_i/L$

Ancho de la prueba: $w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

Medida: $b_f = h + x_{fw}$

**TABLA 7.3.2**

**COEFICIENTE DE ERROR $C^{1/2}_{fw}$ PARA PRUEBAS UNIFORMES**

| Puntaje relativo | Ancho w de la prueba | | | | |
|---|---|---|---|---|---|
| f | 2 | 4 | 6 | 8 | 10 |
| .1 | 3.4 | 3.5 | 3.7 | 3.8 | 4.0 |
| .2 | 2.6 | 2.7 | 2.9 | 3.2 | 3.4 |
| .3 | 2.3 | 2.4 | 2.7 | 3.0 | 3.2 |
| .4 | 2.1 | 2.3 | 2.6 | 2.9 | 3.2 |
| .5 | 2.1 | 2.3 | 2.6 | 2.9 | 3.2 |
| .6 | 2.1 | 2.3 | 2.6 | 2.9 | 3.2 |
| .7 | 2.3 | 2.4 | 2.7 | 3.0 | 3.2 |
| .8 | 2.6 | 2.7 | 2.9 | 3.2 | 3.4 |
| .9 | 3.4 | 3.5 | 3.7 | 3.8 | 4.0 |

Para mayores detalles ver la Tabla B en el Apéndice.

Error estándar : $s_{fw} = C^{1/2}_{fw} / L^{1/2}$

**TABLA 7.3.3**

**ANCHO W ESTIMADO DE LA PRUEBA PARA CINCO FORMAS DE LA PCK**

| Forma de la prueba | Longitud de la prueba | Los dos ítems más fáciles | | Los dos ítems más difíciles | | Ancho de la prueba | |
|---|---|---|---|---|---|---|---|
| | | | | | | w' calculada | w utilizada |
| PCKB | 23 | −6.2* | −4.3 | 5.8 | 6.3 | 12.4* | |
| | 22 | −4.3 | −4.1 | 5.8 | 6.3 | 11.3 | 11 |
| Preescolar | 8 | −6.2 | −4.3 | −2.1 | −2.1 | 4.2 | 4 |
| Primaria | 15 | −2.7 | −2.6 | 2.0 | 2.9 | 5.9 | 6 |
| Adulto | 15 | −1.5 | −1.0 | 5.8 | 6.3 | 8.4 | 8 |
| Piloto | 7 | −6.2 | −4.1 | 4.5 | 6.3 | 14.8 | 15 |

$w' = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

w = w' redondeado el entero más cercano

* El ítem 3 en -6.2 está 2 lógitos por debajo de la cadena más o menos uniforme de 22 ítems del ítem 7 en -4.3 al ítem 24 en 6.3. UFORM es más preciso ante este tipo de no uniformidades, cuando el ancho de la prueba se calcula sin los ítems extremos irregulares.

El primer renglón de la Tabla 7.3.3 se refiere al "banco de 23 ítems" de la PCKB. Con estos 23 ítems hemos construido tres pruebas de rango estrecho focalizadas en tres niveles de habilidad: una forma para preescolar de 8 ítems, una forma para primaria de 15 ítems y una forma para adulto de 15 ítems, así como una forma piloto con 7 ítems de rango amplio. Las calibraciones de los dos ítems más difíciles y los más fáciles de cada una de estas formas están dadas en la Tabla 7.3.3. Con todas estas calibraciones podemos estimar los diferentes anchos de las pruebas usando el método $w_2$ para calcular w' de esta forma:

$$w' = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)] \qquad \textbf{[7.3.1]}$$

y al redondear w' al entero más cercano se tiene el valor de w usado en las tabla 7.3.1 y 7.3.2.

Ahora aplicamos los algoritmos de medición UCON, PROX y UFORM a nuestra muestra de nueve personas y comparamos los resultados. La Tabla 7.3.4 muestra las medidas y errores de UCON para los puntajes de la PCKB de 1 a 22. La Tabla 7.3.5 da los puntajes en la PCKB de las nueve personas, sus medidas de habilidad correspondientes y los errores para cada uno de dichos puntajes calculados con UCON, PROX y UFORM.

La persona 29M, por ejemplo, obtuvo un puntaje de 10 respuestas correctas de los 23 ítems de la PCKB. Su habilidad y error en UCON, consultados en la Tabla 7.3.4, son b = - 0.9 y s = 0.6.

Su habilidad y error calculados con PROX a partir de h = 0, X = 2.2 y L = 23 son

$$\begin{aligned} b &= h + X\ell n[r/(L-r)] \\ &= 0 + 2.2\ell n[10/13] \\ &= -0.6 \end{aligned}$$

y

$$\begin{aligned} s &= X\,[L/r\,(L-r)]^{1/2} \\ &= 2.2\,[23/10(13)]^{1/2} \\ &= 0.9 \end{aligned}$$

El valor del factor de expansión X se obtiene a partir de la varianza de las dificultades de los ítems $s_d^2 = 11.0$ como

$$\begin{aligned} X &= \left(1 + s_d^2/2.89\right)^{1/2} \\ &= (1 + 11/2.89)^{1/2} \\ &= 2.2 \end{aligned}$$

**TABLA 7.3.4**
**HABILIDADES Y ERRORES CON UCON PARA LOS 23 ÍTEMS DE LA PCKB**

| Puntaje r | Habilidad b | Error s |
|---|---|---|
| 1 | −5.8 | 1.2 |
| 2 | −4.6 | 1.0 |
| 3 | −3.9 | 0.8 |
| 4 | −3.3 | 0.7 |
| 5 | −2.8 | 0.7 |
| 6 | −2.4 | 0.6 |
| 7 | −2.0 | 0.6 |
| 8 | −1.6 | 0.6 |
| 9 | −1.3 | 0.6 |
| 10 | −0.9 | 0.6 |
| 11 | −0.5 | 0.6 |
| 12 | −0.1 | 0.7 |
| 13 | 0.3 | 0.7 |
| 14 | 0.8 | 0.7 |
| 15 | 1.4 | 0.7 |
| 16 | 1.9 | 0.8 |
| 17 | 2.4 | 0.8 |
| 18 | 3.0 | 0.8 |
| 19 | 3.6 | 0.8 |
| 20 | 4.3 | 0.9 |
| 21 | 5.2 | 1.0 |
| 22 | 6.3 | 1.2 |

## TABLA 7.3.5

## COMPARACIÓN ENTRE MEDIDAS PARA LA PCKB ESTIMADAS CON UCON, PROX Y UFORM

| Grupo de habilidad | Nombre de la persona | Puntaje PCKB r | Puntaje relativo f=r/L | UCON Habilidad $b_1$ | UCON Error $s_1$ | PROX Habilidad $b_2$ | PROX Error $s_2$ | UFORM Habilidad $b_3$ | UFORM Error $s_3$ | Diferencias PROX-UCON $b_2 - b_1$ | Diferencias UFORM-UCON $b_3 - b_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3M | 1 | .04 | −5.8 | 1.2 | − 6.8 | 2.2 | − 6.1 | 1.2 | −1.0 | − 0.3 |
| Preescolar | 6F | 3 | .13 | −3.9 | 0.8 | − 4.2 | 1.4 | − 4.3 | 0.8 | −0.3 | − 0.4 |
| | 12M | 5 | .22 | −2.8 | 0.7 | − 2.8 | 1.1 | − 3.2 | 0.7 | −0.0 | − 0.4 |
| | 29M | 10 | .43 | −0.9 | 0.6 | − 0.6 | 0.9 | − 0.8 | 0.7 | 0.3 | 0.1 |
| Primaria | 35F | 11 | .48 | −0.5 | 0.6 | − 0.2 | 0.9 | − 0.2 | 0.7 | 0.3 | 0.3 |
| | 69M | 15 | .65 | 1.4 | 0.7 | 1.4 | 1.0 | 1.7 | 0.7 | 0.0 | 0.3 |
| | 88M | 18 | .78 | 3.0 | 0.8 | 2.8 | 1.1 | 3.2 | 0.7 | 0.2 | 0.2 |
| Adulto | 98F | 20 | .87 | 4.3 | 0.9 | 4.2 | 1.4 | 4.3 | 0.8 | 0.1 | 0.0 |
| | 101F | 21 | .91 | 5.2 | 1.0 | 5.1 | 1.6 | 5.0 | 0.9 | 0.1 | 0.2 |

**UCON**

Ver Tabla 7.3.4

**PROX**

Altura h = 0
Expansión X = 2.2
Longitud L = 23

$b_2 = 2.2 \ln [r/(23-r)]$

$s_2 = 2.2 [23/r(23-r)]^{1/3}$

**UFORM**

Altura h = 0
Ancho w = 11
Longitud L = 23

$b_3 = x_{fw}$ Ver Tabla 7.3.1

$s_3 = C^{1/3}_{fw}/23^{1/3}$ Ver Tabla 7.3.2

La habilidad y el error calculados con UFORM con el puntaje relativo f = r / L = 10/23 = 0.43 y los valores para $x_{fw}$ y $C_{fw}{}^{1/2}$ que se encuentran en las Tablas A y B del apéndice con h = 0, w = 11 y L = 23. Por lo tanto

$$b = h + x_{fw}, \qquad x_{fw} = -0.8 \quad \text{de la Tabla A}$$
$$= 0 - 0.8$$
$$= -0.8$$

y

$$s = C_{fw}{}^{1/2}/L^{1/2}, \qquad C_{fw}{}^{1/2} = 3.3 \quad \text{de la Tabla B}$$
$$= 3.3/23^{1/2}$$
$$= 0.7$$

Las últimas columnas de la Tabla 7.3.5 dan la diferencia entre PROX o UFORM y UCON. Con excepción de la medida PROX para la Persona 3M, ninguna diferencia es mayor a 0.4 lógitos. Todas las diferencias son menores a la mitad de los errores estándar asociados con sus medidas de habilidad.

La confianza que puede tenerse en el uso de las Tablas 7.3.1 y 7.3.2 para UFORM o en las Tablas A y B del Apéndice depende del conocimiento de su funcionamiento en una variedad de situaciones de prueba típicas. Wright y Douglas (1975a) investigaron su funcionamiento con un estudio de simulación diseñado para verificar las principales amenazas al uso exitoso de estas tablas para proporcionar útiles medidas.

Los resultados de su estudio se resumen en la Tabla 7.3.6 por los límites que corresponden con la posibilidad de que las dificultades de los ítems de una prueba difieran de la distribución uniforme, antes de que las medidas basadas en dicho supuesto lleguen a ser inaceptables. La Tabla 7.3.6 muestra combinaciones de H – $\beta$, W y L dentro de las cuales el sesgo al estimar $\beta$ causado por la falta de uniformidad en las dificultades de los ítems es inferior a 0.1 lógitos.

El margen o la tolerancia que se muestra en la Tabla 7.3.6 puede parecer sorprendente, ya que permite que la dificultad del ítem adquiera valores tan aleatorios como d = 2.0 cuando la uniformidad requiere $\delta$=1.0. Cuando se calculan h y w a partir de las $d_i$ reales de una prueba, puede demostrarse que el espectro de diseños de pruebas, por más amplio que sea, es excepcionalmente robusto frente a las desviaciones aleatorias en la uniformidad en las dificultades de los ítems.

La Tabla 7.3.6 muestra que a medida que la longitud aumenta por arriba de 30 ítems, ninguna prueba se ve amenazada por un sesgo de medida lo suficientemente grande como para ser importante. Las pruebas de alrededor de 30 ítems, de ancho inferior a 8 lógitos y que se encuentran a menos de 1 lógito de su objetivo $\beta$ están, para todos los propósitos prácticos, libres del sesgo causado por desviaciones aleatorias en la uniformidad de la calibraciones de los ítems, siempre que sean de magnitud inferior a 1 lógito. Solo cuando las pruebas son cortas, menores a 10 ítems, o más amplias que 8 lógitos o más de 2 lógitos fuera del objetivo, el sesgo causado por la falta de uniformidad aleatoria de las dificultades de los ítems llega a exceder 0.2 lógitos. Esto significa que, a pesar de que se basan en la suposición de pruebas perfectamente uniformes, las tablas de medición UFORM se puede utilizar para transformar puntajes en medidas en la mayoría de las aplicaciones prácticas.

**TABLA 7.3.6**

**RENDIMIENTO DEL PROCEDIMIENTO UFORM PARA PRUEBAS MENOS DE 8 LÓGITOS DE ANCHO**

| Máximo sesgo del ítem $\lvert d_i - \delta_i \rvert$ | Máximo valor fuera del objetivo $\lvert H - \beta \rvert$ | Mínima longitud de la prueba L | Máximo sesgo de medida | |
|---|---|---|---|---|
| | | | SESGO | SESGO/SEM |
| 1.0 | 2 | 10 | .2 | .4 |
| | 1 | 30 | .1 | .3 |
| 0.5 | 2 | 10 | .1 | .2 |
| | 1 | 30 | .1 | .2 |

SESGO = El sesgo promedio de medición en 100 réplicas de una prueba en la cual las discrepancias aleatorias respecto de una distribución uniforme de las dificultades de los ítems están acotadas por $\lvert d_i - \delta_i \rvert$.

## 7.4 PRUEBAS INDIVIDUALIZADAS

La necesidad de pruebas individualizadas se vuelve obvia si encontramos que se le han aplicado ítems inapropiados a una persona. La solución a este problema es construir pruebas a la medida de las personas. La construcción de un banco de ítems calibrados hace que sea simple la implementación eficiente de pruebas "confeccionadas a la medida". La uniformidad de la precisión de la medida cerca del centro de las pruebas de altura y ancho típicos muestra que solo necesitamos disponer de ítems seleccionados en valores próximos a un lógito del objetivo previsto para lograr una prueba a la medida "suficientemente buena". Esto se puede hacer de varias maneras.

Confección por estatus. La información sobre la ubicación por grado o por edad a menudo será suficiente para confeccionar una prueba escolar. Se puede utilizar el conocimiento previo de la ubicación aproximada que tiene un grupo "blanco" o un alumno objetivo dentro del grado escolar, junto con la información normalizada de las variables en dicho grado, para determinar un segmento apropiado de ítems. Los datos normativos en una variedad de materias escolares sugieren que las desviaciones estándar dentro del grado son de aproximadamente un lógito. Cuando esto es así, una medida burda de la medida del estudiante, tan burda como del orden de un cuartil en el grado, puede proporcionar información más que suficiente para diseñar una mejor prueba para ese alumno.

Confección por desempeño. Cuando no es suficiente contar con información sobre el grado o la edad, se puede hacer la construcción a la medida con una prueba piloto de 5 a 10 ítems distribuidos a lo largo de las dificultades para cubrir lo más ampliamente el objetivo esperado. Si la prueba piloto se configura para ser auto evaluable, entonces los alumnos pueden usar su número de respuestas correctas para guiarse en una segunda prueba específicamente confeccionada a su nivel de habilidad implícito en su puntaje dentro de la prueba piloto.

Auto confección. Un tercer esquema aún más individualizado puede resultar práctico en muchas circunstancias. La persona a medir recibe un cuadernillo con los ítems presentados en orden uniforme creciente de dificultad y se le pide que encuentre su mejor nivel en el que puede trabajar. La prueba comienza cuando la persona encuentra ítems lo suficientemente difíciles como para interesarle, pero lo suficientemente fáciles como para dominarlos. La prueba continúa con ítems más difíciles hasta que la persona decida que el nivel de dificultad está más allá de su habilidad. La prueba auto confeccionada o personalizada que servirá para medir a esta persona está formada por el segmento continuo de ítems que se intentaron.

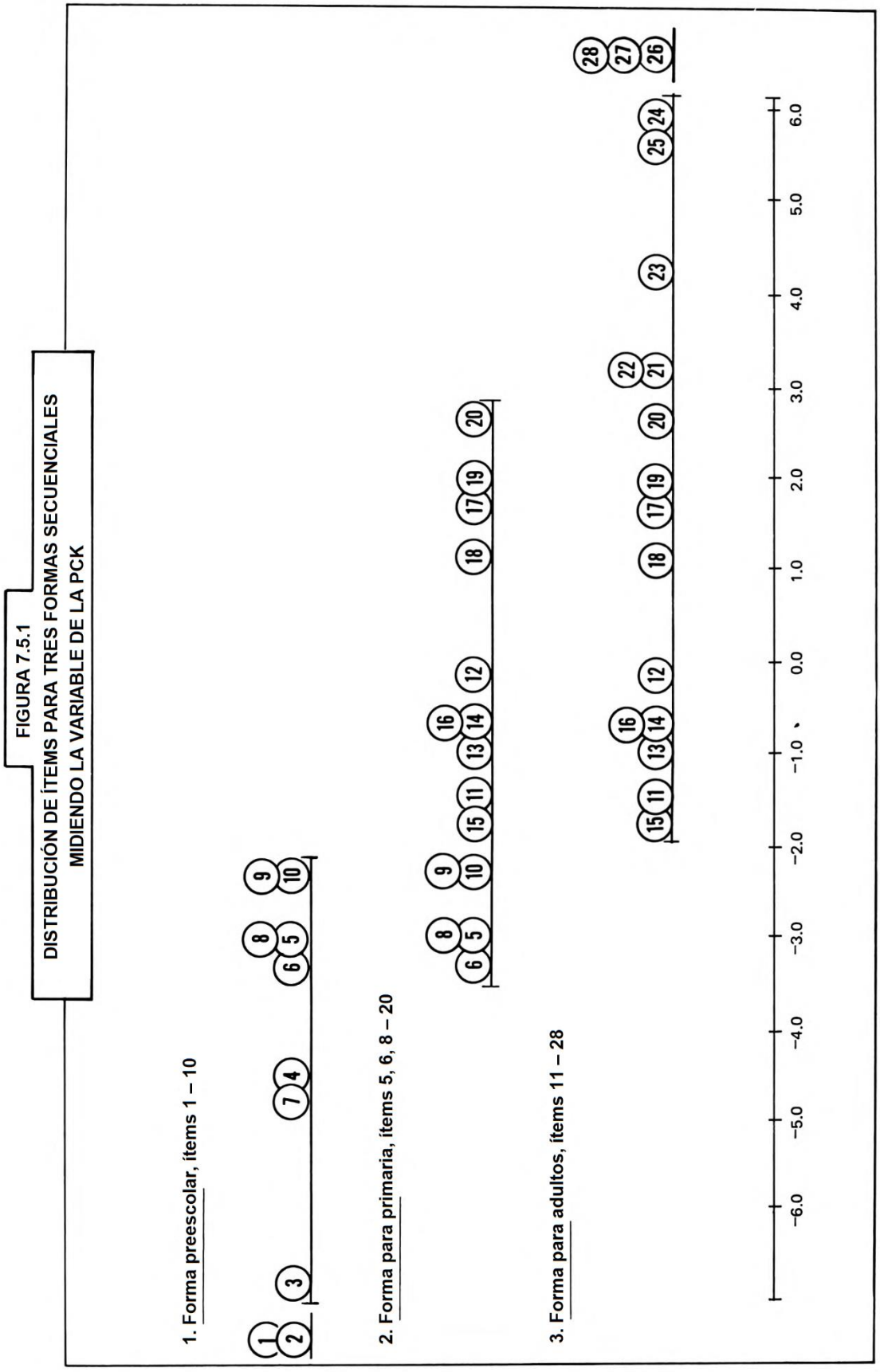


FIGURA 7.5.1
DISTRIBUCIÓN DE ÍTEMS PARA TRES FORMAS SECUENCIALES
MIDIENDO LA VARIABLE DE LA PCK
1. Forma preescolar, ítems 1 – 10
2. Forma para primaria, ítems 5, 6, 8 – 20
3. Forma para adultos, ítems 11 – 28
−6.0
−5.0
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0

Este enfoque se auto-adapta a las variaciones individuales de velocidad, comodidad en la prueba y nivel de desafío o reto productivo. La gran variedad de segmentos de pruebas diferentes que pueden construirse es fácil de manejar. El número secuencial de los ítems más fáciles y más difíciles que se intentaron y la cantidad de respuestas correctas entre ellos, se puede leer en una hoja de respuestas para auto-calificación y convertirse en una medida, en tanto el error estándar se obtiene simplemente buscando estos tres estadísticos en una tabla simple de una página que puede incluirse dentro del cuadernillo de la prueba.

Las pruebas auto confeccionadas corresponden al uso de niveles de piso y techo en pruebas administradas individualmente como la Stanford-Binet. La única diferencia es que, con una prueba auto confeccionada, el segmento de los ítems administrados lo determina la persona que responde la prueba en lugar de que sea definido por un examinador.

## 7.5 CONFECCIÓN DE LA PRUEBA POR ESTATUS

Para ilustrar la confección por estatus, asignamos nuestros ítems de la PCKB a tres Formas secuenciales. La Forma Preescolar está dirigida a niños en edad preescolar. La Forma Primaria está dirigida a niños en el primer grado escolar. La Forma para Adultos es para personas que ya terminaron la educación primaria. Figura 7.5.1 muestra la distribución de los ítems en estas tres Formas.

La Forma preescolar se compone de los primeros 10 ítems. Solo los ítems 3 a 10 están calibrados porque prácticamente todos cuantos han contestado la prueba han acertado en los ítems 1 y 2 .La Forma Primaria se compone de los Ítems 5, 6 y 8 a 20 para cubrir el rango medio de la variable. La Forma para adultos está compuesta por los ítems 11 a 25, que son los ítems calibrados de mayor dificultad, y los ítems 26, 27 y 28 que son tan difíciles que nadie ha podido acertar en ellos.

Debe tenerse en cuenta que podemos incluir estos cinco ítems "fuera de límite" en nuestras formas de prueba sin afectar en absoluto nuestras mediciones. Esto se debe a que podemos enfocar nuestra medición en la parte de la prueba que responde la persona con base en los ítems calibrados mientras dejamos que los ítems extremos sigan trabajando como el límites de la variable PCK. Si eventualmente llegamos a encontrar personas que fallan en los ítems 1 o 2 o que responden los ítems 26, 27 o 28, entonces también los podremos calibrar en la variable PCK y usar sus respuestas en nuestras mediciones.

Los ítems para cada una de las tres formas y sus correspondientes dificultades, donde son conocidas, se dan en la Tabla 7.5.1. Debajo de los ítems en cada Forma están las características de la Forma de prueba: altura h, ancho w y longitud L.

Estas tres Formas de prueba se aplicaron a las nueve personas. La Tabla 7.5.2 muestra las respuestas de las personas en cada una de las Formas. Las personas 3M y 6F pueden medirse solo con la Forma Preescolar, mientras que las personas 98F y 101F solo pueden se miden con la Forma para Adulto. La persona 12M produjo un registro medible en las Formas Preescolar y Primaria. Las personas 69M y 88M produjeron registros medibles en las Formas Primaria y para Adultos. Las personas 29M y 35F produjeron registros medibles en las tres Formas.

Para la persona 29M con puntaje relativo de 0.75 en la Forma Preescolar (h = – 3.3, w = 4, L = 8) buscamos $x_{fw} = 1.4$ y $C_{fw}{}^{½} = 2.6$ en las tablas A y B para encontrar la estimación

$$b = -3.3 + 1.4 = -1.9$$

con error estándar

$$s = 2.6/8^{1/2} = 0.9$$

**TABLA 7.5.1**

**ESTADÍSTICAS DE TRES FORMAS SECUENCIALES DE LA PRUEBA QUE MIDE LA VARIABLE PCK.**

| FORMA PREESCOLAR | | FORMA PRIMARIA | | FORMA PARA ADULTO | |
|---|---|---|---|---|---|
| Nombre del ítem | Dificultad del ítem | Nombre del ítem | Dificultad del ítem | Nombre del ítem | Dificultad del ítem |
| 1 | * | | | | |
| 2 | * | | | | |
| 3 | − 6.2 | | | | |
| 4 | − 4.1 | | | | |
| 5 | − 2.6 | 5 | − 2.6 | | |
| 6 | − 2.7 | 6 | − 2.7 | | |
| 7 | − 4.3 | | | | |
| 8 | − 2.6 | 8 | − 2.6 | | |
| 9 | − 2.1 | 9 | − 2.1 | | |
| 10 | − 2.1 | 10 | − 2.1 | | |
| | | 11 | − 1.0 | 11 | − 1.0 |
| | | 12 | − 0.1 | 12 | − 0.1 |
| | | 13 | − 0.9 | 13 | − 0.9 |
| | | 14 | − 0.5 | 14 | − 0.5 |
| | | 15 | − 1.5 | 15 | − 1.5 |
| | | 16 | − 0.8 | 16 | − 0.8 |
| | | 17 | 1.9 | 17 | 1.9 |
| | | 18 | 1.4 | 18 | 1.4 |
| | | 19 | 2.0 | 19 | 2.0 |
| | | 20 | 2.9 | 20 | 2.9 |
| | | | | 21 | 3.3 |
| | | | | 22 | 3.3 |
| | | | | 23 | 4.5 |
| | | | | 24 | 6.3 |
| | | | | 25 | 5.8 |
| | | | | 26 | ** |
| | | | | 27 | ** |
| | | | | 28 | ** |

**Características de las pruebas con ítems calibrados**

| | Forma preescolar | Forma primaria | Forma para adulto |
|---|---|---|---|
| Altura: | h = −3.3 | h = −0.6 | h = 1.8 |
| Ancho: | w = 4 | w = 6 | w = 8 |
| Longitud: | L = 8* | L = 15 | L = 15** |

* Los ítems 1 y 2 son muy fáciles para poder ser calibrados

** Los ítems 26, 27, 28 son muy difíciles para poder ser calibrados

Los ítems 1, 2, 27, 27 y 28 no pueden usarse para la medición porque sus niveles de dificultad están fuera de rango y no permiten su calibración.

## TABLA 7.5.2

## MEDICIÓN DE LOS TRES GRUPOS DE HABILIDAD CON LAS TRES FORMAS SECUENCIALES

| | | Forma preescolar (h=-3.3, w=4, L=8) | | | | Forma primaria (h=-0.6, w=6, L=15) | | | | Forma para Adultos (h=1.8, w=8, L=15) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grupo de habilidad | Nombre de la persona | Puntaje $r_1$ | Puntaje relativo f=r/8 | Habilidad $b_1$ | Error $s_1$ | Puntaje $r_2$ | Puntaje relativo f=r/15 | Habilidad $b_2$ | Error $s_2$ | Puntaje $r_3$ | Puntaje relativo f=r/15 | Habilidad $b_3$ | Error $s_3$ |
| | 3M | 1 | .13 | −5.7 | 1.1 | 0 | 0 | * | | 0 | 0 | * | |
| Preescolar | 6F | 3 | .38 | −3.9 | 0.8 | 0 | 0 | * | | 0 | 0 | * | |
| | 12M | 5 | .63 | −2.6 | 0.8 | 3 | .20 | −2.8 | 0.8 | 0 | 0 | * | |
| | 29M | 6 | .75 | −1.9 | 0.9 | 7 | .47 | −0.8 | 0.7 | 4 | .27 | −0.2 | 0.8 |
| Primaria | 35F | 7 | .88 | −0.8 | 1.2 | 8 | .53 | −0.4 | 0.7 | 4 | .27 | −0.2 | 0.8 |
| | 69M | 8 | 1.00 | ** | | 12 | .80 | 1.6 | 0.8 | 7 | .47 | 1.6 | 0.8 |
| | 88M | 8 | 1.00 | ** | | 14 | .93 | 3.0 | 1.1 | 10 | .67 | 3.2 | 0.8 |
| Adulto | 98F | 8 | 1.00 | ** | | 15 | 1.00 | ** | | 12 | .80 | 3.4 | 0.8 |
| | 101F | 8 | 1.00 | ** | | 15 | 1.00 | ** | | 13 | .87 | 5.2 | 0.9 |

* El puntaje 0 es demasiado bajo para usarse en la medición

** Los puntajes 8 y 15 son muy altos para usarse en la medición

Para el puntaje relativo de 0.47 obtenido por la Persona 29M en la Forma Primaria (h = –0.6, w = 6, L = 15) buscamos $x_{fw}$ =–0.2 y $C_{fw}{}^{½}$ = 2.6 para encontrar la estimación

$$b = -0.6 - 0.2 = -0.8$$

con error estándar

$$s = 2.6/15^{1/2} = 0.7$$

Para el puntaje relativo de 0.27 obtenido por la Persona 29M en la Forma para Adulto (h = 1.8, w = 8, L = 15) buscamos $x_{fw}$ = –0.2 y $C_{fw}{}^{½}$ = 3.0 estimar

$$b = 1.8 - 2.0 = -0.2$$

con error estándar

$$s = 3.0/15^{1/2} = 0.8$$

Aunque solo una de las formas consideradas se focaliza más en una persona para realizar su "mejor" medida, todavía podemos ver en la Tabla 7.5.2 que, a pesar de la gran variación en el puntaje en diferentes Formas, las medidas para una persona determinada son comparables en su mayoría. La persona 29M produce la mayor variación en las medidas sobre estas tres Formas. Sus tres puntajes relativos de 0.75, 0.47 y 0.27 varían ampliamente como respuesta a la variación en dificultad de las tres formas. De acuerdo con nuestro modelo sus tres medidas de –1.9, –0.8 y –0.2 deberían ser estadísticamente equivalentes, aunque dan la impresión de variar más de lo que pudiéramos desear. Cuando su variación se evalúa a la luz de su errores estándar de 0.9, 0.7 y 0.8 vemos que la estimación más baja de –1.9 más un error estándar en la Forma Preescolar y la estimación más alta de –0.2 menos un error estándar en las Formas para Adulto, se acercan al valor –1.0.

La Tabla 7.5.3 muestra la medida de habilidad para cada persona en toda la PCKB y su medida de habilidad en cada una de las tres formas secuenciales. La diferencia entre cada Forma de prueba y la PCKB se dan a la derecha de la tabla. Cuando estas diferencias se comparan con los errores asociados a ellas se puede ver que todas esas diferencias son menores a medio error estándar, excepto los de las personas 29M y 98F.

Los errores estándar para cada habilidad en la PCKB y las tres Formas de prueba se tienen en la Tabla 7.5.4. Estos valores son estables y consistentes en todas las Formas para las nueve personas.

## 7.6 CONFECCIÓN DE LA PRUEBA POR DESEMPEÑO

Para ilustrar la confección por desempeño desarrollamos una forma piloto de siete ítems a partir de la PCKB. La figura 7.6.1 muestra la distribución de estos siete ítems. Fueron seleccionados para que se distribuyeran lo más uniformemente posible en el rango de 15 lógitos de la variable de la PCK. La tabla 7.6.1 indica las dificultades de los ítems y las características de la Forma Piloto. La altura está centrada en 0.0. El ancho efectivo es de 15 lógitos.

Para demostrar la confección por desempeño con esta Forma Piloto utilizaremos los resultados de nuestras nueve personas en dicha Forma Piloto para señalar la forma secuencial más apropiada para medir cada uno de ellos. Luego, los mediremos en la Forma secuencial y compararemos su medida "confeccionada por desempeño" contra su medida basada en los 23 ítems de la PCKB.

**TABLA 7.5.3**

**COMPARACIÓN DE LAS MEDIDAS DE LAS FORMAS SECUENCIALES CONTRA LAS MEDIDAS EN EL BANCO DE LA PCKB**

| Grupo de habilidad | Nombre de la persona | Habilidad en la PCKB* | Formas secuenciales** | | | Diferencia entre medidas en la PCKB y las Formas secuenciales | | | Error |
|---|---|---|---|---|---|---|---|---|---|
| | | | Preescolar | Primaria | Adulto | | | | |
| | | L=23 $\hat{\beta}$ | L=8 $b_1$ | L=15 $b_2$ | L=15 $b_3$ | $b_1 - \hat{\beta}$ | $b_2 - \hat{\beta}$ | $b_3 - \hat{\beta}$ | $s_{\hat{\beta}}$ |
| Preescolar | 3M | −5.8 | − 5.7 | | | 0.1 | | | 1.2 |
| | 6F | −3.9 | − 3.9 | | | 0.0 | | | 0.8 |
| | 12M | −2.9 | − 2.6 | − 2.8 | | 0.3 | 0.1 | | 0.7 |
| Primaria | 29M | −0.9 | − 1.9 | − 0.8 | − 0.2 | − 1.0 | 0.1 | 0.7 | 0.6 |
| | 35F | −0.5 | − 0.8 | − 0.4 | − 0.2 | − 0.3 | 0.1 | 0.3 | 0.6 |
| | 69M | 1.4 | | 1.6 | 1.6 | | 0.2 | 0.2 | 0.7 |
| Adulto | 88M | 3.0 | | 3.0 | 3.2 | | 0.0 | 0.2 | 0.8 |
| | 98F | 4.3 | | | 3.4 | | | − 0.9 | 0.9 |
| | 101F | 5.2 | | | 5.2 | | | 0.0 | 1.0 |

* Calculada con UCON

** Calculada con UFORM

**TABLA 7.5.4**

**COMPARACIÓN DE LA PRECISIÓN DE LAS MEDIDAS CON LAS FORMAS SECUENCIALES Y LA PRECISIÓN DE LAS MEDIDAS EN EL BANCO DE LA PCK**

| | | | Formas secuenciales | | |
|---|---|---|---|---|---|
| Grupo de habilidad | Nombre de la persona | PCKB L=23 Error $\hat{\sigma}$ | Preescolar L=8 Error $s_1$ | Primaria L=15 Error $s_2$ | Adulto L=15 Error $s_3$ |
| | 3M | 1.2 | 1.1* | | |
| Preescolar | 6F | 0.8 | 0.8 | | |
| | 12M | 0.7 | 0.8 | 0.8 | |
| | 29M | 0.6 | 0.9 | 0.7 | 0.8 |
| Primaria | 35F | 0.6 | 1.2 | 0.7 | 0.8 |
| | 69M | 0.7 | | 0.8 | 0.8 |
| | 88M | 0.8 | | 1.1 | 0.8 |
| Adulto | 98F | 0.9 | | | 0.8 * |
| | 101F | 1.0 | | | 0.9 * |

* Las discrepancias del valor mínimo $\hat{\sigma}$ en la PCKB se deben a la aproximación de UFORM.

En la Tabla 7.6.2 damos el nombre y la habilidad en la PCKB de cada persona. A continuación damos su desempeño en la Forma Piloto. Esto incluye su puntaje piloto r, el puntaje relativo f, la habilidad $b_1$ y el error $s_1$. Después mostramos los intervalos alrededor del objetivo (de $b_1 - s_1$ a $b_1 + s_1$) relativo a cada desempeño en la Forma Piloto. A continuación se tiene la forma secuencial indicada y la habilidad $b_2$ calculada con base en su desempeño. Finalmente, mostramos la diferencia $(b_2 - \hat{\beta})$ entre la medida $\hat{\beta}$ en la PCKB y la medida $b_2$ en la Forma secuencial, junto con el error $s_{\hat{\beta}}$ en la PCKB.

Por ejemplo, la Persona 29M tiene una medida de habilidad de –0.9 en la PCKB. Si usamos la Forma Piloto encontramos que su habilidad es de +1.1 con un error estándar de 1.5, que indica un rango objetivo de –0.4 a 2.6. Dado que la Forma para Adultos está centrada en 1.8 lógitos, puede decirse que es la forma secuencial indicada para medir a 29M. Como resultado de ello, su medida en esta Forma es de –0.2 lógitos.

En la Tabla 7.6.2 vemos que en siete de las nueve personas la diferencia entre su medida en la PCKB y su medida en el rendimiento de la Forma secuencia a la medida difiere en menos de medio error estándar. Sin embargo, las personas 29 M y 98F muestran discrepancias del orden de un error estándar entre sus medidas en la PCKB y la Forma de prueba secuencial que son del orden de un error estándar.

**FIGURA 7.6.1**

**DISTRIBUCIÓN DE LOS ÍTEMS EN UNA FORMA PILOTO PARA UBICAR A LAS PERSONSAS EN LA VARIABLE DE LA PCK**

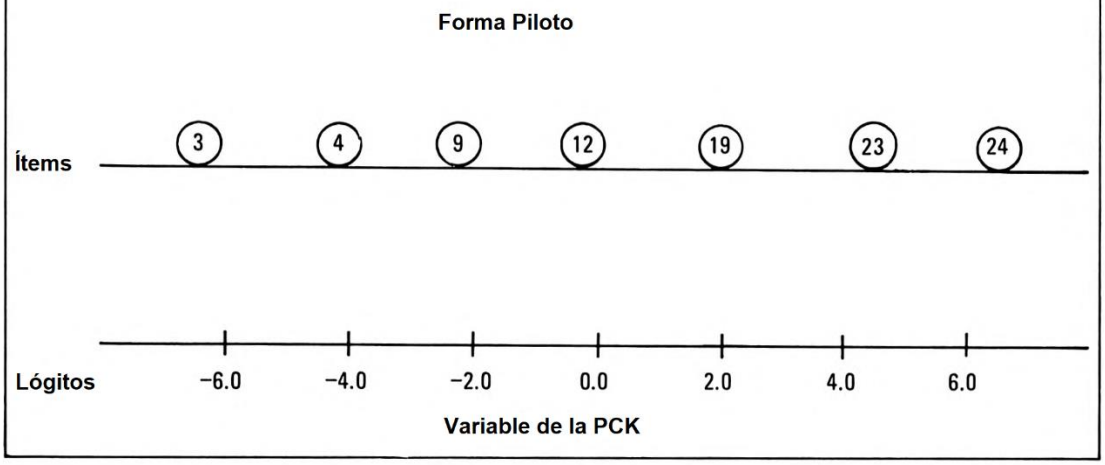



**TABLA 7.6.1**

**VALORES ESTADÍSTICOS DE UNA FORMA PILOTO PARA UBICAR A LAS PERSONSAS EN LA VARIABLE DE LA PCK**

**FORMA PILOTO**

| Nombre del ítem | Dificultad del ítem |
|---|---|
| 3 | −6.2 |
| 4 | −4.1 |
| 9 | −2.1 |
| 12 | −0.1 |
| 19 | 2.0 |
| 23 | 4.5 |
| 24 | 6.3 |

| | |
|---|---|
| **Altura:** | h = 0.0 |
| **Ancho:** | W = 15 |
| **Longitud:** | L = 7 |

**TABLA 7.6.2**

**DISTRIBUCIÓN DE PERSONAS CON UNA PRUEBA PILOTO PARA LA MEDICIÓN DE UNA FORMA SECUENCIAL APROPIADA**

| Grupo de habilidad | Nombre de la persona | Habilidad en la PCKB $\hat{\beta}$ | Forma Piloto: Puntaje r | Forma Piloto: Puntaje relativo f=r/7 | Forma Piloto: Estimado Habilidad $b_1$ | Forma Piloto: Estimado Error $s_1$ | Región objetivo implicada: Bajo $b_1 - s_1$ | Región objetivo implicada: Alto $b_1 + s_1$ | Forma Secuencial Indicada | Medida obtenida por la forma Secuencial indicada $b_2$ | Diferencia entre la PCKB y la Forma Secuencial Indicada $b_2 - \hat{\beta}$ | Error en la PCKB $s_{\hat{\beta}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3M | − 5.8 | 1 | .14 | − 5.5 | 1.6 | − 7.1 | − 3.9 | Preescolar (h=−3.3) | − 5.9 | − 0.1 | 1.2 |
| Preescolar | 6F | − 3.9 | 2 | .29 | − 3.2 | 1.5 | − 4.7 | − 1.7 | Preescolar (h=−3.3) | − 4.0 | − 0.1 | 0.8 |
| | 12M | − 2.8 | 1 | .14 | − 5.5 | 1.6 | − 7.1 | − 3.9 | Preescolar (h=−3.3) | − 2.5 | 0.3 | 0.7 |
| | 29M | − 0.9 | 4 | .57 | 1.1 | 1.5 | − 0.4 | 2.6 | Adulto (h=1.8) | − 0.2 | 0.7 | 0.6 |
| Primaria | 35F | − 0.5 | 4 | .57 | 1.1 | 1.5 | − 0.4 | 2.6 | Adulto (h=1.8) | − 0.2 | 0.3 | 0.6 |
| | 69M | 1.4 | 5 | .71 | 3.2 | 1.5 | 1.7 | 4.7 | Adulto (h=1.8) | 1.6 | 0.2 | 0.7 |
| | 88M | 3.0 | 5 | .71 | 3.2 | 1.5 | 1.7 | 4.7 | Adulto (h=1.8) | 3.2 | 0.2 | 0.8 |
| Adultos | 98F | 4.3 | 5 | .71 | 3.2 | 1.5 | 1.7 | 4.7 | Adulto (h=1.8) | 3.4 | − 0.9 | 0.9 |
| | 101F | 5.2 | 6 | .86 | 5.5 | 1.5 | 4.0 | 7.0 | Adulto (h=1.8) | 5.2 | 0.0 | 1.0 |

## 7.7 PRUEBA PERSONALIZADA

Para desarrollar un ejemplo de prueba personalizada, confeccionada a la medida volvemos a los registros originales de respuestas 1 y 0 de las personas para ver cómo podrían identificarse patrones de respuesta individualizados dentro de una matriz como la de la Tabla 2.3.1. Queremos seleccionar un registro para cada una de nuestras nueve personas verificando que se aproxima a una secuencia personalizada de ítems. Para obtener estos segmentos individualizados hemos establecido un nivel basal, a partir del punto en el que una persona en particular responde los ítems, de forma que tenga tres aciertos antes de la primera falla. Establecimos también un nivel de techo, que se tienen en el punto en que una persona podría dejar de contestar a los ítems teniendo tres fallas sucesivas. Esto produjo una secuencia personalizada única de ítems para cada una de nuestras nueve personas.

En la Tabla 7.7.1 mostramos los patrones de respuesta de estas pruebas personalizadas. Por ejemplo, el primer ítem que falló la Persona 29M es el ítem 6. Continuó contestando con aciertos y fallas en los siguientes ítems hasta que falló los ítems 17, 18 y 19 sucesivamente. Esto definió un segmento a la medida para esta persona, que va desde el Ítem 3 hasta el Ítem 19. En estos 17 ítems alcanzó un puntaje r = 10.

En la Tabla 7.7.2, calculamos una medida para cada persona en función de su segmento de prueba personalizada. Para hacer este cálculo determinamos las características de prueba h, w y L para cada segmento. Las características de prueba del segmento de prueba personalizada para cada persona se encuentran a la izquierda de la Tabla 7.7.2.

Por lo tanto, la Persona 29M tiene un puntaje relativo de 10 en su segmento personalizado de 17 ítems. Tomando en cuenta que su segmento tiene un ancho w = 8, este puntaje de 10 produce una medida de habilidad relativa de 0.7 lógitos, que cuando se ajusta para la altura de su segmento (h = –1.5) produce un estimado de habilidad de –0.8 con un error estándar de 0.7. Esta estimación de habilidad está a solo 0.1 lógitos de distancia de su estimación de habilidad en la PCKB de –0.9 con error de 0.6. Una inspección de las diferencias que figuran en la Tabla 7.7.2 entre las medidas de cada segmento a medida y sus correspondientes medidas en la PCKB muestra que todos los valores obtenidos por construcción de la prueba personalizada están cerca de las medidas de habilidad obtenidas por la PCKB.

En la Tabla 7.7.3 mostramos la eficiencia en el uso del ítem para cada una de nuestras nueve personas, para cada tipo de prueba personalizada. Vemos que se puede ahorrar una cantidad considerable de ítems sin sacrificar mucha precisión de las estimaciones de habilidad.

La Persona 29M con un segmento de 17 ítems personalizados a su medida requiere la mayoría de los ítems, incluso este segmento menor en 6 ítems que el total de 23 ítems de la PCKB y prácticamente no hay pérdida de precisión en su medición. La Persona 3M produce una estimación casi tan precisa con solo 4 ítems personalizados que la que es posible obtener usando los 23 completos. Esto ahorra 19 ítems. La persona 3M, sin embargo, se encuentra en el extremo inferior de la variable de la PCK. Como resultado de ello, solo los cuatro ítems más fáciles son relevantes para medir su habilidad. Si se tuvieran disponibles algunos ítems fáciles adicionales podríamos usarlos ventajosamente para mejorar la precisión de la medida de esta Persona 3M.

El procedimiento a medida siempre consigue la utilización más eficiente de los ítems. Esto es especialmente útil cuando se hacen medidas en los extremos de la escala, en este caso, más allá de ±4 lógitos en la escala de habilidad de la PCKB. Sin embargo, si bien apreciamos esta aparente eficiencia, también debemos darnos cuenta de que los ítems que estamos conservando pueden ser inapropiados para su objetivo. Nuestro verdadero objetivo es hacer mediciones suficientemente precisas para ser útil. La precisión depende del número de ítems utilizados que están lo suficientemente cerca de la persona a medir, con lo que cada ítem contribuye convenientemente a la medida estimada. Esto significa que requerimos que los ítems estén alrededor de un lógito de su objetivo. Una vez que los ítems se acercan a su objetivo, todas las consideraciones adicionales de precisión y, por lo tanto, de eficiencia, se reducen a la pregunta de cuántos de estos ítems "personalizados a la medida" resultan prácticos para que los conteste la persona.

## TABLA 7.7.1

## SECUENCIAS DE RESPUESTAS PERSONALIZADAS DE LA PCKB

| | | Nombre y dificultad del ítem (ordenados por dificultad) | | | | | | | | | | | | | | | | | | | | | | Segmento a medida | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #3 | #7 | #4 | #6 | #5 | #8 | #9 | #10 | #15 | #11 | #13 | #16 | #14 | #12 | #18 | #17 | #19 | #20 | #21 | #22 | #23 | #25 | #24 | | |
| Grupo de habilidad | Nombre de la persona | −6.2 | −4.3 | −4.1 | −2.7 | −2.6 | −2.6 | −2.1 | −2.1 | −1.5 | −1.0 | −0.9 | −0.8 | −0.5 | −0.1 | 1.4 | 1.9 | 2.0 | 2.9 | 3.3 | 3.3 | 4.5 | 5.8 | 6.3 | Longitud L | Puntuación r |
| Preescolar | 3M | 1 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | | | | 4 | 1 |
| | 6F | 1 | 1 | 1 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | | 6 | 3 |
| | 12M | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | | | | | | | | | | | | | | | 9 | 5 |
| Primaria | 29M | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | | | | | | 17 | 10 |
| | 35F | | | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | | | | | | 15 | 9 |
| | 69M | | | | | | | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | | 15 | 9 |
| Adulto | 88M | | | | | | | | | | | | | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 11 | 6 |
| | 98F | | | | | | | | | | | | | | | | | | 1 | 1 | 1 | 0 | 0 | 0 | 6 | 3 |
| | 101F | | | | | | | | | | | | | | | | | | | 1 | 1 | 1 | 0 | 0 | 5 | 3 |

Las respuestas que no se presentan son todas “1” a la izquierda o todas “0” a la derecha.

**TABLA 7.7.2**

## MEDIDAS A PARTIR DE SECUENCIAS DE RESPUESTA PERSONALIZADAS

| | | | Características de las pruebas para los segmentos personalizados | | | Medición con prueba personalizadas | | | | | | Diferencia entre PCKB y prueba personalizada | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grupo de habilidad | Nombre de la persona | Habilidad en la PCKB $\hat{\beta}$ | Altura h | Ancho w | Longitud L | Puntaje r | Puntaje Relativo f = r/L | Habilidad Relativa $x_{fw}$ | Coeficiente de error $C^{1/2}_{fw}$ | Habilidad $b=h+x_{fw}$ | Error $s = C^{1/2}/L^{1/2}$ | Diferencia de Habilidad $b-\hat{\beta}$ | Error en la PCKB $s_{\hat{\beta}}$ |
| | 3M | −5.8 | −4.3 | 4 | 4 | 1 | .25 | −1.4 | 2.6 | −5.7 | 1.3 | 0.1 | 1.2 |
| Preescolar | 6F | −3.9 | −3.8 | 4 | 6 | 3 | .50 | 0.0 | 2.3 | −3.8 | 0.9 | 0.1 | 0.8 |
| | 12M | −2.8 | −3.1 | 4 | 9 | 5 | .56 | 0.3 | 2.3 | −2.8 | 0.8 | 0.0 | 0.7 |
| | 29M | −0.9 | −1.5 | 8 | 17 | 10 | .59 | 0.7 | 2.9 | −0.8 | 0.7 | 0.1 | 0.6 |
| Primaria | 35F | −0.5 | −1.0 | 6 | 15 | 9 | .60 | 0.7 | 2.6 | −0.3 | 0.7 | 0.2 | 0.6 |
| | 69M | 1.9 | −0.7 | 7 | 15 | 9 | .60 | 0.7 | 2.7 | 1.4 | 0.7 | 0.0 | 0.7 |
| | 88M | 3.0 | 2.8 | 8 | 11 | 6 | .55 | 0.4 | 2.9 | 3.2 | 0.9 | 0.2 | 0.8 |
| Adultos | 98F | 4.3 | 4.4 | 4 | 6 | 3 | .50 | 0.0 | 2.3 | 4.4 | 0.9 | 0.1 | 0.9 |
| | 101F | 5.2 | 4.6 | 4 | 5 | 3 | .60 | 0.6 | 2.5 | 5.2 | 1.1 | 0.0 | 1.0 |

$$h = \sum_{i}^{L} d_i/L, \qquad w = [\,(d_L + d_{L-1} - d_2 - d_1)/2\,]\,[\,L/(L-2)\,] \text{ redondeado al entero más cercano}$$

## TABLA 7.7.3

### EFICIENCIA DE MEDIDAS POSIBLES CON TRES TIPOS DE PRUEBAS CONFECCIONADAS A LA MEDIDA

| | | | | Ítems utilizados | | | Ítems guardados | | | Precisión de la medida | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grupo de habilidad | Nombre de la persona | Habilidad en la PCKB $\hat{\beta}$ | PCKB L | Estatus del diseño a la medida $L_1$ | Desempeño de diseño a la media $L_2$ | Personalizada $L_3$ | Estatus del diseño a la medida $L - L_1$ | Desempeño de diseño a la media $L - L_2$ | Personalizada $L - L_3$ | PCKB $\hat{\sigma}$ | Estatus $s_1$ | Desempeño $s_2$ | Auto-confección $s_3$ |
| Preescolar | 3M | −5.8 | 23 | 8 | 12 | 4 | 15 | 11 | 19 | 1.2 | 1.1* | 1.1* | 1.3 |
| | 6F | −3.9 | 23 | 8 | 12 | 6 | 15 | 11 | 17 | 0.8 | 0.8 | 0.8 | 0.9 |
| | 12M | −2.8 | 23 | 8 | 12 | 9 | 15 | 11 | 14 | 0.7 | 0.8 | 0.8 | 0.8 |
| Primaria | 29M | −0.9 | 23 | 15 | 18 | 17 | 8 | 5 | 6 | 0.7 | 0.7 | 0.8* | 0.7 |
| | 35F | −0.5 | 23 | 15 | 18 | 15 | 8 | 5 | 8 | 0.7 | 0.7 | 0.8* | 0.7 |
| | 69M | 1.4 | 23 | 15 | 18 | 15 | 8 | 5 | 8 | 0.7 | 0.8 | 0.8 | 0.7 |
| Adultos | 88M | 3.0 | 23 | 15 | 18 | 11 | 8 | 5 | 12 | 0.7 | 0.8 | 0.8 | 0.9 |
| | 98F | 4.3 | 23 | 15 | 18 | 6 | 8 | 5 | 17 | 0.8 | 0.8 | 0.8 | 0.8 |
| | 101F | 5.2 | 23 | 15 | 18 | 5 | 8 | 5 | 18 | 0.9 | 0.9 | 0.9 | 1.1 |

* Las discrepancias con la PCKB se deben por lo menos a la aproximación UFORM

## 7.8 AJUSTE DE LA PERSONA Y CONTROL DE CALIDAD

Durante la administración de la prueba, puede parecer que un examinado ha respondido a la prueba de acuerdo con la forma planeada. Sin embargo, siempre es necesario examinar el patrón real de respuestas para ver si corresponde de hecho con expectativas razonables.

Consideremos, por ejemplo, una prueba de 10 ítems administrados en orden creciente de dificultad. La Tabla 7.8.1 muestra cinco maneras diferentes en que se puede lograr un puntaje de cinco en tal prueba. Los patrones de respuesta de las Personas A y B, e incluso la C, parecen razonables. Los aciertos ocurren en los ítems más fáciles de la izquierda y las fallas se presentan en los ítems más difíciles ubicados a la derecha. Sin embargo, los patrones de las Personas D y E son bastante inverosímiles. ¿Cómo puede suceder que la Persona D obtiene un puntaje de cinco al responder acertadamente los cinco ítems más difíciles mientras al mismo tiempo falla en los cinco ítems más fáciles? Eso es tan contradictorio con nuestras expectativas para interpretar un registro de prueba que no podemos tomar el puntaje de cinco, obtenido por la persona D, como base para una medida válida de su habilidad. La persona D puede ser inteligente y descuidada o tonta y afortunada, pero una cosa es cierta, la Persona D no tiene la habilidad intermedia que implica obtener un puntaje de cinco.

El registro de respuestas de la Persona E también nos genera preguntas. Si la persona E pudiera responder correctamente los ítems en el rango medio de dificultad, incluyendo también cuatro o cinco de los ítems más difíciles, ¿cómo es posible que falle en los tres ítems más fáciles?

El modelo de medición de Rasch conduce a un procedimiento integral pero de fácil aplicación para evaluar la validez del registro de respuestas de cada examinado. Este procedimiento compara el registro de respuesta de la persona contra nuestra expectativa de lo que debería suceder de acuerdo con el modelo de respuesta. El procedimiento usa esta comparación para calcular un estadístico de "ajuste" que indica el grado en que el desempeño de la persona en la prueba está de acuerdo con las expectativas del modelo.

Si $x_{vi}$ es la respuesta de la persona v con la medida tentativa $b_v$ al ítem i que tiene una calibración $d_i$ en el banco y si $x_{vi} = 0$ para una respuesta incorrecta o $x_{vi} = 1$ para una respuesta correcta, entonces, de acuerdo con nuestro modelo de medición:

$$z_{vi}^2 = \exp\left[(2x_{vi} - 1)(d_i - b_v)\right] \qquad \textbf{[7.8.1]}$$

es un residuo cuadrático estandarizado para evaluar la relación entre la respuesta observada $x_{vi}$ y sus expectativas de acuerdo con el modelo, dados $b_v$ y $d_i$. Según las expectativas este $z_{vi}^2$ debería distribuirse aproximadamente como ji-cuadrado con cerca de (L – 1) / L grados de libertad, donde L es el número de ítems en la prueba utilizada para estimar $b_v$. Si el conjunto de $\{z_{vi}^2\}$ se distribuye de esta manera, entonces no tenemos ninguna razón interna para invalidar $b_v$. En caso contrario, debemos reconocer una desviación en los datos de nuestras expectativas y debemos ver qué podemos hacer al respecto.

Cada respuesta $x_{vi}$ en el conjunto de i = 1 a L proporcionada por la persona v produce su propio valor prácticamente independiente $z_{vi}^2$. Podemos sumar este conjunto de L residuos $\{z_{vi}^2\}$ en un estadístico ji-cuadrado con aproximadamente (L – 1) grados de libertad, y por conveniencia expresamos este estadístico estandarizado ji-cuadrado de esta forma:

$$t_v = \left[\ell n(v_v) + (v_v - 1)\right]\left[(L-1)/8\right]^{1/2} \qquad \sim N(0,1) \qquad \textbf{[7.8.2]}$$

**TABLA 7.8.1**

**CINCO MANERAS DE OBTENER CINCO EN UNA PRUEBA DE CINCO ÍTEMS**

| | Ítems en orden creciente de dificultad | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ítem más fácil | | | | | | | | | Ítem más difícil | |
| Persona | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Puntaje |
| A | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 5 |
| B | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 5 |
| C | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 5 |
| D | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| E | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 5 |

Donde $v_v$ es la media cuadrática:

$$v_v = \sum_{i}^{L} z_{vi}^2 / (L - 1) \qquad \textbf{[7.8.3]}$$

El divisor 8 que aparece en la Ecuación 7.8.2 proviene de promediar dos valores estandarizados opuestos de la media cuadrática v. Entonces, si:

$$t_1 = (v - 1)[(L - 1)/2]^{1/2} \qquad \sim N(0,1)$$

y

$$t_2 = [\ell n(v)][(L - 1)/2]^{1/2} \qquad \sim N(0,1)$$

entonces:

$$t = (t_1 + t_2)/2 = [\ell n(v) + (v - 1)][(L - 1)/8]^{1/2} \qquad \sim N(0,1)$$

En la Tabla 7.8.2 hemos realizado el análisis de ajuste para los patrones de respuestas de las personas 12M, 35F y 88M. La persona 12M tiene una medida preliminar o tentativa de b = –2.8. Para su primer ítem, d= –6.2, su respuesta es x =0. Esto le genera una diferencia (d – b) de

$$(d - b) = [-6.2 - (-2.8)] = -3.4$$

Dado que $(2x - 1) = -1$

Se tiene que $z^2 = \exp[(2x - 1)(d - b)] = \exp[-(-3.4)] = \exp[(3.4)] = 30$

## TABLA 7.8.2
## CÁLCULO DEL AJUSTE DE LAS PERSONAS

| Nombre de la persona | Habilidad tentativa* | Estadístico de respuesta | Patrones de respuesta de las personas ** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | b | d | −6.2 | −4.3 | −4.1 | −2.7 | −2.6 | −2.6 | −2.1 | −2.1 | −1.5 | | | | | | |
| 12M | −2.8 | x | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | | | | | | |
| | | (d−b) | −3.4 | −1.5 | −1.3 | 0.1 | 0.2 | 0.2 | 0.7 | 0.7 | 1.3 | | | | | | |
| | | $z^2$ | 30.0 | 0.2 | 0.3 | 1.1 | 1.2 | 1.2 | 0.5 | 0.5 | 0.3 | | | | | | |
| | | d | −4.1 | −2.7 | −2.6 | −2.6 | −2.1 | −2.1 | −1.5 | −1.0 | −0.9 | −0.8 | −0.5 | −0.1 | 1.4 | 1.9 | 2.0 |
| 35F | −0.3 | x | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| | | (d−b) | −3.8 | −2.4 | −2.3 | −2.3 | −1.8 | −1.8 | −1.2 | −0.7 | −0.6 | −0.5 | −0.2 | 0.2 | 1.7 | 2.2 | 2.3 |
| | | $z^2$ | 0.0 | 0.1 | 0.1 | 10.0 | 0.2 | 0.2 | 0.3 | 2.0 | 0.5 | 0.6 | 1.2 | 1.2 | 0.2 | 0.1 | 0.1 |
| | | d | −0.5 | −0.1 | 1.4 | 1.9 | 2.0 | 2.9 | 3.3 | 3.3 | 4.5 | 5.8 | 6.3 | | | | |
| 88M | 3.2 | x | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | | | |
| | | (d−b) | −3.7 | −3.3 | −1.8 | −1.3 | −1.2 | −0.3 | 0.1 | 0.1 | 1.3 | 2.6 | 3.1 | | | | |
| | | $z^2$ | 0.0 | 0.0 | 0.2 | 3.7 | 0.3 | 0.7 | 0.9 | 1.1 | 0.3 | 0.1 | 0.0 | | | | |

* Las habilidades son las medidas de los segmentos personalizados incluidos en la columna 10 de la Tabla 7.7.2

** Los patrones de respuesta se encuentran en la Tabla 7.7.1

$$z^2 = \exp\,[(2x - 1)(d - b)]$$

**TABLA 7.8.3**

**CÁLCULO DEL AJUSTE DE LA PERSONA: ANÁLISIS RESIDUAL**

| Nombre de la persona | Habilidad tentativa b | Suma de cuadrados $z^2$ | Grados de libertad (L–1) | Media cuadrática v | Estadístico de ajuste t |
|---|---|---|---|---|---|
| 12M | – 2.8 | 35.0 | 8 | 4.4 | 4.9* |
| 35F | – 0.3 | 16.8 | 14 | 1.2 | 0.5 |
| 88M | 3.2 | 7.3 | 10 | 0.7 | –0.7 |

* Indica desajuste

$$z^2 = \exp[(2x-1)(d-b)]$$

$$v = \Sigma z^2/(L-1)$$

$$t = [\ell n(v) + (v-1)]\,[(L-1)/8]^{1/2}$$

Para cualquier otra respuesta en el segmento personalizado de 9 ítems de la Persona 12M, hemos dado su x, (d – b) y $z^2$. El análisis residual basado en esta fila de $z^2$ para la Persona 12M conduce a

$$\sum_{i}^{9} z_i^2 = 35,$$

que, para 8 grados de libertad, produce una media cuadrática de v = 4.4 y una desviación aproximadamente normal t = 4.9. Los análisis residuales para estas tres personas se resumen en la Tabla 7.8.3.

Observe que en la Tabla 7.8.2 hemos usado (d – b) en lugar del (b – d) que se utilizó en el Capítulo 4. Esto se debe a que la forma (d – b) es conveniente para el cálculo de $z^2$. Cada vez que una respuesta es 0, se debe agregar un signo menos a la diferencia (d – b) que lo convierte en (b – d). Sin embargo, si tenemos en cuenta este cambio de signo, podemos usar la Tabla 4.3.3 para determinar los valores en la Tabla 7.8.2. Sin embargo, si se usa la Tabla 4.3.3, encontraremos que los valores en la Tabla 7.8.2 son ligeramente más exactos que los valores determinados en la Tabla 4.3.3. La diferencia es más grande en las respuestas que se ajustan bien, pero estas respuestas son las que intervienen menos en el análisis del desajuste. La suma de cuadrados $\Sigma z^2$ de 12M con base en la Tabla 4.3.3 debería ser 33 en lugar de 35 que aparece en la Tabla 7.8.3. El valor de t resultante sería 4.5 en lugar de 4.9.

El estadístico de ajuste t se distribuye más o menos normalmente pero con colas más largas. En nuestra experiencia práctica, el nivel de rechazo de 2 aceptado generalmente es innecesariamente conservador. Las pautas generales que usamos actualmente para interpretar t como un indicador de desajuste son:

Si t <3 aceptamos la medición de la persona como probablemente válida.

Si 3 <t <5 hacemos un examen cuidadoso del patrón de respuesta para identificar y considerar posibles fuentes de desajuste.

Si t> 5 rechazamos la medida tal como está y realizamos los pasos que podamos para extraer una medida "corregida" de un segmento aceptable del registro de respuesta, en caso de que exista.

El estudio detallado de personas que desajustan al modelo depende, por supuesto, de un estudio detallado de las desviaciones aproximadamente normales en los registros de respuestas:

$$z_{vi} = (2\,x_{vi} - 1)\{\exp[(2\,x_{vi} - 1)(d_i - b_v)/2]\}$$

con ese estudio se rastrean las posibles fuentes de irregularidad.

Dado que las partes de $\Sigma\, z^2$ que más contribuyen en el valor de t son los términos positivos más grandes, podemos simplificar la determinación de la validez de registro con ayuda de un estadístico que se calcule rápidamente enfocándonos en las respuestas más inesperadas. La Tabla 4.3.3 (también dada en la Tabla C del Apéndice) muestra que la diferencia entre la medida b de la persona y la dificultad d del ítem debe ser del orden de ±2.0 antes de que $z^2$ se haga mayor a 7 o que su probabilidad sea menor que 0.12. Para poder alcanzar una probabilidad de 0.05 menos en una respuesta dada, debemos relajar nuestro estándar para que la diferencia (d – b) sea de ±3 produciendo una $z^2$ de 20.

Si ahora nos concentramos en ver las respuestas inesperadas para las cuales $|d - b| > 3$, entonces los valores de $z^2$ pueden consultarse en la Tabla 4.3.3 (o en la Tabla C del Apéndice) y combinarse con un valor promedio de 1 para todos los ítems restantes en el segmento de respuesta, con lo que se produce un valor burdo (o aproximado) de $\Sigma\, z^2$ para el cual, a su vez, puede calcularse un valor burdo (o aproximado) del estadístico t.

Por ejemplo, sobre las 9 respuestas de la Persona 12M, solo hay una sorpresa que ocurre donde $(d - b) = -3.4$ y $z^2 = 30$. Combinando este valor de 30 con ocho veces 1 para los ocho ítems restantes de la prueba, nos dan un valor burdo de $\Sigma\, z^2 = 38$ y un valor burdo de $t = 5.3$. Este valor para t no está lejos de 4.9, que es el valor más exacto que calculamos en la Tabla 7.8.2 y que nos lleva a la misma conclusión de desajuste significativo.

En la Tabla 7.8.4 resumimos el análisis residual para las nueve personas. Presentamos la medida de la habilidad y el error estándar de cada persona en su segmento de ítems confeccionado a la medida. A continuación damos la suma de cuadrados, grados de libertad, medias cuadráticas y estadísticos de ajuste para el registro de cada persona. En ocho casos no encontramos evidencia de desajuste, por lo que aceptamos que sus medidas son plausibles. Solo el segmento personalizado a la medida de la Persona 12M muestra un desajuste significativo. Como vimos en la Tabla 7.8.2, este desajuste se debe completamente a que tiene una respuesta incorrecta en el primer ítem que es el más fácil de este registro. El origen de esta respuesta incorrecta podría ser una falla al responder la prueba o un error de funcionamiento personal. En cualquier caso, todavía estamos interesados en determinar la mejor estimación posible de la habilidad de la Persona 12M. El problema de extraer la mejor medida posible con base en un registro defectuoso se analizará en la Sección 7.10.

**TABLA 7.8.4**

**AJUSTE DEL ANÁLISIS DE LAS NUEVE PERSONAS MEDIDO MEDIANTE PRUEBAS PERSONALIZADAS**

| Grupo de habilidad | Nombre de la persona | Medida | | Análisis Residual | | | |
|---|---|---|---|---|---|---|---|
| | | Habilidad b | Error s | Suma de cuadrados $\Sigma z^2$ | Grados de libertad (L–1) | Media cuadrática v | Estadístico de ajuste t |
| Preescolar | 3M | −5.7 | 1.3 | 1.1 | 3 | 0.4 | −1.0 |
| | 6F | −3.8 | 0.9 | 2.4 | 5 | 0.5 | −1.0 |
| | 12M | −2.8 | 0.8 | 35.0 | 8 | 4.4 | 4.9 * |
| Primaria | 29M | −0.8 | 0.7 | 17.6 | 16 | 1.1 | 0.3 |
| | 35F | −0.3 | 0.7 | 16.8 | 14 | 1.2 | 0.5 |
| | 69M | 1.4 | 0.6 | 20.0 | 14 | 1.4 | 1.0 |
| Adultos | 88M | 3.2 | 0.9 | 7.3 | 10 | 0.7 | −0.6 |
| | 98F | 4.4 | 0.9 | 2.1 | 5 | 0.4 | −1.1 |
| | 101F | 5.2 | 1.1 | 1.7 | 4 | 0.4 | −1.0 |

* Indica desajuste

## 7.9 DIAGNOSTICO DEL DESAJUSTE

Consideremos nuevamente la prueba de 10 ítems ordenados por dificultad creciente como se concibió en la Tabla 7.8.1. Si encontráramos que el patrón producido por la Persona E fuera el siguiente:

| | Puntaje |
|---|---|
| 0 0 0 1 0 1 1 0 1 1 | 5 |

estaríamos desconcertados y nos preguntaríamos cómo pudo responder correctamente esta persona las preguntas difíciles fallando en tres primeras preguntas más fáciles. ¿Será que estaba "durmiendo" en la parte más fácil de la prueba? Sería un patrón de “dormilón”[10].

Por otro lado, si encontráramos el patrón de respuesta

| | Puntaje |
|---|---|
| 1 0 1 0 0 0 0 1 1 1 | 5 |

nuestra sorpresa sería tan grande como la anterior, pero ahora podríamos inclinarnos a explicar la irregularidad como el resultado de tener la buena suerte de "adivinar" la respuesta correcta en los tres ítems más difíciles.

Tanto la naturaleza probabilística del modelo como nuestra experiencia cotidiana con los patrones típicos de respuesta, nos llevan a esperar patrones mixtos con una región central de respuestas correctas e incorrectas. Si ahora nos encontramos con un patrón como este:

| | Puntaje |
|---|---|
| 1 1 1 1 1 0 0 0 0 0 | 5 |

[10] Sleeping se tradujo como “dormilón”.

nos parecería "demasiado bueno para ser verdad". Este patrón inesperadamente regular es a veces producido por personas que trabajan muy lentamente y con sumo cuidado, de tal modo que se niegan a pasar al siguiente ítem hasta que han hecho todo lo posible por responder correctamente al presente ítem. Nos referiremos a este patrón como "cauteloso"[11].

Finalmente, también identificamos una variante especial del "dormilón" que podría ser mejor llamado "descuidado"[12], en el que las respuestas incorrectas se agrupan al comienzo de la prueba, sugiriendo que la persona tuvo problemas al inicio de la prueba.

Para resumir, identificamos los siguientes tipos de patrones de respuesta:

| | | Puntaje |
|---|---|---|
| "normal" | 1 1 1 1 0 1 0 0 0 0 | 5 |
| | 1 1 1 0 1 0 1 0 0 0 | 5 |
| "dormilón" o "descuidado" | 0 0 0 1 0 1 1 0 1 1 | 5 |
| "adivinador" | 1 0 1 0 0 0 0 1 1 1 | 5 |
| "cauteloso" | 1 1 1 1 1 0 0 0 0 0 | 5 |

En la Sección 7.8 identificamos un patrón de respuesta de desajuste para la Persona 12M. Ahora investigaremos los registros que desajustan, como el de la Persona 12M, para ver la manera de diagnosticar patrones de respuesta irregulares. El patrón de respuesta personalizado para la persona 12M, con ítems en orden creciente de dificultad, es

| | Puntaje |
|---|---|
| 0 1 1 1 1 1 0 0 0 | 5 |

La evaluación de este patrón de respuesta en la Tabla 7.8.3 muestra un desajuste significativo, t = 4.9. En la Tabla 7.9.1 mostramos nuevamente el patrón de respuesta para la Persona 12M y agregamos para cada respuesta la probabilidad p de ocurrencia de acuerdo con el modelo. También damos su patrón de respuesta en términos de z además de $z^2$. Cuando presentamos en forma gráfica los valores de z para la Persona 12M en la Figura 7.9.1, vemos qué se puede parecer a un patrón de respuesta "dormilón" o "descuidado". Esta figura muestra el segmento de ítems respondidos por la persona. Cada ítem está espaciado horizontalmente a lo largo la variable PCK según su dificultad en la escala de lógitos. Su posición vertical está determinada por el residuo estandarizado z de la persona al responder ese ítem.

El patrón de respuesta de la Persona 12M que se observa en la Figura 7.9.1 muestra cómo el estadístico z indica desajuste. El ítem 3 tiene una z = –5.5 mientras que los otros ítems tienen z cerca de su valor esperado de cero. El efecto del ítem 3 sobre el patrón de respuesta de la persona 12M se puede resaltar al considerar los dos patrones alternativos dados en la Tabla 7.9.1 y en la Figura 7.9.1.

En la misma figura, el patrón alternativo A conduce a un puntaje de cinco al intercambiar la respuesta correcta de "1" en el Ítem 8, que es un ítem relativamente difícil, por la respuesta incorrecta de "0" en el ítem 3, que es el ítem más fácil de la prueba. Ahora tenemos el patrón:

| | Puntaje |
|---|---|
| 1 1 1 1 1 0 0 0 0 | 5 |

[11] Plodding se tradujo como "cauteloso", de preferencia a "lento".
[12] Fumbling se tradujo como "descuidado" de preferencia a "torpe" que parece peyorativo.

**TABLA 7.9.1**

**DIAGNOSTICANDO AL "DORMILÓN"**

| Descripción del caso | Estadístico de respuesta | Nombres y dificultades de los ítems (en orden creciente de dificultad) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | #3 | #7 | #4 | #6 | #5 | #8 | #9 | #10 | #15 |
| | | −6.2 | −4.3 | −4.1 | −2.7 | −2.6 | −2.6 | −2.1 | −2.1 | −1.5 |
| | | Patrones de respuesta | | | | | | | | |
| | x | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| Persona 12M | (2x−1)(d−b) | 3.4 | −1.5 | −1.3 | 0.1 | 0.2 | 0.2 | −0.7 | −0.7 | −1.3 |
| (b = − 2.8) | $z^2$ | 30.0 | 0.2 | 0.3 | 1.1 | 1.2 | 1.2 | 0.5 | 0.5 | 0.3 |
| | p | .03 | .83 | .77 | .48 | .45 | .45 | .67 | .67 | .77 |
| | z | −5.5 | 0.5 | 0.5 | 1.1 | 1.1 | 1.1 | −0.7 | −0.7 | −0.5 |
| | x | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Alternativa A | (2x−1)(d−b) | −3.4 | −1.5 | −1.3 | 0.1 | 0.2 | −0.2 | −0.7 | −0.7 | −1.3 |
| | $z^2$ | 0.0 | 0.2 | 0.3 | 1.1 | 1.2 | 0.8 | 0.5 | 0.5 | 0.3 |
| | p | .99 | .83 | .77 | .48 | .45 | .56 | .67 | .67 | .77 |
| | z | 0.2 | 0.5 | 0.5 | 1.1 | 1.1 | −0.9 | −0.7 | −0.7 | −0.5 |
| | x | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| Alternativa B | (2x−1)(d−b) | −3.4 | −1.5 | −1.3 | 0.1 | −0.2 | −0.2 | −0.7 | −0.7 | 1.5 |
| | $z^2$ | 0.0 | 0.2 | 0.3 | 1.1 | 0.8 | 0.8 | 0.5 | 0.5 | 4.5 |
| | p | .99 | .83 | .77 | .48 | .56 | .56 | .67 | .67 | .18 |
| | z | 0.2 | 0.5 | 0.5 | 1.1 | −0.9 | −0.9 | −0.7 | −0.7 | 2.1 |

$z^2 = \exp[(2x-1)(d-b)]$ $\quad p = 1/(1+z^2)$ $\quad z = (2x-1)\exp[(2x-1)(d-b)/2]$

**FIGURA 7.9.1**
**DIAGNOSTICANDO AL "DORMILÓN"**

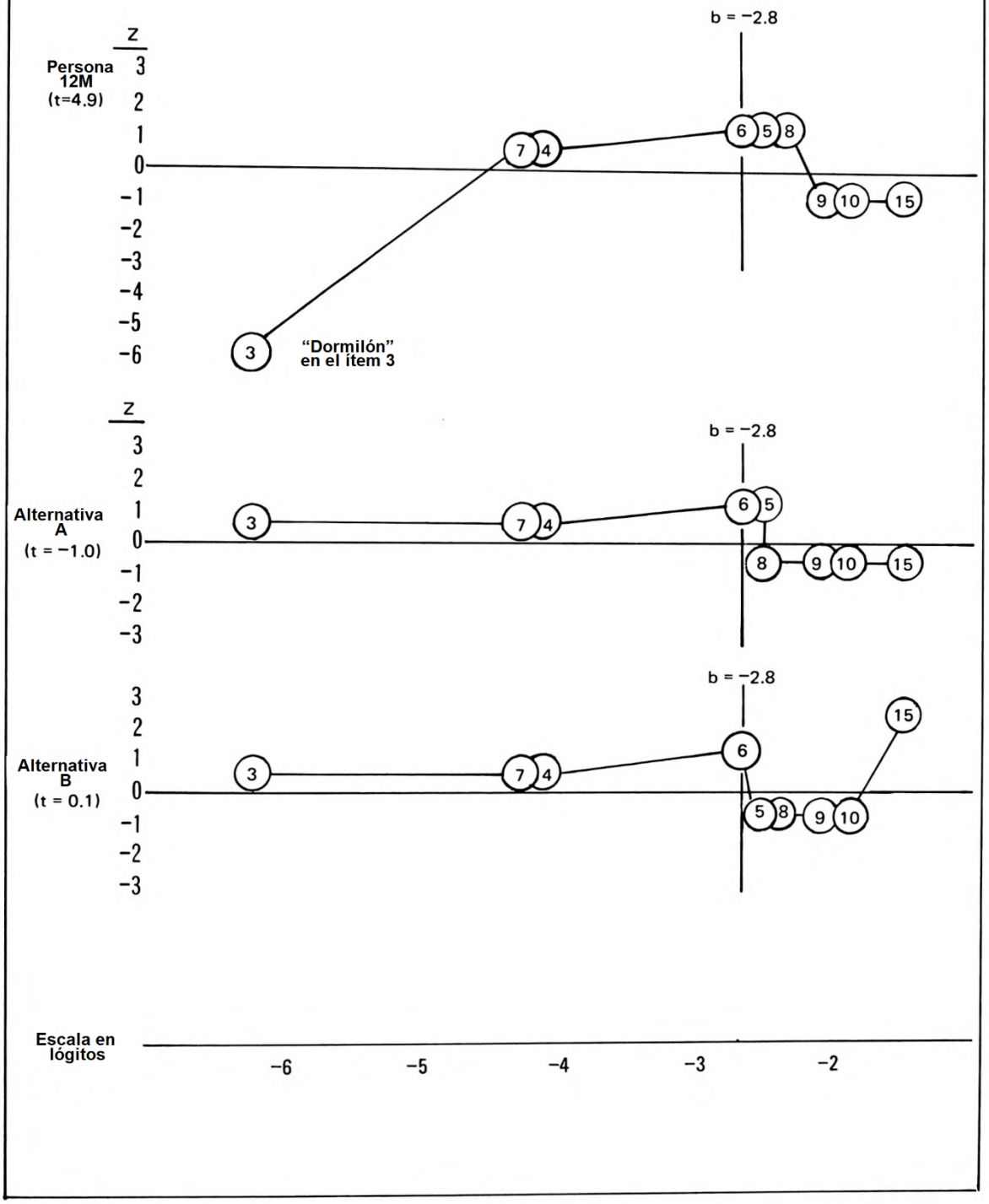


**TABLA 7.9.2**

**ANÁLISIS RESIDUAL PARA EL PATRÓN "DORMILÓN" DE LA PERSONA 12M**

| Descripción del caso | Suma de Cuadrados $\Sigma z^2$ | Media cuadrática v | Estadístico de ajuste t |
|---|---|---|---|
| Persona 12M (b=-2.8) | 35.0 | 4.4 | 4.9 * |
| Alternativa A | 4.6 | 0.6 | −1.0 |
| Alternativa B | 8.5 | 1.1 | 0.1 |

* Indica desajuste

$$z^2 = \exp[(2x - 1)(d - b)]$$

$$v = \Sigma z^2/(L - 1)$$

$$t = [\ell n\,(v) + (v - 1)]\,[L - 1)/8]^{1/2}$$

Los estadísticos de desajuste para estos tres patrones se resumen en la Tabla 7.9.2. Ahí vemos que la Alternativa A tiene una t = −1.0 en lugar de t = 4.9 para la persona de 12M.

En el patrón alternativo B, intercambiamos la respuesta correcta de "1" al ítem 5 con la respuesta incorrecta de "0" al ítem 15, que es el ítem más difícil en el segmento. Esto produce el patrón de respuesta alternativa

| | Puntaje |
|---|---|
| 1 1 1 1 0 0 0 0 1 | 5 |

Curiosamente, el desajuste debido al intercambio en el patrón B es pequeño, tan solo t = 0.1. Esto es debido a que el ítem 15, con dificultad d = −1.5, no es tan difícil en relación con la habilidad b = −2.8 de la persona 12M porque el Ítem 3 es demasiado fácil, con dificultad −6.2.

En las Tablas 7.9.3 y 7.9.4 y en la figura 7.9.2 ilustramos los patrones de respuesta del "dormilón" y del "adivinador" usando el registro observado de la Persona 88M. Para cambiar su patrón de respuesta a un patrón dormilón tomamos dos de los ítems fáciles y reemplazamos sus respuestas correctas por incorrectas y pasamos estas dos respuestas correctas a los ítems 17 y 21, manteniendo constante su puntaje r = 6. Ahora tenemos el patrón de respuesta

| | Puntaje |
|---|---|
| 0 0 1 1 1 1 1 1 0 0 0 | 6 |

que es característico del dormilón. Este patrón alcanza t = 9.1 en la Tabla 7.9.4.

## TABLA 7.9.3

## PATRONES DE RESPUESTA "DORMILÓN" Y "ADIVINADOR"

| | | Nombres y dificultades de los ítems (en orden creciente de dificultad) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #14 | #12 | #18 | #17 | #19 | #20 | #21 | #22 | #23 | #25 | #24 |
| | | −0.5 | −0.1 | 1.4 | 1.9 | 2.0 | 2.9 | 3.3 | 3.3 | 4.5 | 5.8 | 6.3 |
| **Descripción del caso** | **Estadístico de respuesta** | **Patrones de respuesta** | | | | | | | | | | |
| | x | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| **Persona 88M (b=3.2)** | (2x−1)(d−b) | −3.7 | −3.3 | −1.8 | 1.3 | −1.2 | −0.3 | −0.1 | 0.1 | −1.3 | −2.6 | −3.1 |
| | $z^2$ | 0.0 | 0.0 | 0.2 | 3.7 | 0.3 | 0.7 | 0.9 | 1.1 | 0.3 | 0.1 | 0.0 |
| | p | .98 | .96 | .86 | .21 | .77 | .57 | .52 | .48 | .79 | .93 | .96 |
| | z | 0.2 | 0.2 | 0.4 | −1.9 | 0.5 | 0.9 | −1.0 | 1.1 | −0.5 | −0.3 | −0.2 |
| **Patrón "dormilón"** | x | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| | (2x−1)(d−b) | 3.7 | 3.3 | −1.8 | −1.3 | −1.2 | −0.3 | 0.1 | 0.1 | −1.3 | −2.6 | −3.1 |
| | $z^2$ | 40.4 | 27.1 | 0.2 | 0.3 | 0.3 | 0.7 | 1.1 | 1.1 | 0.3 | 0.1 | 0.0 |
| | p | .02 | .04 | .86 | .79 | .77 | .57 | .48 | .48 | .79 | .93 | .96 |
| | z | −6.4 | −5.2 | 0.4 | 0.5 | 0.5 | 0.9 | 1.1 | 1.1 | −0.5 | −0.3 | −0.2 |
| **Patrón "adivinador"** | x | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | (2x−1)(d−b) | −3.7 | −3.3 | −1.8 | −1.3 | 1.2 | 0.3 | −0.1 | −0.1 | −1.3 | 2.6 | 3.1 |
| | $z^2$ | 0.0 | 0.0 | 0.2 | 0.3 | 3.3 | 1.4 | 0.9 | 0.9 | 0.3 | 13.5 | 22.2 |
| | p | .98 | .96 | .86 | .79 | .23 | .43 | .52 | .52 | .79 | .07 | .04 |
| | z | 0.2 | 0.2 | 0.4 | 0.5 | −1.8 | −1.2 | −1.0 | −1.0 | −0.5 | 3.7 | 4.7 |

$z^2 = \exp[(2x - 1)(d - b)]$ $\quad p = 1/(1 + z^2)$ $\quad z = (2x - 1)\exp[(2x - 1)(d - b)/2]$

**TABLA 7.9.4**
**ANÁLISIS RESIDUAL PARA PATRONES DE RESPUESTA "DORMILÓN" Y "ADIVINADOR"**

| Descripción del caso | Suma de cuadrados $\Sigma z^2$ | Media cuadrática v | Estadístico de ajuste t |
|---|---|---|---|
| Persona 88M (b=3.2) | 7.3 | 0.7 | −0.7 |
| Patrón "dormilón" | 71.6 | 7.2 | 9.1* |
| Patrón "adivinador" | 43.0 | 4.3 | 5.3* |

*Indica desajuste

$$z^2 = \exp[(2x - 1)(d - b)]$$

$$v = \Sigma z^2/(L - 1)$$

$$t = [\ell n(v) + (v - 1)][(L - 1)/8]^{1/3}$$

Para hacer un patrón de adivinación, reorganizamos las respuestas para formar

1 1 1 1 0 0 0 0 0 1 1 $\qquad$ $\frac{\text{Puntaje}}{6}$

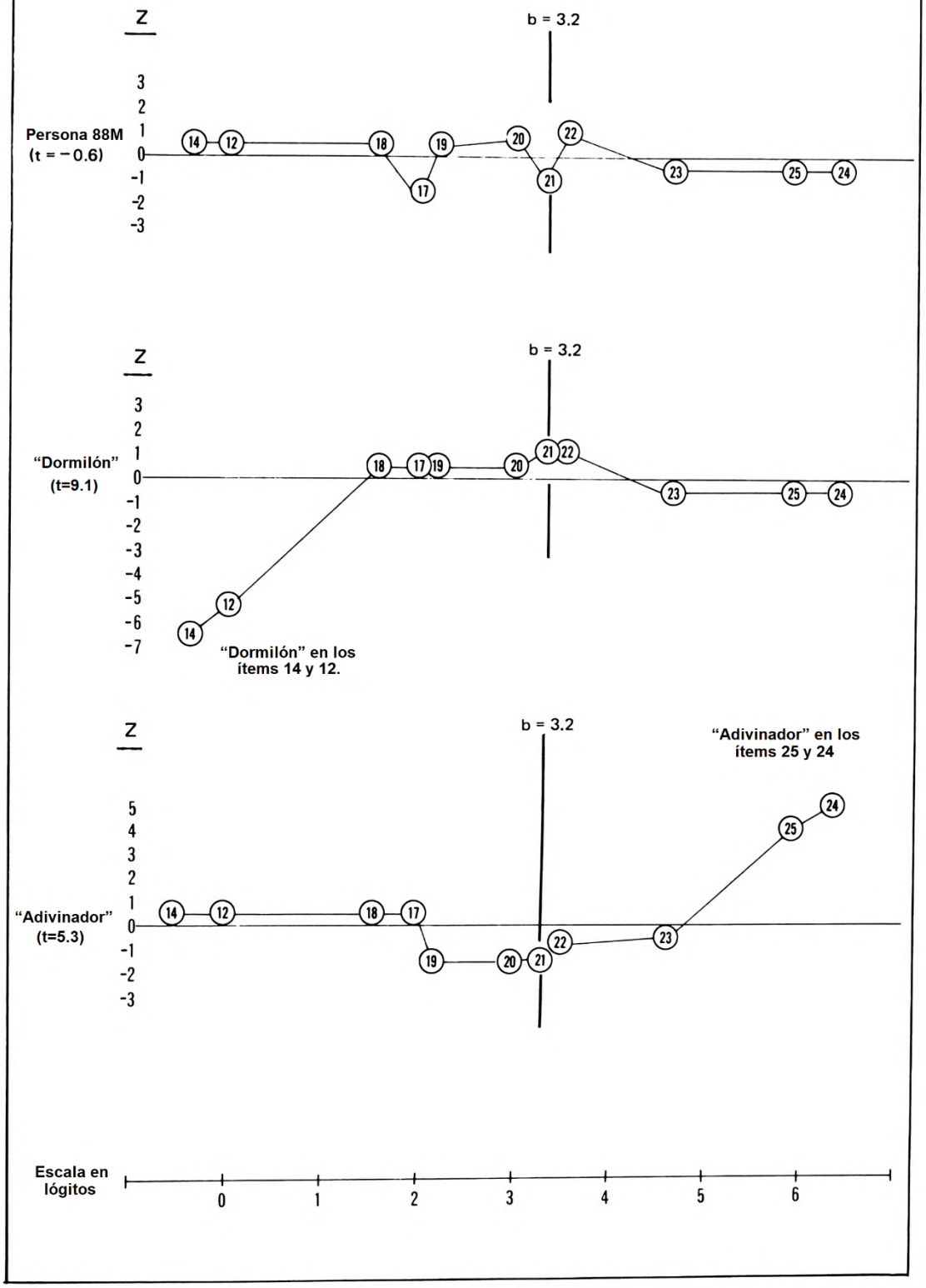
para el cual t = 5.3 en la Tabla 7.9.4. la Figura 7.9.2 compara el patrón de respuestas de la persona 88M aceptado previamente, con este patrón alternativo inaceptable del dormilón y del adivinador.

Un patrón característico de "descuidado" se ilustra en Tablas 7.9.5 y 7.9.6 y en la Figura 7.9.3 donde el patrón de respuesta aceptable de la Persona 29M ha sido modificado para mostrar respuestas incorrectas en los primeros cuatro ítems de la prueba, es decir, en los Ítems 3 a 6. Mientras que el efecto de una respuesta incorrecta entre estos primeros cuatro ítems no producen un desajuste significativo, como se ve en el patrón observado para la Persona 29M, si hacemos que los cuatro ítems sean incorrectos para ilustrar "descuido", el desajuste se vuelve significativo con t = 24.1.

El segundo patrón ilustrado en la figura 7.9.3 es "cauteloso". En este patrón la persona responde acertadamente todos los ítems correctos hasta donde puede y todos los ítems restantes son incorrectos. Este patrón puede presentase en función de un estilo de responder a la prueba gobernado por hábitos de trabajo lentos y deliberados. Mientras que el dormilón, el adivinador y el descuidado conducen a valores positivos de t, el patrón cauteloso, en cambio, produce un valor de t negativo. El valor negativo indica que el patrón de respuestas observadas se ajusta mejor de lo que esperamos. ¡Indica que incluso falta la variabilidad aleatoria esperada por el modelo!

**FIGURA 7.9.2**

**PATRONES DE RESPUESTA "DORMILÓN" Y "ADIVINADOR"**



## TABLA 7.9.5

## PATRONES DE RESPUESTA "DESCUIDADO" Y "CAUTELOSO"

| Descripción del caso | Estadístico de respuesta | Nombres y dificultades de los ítems (en orden creciente de dificultad) | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | #11 | #12 | #13 | #14 | #15 | #16 | #17 | #18 | #19 |
| | | −6.2 | −4.1 | −2.6 | −2.7 | −4.3 | −2.6 | −2.1 | −2.1 | −1.0 | −0.1 | −0.9 | −0.5 | −1.5 | −0.8 | 1.9 | 1.4 | 2.0 |
| | | Patrones de respuesta | | | | | | | | | | | | | | | | |
| | x | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| | (2x−1)(d−b) | −5.3 | −3.2 | −1.7 | 1.8 | −3.4 | −1.7 | −1.2 | 1.2 | 0.1 | 0.8 | 0.0 | −0.4 | −0.6 | 0.1 | −2.8 | −2.3 | −2.9 |
| | $z^2$ | 0.0 | 0.0 | 0.2 | 6.0 | 0.0 | 0.2 | 0.3 | 3.3 | 1.1 | 2.2 | 1.0 | 0.7 | 0.6 | 1.1 | 0.1 | 0.1 | 0.1 |
| Persona 29M | p | 1.00 | .96 | .85 | .14 | .97 | .85 | .77 | .23 | .48 | .31 | .50 | .60 | .65 | .48 | .94 | .91 | .95 |
| (b=-0.9) | z | 0.1 | 0.2 | 0.4 | −2.5 | 0.2 | 0.4 | 0.6 | −1.8 | −1.0 | 1.5 | 1.0 | −0.8 | 0.7 | 1.0 | −0.3 | −0.3 | −0.2 |
| | x | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| Patrón | (2x−1)(d−b) | 5.3 | 3.2 | 1.7 | 1.8 | −3.4 | −1.7 | −1.2 | −1.2 | −0.1 | 0.8 | 0.0 | 0.4 | −0.6 | 0.1 | −2.8 | −2.3 | −2.9 |
| "descuidado" | $z^2$ | 200.3 | 24.5 | 5.5 | 6.0 | 0.0 | 0.2 | 0.3 | 0.3 | 0.9 | 2.2 | 1.0 | 1.5 | 0.6 | 1.1 | 0.1 | 0.1 | 0.1 |
| | p | .00 | .04 | .15 | .14 | .97 | .85 | .77 | .77 | .52 | .31 | .50 | .40 | .65 | .48 | .94 | .91 | .95 |
| | z | −14.2 | −5.0 | −2.3 | −2.5 | 0.2 | 0.4 | 0.6 | 0.6 | 1.0 | 1.5 | 1.0 | 1.2 | 0.7 | 1.0 | −0.3 | −0.3 | −0.2 |
| | x | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Patrón | (2x−1)(d−b) | −5.3 | −3.2 | −1.7 | −1.8 | −3.4 | −1.7 | −1.2 | −1.2 | −0.1 | 0.8 | 0.0 | −0.4 | 0.6 | −0.1 | −2.8 | −2.3 | −2.9 |
| "cauteloso" | $z^2$ | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.2 | 0.3 | 0.3 | 0.9 | 2.2 | 1.0 | 0.7 | 1.8 | 0.9 | 0.1 | 0.1 | 0.1 |
| | p | 1.00 | .96 | .85 | .86 | .97 | .85 | .77 | .77 | .52 | .31 | .50 | .60 | .35 | .52 | .94 | .91 | .95 |
| | z | 0.1 | 0.2 | 0.4 | 0.4 | 0.2 | 0.4 | 0.6 | 0.6 | 1.0 | 1.5 | −1.0 | −0.8 | −1.4 | −1.0 | −0.3 | −0.3 | −0.2 |

$z^2 = \exp[(2x-1)(d-b)]$ $\quad p = 1/(1+z^2)$ $\quad z = (2x-1)\exp[(2x-1)(d-b)/2]$

**TABLA 7.9.6**

**ANÁLISIS DE LOS RESIDUOS PARA PATRONES DE RESPUESTA "DESCUIDADO" Y "CAUTELOSO"**

| Descripción del caso | Suma de cuadrados $\Sigma z^2$ | Media cuadrática v | Estadístico de ajuste t |
|---|---|---|---|
| Persona 29M (b=-0.9) | 17.6 | 1.1 | 0.3 |
| Patrón "descuidado" | 244.7 | 15.3 | 24.1* |
| Patrón "cauteloso" | 9.0 | 0.6 | –1.3 |

* Indica desajuste

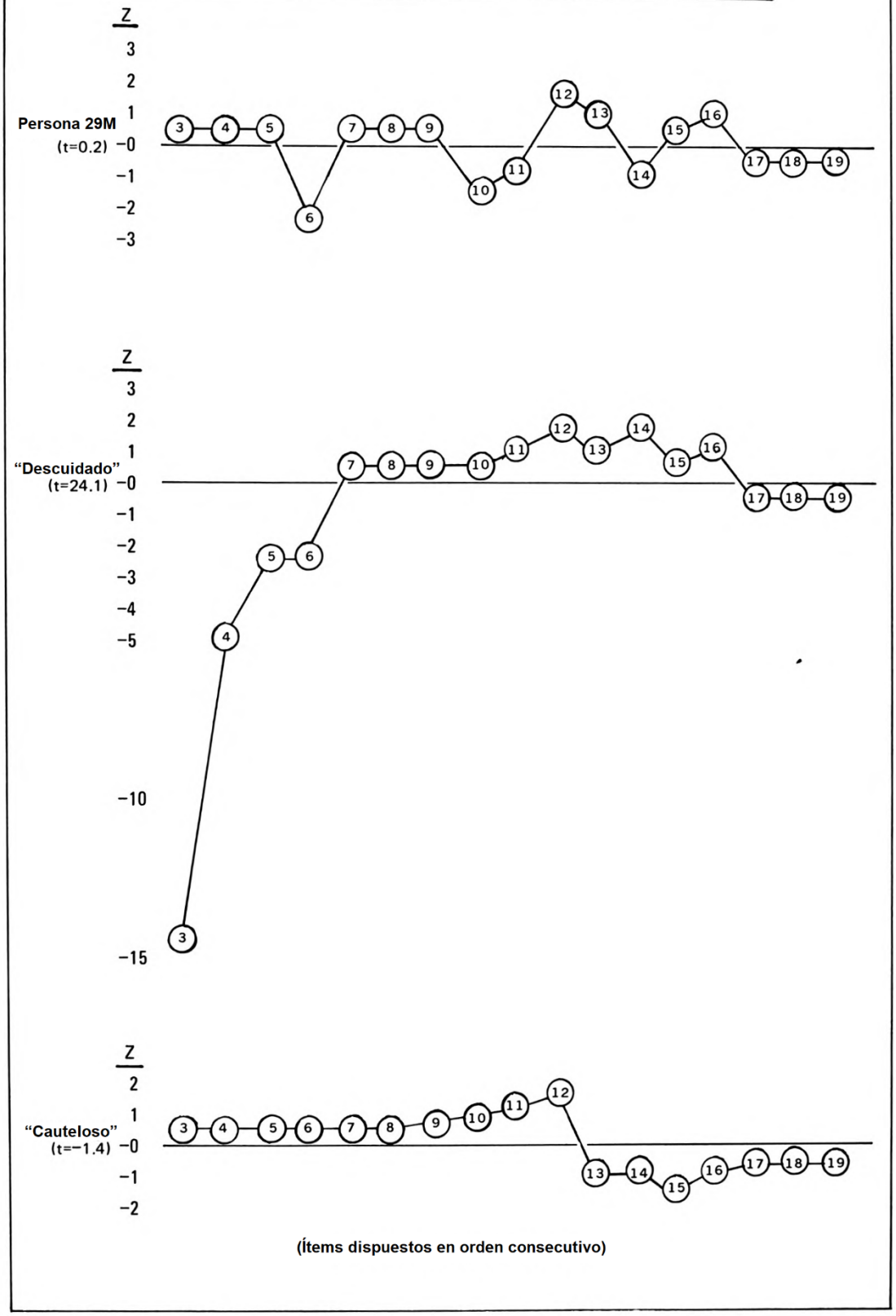
$$z^2 = \exp[(2x - 1)(d - b)]$$

$$v = \sum_{i}^{L} z^2/(L - 1)$$

$$t = [\ln(v) + (v - 1)][(L - 1)/8]^{1/3}$$

**FIGURA 7.9.3**

**PATRONES DE RESPUESTA "DESCUIDADO" Y "CAUTELOSO"**



## 7.10 CORRIGIENDO UNA MEDIDA

Cuando detectamos un desajuste significativo en un registro de respuesta, pasamos a diagnosticar el patrón de respuesta e identificar las posibles razones por las cuales puede ocurrir, llevándonos finalmente a la necesidad de decidir si puede determinarse alguna medida mejorada. Decidir si esta medida estadísticamente "corregida" es justa para la persona o apropiada a ciertas circunstancias es algo que no puede resolverse por simple estadística. Sin embargo, el hecho de saber cómo corregir objetivamente una medida puede darnos una mejor comprensión de su posible significado en el desempeño de una persona.

Hemos identificado que la respuesta de la persona 12M al primer ítem es poco verosímil en el primer segmento de prueba dado en las Tablas 7.9.1 y 7.9.2. Si llegamos a decidir que esta respuesta en particular no es típica de la Persona 12M, entonces deberíamos eliminar la respuesta incorrecta al ítem 3 y calcular una nueva estimación de habilidad basada en sus respuestas a los ocho ítems restantes. Este nuevo cálculo de su medida de habilidad se da en las Tablas 7.10.1 y 7.10.2. La medida corregida b' = –2,2 coloca a la Persona 12M aproximadamente 0.6 lógitos más arriba en la variable de la PCK. La figura 7.10.1 muestra el efecto de esta corrección en el ajuste de la Persona 12M con t' = –0.8 en lugar de t = 4.9.

Ahora tenemos dos estimaciones de la habilidad de la Persona 12M, una en b = –2.8 y otra en b' = –2.2. Para decidir cuál de las dos es una mejor estimación depende de cómo evaluamos la respuesta de la Persona 12M al Ítem 3. Si pensamos que esta respuesta es inverosímil y que es muy probable que obtenga el ítem 3 correcto, en caso de que lo intentara de nuevo, entonces podríamos usar como su medida el valor corregido b' = –2.2. Por el contrario, si pensamos que la Persona 12M obtuvo el ítem 3 incorrecto debido a un error significativo al responder, entonces podríamos considerar que b = –2.8 es mejor para reflejar su posición en la variable de la PCK. La experiencia clínica con la variable de la PCK apoya la probabilidad de que un error de este tipo es de hecho un indicador de una deficiencia y que su respuesta incorrecta al ítem 3 podría ser un elemento importante a tomar en cuenta en su evaluación. En consecuencia, en este caso, podríamos elegir la medida sin corregir b = –2.8.

En las Tablas 7.10.3 y 7.10.4 y en la figura 7.10.2 mostramos la corrección de un patrón típico de "adivinador". Las respuestas de la persona a ítems sucesivamente más difíciles muestran cuatro respuestas correctas seguidas de cinco respuestas incorrectas y luego dos respuestas correctas. Este patrón de respuesta tiene un desajuste significativo de t = 5.3. Debemos preguntarnos si el estimador de habilidad b = 3.2 es un buen indicador de la posición de esta persona en la variable de la PCK. Dada la cadena de respuestas incorrectas de esta persona antes de sus dos últimas respuestas correctas, podríamos calcular una nueva estimación con estas dos últimas respuestas sorprendentes eliminadas del registro. Con este nuevo patrón truncado se tiene b' = 1.7 y t' = –1.2. El análisis estadístico por sí solo no nos puede decir cuál estimación es la más apropiada, pero sí puede detectar y organizar la información disponible en un resumen conciso y objetivo para que nosotros lo usemos como parte de nuestra evaluación de la persona.

Las personas que adivinan pueden responder acertadamente a ítems difíciles con más frecuencia de lo que sus habilidades podrían predecir, especialmente en ítems de opción múltiple. Esto los hace parecer más capaces, especialmente cuando muchos ítems son demasiado difíciles para ellos, porque su frecuencia de éxito no disminuye a medida que aumenta la dificultad del ítem. Un efecto similar pero opuesto ocurre cuando las personas capaces se vuelven descuidadas con ítems fáciles que hacen que estas personas parezcan menos capaces.

Las respuestas de los ítems que se ven afectadas por la adivinación o el descuido en realidad reflejan la influencia simultánea de dos variables. Por un lado, existe la habilidad que se está midiendo y además, está la tendencia a adivinar o a ser descuidado. La posibilidad de "adivinación" ante un ítem puede o no ser una función simple de su dificultad en la variable principal o de los distractores en el caso de ítems de opción múltiple, sin embargo, para

## TABLA 7.10.1

## CORRIGIENDO LA MEDIDA DE LA PERSONA 12M POR "DORMILONA"

| Descripción del caso | Estadístico de respuesta | Nombres y dificultades de los ítems (en orden creciente de dificultad) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | #3 | #7 | #4 | #6 | #5 | #8 | #9 | #10 | #15 |
| | | −6.2 | −4.3 | −4.1 | −2.7 | −2.6 | −2.6 | −2.1 | −2.1 | −1.5 |
| | | Patrones de respuesta | | | | | | | | |
| | | desechar | | | b = − 2.8 | | | | | |
| Patrón observado para la Persona 12M | x | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| | (2x−1)(d−b) | 3.4 | −1.5 | −1.3 | 0.1 | 0.2 | 0.2 | −0.9 | −0.9 | −1.5 |
| | $z^2$ | 30.0 | 0.2 | 0.3 | 1.1 | 1.2 | 1.2 | 0.4 | 0.4 | 0.2 |
| | z | −5.5 | 0.5 | 0.5 | 1.1 | 1.1 | 1.1 | −0.6 | −0.6 | −0.5 |
| | | | | | | | b′ = − 2.2 | | | |
| Patrón corregido para la Persona 12M | x | | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| | (2x−1)(d−b′) | | −2.1 | −1.9 | −0.5 | −0.4 | −0.4 | −0.1 | −0.1 | −0.7 |
| | $z^2$ | | 0.1 | 0.2 | 0.6 | 0.7 | 0.7 | 0.9 | 0.9 | 0.5 |
| | z | | 0.4 | 0.4 | 0.8 | 0.8 | 0.8 | −1.0 | −1.0 | −0.7 |

Regla de corrección para "Dormilón": desechar d < ( b - 2), es decir, d < - 2.8 - 2 = - 4.8

## TABLA 7.10.2

## ANÁLISIS RESIDUAL DEL PATRON "DORMILÓN" CORREGIDO

| Caso | Diseño de prueba | | | Medición | | | | | | Análisis residual | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Altura h | Ancho w | Longitud L | Puntaje r | Puntaje relativo f=r/L | Habilidad relativa $x_{fw}$ | Coeficiente de error $C_{fw}^{½}$ | Habilidad b | Error s | Suma de cuadrados $\Sigma z^2$ | Media cuadrática v | Estadístico de ajuste t |
| Persona 12 M | –3.1 | 4 | 9 | 5 | .56 | 0.3 | 2.3 | – 2.8 | 0.8 | 35.0 | 4.4 | 4.9 * |
| Patrón corregido | – 2.8 | 3 | 8 | 5 | .63 | 0.6 | 2.2 | – 2.2 | 0.8 | 4.6 | 0.7 | – 0.8 |

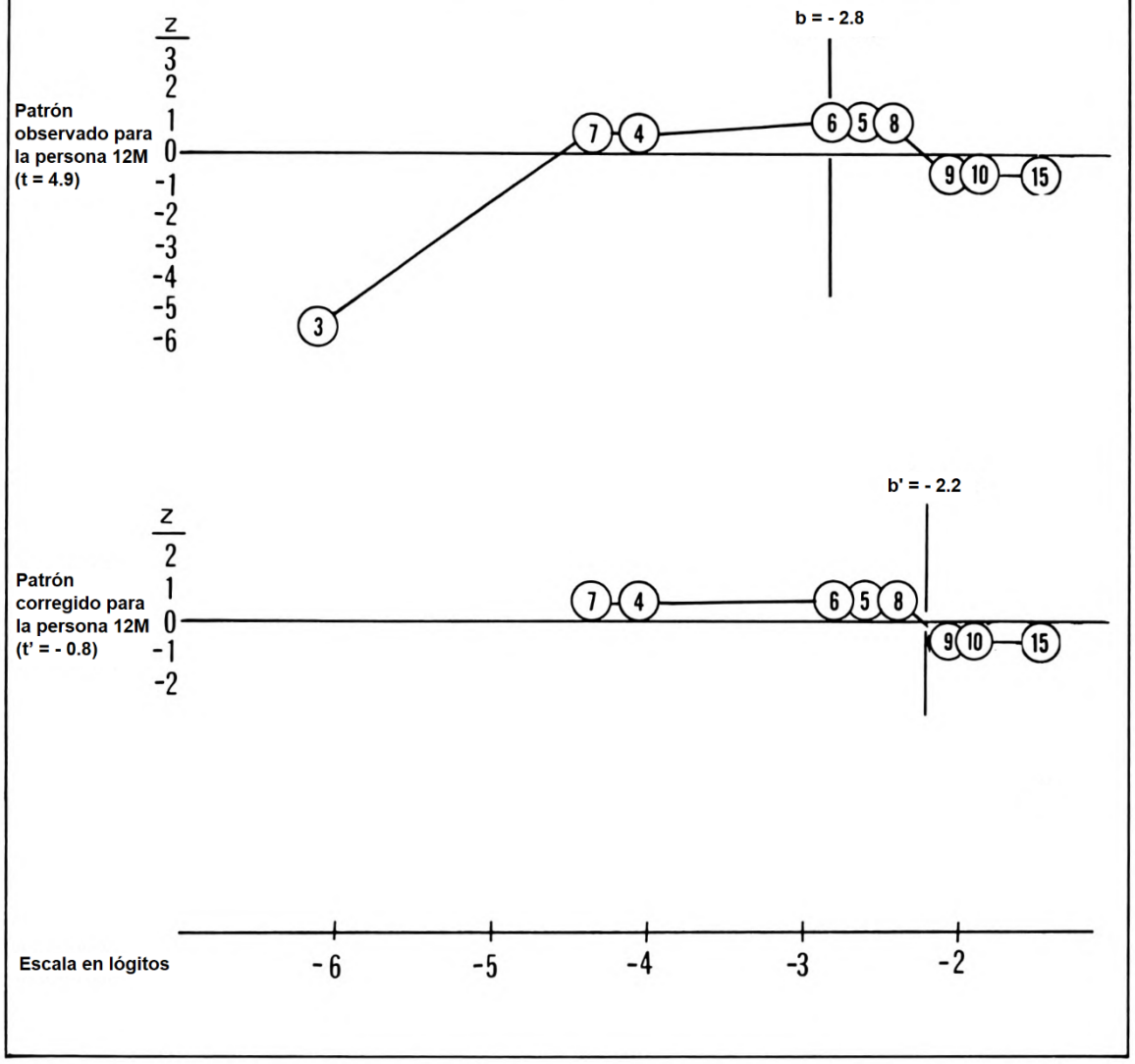
* Indica desajuste

**Corrección en la medida: (- 2.2) - (- 2.8) = 0.6**

$$v = \Sigma\, z^2/(L-1) \qquad t = [\ell n(v) + (v-1)]\,[(L-1)/8]^{½}$$

**FIGURA 7.10.1**

**CORRECCIÓN DE LA MEDIDA DE LA PERSONA 12 M POR "DORMILONA"**



## TABLA 7.10.3

## CORRIGIENDO EL PATRÓN "ADIVINADOR"

| Descripción del caso | Estadístico de respuesta | Nombres y dificultades de los ítems (en orden creciente de dificultad) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #14 | #12 | #18 | #17 | #19 | #20 | #21 | #22 | #23 | #25 | #24 |
| | | −0.5 | −0.1 | 1.4 | 1.9 | 2.0 | 2.9 | 3.3 | 3.3 | 4.5 | 5.8 | 6.3 |
| | | Patrones de respuesta | | | | | | | | | | |
| | | | | | | | b = 3.2 | | | | desechar | |
| Patrón "Adivinador" (b = 3.2) | x | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | (2x−1)(d−b) | −3.7 | −3.3 | −1.8 | −1.3 | 1.2 | 0.3 | −0.1 | −0.1 | −1.3 | 2.6 | 3.1 |
| | $z^2$ | 0.0 | 0.0 | 0.2 | 0.3 | 3.3 | 1.4 | 0.9 | 0.9 | 0.3 | 13.5 | 22.2 |
| | z | 0.2 | 0.2 | 0.4 | 0.5 | −1.8 | −1.2 | −1.0 | −1.0 | −0.5 | 3.7 | 4.7 |
| | | | | | b' = 1.7 | | | | | | | |
| Patrón corregido (b' = 1.7) | x | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | |
| | (2x−1)(d−b') | −2.2 | −1.8 | −0.3 | −0.2 | −0.3 | −1.2 | −1.6 | −1.6 | −2.8 | | |
| | $z^2$ | 0.1 | 0.2 | 0.7 | 1.2 | 0.7 | 0.3 | 0.2 | 0.2 | 0.1 | | |
| | z | 0.3 | 0.4 | 0.9 | 1.1 | −0.9 | −0.5 | −0.4 | −0.4 | −0.2 | | |

**Regla de corrección para "Adivinador": para m ítems elegidos desechar d > [ b + ln ( m - 1 )]**

**Es decir, si m=5, entonces desechar d > 3.2 + ln (4) = 3.2 + 1.4 = 4.6**

## TABLA 7.10.4

## RESIDUAL ANALYSIS OF A CORRECTED "GUESSING" PATTERN

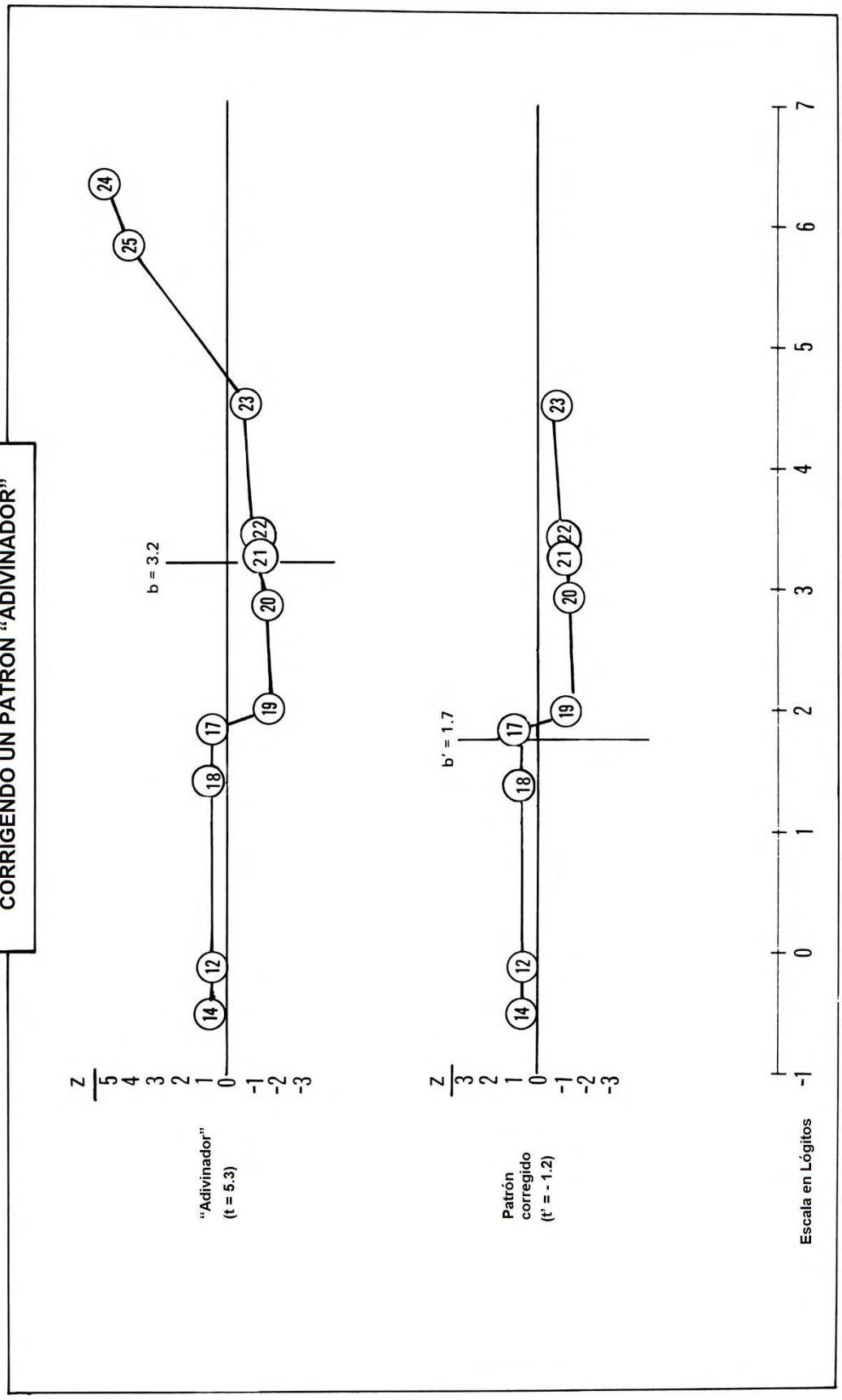
| Descripción del caso | Diseño de prueba | | | Medición | | | | | | Análisis residual | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Altura h | Ancho w | Longitud L | Puntaje r | Puntaje relativo f=r/L | Habilidad relativa $x_{fw}$ | Coeficiente de error $C_{fw}^{½}$ | Habilidad b | Error s | Suma de cuadrados $\Sigma z^2$ | Media cuadrática v | Estadístico de ajuste t |
| Patrón "Adivinador" | 2.8 | 7 | 11 | 6 | .55 | 0.4 | 2.9 | 3.2 | 0.9 | 43.0 | 4.3 | 5.3* |
| Patrón corregido | 2.1 | 5 | 9 | 4 | .44 | −0.4 | 2.4 | 1.7 | 0.8 | 3.7 | 0.5 | −1.2 |

* Indica desajuste

Corrección en la medida: 1.7 - 3.2 = -1.5

$$v = \Sigma z^2/(L - 1) \qquad t = [\ln(v) + (v - 1)]\,[(L - 1)/8]^{½}$$

**FIGURA 7.10.2**
**CORRIGENDO UN PATRÓN "ADIVINADOR"**



poder medir a la persona, hay dos variables involucradas: una es su habilidad, la otra es su inclinación a adivinar o a ser descuidado. La medición de cualquiera de las variables está amenazada por la presencia de la otra.

En las situaciones donde pensemos que la adivinación puede verse influida por el formato de la prueba como, por ejemplo, cuando pensamos que una persona puede adivinar al azar sobre m alternativas de opción múltiple, podríamos usar la probabilidad de adivinar de 1/m como un umbral por debajo del cual suponemos que puede ocurrir la adivinación. Para proteger nuestras medidas contra este tipo de adivinación, podemos entonces eliminar todos los ítems de un registro de respuesta cuya dificultad es mayor que $b + \ell n\ (m - 1)$ donde b es la habilidad inicial estimada de la persona. Después de esta eliminación, volvemos a estimar la habilidad de la persona con los ítems que le faltaron de contestar. Si hacemos esto, tomaremos la posición donde los ítems son tan difíciles que propician que una persona pueda obtener una mejor calificación si adivina contra la posibilidad de intentar contestarlos. De hecho tales ítems no deberían usarse para estimar la habilidad de la persona.

En las Tablas 7.10.5 y 7.10.6 mostramos un patrón "descuidado" y su corrección. Aquí tenemos un segmento de 17 ítems de dificultad creciente y el patrón de respuesta inicia con cuatro respuestas incorrectas seguidas por diez respuestas correctas y luego tres incorrectas. El patrón parece poco verosímil y se identifica un desajuste importante de t = 23.1. Algún factor extraño parece estar influyendo en las primeras cuatro respuestas. Podría ser un problema causado por el procedimiento de administración de la prueba, o por el comportamiento de la persona que se está examinando. Un patrón de respuestas corregidas podría formarse eliminando las primeras cuatro respuestas incorrectas y considerando solo el segmento continuo de respuestas correctas y las tres incorrectas que los siguen.

Las respuestas corregidas que resultan de este cambio muestran un patrón "cauteloso" con t = –3.0. Este patrón produce una medida de habilidad b' = 1.1, considerablemente mayor que la original b = –0.9. No se puede tomar una decisión final sobre este problema, sin embargo, hasta que haya suficiente información clínica o de comportamiento que pueda recopilarse para aclarar el significado de esas primeras cuatro respuestas incorrectas.

En resumen, los aspectos estadísticos de nuestra estrategia de corrección son:

a. Cuando la mayoría de las respuestas inesperadas son "incorrectas" y $t > 3$, deben desecharse todos los ítems "demasiado fáciles" de dificultad $d_i < (b_v - 2)$
   1. Calcule la nueva estimación de habilidad después de eliminar los ítems "demasiado fáciles".
   2. Haga un nuevo análisis de ajuste.

b. Cuando la mayoría de las respuestas inesperadas son "correctas" y $t > 3$, elimine todos los ítems "demasiado difíciles" de medida $d_i > [b_v + \ell n\ (m - 1)]$ donde m es el número de alternativas.
   1. Calcule la nueva estimación de habilidad después de eliminar los ítems "demasiado difíciles".
   2. Haga un nuevo análisis de ajuste.

## TABLA 7.10.5

## CORRECCIÓN DEL PATRÓN "DESCUIDADO"

| | | Nombres y dificultades de los ítems (en orden creciente de dificultad) | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | #11 | #12 | #13 | #14 | #15 | #16 | #17 | #18 | #19 |
| | | –6.2 | –4.1 | –2.6 | –2.7 | –4.3 | –2.6 | –2.1 | –2.1 | –1.0 | –0.1 | –0.9 | –0.5 | –1.5 | –0.8 | 1.9 | 1.4 | 2.0 |
| Descripción del caso | Estadístico de respuesta | Patrones de respuesta | | | | | | | | | | | | | | | | |
| | | desechar | | | | | | | | | | | | | | | | |
| Patrón | x | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| "descuidado" | (2x–1)(d–b) | 5.3 | 3.2 | 1.7 | 1.8 | –3.4 | –1.7 | –1.2 | –1.2 | –0.1 | 0.8 | 0.0 | 0.4 | –0.6 | 0.1 | –2.8 | –2.3 | –2.9 |
| (b= -0.9) | $z^2$ | 200.3 | 24.5 | 5.5 | 6.0 | 0.0 | 0.2 | 0.3 | 0.3 | 0.9 | 2.2 | 1.0 | 1.5 | 0.6 | 1.1 | 0.1 | 0.1 | 0.1 |
| | z | –14.2 | –5.0 | –2.3 | –2.5 | 0.2 | 0.4 | 0.6 | 0.6 | 1.0 | 1.5 | 1.0 | 1.2 | 0.7 | 1.0 | –0.3 | –0.3 | –0.2 |
| Patrón | x | | | | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| corregido | (2x–1)(d–b') | | | | | –5.4 | –3.7 | –3.2 | –3.2 | –2.1 | –1.2 | –2.0 | –1.6 | –2.6 | –1.9 | –0.8 | –0.3 | –0.9 |
| (b'= 1.1) | $z^2$ | | | | | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.1 | 0.1 | 0.1 | 0.1 | 0.4 | 0.7 | 0.4 |
| | z | | | | | 0.1 | 0.2 | 0.2 | 0.2 | 0.3 | 0.5 | 0.4 | 0.4 | 0.3 | 0.4 | –0.7 | –0.9 | –0.6 |

**Regla de corrección para "Descuidado": desechar todos los segmentos continuos de respuestas "incorrectas" a ítems fáciles que comienzan un patrón de repuestas seguidos de un segmento continuo de al menos tantas respuestas "correctas" a ítems más difíciles.**

## TABLA 7.10.6

## ANÁLISIS RESIDUAL DEL PATRÓN DE RESPUESTAS "DESCUIDADO"

| Descripción del caso | Diseño de prueba | | | Medición | | | | | | Análisis residual | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Altura h | Ancho w | Longitud L | Puntaje r | Puntaje relativo f=r/L | Habilidad relativa $x_{fw}$ | Coeficiente de error $C_{fw}^{½}$ | Habilidad b | Error s | Suma de cuadrados $\Sigma z^2$ | Media cuadrática v | Estadístico de ajuste t |
| Patrón "Descuidado" | −1.6 | 8 | 17 | 10 | .59 | 0.7 | 2.9 | −0.9 | 0.7 | 244.7 | 15.3 | 24.1* |
| Patrón corregido | − 0.8 | 6 | 13 | 10 | .77 | 1.9 | 2.8 | 1.1 | 0.8 | 2.4 | 0.2 | − 3.0** |

Corrección en la medida: 1.1 - ( - 0.9) = 2.0

* Indica desajuste

** Indica descuido

$$v = \Sigma z^2/(L-1) \qquad t = [\ell n(v) + (v-1)]\,[(L-1)/8]^{½}$$

# 8. ELECCIÓN DE UNA ESCALA

## 8.1 INTRODUCCIÓN

Los lógitos son las unidades de medida que hemos usado hasta ahora. Estas unidades se obtienen directamente del modelo logístico de respuesta que especifica la estimación de la probabilidad de que la persona v responda correctamente al ítem i, como

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)]$$

donde $b_v$ es la habilidad estimada para la persona v y $d_i$ es la dificultad estimada para el ítem i. Se deduce que la probabilidad de una respuesta correcta se expresan en función del momio:

$$p_{vi}/(1 - p_{vi}) = \exp(b_v - d_i)$$

de la cual el logaritmo natural del momio para una respuesta correcta es:

$$\ell n\,[p_{vi}/(1 - p_{vi})] = (b_v - d_i).$$

Los logaritmos de los momios se denominan "lógitos" y las diferencias entre ítems y personas se expresan en unidades de lógitos.

La elección de una unidad es completamente arbitraria, pero es absolutamente necesario elegir alguna unidad de medida. Es posible continuar utilizando los lógitos iniciales como las unidades de medida, pero esto tiene dos desventajas. Los lógitos involucran valores negativos y decimales que son características numéricas que pueden hacerlos innecesariamente confusos.

Por ejemplo, la escala de la PCK en lógitos se extiende de –5.8 a +5.2. En las longitudes de prueba habituales, los valores del error estándar de medida pueden ser tan bajos como 0.6 lógitos. Podríamos agregar una constante, del orden de 10, para eliminar los valores negativos, pero esto no permitiría evitar los decimales simplemente redondeando al entero más cercano de las medidas en lógitos obtenidas en la PCK. El redondeo produciría una diferencia menos apreciable de casi dos errores estándar y, por lo tanto, podría anular diferencias significativas. Si transformamos el valor a una escala de lógitos multiplicando por 10 y sumando 100, obtendríamos medidas en una nueva escala de medidas de 42 a 152, que transmitiría la misma información que la escala inicial en lógitos, pero que estaría libre de valores negativos y decimales.

Para crear una nueva escala sin la inconveniencia de los decimales, debemos multiplicar los lógitos por un factor de escala o "espaciamiento" suficientemente grande para no perder información útil al redondear las nuevas unidades al entero más cercano. Una vez elegido el factor de espaciamiento y determinada la unidad de nuestra nueva escala, agregaremos un valor de corrimiento o "ubicación" que hará que las nuevas unidades enteras sean suficientemente grandes para que el valor más bajo posible sea mayor que cero. La nueva escala se define determinando dos valores: el factor multiplicativo establece el espaciado, o unidades de la escala y el valor aditivo que establece la ubicación u origen de la escala.

Generalmente es fácil elegir el valor aditivo que localiza todos los valores posibles por encima de cero. Sin embargo, la elección de un factor multiplicativo requiere más consideraciones. Si queremos trabajar en unidades enteras, debemos organizar todo para que tenga sentido

cualquier diferencia menor a un entero en nuestra nueva escala. Esto requiere que investiguemos el tamaño de la diferencia menos significativa.

Además de determinar una diferencia mínima significativa, deberíamos disponer de puntos fáciles de recordar, como 50, 100 o 500 en la nueva escala, ubicados en valores asociados con criterios importantes a lo largo de la variable o en ubicaciones típicas de una referencia grupal o normativa. Es posible que encontremos de utilidad relacionar nuestra nueva unidad de medida con las probabilidades de respuesta que puede predecir el modelo de respuesta. También podemos desear movernos entre probabilidades de respuesta fáciles de recordar como 0.10, 0.25, 0.50, 0.75 y 0.90 con incrementos regulares de 5, 10, 20 o 25 unidades a lo largo de nuestra nueva escala.

Por lo tanto, además de eliminar negativos y decimales innecesarios, al agregar una constante y establecer una unidad menos significativa mayor que uno, podremos organizar nuestra escala alrededor de las consideraciones normativas, sustantivas o de probabilidad de respuesta.

## 8.2 FÓRMULAS PARA CONSTRUIR NUEVAS ESCALAS

Para explicar cómo se construye una nueva escala, haremos su definición como la transformación lineal $y = \alpha + \gamma x$ en la cual x es la escala en lógitos, y es la nueva escala, $\alpha$ es el factor de ubicación para determinar el origen de la nueva escala, γ es el factor de espaciado para determinar la nueva unidad de la escala. Hacemos esta transformación lineal porque deseamos conservar las características de intervalo en lógitos que produce el modelo de Rasch.

Nuestras nuevas medidas B y las nuevas calibraciones D pueden expresarse en términos de sus contrapartes en lógitos b y d como:

$$B = \alpha + \gamma b \quad \text{para las personas} \qquad \textbf{[8.2.1]}$$

$$D = \alpha + \gamma d \quad \text{para los ítems} \qquad \textbf{[8.2.2]}$$

Los nuevos errores estándar de medida y calibración son:

$$ES(B) = \gamma ES(b) \qquad \textbf{[8.2.3]}$$

$$ES(D) = \gamma ES(d) \qquad \textbf{[8.2.4]}$$

Esto muestra cómo las características de nuestra nueva escala dependen de los valores de $\alpha$ y $\gamma$ elegidos para definirla.

De paso debemos revisar de nuevo que las habilidades de las personas y las dificultades de los ítems permiten ubicar posiciones de interés en una variable común. Construimos esta variable trabajando necesariamente con las calibraciones de los ítems que la definen. Sin embargo, cuando usamos la variable para medir a las personas debemos trabajar con las medidas a lo largo de la variable definida por los ítems, no por las personas. Lo que una medida nos dice acerca de la persona es cuál es el nivel de dificultad de los ítems que esa persona puede responder correctamente la mitad de las veces. De la misma manera, la calibración de un ítem nos habla del nivel de habilidad de las personas que probablemente lo respondan correctamente la mitad del tiempo. De hecho si no dijéramos "medida" para referirnos a la ubicación de las personas y "calibración "para referirnos a la ubicación de los ítems, podríamos intercambiar los términos para hablar de la dificultad del ítem como su medida y de la habilidad de la persona como su calibración.

## 8.3 LA MENOR DIFERENCIA MEDIBLE

Queremos liberar de decimales a nuestra nueva escala, pero no queremos cancelar el uso de información útil. Como resultado, tenemos que determinar la menor diferencia medible MDM (o LMD por sus siglas en inglés) en nuestra escala de lógitos para que podamos elegir un factor de

espaciamiento $\gamma$ que hace que la MDM en lógitos se traduzca al menos por un entero en nuestra nueva escala. Aunque dos personas estén muy cercanas sus puntajes observados deben estar suficientemente separados, sin ser el mismo valor. Esta es la mínima diferencia observable MDO (o LOD por sus siglas en inglés). Necesitamos transformar esta MDO en su correspondiente MDM en lógitos de habilidad.

La habilidad b en lógitos se obtiene del puntaje r a través de la expectativa del modelo de respuesta:

$$r = \sum_{i}^{L} \{\exp(b_r - d_i)/[1 + \exp(b_r - d_i)]\}$$

Como resultado, el MDM debe seguir al MDO por una razón de cambio $\partial b/\partial r$ por la cual los puntajes de r producen medidas de b, es decir:

$$MDM \simeq \frac{\partial b}{\partial r}\ MDO$$

Para estandarizar las observaciones respecto de la longitud de prueba L generalizamos el puntaje bruto r al puntaje relativo f = r / L. Luego, con MDO = 1 en el puntaje r y f = r / L, tenemos que MDM = 1 / L en puntaje relativo f, con lo que se tiene:

$$MDM \simeq \frac{\partial b}{\partial f}\ MDO = \frac{\partial b}{\partial f}\ (1/L) \qquad \textbf{[8.3.1]}$$

El modelo de respuesta de Rasch nos da la relación esperada entre el puntaje relativo f y la probabilidad estimada de respuesta $p_{fi}$:

$$f = \sum_{i}^{L} p_{fi}/L$$

En donde: $p_{fi} = \exp(b_f - d_i)/[1 + \exp(b_f - d_i)]$.

Por lo tanto, la razón de cambio a la cual el puntaje relativo f produce b es

$$\frac{\partial b}{\partial f} = \left[\sum_{i}^{L} p_{fi}(1 - p_{fi})/L\right]^{-1} = C_{fw}$$

que resulta ser el coeficiente de error $C_{fw}$ discutido en los Capítulos 6 y 7.

Los subíndices de este coeficiente indican el ancho w y el puntaje relativo f porque, como aprendimos en el Capítulo 6, los valores exactos de este coeficiente no solo dependen de la relación entre el nivel de dificultad de la prueba y la habilidad de la persona expresada en puntaje relativo f = r / L, sino también depende del rango de dificultad cubierto por la prueba. Esto nos da al menos una diferencia medible:

$$MDM \simeq \frac{\partial b}{\partial f}\ (1/L) = C_{fw}/L \qquad \textbf{[8.3.2]}$$

La forma en que $C_{fw}$ y, por lo tanto, MDM varía con b, se muestra en la figura 8.3.1. Conforme la medida de habilidad se aleja del centro de prueba y / o la curva de operación de la prueba se vuelve más plana, MDM se hace más grande. Afortunadamente, el rango de valores que $C_{fw}$ tendrá en la práctica es limitado.

Cuando $1/8 < p_{fi} < 7/8 \qquad i=1, L$

entonces $4 < C_{fw} < 9$

y $C_{fw} = 6$

se puede usar como un valor único de trabajo conveniente para $C_{fw}$ (consulte la Tabla 6.8.1 para obtener más información).

Esto nos da una definición de trabajo de la menor diferencia medible:

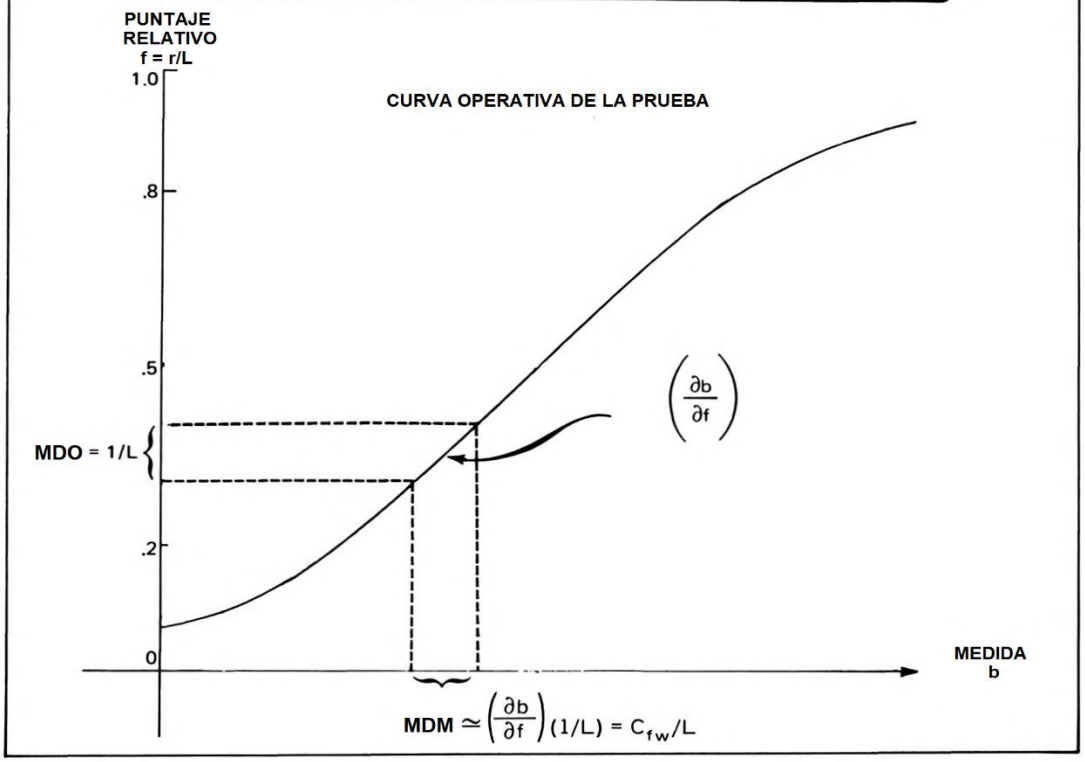
$$MDM = 6/L \qquad \textbf{[8.3.3]}$$

que implica un factor de espaciamiento:

$$\gamma_{MDM} > L/6$$

**FIGURA 8.3.1**

**DETERMINACIÓN DE LA MENOR DIFERENCIA SIGNIFICATIVA**




El MDM se aproxima a la unidad significativa más pequeña posible, ya que se deriva de la Mínima Diferencia Observable. Sin embargo, desde el punto de vista de la estimación, podríamos considerar que un error estándar de medida ESM (o SEM por sus siglas en inglés) es en realidad la menor diferencia "creíble". En lógitos, el ESM se relaciona con el MDM como:

$$ESM = (MDM)^{1/2} = C_{fw}{}^{1/2}/L^{1/2}$$

que sugiere el valor de trabajo

$$ESM = 2.5/L^{1/2} \qquad \textbf{[8.3.4]}$$

como una base alterna para determinar el factor de espaciamiento

$$\gamma_{ESM} > L^{1/2}/2.5$$

Siempre que haya más de seis ítems en nuestra prueba, el ESM determina un $\gamma$ menor que el MDM, porque

$$L^{1/2}/2.5 < L/6 \qquad \text{cuando} \quad L > 6.$$

Una escala basada en el ESM, podría ser más simple numéricamente, sin embargo, también sería menos discriminante en su incremento entero que otra escala basada en el MDM. Cuál elección es preferible en cualquier situación particular no puede resolverse con base en consideraciones estadísticas. La elección dependerá inevitablemente del uso al que se destinen las medidas.

Finalmente, podríamos considerar la diferencia menos significativa entre las medidas independientes, ya sean réplicas de respuestas de la misma persona o comparaciones entre diferentes personas, como un límite superior de cuán burda podríamos permitir que se convirtiera nuestra nueva escala. Para determinar esta diferencia menos significativa MDS (o LSD por sus siglas en inglés) tomamos

$$MDS_{ab} = \left(ESM_a^{\,2} + ESM_b^{\,2}\right)^{1/2} \simeq \left(2\,ESM^2\right)^{1/2}$$

y llegar al valor de trabajo

$$MDS = 1.4\,ESM = 3.5/L^{1/2} \qquad \textbf{[8.3.5]}$$

que produce un factor de espaciamiento mínimo

$$\gamma_{MDS} > L^{1/2}/3.5$$

Siempre que el número de ítems en nuestra prueba sea mayor a seis, las magnitudes relativas de las bases para determinar un límite inferior para el factor de espaciamiento son

$$MDM < ESM < MDS$$

y entonces los factores de espaciamiento que determinan están ordenados

$$\gamma_{MDM} > \gamma_{ESM} > \gamma_{MDS}$$

La figura 8.3.2 muestra las relaciones entre MDM, ESM y MDS en lógitos para la Prueba PCKB de 23 ítems. Los ítems, sus valores en lógitos, puntajes equivalentes y MDO de los ítems 3, 4, 12 a 13, 18 a 19 y 20, 21 junto con sus respectivos MDM, ESM y MDS.

Podemos comparar los valores exactos en la Figura 8.3.2 con las aproximaciones dadas por las ecuaciones 8.3.3, 8.3.4 y 8.3.5.

| | Mínimo valor de $\gamma$ |
|---|---|
| $MDM = 6/L = 6/23 = 0.26$ | 4 |
| $ESM = 2.5/L^{1/2} = 2.5/4.8 = 0.52$ | 2 |
| $MDS = 3.5/L^{1/2} = 3.5/4.8 = 0.73$ | 1.5 |

Debido a que el rango de 13 lógitos de la PCKB es inusualmente amplio, estas aproximaciones son más pequeñas que los valores exactos dados en la figura 8.3.2. El mínimo factor $\gamma$ de espaciamiento de MDM indicado por los valores exactos sería aproximadamente de 2 mientras que las aproximaciones podrían conducir a un mínimo de $\gamma$ de 4. Dado que solo estaríamos en riesgo de perder información si las aproximaciones nos llevaran a un valor de $\gamma$ menor a 2, vemos que incluso en esta situación extrema las aproximaciones no nos engañan.

## 8.4 DEFINICIÓN DEL FACTOR DE ESPACIAMIENTO

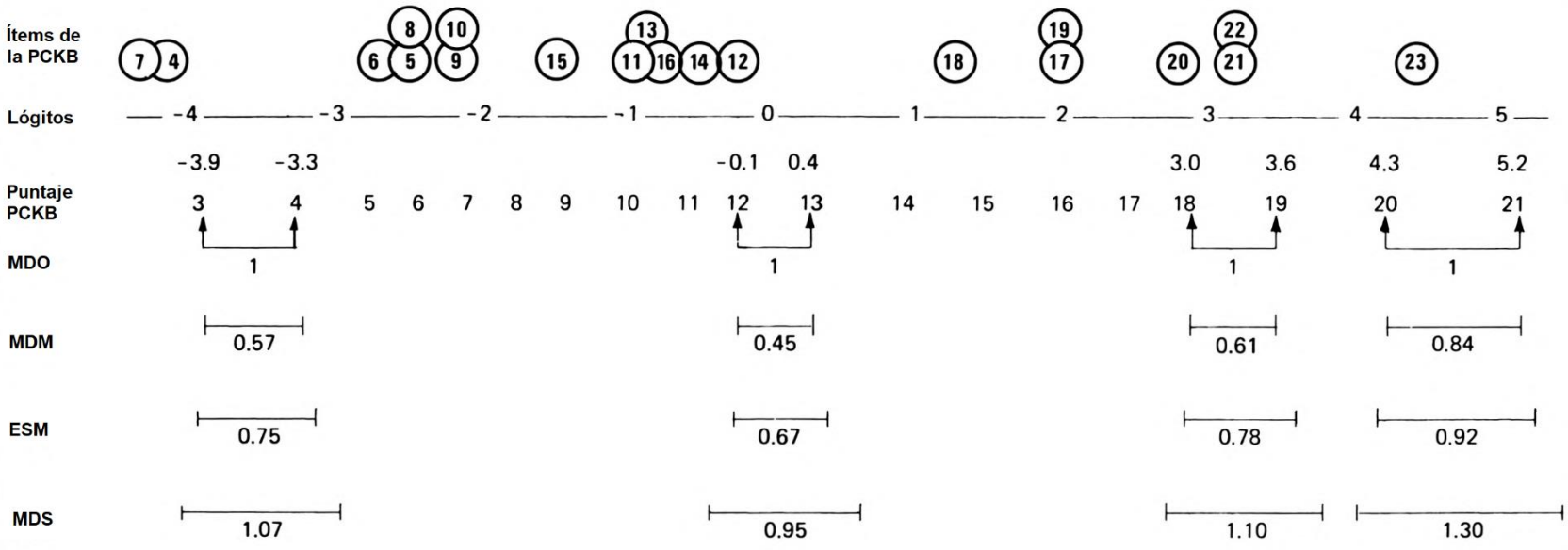
Una vez definida la diferencia menos significativa en lógitos, ya sea como la mínima diferencia medible MDM(b) = 6/L para mantener la máxima amplitud de observación o con base en el error estándar de medida ESM(b) = 2.5/L½ y su mínima diferencia significativa MDS(b) = 3.5/L½ para mantener la confiabilidad estadística, de esta forma podemos usar esta diferencia menos significativa para establecer un factor de espaciamiento, que hará que todas las diferencias interpretables en nuestra nueva escala sean mayores que uno.

Si nuestro objetivo es hacer la diferencia menos medible en la nueva escala MDM (B) > 1, entonces, dado que $\gamma$ = MDM (B) / MDM (b), se sigue, como en la Ecuación 8.3.3,

$$\gamma_{MDM} > L/6 \qquad \textbf{[8.4.1]}$$

**FIGURA 8.3.2**

**MÍNIMA DIFERENCIA SIGNIFICATIVA EN LÓGITOS PARA LA PCKB**




**La escala utilizada aquí se definió por 23 ítems. El ítem 3 en -6,2, el ítem 25 en 5.8 y el ítem 24 en 6.3, sin embargo está ubicados fuera del alcance de esta figura.**

**Los valores de la figura son exactos. Su aproximación burda con las ecuaciones 8.3.3, 8.3.4 y 8.3.5 son MDM = 0.26, ESM = 0.52, MDS = 0.73.**

$MDO = (r+1) - r = 1$ $\quad MDM = b_{r+1} - b_r$ $\quad ESM = MDM \times 1.4$ $\quad MDS = 1.4\ ESM$

es el factor de espaciamiento que garantiza que no se cancelen las diferencias observables al tomar el redondeo al entero más cercano.

En cambio, si estuviéramos interesados en mantener el factor de espaciamiento $\gamma$ lo más pequeño posible, para evitar la presentación de diferencias de escala que son estadísticamente poco confiables, podríamos establecer $\gamma$ en 1/ESM (b) o incluso 1/MDS (b) que es

$$\gamma_{ESM} = L^{1/2}/2.5 \qquad \textbf{[8.4.2]}$$

o

$$\gamma_{MDS} = L^{1/2}/3.5 \qquad \textbf{[8.4.3]}$$

A menudo, sin embargo, habrá otras consideraciones que nos llevarán a permitir que $\gamma$ pueda ser incluso mayor que L/6 para alcanzar intervalos de escala de referencia como 5, 10, 20, 25, 50 o 100.

Para tener una idea aproximada de los valores típicos útiles de $\gamma$, enumeramos en la Tabla 8.4.1 algunos valores para las diferencias menos significativas en función de varias longitudes de prueba. En la tabla 8.4.1 vemos que raramente estaríamos satisfechos con un factor de espaciamiento menor que 5 aunque rara vez necesitamos uno más grande que 100. La Tabla 8.4.1 sugiere que podríamos trabajar satisfactoriamente con estos valores:

$\gamma$ = 5 para pruebas cortas en el aula de 20 o 30 ítems,
$\gamma$ = 10 para pruebas con unidades típicas de 50 a 60 ítems
$\gamma$ = 20 o 25 para pruebas largas de 120 a 150 ítems.

y

Solo en pruebas de duración poco convencional, tales como exámenes de 1,000 ítems, es que podríamos necesitar $\gamma$ = 100.

**TABLA 8.4.1**

**RELACIÓN ENTRE FACTORES DE ESPACIAMIENTO Y LONGITUD DE PRUEBA**

| | Mínima diferencia significativa | | | Factor de espaciamiento aproximado $\gamma$ para alcanzar una escala entera | | |
|---|---|---|---|---|---|---|
| Longitud de la prueba L | Mínimo MDM 6/L | ESM 2.5/L½ | Máximo MDS 3.5/L½ | Máximo $\gamma_{MDM}$ | $\gamma_{ESM}$ | Mínimo $\gamma_{MDS}$ |
| 30 | 0.20 | 0.46 | 0.64 | [5] | 2 | 2 |
| 60 | 0.10 | 0.32 | 0.45 | (10) | 3 | 2 |
| 120 | 0.05 | 0.23 | 0.32 | 20 | 4 | 3 |
| 150 | 0.04 | 0.20 | 0.29 | 25 | [5] | 3 |
| 300 | 0.02 | 0.14 | 0.20 | 50 | 7 | [5] |
| 600 | 0.01 | 0.10 | 0.14 | 100 | (10) | 7 |
| 1200 | 0.005 | 0.07 | 0.10 | 200 | 15 | (10) |

Una vez decidido el valor de $\gamma$ lo aplicamos a nuestras medidas de personas b y dificultades de los ítems d para ubicarlos en nuestra nueva escala como valores B y D. La relación entre MDM y ESM en lógitos para MDM(b) = [ESM(b)]² es fácil de recordar, mientras que su relación en la nueva escala también involucra a $\gamma$. Ya que

$$\mathrm{MDM(B)} = \gamma\mathrm{MDM(b)}$$

$$\mathrm{ESM(B)} = \gamma\mathrm{ESM(b)}$$

Pero $$\mathrm{MDM(b)} = [\mathrm{ESM(b)}]^2$$

de lo que se sigue en nuestra nueva escala

$$\mathrm{MDM(B)} = [\mathrm{ESM(B)}]^2/\gamma \quad \textbf{[8.4.4]}$$

$$\mathrm{ESM(B)} = [\gamma\mathrm{MDM(B)}]^{1/2} \quad \textbf{[8.4.5]}$$

Por ejemplo, supongamos que para re-escalar la PCKB que se muestra en la figura 8.3.2 elegimos como valor de $\gamma = 5$. Entonces se tiene esta relación en lógitos

$$\mathrm{MDM(B)} = [\mathrm{ESM(b)}]^2$$

en nuestra nueva escala

$$\mathrm{MDM(B)} = [\mathrm{ESM(B)}]^2/5$$

y $$\mathrm{ESM(B)} = [5\mathrm{MDM(B)}]^{1/2}$$

Así, mientras que un ESM(b) de $0.75 = 0.57^{1/2}$ se asocia con un MDM(b) de 0.57, cuando $\gamma = 5$, entonces

$$\mathrm{MDM(B)} = 5 \times 0.57 = 2.81$$

pero $$\mathrm{ESM(B)} = 5 \times 0.75 = 3.75 = (5 \times 2.81)^{1/2}$$

## 8.5. UNIDADES DE ESCALAMIENTO NORMATIVO: NITS

Si queremos que nuestra escala se fundamente en una referencia normativa, podemos usar como factores la media observada en lógitos m y la desviación estándar s en lógitos de la muestra normativa elegida. Estos factores entran en una transformación preliminar d´= (d – m)/s y b´= (b – m)/s, que pone la media del grupo normativo en cero y la unidad de la escala en una desviación estándar normativa.

Después de este paso preliminar, elegimos un factor de espaciamiento lo suficientemente grande para que las diferencias significativas se vuelvan más grandes que uno y en un valor que fija la desviación estándar normativa en alguna unidad fácil de recordar, como 10, 20, 50 o incluso 100. Al mismo tiempo, elegimos el factor de ubicación $\alpha$ para que la media del grupo normativo sea también fácil de recordar, por ejemplo, en 50, 100 o 500.

Por lo tanto, para crear una escala basada en unidades normalizadas (o normativas), llamadas NIT, utilizamos esta expresión para las personas:

$$\mathrm{B} = \alpha + \gamma(\mathrm{b} - \mathrm{m})/\mathrm{s} \quad \textbf{[8.5.1]}$$

y esta para los ítems:

$$\mathrm{D} = \alpha + \gamma(\mathrm{d} - \mathrm{m})/\mathrm{s}$$

Luego en la nueva escala de NIT tenemos la media normativa M = $\alpha$ y la desviación estándar normativa S = $\gamma$.

Usando la administración de la PCKB a las 68 personas mayores de 8 años como datos normativos, tenemos m = 1.3 y s = 1.9, ambos en lógitos. Si ahora elegimos $\alpha = 50$, $\gamma = 10$, obtenemos una nueva escala en NIT, en la cual

$$B = 50 + 10\,(b - 1.3)/1.9 \qquad \textbf{[8.5.2]}$$
$$= 50 - 10\,(1.3/1.9) + 10b/1.9$$
$$= 43.2 + 5.3b$$

y

$$D = 43.2 + 5.3d$$

Observe que con esta definición de escala, la media normativa de m = 1.3 lógitos se convierte en

$$M = 43.2 + 5.3(1.3) = 50 \text{ NIT}$$

Si ahora establecemos b en

$$m + s = 1.3 + 1.9 = 3.2$$

entonces

$$M + S = 43.2 + 5.3(3.2) = 60 \text{ NIT}$$

así que se tiene

$$M + S - M = S = 60 - 50 = 10 \text{ NIT}$$

es la desviación estándar normativa en la nueva escala de NITs.

La figura 8.5.1 muestra la distribución de las 68 personas normativas. Debajo de su distribución se tienen las medidas de habilidad para cada puntaje en lógitos y en NITs, y en la parte inferior están los ítems de la PCKB que definen la variable. En la figura 8.5.1 vemos que

$$30 \text{ NITs} \rightarrow m - 2s,\ 40 \text{ NITs} \rightarrow m - s,\ 50 \text{ NITs} \rightarrow m,\ 60 \text{ NITs} \rightarrow m + s \ y\ 70 \text{ NITs} \rightarrow m + 2s$$

## 8.6. UNIDADES DE ESCALAMIENTO SUSTANTIVO: SITS

Con otro propósito, podríamos optar por referir nuestra nueva escala con base en consideraciones sustantivas como por ejemplo un nivel basal y otro nivel de competencia, un nivel de entrada y un nivel de salida o cualquier jerarquía sobre dos niveles de dominio o de desempeño. Para lograr esto, encontramos los niveles de dificultad $d_1$ y $d_2$ en nuestra escala de lógitos que marcan nuestra elección de dos posiciones de criterio. A continuación transformamos estos lógitos a los valores $D_1$ y $D_2$ en una nueva escala sustantiva o SIT, que permiten ubicar nuestros criterios en posiciones fáciles de recordar como 50, 100 o 200.

Si $d_1$ y $d_2$ identifican las posiciones de los criterios en lógitos y $D_1$ y $D_2$ representan esas posiciones fáciles de recordar para dichos criterios en la nueva escala, entonces

$$\alpha = (D_1\, d_2 - D_2\, d_1)/(d_2 - d_1)$$
$$\gamma = (D_2 - D_1)/(d_2 - d_1)$$

y entonces

$$B = \alpha + \gamma b$$
$$D = \alpha + \gamma d$$

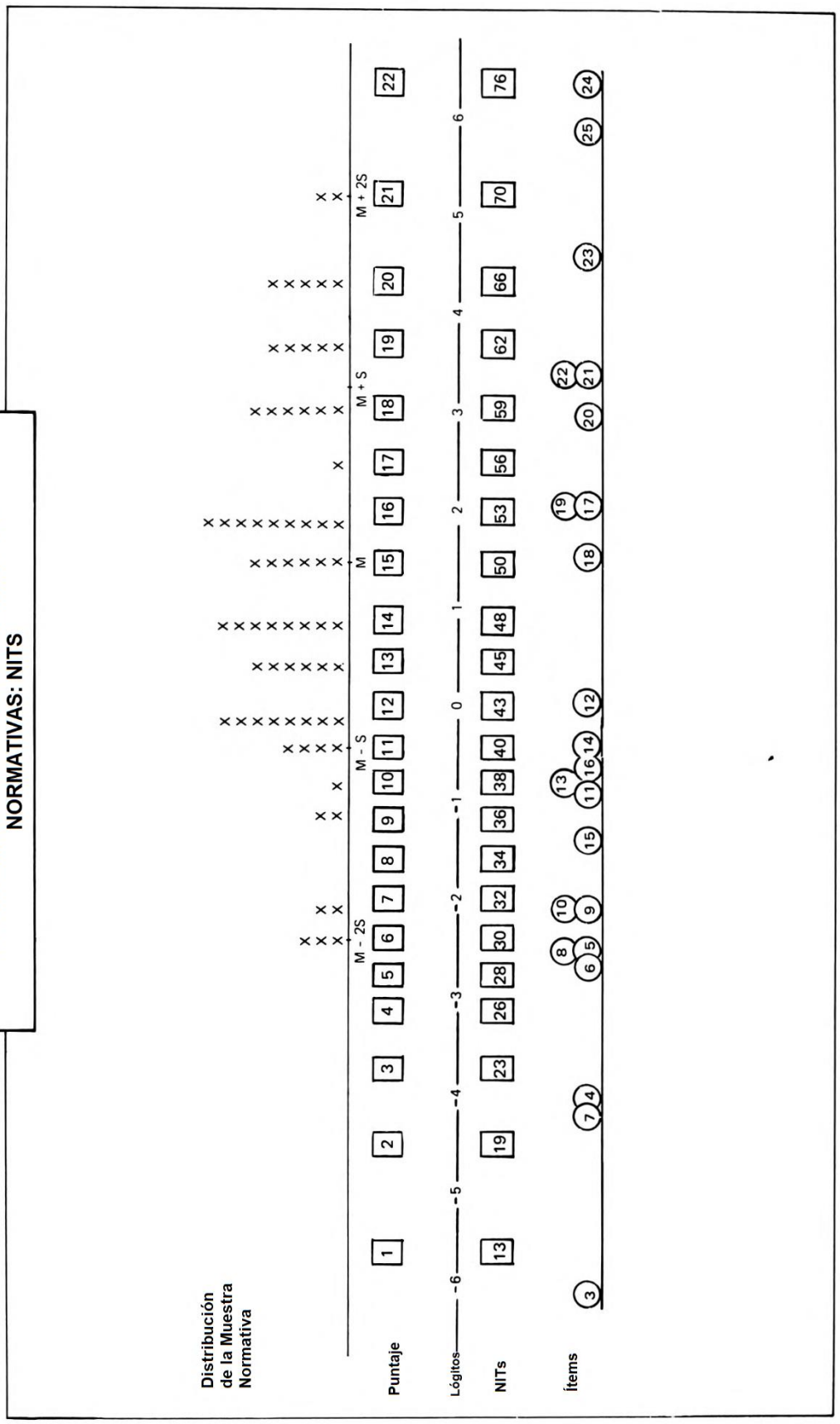
Se vuelven

$$B = [(D_1\, d_2 - D_2\, d_1) + (D_2 - D_1)b]/(d_2 - d_1) \qquad \textbf{[8.6.1]}$$

y

$$D = [(D_1\, d_2 - D_2\, d_1) + (D_2 - D_1)d]/(d_2 - d_1)$$

**FIGURA 8.5.1**
**ESCALAMIENTO DE LA PKCB EN UNIDADES NORMATIVAS: NITS**



Para aplicar este método a nuestro ejemplo de la PCKB, designaremos un nivel basal en la mediana de 3 golpes donde $d_1 = -3.4$ lógitos y un nivel de competencia en la mediana de 5 golpes donde $d_2 = 1.4$ lógitos. Luego, nos organizaremos para reportar ambos criterios como $D_1 = 30$ para el nivel basal y $D_2 = 50$ para el nivel de competencia usando:

$$\alpha = [30(1.4) - 50(-3.4)]/[1.4 - (-3.4)]$$
$$= (42 + 170)/4.8$$
$$= 44.2$$

y

$$\gamma = (50 - 30)/[1.4 - (-3.4)]$$
$$= 20/4.8$$
$$= 4.2$$

con lo que se define nuestra nueva escala sustantiva en SITs como

$$B = 44.2 + 4.2b \qquad \textbf{[8.6.2]}$$

y

$$D = 44.2 + 4.2d$$

Esta escala transforma la mediana de 3 golpes en $d_1 = -3.4$ lógitos a $D_1 = 44.2 + 4.2\ (-3.4) = 30$ SIT y la mediana de 5 golpes en $d_2 = 1.4$ lógitos a $D_2 = 44.2 + 4.2\ (1.4) = 50$ SIT.

En la figura 8.6.1 mostramos los ítems de la PCKB y una definición sustantiva de la variable de la PCK, marcando las posiciones de la escala de cada número en la mediana de golpes. Los puntajes de habilidad en lógitos y en SIT aparecen debajo de esta definición sustantiva.

## 8.7 UNIDADES DE ESCALAMIENTO DE PROBABILIDAD DE RESPUESTA: CHIPS

Si estamos interesados en utilizar nuestra prueba para predecir tasas de respuesta de desempeño exitoso, entonces una "escala con referencia a un desempeño" puede ser útil. Las unidades de probabilidad de respuesta o CHlPs (por Chicago Probability Units) sirven para identificar los cambios a través de probabilidades de respuesta de 0.10, 0.25, 0.50, 0.75 y 0.90, con múltiplos fáciles de recordar en la variable, como 5, 10, 20 o 25.

A partir del modelo de respuesta

$$p = \exp(b - d)/[1 + \exp(b - d)]$$

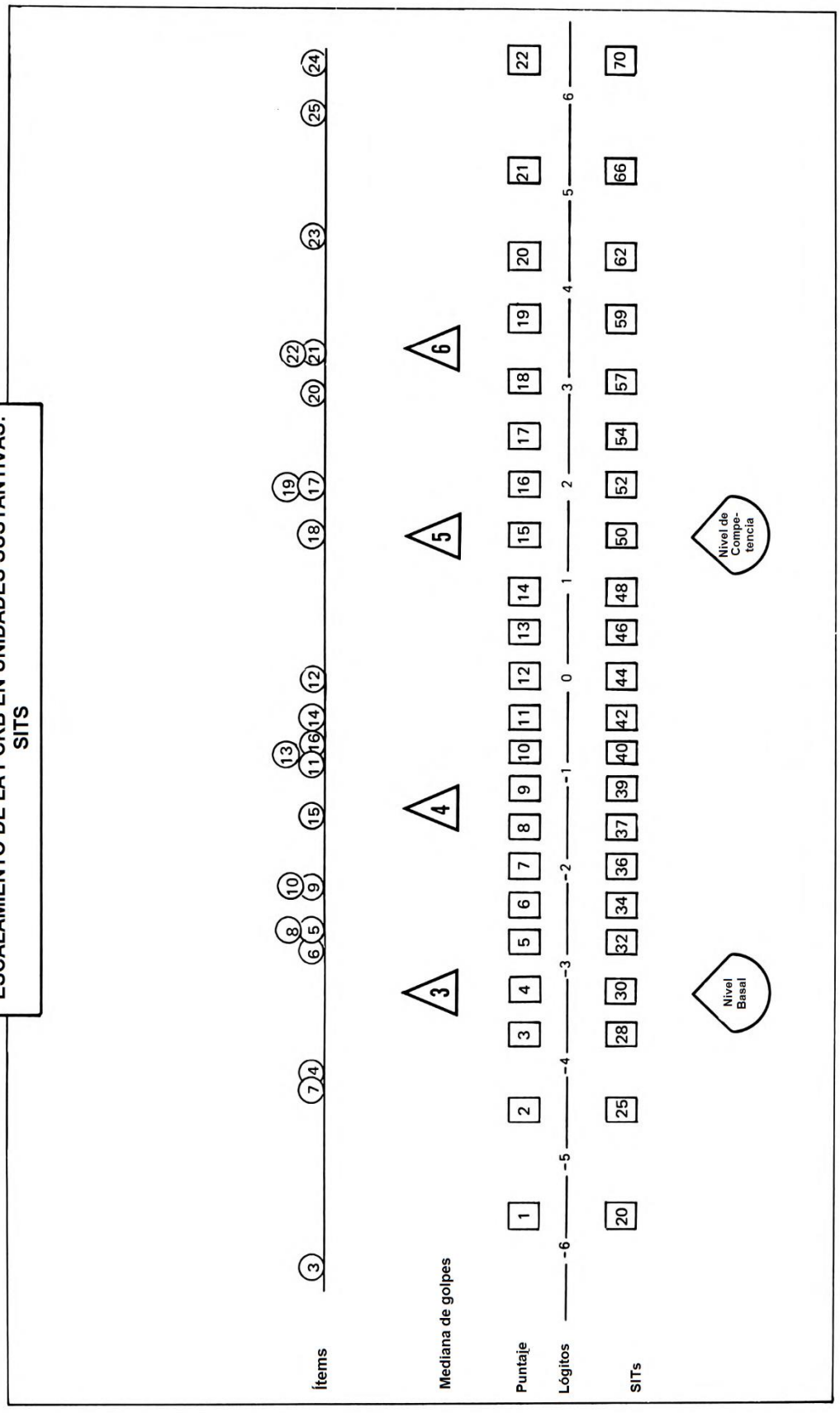
podemos determinar las diferencias (b – d) entre la habilidad de la persona y la dificultad del ítem que corresponden con las probabilidades de respuesta 0.10, 0.25, 0.50, 0.75 y 0.90. Al resolver para (b – d) en lógitos tenemos:

$$(b - d) = \ell n[p/(1 - p)]$$

con lo que se llega a:

| Probabilidad de éxito p | Diferencia entre la habilidad de la persona y la dificultad del ítem en lógitos. b - d |
|---|---|
| 0.10 | -2.2 |
| 0.25 | -1.1 |
| 0.50 | 0.0 |
| 0.75 | 1.1 |
| 0.90 | 2.2 |

**FIGURA 8.6.1**
**ESCALAMIENTO DE LA PCKB EN UNIDADES SUSTANTIVAS: SITS**



Para determinar una nueva escala de esta manera usamos

$$B = \alpha + \gamma(b - c) \quad \textbf{[8.7.1]}$$

en el cual

c = puede elegirse como normativa o sustantiva para la ubicación en la escala de lógitos

$\gamma$ = un múltiplo atractivo de 1 / 1.1 = 0.91 de forma de que 5, 10, 20 o 25 lleven a los valores de $\gamma$ en 4.55, 9.1, 18.2 o 22.75

y

$\alpha$ = 50, 100 o 500

Para la PCKB podríamos hacer una elección normativa de c = 1.3 lógitos en la media de lógitos del grupo normativo de 68 personas. También podríamos establecer $\gamma$ =4.55 que nos da un espaciamiento CHIP de 5 y elegir $\alpha$ para ubicar la media normativa en 50. En consecuencia nuestra formulación de escala CHIP se convierte de esta forma:

$$B = 50 + 4.55(b - 1.3) \quad \textbf{[8.7.2]}$$
$$= 44.09 + 4.55b$$
$$= 44.1 + 4.6b$$

Tenga en cuenta que cuando b se ubica en la media del grupo normativo, entonces

$$B = 44.1 + 4.6(1.3) = 50 \text{ CHIPs}$$

Nuestra elección de $\gamma$=4.55 produce las siguientes relaciones entre las posiciones relativas de una persona en B y un ítem en D:

| Probabilidad de éxito p | Diferencia entre la habilidad de la persona y la dificultad del ítem en CHIPS. B - D |
|---|---|
| 0.10 | -10 |
| 0.25 | -5 |
| 0.50 | 0 |
| 0.75 | 5 |
| 0.90 | 10 |

Por lo tanto, esperamos que cuando una persona confronte a un ítem que esté 10 CHIPs por debajo de su habilidad, entonces la probabilidad de una respuesta correcta es aproximadamente de 0.90. Cuando está 5 CHIPs por debajo, la tasa de éxito pronosticado es de 0.75. Por otro lado, si un ítem es 5 CHIPs más difícil que para la habilidad de la persona, esperamos que la tasa de éxito se reduzca a 0.25 y cuando la persona está en desventaja de 10 CHIPs, esperamos éxito solamente 0.10 de las veces.

Si tuviéramos que hacer una elección sustantiva de una ubicación en la escala, podríamos usar la mediana de 5 golpes en la PCK, que se encuentra en 1.4 lógitos como nuestra ubicación de referencia en lugar de la media de la muestra normativa en 1.3 lógitos. Entonces nuestra formulación de escala CHIP se convertiría en:

$$D = 50 + 4.55(d - 1.4) \quad \textbf{[8.7.3]}$$
$$= 43.63 + 4.55d$$
$$= 43.6 + 4.6d$$

y entonces

$$B = 43.6 + 4.6b$$

De esta forma, cuando b se encuentre en la mediana de 5 golpes en 1.4 lógitos, se tiene:

$$B = 43.6 + 4.6b(1.4) = 50 \text{ CHIPs}$$

La tabla 8.7.1 pone juntas las escalas en lógitos, NIT, SIT y CHIP para la PCKB.

**TABLA 8.7.1**

**PUNTAJE DE PCKB, MEDIDAS Y ERRORES EN LÓGITOS, NITS, SITS Y CHIPS**

| Resultado de la prueba | Habilidad de la persona | | | | Error de Medida | | | |
|---|---|---|---|---|---|---|---|---|
| | Lógitos | NITs | SITs | CHIPs | Lógitos | NITs | SITs | CHIPs |
| 22 | 6.26 | 76 | 70 | 73 | 1.20 | 6 | 5 | 6 |
| 21 | 5.15 | 70 | 66 | 68 | 0.99 | 5 | 4 | 5 |
| 20 | 4.31 | 66 | 62 | 64 | 0.89 | 5 | 4 | 4 |
| 19 | 3.60 | 62 | 59 | 61 | 0.83 | 4 | 3 | 4 |
| 18 | 2.99 | 59 | 57 | 58 | 0.78 | 4 | 3 | 4 |
| 17 | 2.42 | 56 | 54 | 55 | 0.76 | 4 | 3 | 3 |
| 16 | 1.88 | 53 | 52 | 53 | 0.75 | 4 | 3 | 3 |
| 15 | 1.35 | 50 | 50 | 50 | 0.74 | 4 | 3 | 3 |
| 14 | 0.84 | 48 | 48 | 48 | 0.73 | 4 | 3 | 3 |
| 13 | 0.35 | 45 | 46 | 46 | 0.70 | 4 | 3 | 3 |
| 12 | −0.10 | 43 | 44 | 44 | 0.67 | 4 | 3 | 3 |
| 11 | −0.51 | 40 | 42 | 42 | 0.65 | 3 | 3 | 3 |
| 10 | −0.90 | 38 | 40 | 40 | 0.63 | 3 | 3 | 3 |
| 9 | −1.28 | 36 | 39 | 38 | 0.62 | 3 | 3 | 3 |
| 8 | −1.65 | 34 | 37 | 37 | 0.62 | 3 | 3 | 3 |
| 7 | −2.02 | 32 | 36 | 35 | 0.63 | 3 | 3 | 3 |
| 6 | −2.42 | 30 | 34 | 33 | 0.65 | 3 | 3 | 3 |
| 5 | −2.85 | 28 | 32 | 31 | 0.69 | 4 | 3 | 3 |
| 4 | −3.33 | 26 | 30 | 29 | 0.74 | 4 | 3 | 3 |
| 3 | −3.90 | 23 | 28 | 26 | 0.82 | 4 | 3 | 4 |
| 2 | −4.65 | 19 | 25 | 23 | 0.95 | 5 | 4 | 4 |
| 1 | −5.75 | 13 | 20 | 18 | 1.23 | 7 | 5 | 6 |

[8.5.2] NITs $B = 43.2 + 5.3b$ , $SE(B) = 5.3\ SE(b)$

[8.6.2] SITs $B = 44.2 + 4.2b$ , $SE(B) = 4.2\ SE(b)$

[8.7.2] CHIPs $B = 44.1 + 4.6b$ , $SE(B) = 4.6\ SE(b)$

## 8.8 FORMULARIOS DE INFORMES

El uso del modelo de Rasch en la construcción de una prueba puede facilitar su interpretación. Vamos a ilustrar esto con un formulario de informe desarrollado para la variable PCK.

La Figura 8.8.1 proporciona un mapa de la variable PCK. Este mapa muestra todos los datos reunidos hasta ahora: los ítems PCKB posicionados a lo largo de la variable según sus niveles de dificultad, los criterios sustantivos del número de golpes, retrocesos y distancias entre bloques y la información normativa de la mediana de edades para los niños, así como la media y la desviación estándar para los adultos. El mapa muestra hasta qué punto se ha definido la variable PCK y cómo varias medidas posibles de PCK se relacionan con consideraciones sustantivas y normativas. Un formulario de informe se puede desarrollar a partir de este mapa.

La figura 8.8.2 es un formulario de informe para interpretar el desempeño individual en el PCKB. Esta forma, que podría usarse para un solo individuo o para un grupo completo, muestra el rendimiento de las personas 12M y 88M, así como un registro de respuesta idéntico al puntaje de 88M pero diseñado para mostrar un patrón "dormilón" de varias fallas inesperadas.

Observe que Persona 12M con su puntaje de 5 se ubica en -2.8 lógitos en la variable de la PCK. Esto lo pone a medio camino entre 3 y 4 golpes definidos sustantivamente y en la posición de la mediana normativa de 5 años de edad. Sin embargo, la Persona 88M ubicada en 3.0 lógitos funciona sustantivamente en 6 golpes y aproximadamente a una desviación estándar por arriba de la media normativa del adulto.

En muchos casos, será útil detectar el desajuste inmediatamente después de registrar las respuestas de la persona. El formulario de informe de la figura 8.8.2 es ideal para este propósito. Para ello debemos registrar la respuesta correcta o incorrecta a cada ítem en su posición sobre la variable y también la posición resultante de la persona en la misma variable. El desajuste puede ser estimado directamente de este formulario de respuesta por medio de una Regla de Desajuste.

La figura 8.8.3 muestra una Regla de Desajuste escalada en lógitos. Está marcada para indicar las desviaciones en lógitos y el correspondiente índice de desajuste $y^2 = (z^2 - 1)$ a izquierda y derecha de su valor central. Observe que las desviaciones de las respuestas inesperadas de 1, 2, 3 y 4 lógitos corresponden con una $y^2$ de 2, 6, 19 y 54 respectivamente. Si colocamos el centro de la regla en el punto de la variable donde se encuentra la persona y comparamos las marcas de la regla con la respuesta que hace la persona a cada ítem, podremos calcular, de un vistazo, el desajuste del registro de la persona.

Siempre que observemos una respuesta inesperada, es decir, un "0" para una respuesta incorrecta en la región fácil a la izquierda o un "1" para una respuesta correcta en la región de difícil a la derecha, entonces los $y^2$ correspondientes en la regla se adicionan para formar la suma $Q = \Sigma\, y^2$ solo por las respuestas inesperadas del registro. Esta suma Q dividida por la raíz cuadrada del número total de ítems L en la prueba produce el estadístico de desajuste

$$U = Q/L^{1/2} \qquad \textbf{[7.8.1]}$$

si el registro se ajusta al modelo de respuesta, entonces se distribuye aproximadamente como

$$U \sim N(0,2)$$

Este estadístico es fácil de calcular y puede utilizarse para evaluar el desajuste. Si $U > 5$ la probabilidad de que el registro sea aceptable se reduce a menos de 0.01 y parece razonable cuestionar la validez de dicho registro.

En una aplicación práctica con un lote de registros para evaluar, es más razonable comenzar con el registro para el que U es máximo y ver si la fuente de invalidez puede ser identificada y

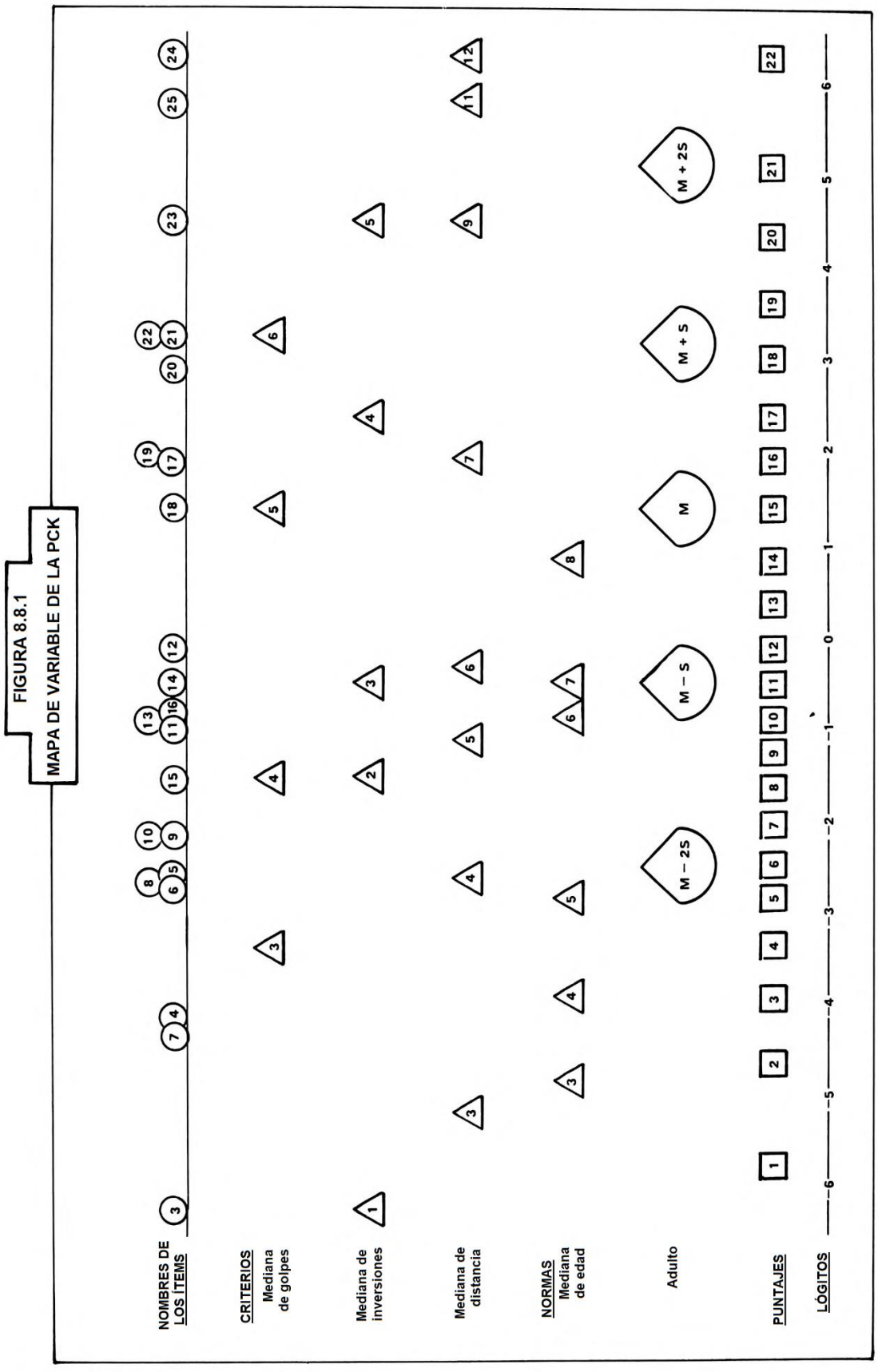


FIGURA 8.8.1
MAPA DE VARIABLE DE LA PCK
NOMBRES DE LOS ÍTEMS
CRITERIOS
Mediana de golpes
Mediana de inversiones
Mediana de distancia
NORMAS
Mediana de edad
Adulto
PUNTAJES
LÓGITOS
M − 2S
M − S
M
M + S
M + 2S

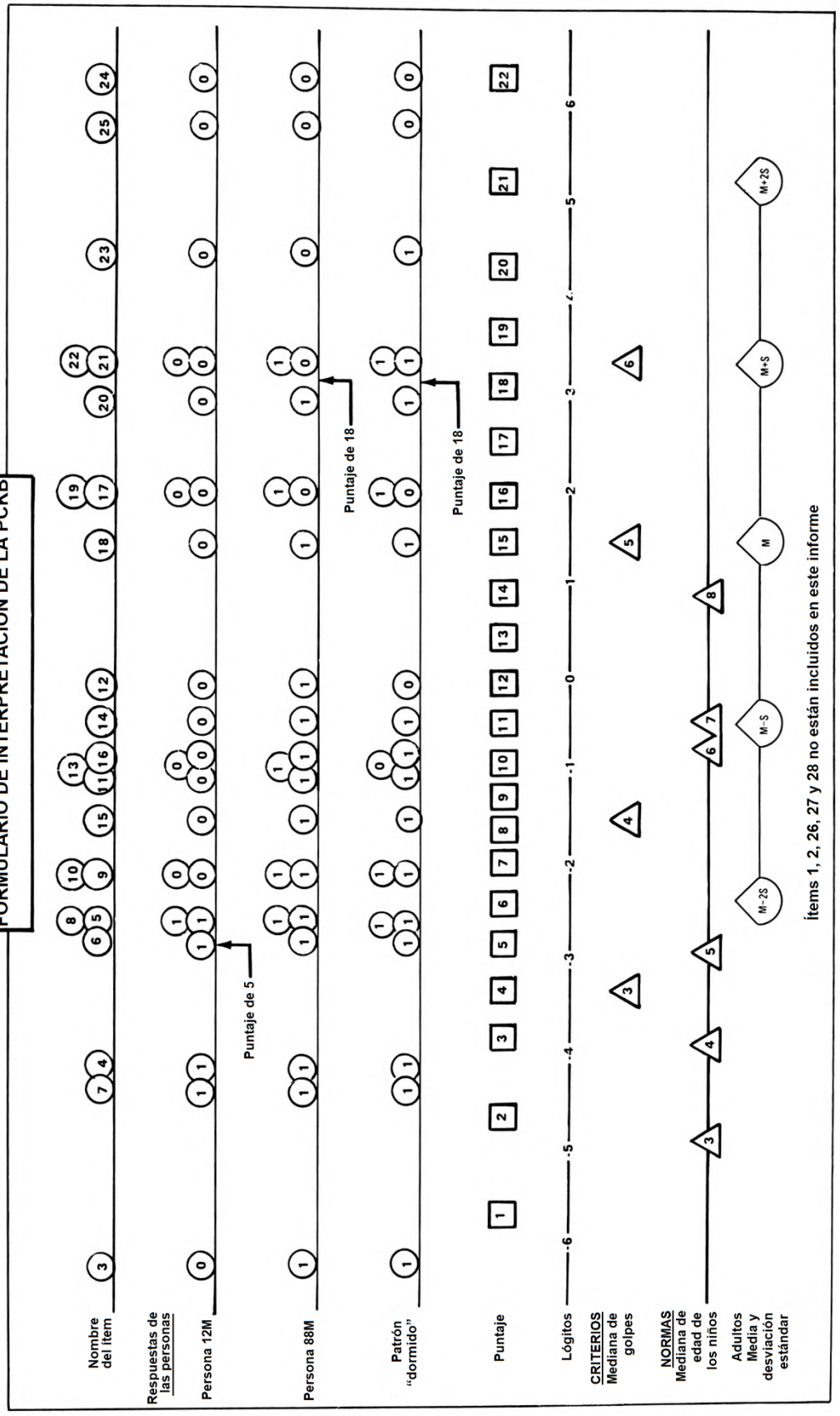


FIGURA 8.8.2
FORMULARIO DE INTERPRETACIÓN DE LA PCKB
Nombre del ítem
Respuestas de las personas
Persona 12M
Puntaje de 5
Persona 88M
Puntaje de 18
Patrón "dormido"
Puntaje de 18
Puntaje
Lógitos
CRITERIOS Mediana de golpes
NORMAS Mediana de edad de los niños
Adultos Media y desviación estándar
M−2S
M−S
M
M+S
M+2S
Ítems 1, 2, 26, 27 y 28 no están incluidos en este informe

buscar la forma de enfrentarla. Los casos siguientes pueden ser manejados en función de los valores de U hasta descubrir todas las explicaciones útiles sobre la invalidez que implica alcanzar U > 5.

La figura 8.8.2 muestra el segmento de prueba de 23 ítems de la PCKB en orden creciente de dificultad, junto con tres registros de respuesta. Estos registros muestran las respuestas correctas e incorrectas en los 23 ítems. Para evaluar el ajuste del registro de la persona 12M, colocamos el centro de la Regla de Desajuste en la flecha que marca su posición en –2.8 lógitos determinados por su puntaje de 5. Su respuesta incorrecta al Ítem 3 en –6.2 lógitos produce una $y^2$ de aproximadamente 30 para Q= 30 y U=$30/23^{½}$=6.3. Esto corresponde a t = 4.5 que está muy próximo del valor más exacto de t dado en la Tabla 7.8.3. Nuevamente, vemos que es muy improbable que esta respuesta llegue a aceptarse como parte de una medida válida para la Persona 12M.

La Regla de Desajuste también se aplicó a la Persona 88M y al patrón "dormilón" con el mismo puntaje de 18. El patrón para la Persona 88M produce un registro de respuesta que conduce a $\Sigma y^2 = 2$ y U= 0.4. Los patrones "dormilones", sin embargo, producen esto:

$$\sum y^2 = 46 + 20 + 2 + 3 = 71$$

y entonces

$$U = 71/23^{1/2} = 14.8$$

Todos estos resultados se concentran en la Tabla 8.8.1

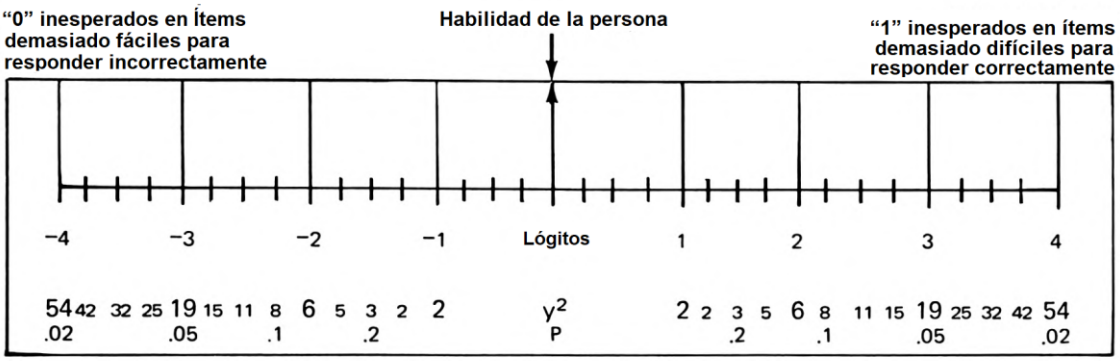
**TABLA 8.8.1**

**ANALISIS RÁPIDO DE VALIDEZ DEL REGISTRO DE RESPUESTA**

| Persona | Puntaje r | Suma de respuestas inesperadas $Q = \Sigma y^2$ | Estadístico de ajuste $U = Q/L^{½}$ |
|---|---|---|---|
| 12M | 5 | 30 | 6.3* |
| 88M | 18 | 2 | 0.4 |
| Patrón "Dormilón" | 18 | 46 + 20 + 2 + 3 = 71 | 14.8* |
| | L = 23 | | * Desajuste |

**FIGURA 8.8.3**

**REGLA DE DESAJUSTE**




**CÓMO UTILIZAR LA REGLA DE DESAJUSTE**

1. Coloque los elementos en el formulario de registro métrico correspondiente a la métrica de la regla.
2. Registre las respuestas de la persona a los ítems en el formulario de registro.
3. Ubique la posición de habilidad de la persona en el formulario de registro contando el puntaje r y ubicándolo entre las posiciones r y (r + 1) de los ítems.
4. Coloque el centro de la Regla de Desajuste en la posición de la habilidad de la persona.
5. Sume $y^2$ para todas las respuestas inesperadas,"0"a la izquierda y "1"a la derecha para obtener Q.
6. Haga L igual a la cantidad total de elementos.
7. Calcule el estadístico de desajuste $U = Q/L^{½}$.
8. Si U> 5 vuelva a revisar el registro de la persona en busca de fuentes de invalidez.

P es la improbabilidad de cada respuesta.

# APÉNDICES


| | |
|---|---|
| TABLA A | 212 y 213 |
| TABLA B | 214 y 215 |
| TABLA C | 216 |

## TABLA A

## HABILIDAD RELATIVA $x_{fw}$ PARA PRUEBAS UNIFORMES EN LÓGITOS

| Puntaje relativo f > 0.50 | Ancho de la prueba w | | | | | | | | Puntaje relativo f < 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| .50 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | .50 |
| .51 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | .49 |
| .52 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | .48 |
| .53 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | .47 |
| .54 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.3 | .46 |
| .55 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.3 | 0.4 | 0.4 | .45 |
| .56 | 0.2 | 0.3 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | .44 |
| .57 | 0.3 | 0.3 | 0.3 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | .43 |
| .58 | 0.3 | 0.3 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.7 | .42 |
| .59 | 0.4 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.7 | 0.7 | .41 |
| .60 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.7 | 0.7 | 0.8 | .40 |
| .61 | 0.5 | 0.5 | 0.5 | 0.6 | 0.7 | 0.7 | 0.8 | 0.9 | .39 |
| .62 | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | .38 |
| .63 | 0.5 | 0.6 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | 1.1 | .37 |
| .64 | 0.6 | 0.6 | 0.7 | 0.7 | 0.8 | 0.9 | 1.1 | 1.2 | .36 |
| .65 | 0.6 | 0.7 | 0.7 | 0.8 | 0.9 | 1.0 | 1.1 | 1.3 | .35 |
| .66 | 0.7 | 0.7 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 | .34 |
| .67 | 0.7 | 0.8 | 0.8 | 0.9 | 1.0 | 1.2 | 1.3 | 1.4 | .33 |
| .68 | 0.8 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 | 1.4 | 1.5 | .32 |
| .69 | 0.8 | 0.9 | 0.9 | 1.0 | 1.2 | 1.3 | 1.4 | 1.6 | .31 |
| .70 | 0.9 | 0.9 | 1.0 | 1.1 | 1.2 | 1.4 | 1.5 | 1.7 | .30 |
| .71 | 0.9 | 1.0 | 1.0 | 1.2 | 1.3 | 1.4 | 1.6 | 1.8 | .29 |
| .72 | 1.0 | 1.0 | 1.1 | 1.2 | 1.4 | 1.5 | 1.7 | 1.9 | .28 |
| .73 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | 1.6 | 1.8 | 2.0 | .27 |
| .74 | 1.1 | 1.1 | 1.2 | 1.3 | 1.5 | 1.7 | 1.9 | 2.1 | .26 |
| .75 | 1.1 | 1.2 | 1.3 | 1.4 | 1.6 | 1.7 | 1.9 | 2.1 | .25 |
| .76 | 1.2 | 1.2 | 1.3 | 1.5 | 1.6 | 1.8 | 2.0 | 2.2 | .24 |
| .77 | 1.2 | 1.3 | 1.4 | 1.5 | 1.7 | 1.9 | 2.1 | 2.3 | .23 |
| .78 | 1.3 | 1.4 | 1.5 | 1.6 | 1.8 | 2.0 | 2.2 | 2.4 | .22 |
| .79 | 1.3 | 1.4 | 1.5 | 1.7 | 1.9 | 2.1 | 2.3 | 2.5 | .21 |
| .80 | 1.4 | 1.5 | 1.6 | 1.8 | 1.9 | 2.2 | 2.4 | 2.6 | .20 |
| .81 | 1.5 | 1.6 | 1.7 | 1.8 | 2.0 | 2.2 | 2.5 | 2.7 | .19 |
| .82 | 1.5 | 1.6 | 1.7 | 1.9 | 2.1 | 2.3 | 2.6 | 2.8 | .18 |
| .83 | 1.6 | 1.7 | 1.8 | 2.0 | 2.2 | 2.4 | 2.7 | 2.9 | .17 |
| .84 | 1.7 | 1.8 | 1.9 | 2.1 | 2.3 | 2.5 | 2.8 | 3.0 | .16 |
| .85 | 1.8 | 1.8 | 2.0 | 2.2 | 2.4 | 2.6 | 2.9 | 3.2 | .15 |
| .86 | 1.8 | 1.9 | 2.1 | 2.3 | 2.5 | 2.7 | 3.0 | 3.3 | .14 |
| .87 | 1.9 | 2.0 | 2.2 | 2.4 | 2.6 | 2.8 | 3.1 | 3.4 | .13 |
| .88 | 2.0 | 2.1 | 2.3 | 2.5 | 2.7 | 2.9 | 3.2 | 3.5 | .12 |
| .89 | 2.1 | 2.2 | 2.4 | 2.6 | 2.8 | 3.1 | 3.3 | 3.7 | .11 |
| .90 | 2.2 | 2.3 | 2.5 | 2.7 | 2.9 | 3.2 | 3.5 | 3.8 | .10 |
| .91 | 2.3 | 2.4 | 2.6 | 2.8 | 3.1 | 3.3 | 3.6 | 3.9 | .09 |
| .92 | 2.5 | 2.6 | 2.7 | 2.9 | 3.2 | 3.5 | 3.8 | 4.1 | .08 |
| .93 | 2.6 | 2.7 | 2.9 | 3.1 | 3.4 | 3.6 | 4.0 | 4.3 | .07 |
| .94 | 2.8 | 2.9 | 3.1 | 3.3 | 3.5 | 3.8 | 4.1 | 4.5 | .06 |
| .95 | 3.0 | 3.1 | 3.3 | 3.5 | 3.7 | 4.0 | 4.4 | 4.7 | .05 |
| .96 | 3.2 | 3.3 | 3.5 | 3.7 | 4.0 | 4.3 | 4.6 | 5.0 | .04 |
| .97 | 3.5 | 3.6 | 3.8 | 4.0 | 4.3 | 4.6 | 5.0 | 5.3 | .03 |
| .98 | 3.9 | 4.0 | 4.2 | 4.5 | 4.7 | 5.1 | 5.4 | 5.8 | .02 |
| .99 | 4.6 | 4.8 | 4.9 | 5.2 | 5.5 | 5.8 | 6.1 | 6.5 | .01 |
| f > 0.50 Medida $b_f = h + x_{fw}$ | | | | | | | | | f < 0.50 Medida $b_f = h - x_{fw}$ |

**Puntaje de la prueba: r** **Longitud de la prueba: L** **Puntaje relativo: f=r/L**

**Altura de la prueba:** $h = \sum_{1}^{L} d_i/L$ **Ancho de la prueba:** $w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)]$

## TABLA A
## HABILIDAD RELATIVA $x_{fw}$ PARA PRUEBAS UNIFORMES EN LÓGITOS
## (Continuación)

| Puntaje relativo f > 0.50 | Ancho de la prueba w | | | | | | | | Puntaje relativo f < 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | |
| .50 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | .50 |
| .51 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | .49 |
| .52 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.3 | .48 |
| .53 | 0.2 | 0.3 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | .47 |
| .54 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.6 | .46 |
| .55 | 0.4 | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.7 | 0.8 | .45 |
| .56 | 0.5 | 0.6 | 0.6 | 0.7 | 0.7 | 0.8 | 0.8 | 0.9 | .44 |
| .57 | 0.6 | 0.6 | 0.7 | 0.8 | 0.8 | 0.9 | 1.0 | 1.1 | .43 |
| .58 | 0.7 | 0.7 | 0.8 | 0.9 | 1.0 | 1.0 | 1.1 | 1.2 | .42 |
| .59 | 0.7 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | .41 |
| .60 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | .40 |
| .61 | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.7 | .39 |
| .62 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | 1.6 | 1.7 | 1.8 | .38 |
| .63 | 1.1 | 1.2 | 1.3 | 1.4 | 1.6 | 1.7 | 1.8 | 2.0 | .37 |
| .64 | 1.2 | 1.3 | 1.4 | 1.6 | 1.7 | 1.8 | 2.0 | 2.1 | .36 |
| .65 | 1.3 | 1.4 | 1.5 | 1.7 | 1.8 | 2.0 | 2.1 | 2.3 | .35 |
| .66 | 1.3 | 1.5 | 1.6 | 1.8 | 1.9 | 2.1 | 2.2 | 2.4 | .34 |
| .67 | 1.4 | 1.6 | 1.7 | 1.9 | 2.1 | 2.2 | 2.4 | 2.6 | .33 |
| .68 | 1.5 | 1.7 | 1.8 | 2.0 | 2.2 | 2.4 | 2.5 | 2.7 | .32 |
| .69 | 1.6 | 1.8 | 1.9 | 2.1 | 2.3 | 2.5 | 2.7 | 2.9 | .31 |
| .70 | 1.7 | 1.9 | 2.1 | 2.2 | 2.4 | 2.6 | 2.8 | 3.0 | .30 |
| .71 | 1.8 | 2.0 | 2.2 | 2.4 | 2.6 | 2.8 | 3.0 | 3.2 | .29 |
| .72 | 1.9 | 2.1 | 2.3 | 2.5 | 2.7 | 2.9 | 3.1 | 3.3 | .28 |
| .73 | 2.0 | 2.2 | 2.4 | 2.6 | 2.8 | 3.0 | 3.2 | 3.5 | .27 |
| .74 | 2.1 | 2.3 | 2.5 | 2.7 | 2.9 | 3.2 | 3.4 | 3.6 | .26 |
| .75 | 2.1 | 2.4 | 2.6 | 2.8 | 3.1 | 3.3 | 3.5 | 3.8 | .25 |
| .76 | 2.2 | 2.5 | 2.7 | 2.9 | 3.2 | 3.4 | 3.7 | 3.9 | .24 |
| .77 | 2.3 | 2.6 | 2.8 | 3.1 | 3.3 | 3.6 | 3.8 | 4.1 | .23 |
| .78 | 2.4 | 2.7 | 2.9 | 3.2 | 3.4 | 3.7 | 4.0 | 4.2 | .22 |
| .79 | 2.5 | 2.8 | 3.0 | 3.3 | 3.6 | 3.8 | 4.1 | 4.4 | .21 |
| .80 | 2.6 | 2.9 | 3.1 | 3.4 | 3.7 | 4.0 | 4.3 | 4.6 | .20 |
| .81 | 2.7 | 3.0 | 3.3 | 3.5 | 3.8 | 4.1 | 4.4 | 4.7 | .19 |
| .82 | 2.8 | 3.1 | 3.4 | 3.7 | 4.0 | 4.3 | 4.6 | 4.9 | .18 |
| .83 | 2.9 | 3.2 | 3.5 | 3.8 | 4.1 | 4.4 | 4.7 | 5.0 | .17 |
| .84 | 3.0 | 3.3 | 3.6 | 3.9 | 4.2 | 4.6 | 4.9 | 5.2 | .16 |
| .85 | 3.2 | 3.4 | 3.8 | 4.1 | 4.4 | 4.7 | 5.0 | 5.4 | .15 |
| .86 | 3.3 | 3.6 | 3.9 | 4.2 | 4.5 | 4.9 | 5.2 | 5.5 | .14 |
| .87 | 3.4 | 3.7 | 4.0 | 4.3 | 4.7 | 5.0 | 5.4 | 5.7 | .13 |
| .88 | 3.5 | 3.8 | 4.2 | 4.5 | 4.8 | 5.2 | 5.5 | 5.9 | .12 |
| .89 | 3.7 | 4.0 | 4.3 | 4.6 | 5.0 | 5.3 | 5.7 | 6.1 | .11 |
| .90 | 3.8 | 4.1 | 4.5 | 4.8 | 5.2 | 5.5 | 5.9 | 6.3 | .10 |
| .91 | 3.9 | 4.3 | 4.6 | 5.0 | 5.3 | 5.7 | 6.1 | 6.5 | .09 |
| .92 | 4.1 | 4.4 | 4.8 | 5.2 | 5.5 | 5.9 | 6.3 | 6.7 | .08 |
| .93 | 4.3 | 4.6 | 5.0 | 5.4 | 5.7 | 6.1 | 6.5 | 6.9 | .07 |
| .94 | 4.5 | 4.8 | 5.2 | 5.6 | 5.9 | 6.3 | 6.7 | 7.1 | .06 |
| .95 | 4.7 | 5.1 | 5.4 | 5.8 | 6.2 | 6.6 | 7.0 | 7.4 | .05 |
| .96 | 5.0 | 5.3 | 5.7 | 6.1 | 6.5 | 6.9 | 7.3 | 7.7 | .04 |
| .97 | 5.3 | 5.7 | 6.1 | 6.4 | 6.8 | 7.2 | 7.7 | 8.1 | .03 |
| .98 | 5.8 | 6.1 | 6.5 | 6.9 | 7.3 | 7.7 | 8.1 | 8.6 | .02 |
| .99 | 6.5 | 6.9 | 7.3 | 7.7 | 8.1 | 8.5 | 8.9 | 9.3 | .01 |
| f > 0.50 Medida $b_f = h + x_{fw}$ | | | | | | | | | f < 0.50 Medida $b_f = h + x_{fw}$ |

**Puntaje de la prueba: r** **Longitud de la prueba: L** **Puntaje relativo: f=r/L**

**Altura de la prueba:** $h = \sum_{i}^{L} d_i/L$ **Ancho de la prueba:** $w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)]$

## TABLA B

## COEFICIENTE DE ERROR $C_{fw}^{1/2}$ PARA PRUEBAS UNIFORMES EN LÓGITOS

| Puntaje relativo f > 0.50 | Ancho de la prueba w 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Puntaje relativo f < 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| .50 | 2.0 | 2,1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .50 |
| .51 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .49 |
| .52 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .48 |
| .53 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .47 |
| .54 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .46 |
| .55 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .45 |
| .56 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .44 |
| .57 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .43 |
| .58 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.9 | .42 |
| .59 | 2.1 | 2.1 | 2.2 | 2.3 | 2.5 | 2.6 | 2.7 | 2.9 | .41 |
| .60 | 2.1 | 2.1 | 2.2 | 2.3 | 2.5 | 2.6 | 2.7 | 2.9 | .40 |
| .61 | 2.1 | 2.1 | 2.2 | 2.3 | 2.5 | 2.6 | 2.8 | 2.9 | .39 |
| .62 | 2.1 | 2.1 | 2.2 | 2.3 | 2.5 | 2.6 | 2.8 | 2.9 | .38 |
| .63 | 2.1 | 2.1 | 2.2 | 2.4 | 2.5 | 2.6 | 2.8 | 2.9 | .37 |
| .64 | 2.1 | 2.2 | 2.2 | 2.4 | 2.5 | 2.6 | 2.8 | 2.9 | .36 |
| .65 | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.8 | 2.9 | .35 |
| .66 | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.8 | 2.9 | .34 |
| .67 | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.7 | 2.8 | 2.9 | .33 |
| .68 | 2.2 | 2.2 | 2.3 | 2.4 | 2.5 | 2.7 | 2.8 | 3.0 | .32 |
| .69 | 2.2 | 2.2 | 2.3 | 2.4 | 2.6 | 2.7 | 2.8 | 3.0 | .31 |
| .70 | 2.2 | 2.3 | 2.3 | 2.4 | 2.6 | 2.7 | 2.8 | 3.0 | .30 |
| .71 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 3.0 | .29 |
| .72 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.9 | 3.0 | .28 |
| .73 | 2.3 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.9 | 3.0 | .27 |
| .74 | 2.3 | 2.3 | 2.4 | 2.5 | 2.6 | 2.8 | 2.9 | 3.0 | .26 |
| .75 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 | .25 |
| .76 | 2.4 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.1 | .24 |
| .77 | 2.4 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 3.0 | 3.1 | .23 |
| .78 | 2.4 | 2.5 | 2.6 | 2.6 | 2.8 | 2.9 | 3.0 | 3.1 | .22 |
| .79 | 2.5 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 | 3.1 | .21 |
| .80 | 2.5 | 2.6 | 2.6 | 2.7 | 2.8 | 2.9 | 3.1 | 3.2 | .20 |
| .81 | 2.6 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 | 3.1 | 3.2 | .19 |
| .82 | 2.6 | 2.7 | 2.7 | 2.8 | 2.9 | 3.0 | 3.1 | 3.2 | .18 |
| .83 | 2.7 | 2.7 | 2.8 | 2.9 | 3.0 | 3.1 | 3.2 | 3.3 | .17 |
| .84 | 2.7 | 2.8 | 2.9 | 2.9 | 3.0 | 3.1 | 3.2 | 3.3 | .16 |
| .85 | 2.8 | 2.9 | 2.9 | 3.0 | 3.1 | 3.2 | 3.3 | 3.4 | .15 |
| .86 | 2.9 | 2.9 | 3.0 | 3.1 | 3.2 | 3.3 | 3.4 | 3.4 | .14 |
| .87 | 3.0 | 3.0 | 3.1 | 3.2 | 3.2 | 3.3 | 3.4 | 3.5 | .13 |
| .88 | 3.1 | 3.1 | 3.2 | 3.3 | 3.3 | 3.4 | 3.5 | 3.6 | .12 |
| .89 | 3.2 | 3.2 | 3.3 | 3.4 | 3.4 | 3.5 | 3.6 | 3.7 | .11 |
| .90 | 3.3 | 3.4 | 3.4 | 3.5 | 3.6 | 3.7 | 3.7 | 3.8 | .10 |
| .91 | 3.5 | 3.5 | 3.6 | 3.7 | 3.7 | 3.8 | 3.9 | 3.9 | .09 |
| .92 | 3.7 | 3.7 | 3.8 | 3.8 | 3.9 | 4.0 | 4.0 | 4.1 | .08 |
| .93 | 3.9 | 4.0 | 4.0 | 4.1 | 4.1 | 4.2 | 4.3 | 4.3 | .07 |
| .94 | 4.2 | 4.2 | 4.3 | 4.3 | 4.4 | 4.5 | 4.5 | 4.6 | .06 |
| .95 | 4.6 | 4.6 | 4.7 | 4.7 | 4.8 | 4.8 | 4.9 | 4.9 | .05 |
| .96 | 5.1 | 5.1 | 5.2 | 5.2 | 5.3 | 5.3 | 5.4 | 5.4 | .04 |
| .97 | 5.9 | 5.9 | 5.9 | 6.0 | 6.0 | 6.0 | 6.1 | 6.1 | .03 |
| .98 | 7.1 | 7.2 | 7.2 | 7.2 | 7.3 | 7.3 | 7.3 | 7.4 | .02 |
| .99 | 10.1 | 10.1 | 10.1 | 10.1 | 10.1 | 10.2 | 10.2 | 10.2 | .01 |
| f > .50 | | | | | | | | | f < .50 |

**Puntaje de la prueba: r** **Longitud de la prueba: L** **Puntaje relativo: f=r/L**

**Altura de la prueba:** $h = \sum_{1}^{L} d_i/L$ **Ancho de la prueba:** $w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)]$

**Error estándar:** $s_{fw} = C_{fw}^{1/2} / L^{1/2}$

## TABLA B

## COEFICIENTE DE ERROR $C_{fw}^{1/2}$ PARA PRUEBAS UNIFORMES EN LÓGITOS (Continuación)

| Puntaje relativo f > 0.50 | Ancho de la prueba w | | | | | | | | Puntaje relativo f < 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | |
| .50 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .50 |
| .51 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .49 |
| .52 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .48 |
| .53 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .47 |
| .54 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .46 |
| .55 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .45 |
| .56 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .44 |
| .57 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .43 |
| .58 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .42 |
| .59 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .41 |
| .60 | 2.9 | 3.0 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | .40 |
| .61 | 2.9 | 3.1 | 3.2 | 3.3 | 3.5 | 3.6 | 3.8 | 3.9 | .39 |
| .62 | 2.9 | 3.1 | 3.2 | 3.3 | 3.5 | 3.6 | 3.8 | 3.9 | .38 |
| .63 | 2.9 | 3.1 | 3.2 | 3.3 | 3.5 | 3.6 | 3.8 | 3.9 | .37 |
| .64 | 2.9 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .36 |
| .65 | 2.9 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .35 |
| .66 | 2.9 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .34 |
| .67 | 2.9 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .33 |
| .68 | 3.0 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .32 |
| .69 | 3.0 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .31 |
| .70 | 3.0 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .30 |
| .71 | 3.0 | 3.1 | 3.3 | 3.4 | 3.5 | 3.6 | 3.8 | 3.9 | .29 |
| .72 | 3.0 | 3.1 | 3.3 | 3.4 | 3.5 | 3.7 | 3.8 | 3.9 | .28 |
| .73 | 3.0 | 3.1 | 3.3 | 3.4 | 3.5 | 3.7 | 3.8 | 3.9 | .27 |
| .74 | 3.0 | 3.2 | 3.3 | 3.4 | 3.5 | 3.7 | 3.8 | 3.9 | .26 |
| .75 | 3.0 | 3.2 | 3.3 | 3.4 | 3.6 | 3.7 | 3.8 | 3.9 | .25 |
| .76 | 3.1 | 3.2 | 3.3 | 3.4 | 3.6 | 3.7 | 3.8 | 3.9 | .24 |
| .77 | 3.1 | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | .23 |
| .78 | 3.1 | 3.2 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | .22 |
| .79 | 3.1 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 4.0 | .21 |
| .80 | 3.2 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.9 | 4.0 | .20 |
| .81 | 3.2 | 3.3 | 3.4 | 3.5 | 3.7 | 3.8 | 3.9 | 4.0 | .19 |
| .82 | 3.2 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | .18 |
| .83 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | .17 |
| .84 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.9 | 4.0 | 4.1 | .16 |
| .85 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | 4.1 | .15 |
| .86 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | 4.1 | .14 |
| .87 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | 4.1 | 4.2 | .13 |
| .88 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | 4.1 | 4.1 | 4.2 | .12 |
| .89 | 3.7 | 3.8 | 3.9 | 4.0 | 4.0 | 4.1 | 4.2 | 4.3 | .11 |
| .90 | 3.8 | 3.9 | 4.0 | 4.1 | 4.1 | 4.2 | 4.3 | 4.4 | .10 |
| .91 | 3.9 | 4.0 | 4.1 | 4.2 | 4.3 | 4.3 | 4.4 | 4.5 | .09 |
| .92 | 4.1 | 4.2 | 4.3 | 4.3 | 4.4 | 4.5 | 4.6 | 4.6 | .08 |
| .93 | 4.3 | 4.4 | 4.5 | 4.5 | 4.6 | 4.7 | 4.7 | 4.8 | .07 |
| .94 | 4.6 | 4.6 | 4.7 | 4.8 | 4.8 | 4.9 | 5.0 | 5.0 | .06 |
| .95 | 4.9 | 5.0 | 5.0 | 5.1 | 5.2 | 5.2 | 5.3 | 5.3 | .05 |
| .96 | 5.4 | 5.5 | 5.5 | 5.6 | 5.6 | 5.7 | 5.7 | 5.8 | .04 |
| .97 | 6.1 | 6.2 | 6.2 | 6.3 | 6.3 | 6.3 | 6.4 | 6.4 | .03 |
| .98 | 7.4 | 7.4 | 7.4 | 7.5 | 7.5 | 7.5 | 7.6 | 7.6 | .02 |
| .99 | 10.2 | 10.2 | 10.3 | 10.3 | 10.3 | 10.3 | 10.4 | 10.4 | .01 |
| f > .50 | | | | | | | | | f < .50 |

Puntaje de la prueba: r

Longitud de la prueba: L

Puntaje relativo: f=r/L

Altura de la prueba: $h = \sum_{i}^{L} d_i/L$

Ancho de la prueba: $w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)]$

Error estándar: $s_{fw} = C_{fw}^{1/2} / L^{1/2}$

## TABLA C
## ESTADÍSTICAS DE DESAJUSTE

| Diferencia entre la habilidad de la persona y la dificultad del ítem (b - d) | Residuo cuadrático estandarizado $z^2 = \exp(b - d)$ | Improbabilidad de respuesta $p = 1 / (1 + z^2)$ | Eficiencia relativa de la observación $I = 400p(1 - p)$ | Número de ítems necesarios para mantener igual precisión $L = 1000 / I$ |
|---|---|---|---|---|
| −0.6, 0.3 | 1 | .50 | 100 | 10 |
| 0.4, 0.8 | 2 | .33 | 90 | 11 |
| 0.9, 1.2 | 3 | .25 | 75 | 13 |
| 1.3, 1.4 | 4 | .20 | 65 | 15 |
| 1.5, 1.4 | 5 | .17 | 55 | 18 |
| 1.7, 1.8 | 6 | .14 | 50 | 20 |
| 1.9, 2.0 | 7 | .12 | 45 | 22 |
| 2.1 | 8 | .11 | 40 | 25 |
| 2.2 | 9 | .10 | 36 | 28 |
| 2.3 | 10 | .09 | 33 | 30 |
| 2.4 | 11 | .08 | 31 | 32 |
| 2.5 | 12 | .08 | 28 | 36 |
| 2.6 | 14 | .07 | 25 | 40 |
| 2.7 | 15 | .06 | 23 | 43 |
| 2.8 | 17 | .06 | 21 | 48 |
| 2.9 | 18 | .05 | 20 | 50 |
| 3.0 | 20 | .05 | 18 | 55 |
| 3.1 | 22 | .04 | 16 | 61 |
| 3.2 | 25 | .04 | 15 | 66 |
| 3.3 | 27 | .04 | 14 | 73 |
| 3.4 | 30 | .03 | 12 | 83 |
| 3.5 | 33 | .03 | 11 | 91 |
| 3.6 | 37 | .03 | 10 | 100 |
| 3.7 | 41 | .02 | 9 | 106 |
| 3.8 | 45 | .02 | 9 | 117 |
| 3.9 | 50 | .02 | 8 | 129 |
| 4.0 | 55 | .02 | 7 | 142 |
| 4.1 | 60 | .02 | 6 | 156 |
| 4.2 | 67 | .02 | 6 | 172 |
| 4.3 | 74 | .01 | 5 | 189 |
| 4.4 | 81 | .01 | 5 | 209 |
| 4.5 | 90 | .01 | 4 | 230 |
| 4.6 | 99 | .01 | 4 | 254 |

# BIBLIOGRAFÍA


Allerup, P. and Sorber, G. *The Rasch Model for Questionnaires With a Computer Program*. Copenhagen: Danmarks Paedagogiske Institut, 1977.

Andersen, E. B. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society B,* 1970, *32,* 283-301.

Andersen, E. B. Asymptotic properties of conditional likelihood ratio tests. *Journal of the American Statistical Association,* 1971, *66,* 630-633.

Andersen, E. B. The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society B,* 1972a, *34,* 42-54.

Andersen, E. B. A computer program for solving a set of conditional maximum likelihood equations arising in the Rasch model for questionnaires. *Research Memorandum 72-6.* Princeton, N.J.: Educational Testing Service, 1972b.

Andersen, E. B. *Conditional Inference and Models for Measuring.* Copenhagen: Mentalhygiejnisk Forlag, 1973.

Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika,* 1977, *42,* 69-81.

Andrich, D. Latent trait psychometric theory in the measurement and evaluation of essay writing ability. Doctoral dissertation, University of Chicago, 1973.

Andrich, D. The Rasch multiplicative binomial model: Applications to attitude data. *Research Report No. 1.* Measurement and Statistics Laboratory, Department of Education, University of Western Australia, 1975.

Angoff, W. H. Measurement and scaling. In C. W. Harris (Ed.), *Encyclopedia of Educational Research.* New York: Macmillan, 1960.

Arthur, Grace. *A Point Scale of Performance Tests.* New York: Psychological Corporation, 1947.

Baker, F. B. UNIVAC scientific computer program for test scoring and item analysis. *Behavioral Sciences,* 1959,4, 254-255.

Baker, F. B. Empirical comparison of item parameters based on the logistic and normal functions, *Psychometrika,* 1961, *26,* 235-246.

Baker, F. B. Generalized item and test analysis program — a program for the Control Data 1604 Computer. *Educational and Psychological Measurement,* 1963, *23,* 187-190.

Bamdorff-Nielsen, O. *Information and Exponential Families in Statistical Theory.* New York: John Wiley and Sons, 1978.

Bimbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick, *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley, 1968.

Choppin, B. An item bank using sample-free calibration. *Nature,* 1968, *219,* 870-872.

Choppin,B. *Item Banking and the Monitoring of Achievement.* Slough, England: National Foundation for Educational Research, 1978.

Cohen, L. A modified logistic response model for item analysis. Manuscript, 1976.

Connolly, A. J., Nachtman, W. and Pritchett, E. M. *Keymath: Diagnostic Arithmetic Test.* Circle Pines, Minn.: American Guidance Service, 1971.

Cornish, G. and Wines, R. *ACER Mathematics Profile Series (MAPS).* Hawthorn, Victoria: Australian Council for Educational Research, 1977.

Douglas, Graham A. Test design strategies for the Rasch psychometric model. Doctoral dissertation, University of Chicago, 1974.

Draba, R. E. The Rasch model and legal criteria of a "reasonable" classification. Doctoral dissertation, University of Chicago, 1978.

Elliott, C. D., Murray, D. C., and Pearson, L. S. *The British Ability Scales*. Slough, England: National Foundation for Educational Research, 1977.

Fischer, G. H. and Scheiblechner, H. H. Two simple methods for asymptotically unbiased estimation in Rasch's model with two categories of answers. *Research Bulletin No.1,* Psychological Institute, University of Vienna, 1970.

Gulliksen, H. *Theory of Mental Tests*. New York: John Wiley & Sons, 1950.

Gustafsson, J. E. The Rasch model for dichotomous items: Theory, applications and a computer program. *Report No. 63.* Institute of Education, University of Goteberg, 1977.

Habermann, S. Maximum likelihood estimates in exponential response models. *The Annals of Statistics,* 1977,5, 815-841.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs,* 1947, *61.*

Loevinger, J. Person and population as psychometric concepts. *Psychological Review,* 1965, *72,* 143-155.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three parameter logistic model. *Educational and Psychological Measurement,* 1968, *28,* 989-1020.

Mason, G. P. and Odeh, R. E. A short-cut formula for standard deviation. *Journal of Educational Measurement,* 1968, 5, 319-320.

Mead, R. J. Analysis of fit to the Rasch model. Doctoral dissertation, University of Chicago, 1975.

Panchapakesan, N. The simple logistic model and mental measurement. Doctoral dissertation, University of Chicago, 1969.

Rasch, G. On simultaneous factor analysis in several populations. In *Uppsala Symposium on Psychological Factor Analysis.* Stockholm: Almquist and Wiksells, 1953, 65-71.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut, 1960 (To be reprinted by University of Chicago Press, 1980).

Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1961, *4,* 321-333.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in Mathematical Social Science.* Chicago: Science Research Associates, 1966a, 89-108.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology.* 1966b, *19,* 49-57.

Rasch, G. An informal report on the present state of a theory of objectivity in comparisons. In L. J. van der Kamp and C. A. J. Viek (Eds.), *Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof."* Leiden, 1967.

Rasch, G. A mathematical theory of objectivity and its consequences for model construction. In *Report from European Meeting on Statistics, Econometrics and Management Sciences,* Amsterdam, 1968.

Rentz, R. R. and Bashaw, W. L. *Equating Reading Tests with the Rasch Model.* Athens, Georgia: Educational Resource Laboratory, 1975.

Rentz, R. R. and Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement,* 1977, *14,* 161-180.

Thorndike, E. L.*,et al. The Measurement of Intelligence.* New York: Columbia University, Teachers College, 1926.

Thurstone, L. L. A method of scaling psychological and educational tests. *Journal of Educational Psychology,* 1925,*16,* 433-451.

Thurstone, L. L. The unit of measurement in educational scales. *Journal of Educational Psychology,* 1927, *18,* 505-524.

Tucker, L. R. Scales minimizing the importance of reference groups. In *Proceedings of the 1952 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1953.

Willmott, A. and Fowles, D. *The Objective Interpretation of Test Performance: The Rasch Model Applied*. Atlantic Highlands, N.J.: NFER Publishing Co., Ltd., 1974.

Woodcock, R. W. *Woodcock Reading Mastery Tests.* Circle Pines, Minnesota: American Guidance Service, 1974.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement,* 1 97 7,*14,* 97-116.

Wright, B. D. and Douglas, G. A. Best test design and self-tailored testing. *Research Memorandum No. 19,* Statistical Laboratory, Department of Education, University of Chicago, 1975a.

Wright, B. D. and Douglas, G. A. Better procedures for sample-free item analysis. *Research Memorandum No. 20,* Statistical Laboratory, Department of Education, University of Chicago, 1975b.

Wright, B. D. and Douglas, G. A. Rasch item analysis by hand. *Research Memorandum No. 21,* Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Douglas, G. A. Best procedures for sample-free item analysis. *Applied Psychological Measurement,* 1977a, *1,* 281-294.

Wright, B. D. and Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement,* 1977b, *37,* 573-586.

Wright, B. D. and Mead, R. J. CA LF IT: Sample-free calibration with a Rasch measurement model. *Research Memorandum No. 18*. Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B. D. and Mead, R. J. BICAL: Calibrating items with the Rasch model. *Research Memorandum No. 23,* Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Mead, R. J. *The Use of Measurement Models in the Definition and Application of Social Science Variables*. Arlington, VA: U.S. Army Research Institute Technical Report DAHC19-76-G-0011,1977.

Wright, B. D., Mead, R. J. and Draba, R. E. Detecting and correcting test item bias with a logistic response model. *Research Memorandum No. 22,* Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis, *Educational and Psychological Measurement,* 1969, *29,* 23-48.

# ÍNDICE


Adivinador (ver Patrón: respuesta)
Ajuste burdo (ver Ajuste)
Ajuste:
  análisis, 2 – 4, 23 – 24, 66 – 82
  corrección del desajuste, 181 – 190
  burdo, 124 – 125
  diagnóstico de desajuste, 170 – 180
  ejemplo a mano, 69 – 79
  ítem, 52 – 55, 58 – 59, 77 – 79, 121 – 125
  enlace (eslabón), 93 – 96, 98
  circuito (bucle), 100
  persona, 2 – 4, 76 – 77, 121 – 125, 165 – 180, 205 – 209
  respuesta, 69 – 77, 121 – 125, 165 – 180
  regla para el análisis de ajuste, 208 – 209
  resumen del análisis de ajuste, 79 – 80
  tabla para análisis de ajuste, 73, 216
  ejemplo por computadora, 52 – 55, 58 – 59, 80 – S2
Algoritmos de cálculo:
  BICAL, 46 – 54
  PROX, 61 – 62
    ejemplo a mano, 30 7 44
    ejemplo por computadora, 46 – 55
  UCON, 62 – 65
  UFORM, 143 – 151
    tablas, 146, 212, 214
Análisis de ajuste (ver Ajuste)

Banco (ver Construcción del banco de ítems)
Beta β (ver Persona: medida de habilidad)
BICAL, 46 – 54
  control, 46 – 47
  salida de resultados, 46, 48 – 52, 54
Bucle (ver Circuito)

Cadena, 99
Calibración δ y d (ver Ítem: calibración de la dificultad)
Calibración del ítem independiente de la muestra, vii – xiii, 15, 20, 25– 26
Características métricas de la prueba, 131 –133
Cauteloso (ver Respuesta: patrón)
CHIP (ver Escala: CHIP)
Circuito (bucle), 100 (ver Construcción del banco de ítems)
  ajuste, 100
Coeficiente de error $C_f$, 135 – 140, 146, 193 – 194, 214
Conexión de dos pruebas, 96 – 98 (ver Enlace: formas de pruebas)
Confiabilidad de la calibración (ver Error estándar de la calibración del ítem)
Confiabilidad de la medida (ver Error estándar de la medida de la persona)
Construcción del banco de ítems, 98 – 118
  armadura (web), 102 – 106
  cadena, 99
  circuito (bucle), 100
  enlace (eslabón), 96 – 106
  PCK ejemplo, 106 – 118
  red (network), 101 – 103
Control de calidad de la calibración del ítem, 121 – 125 (ver Ítem: ajuste)
Control de calidad de la medida de la persona, 155 – 170 (ver Ajuste: persona)
Control de calidad, 121 – 125, 165 – 170 (ver Ajuste)
Corrección de una medida, 181 – 190
Criterio, referencia a, 118 – 121, 199 – 202, 204, 206 – 207

Delta δ (ver Ítem: calibración de la dificultad)
Desajuste, diagnóstico de, 170 – 180
Desajuste (ver Ajuste)
Descuidado (ver Respuesta: patrón)
Diagnóstico de desajuste, 170 – 180
Dificultad δ y d (ver Ítem: calibración de la dificultad)
Discriminación (ver Ítem: índice de discriminación)
Diseño óptimo de prueba (ver Prueba: diseño)
Diseño o confección a la medida, 151 – 164
  desempeño (rendimiento), 156 – 160, 164
  estatus, 153 – 156, 164
  personalizado, 161 – 164
Diseño de pruebas, 131 – 140
  altura H y h, 132 – 133, 137 – 140
  ancho (o rango) W y w, 132 – 133, 136 – 140
  curva operativa, 132 – 133, 138
  distribución de ítems o personas, 130 – 139
  forma, 132 – 140
  longitud L, 132 – 133, 136 – 140
Distribución normal de ítems o personas, 21, 130 – 131, 133 – 134, 137, 139
Distribución de intervalos de ítems o personas, 130 – 131, 133 – 134, 137, 139
Dormilón (ver Respuesta: patrón)

Edición de datos, 31 – 34, 47 – 49
Eficiencia, 74 – 75, 139, 161, 164
Enlace:
  formas de pruebas, 96 – 106
  ítems comunes, 108 – 109, 112 – 118
  personas comunes, 107 – 112
Enlace (eslabón), 96 – 98 (ver Construcción de banco de ítems)
  ajuste, 94 – 96

Equiparación con ítems comunes, 108 – 109, 112 – 118
Equiparación con personas comunes, 106 – 112

Error estándar:
calibración del ítem PROX, 21 – 22
calibración del ítem SE($d_i$), 21 22, 25 – 26, 61 – 65, 143 – 146, 192
circuito (bucle), 100
coeficiente $C_f$, 135 – 140, 146, 193 – 194, 214
eslabón (enlace), 96 – 98
factor de espaciamiento, 195
línea de identidad, 89, 92 – 95
medida de la persona PROX, 21 – 22
medida de la persona SE($b_v$), Sv, ESM y S, 21 – 22, 27, 61 – 65, 132 – 136, 140, 192, 194 – 198
PROX calibración del ítem, 21 – 22
PROX medida de la persona, 21 – 22
Escala, 191 - 204
CHIP, 201 - 204
factor de espaciamiento γ, 191 - 198
lineal, vii, 7 - 9, 25, 27, 191 - 192
lógito, 16 - 17, 25, 27, 30, 34, 36, 191 - 192
MDM, 132, 135, 192 – 198
NIT, 198 – 200, 204
SIT, 199 – 202, 204
valor aditivo α, 191 – 192
Estadísticas tradicionales de las pruebas, 24 – 27
Estimación de aproximación uniforme UFORM, 143 – 151, 212, 214
Estimación aproximadamente normal PROX, 21 – 22, 28 – 45, 50 – 55, 60 – 62, 143, 149 – 150
algoritmo informático, 61 – 62
algoritmo manual, 21 – 22, 34, 38 – 40, 42, 44
ejemplo de computadora, 46 – 55
ejemplo manual, 30 – 44
manual vs computadora, 55 – 56
PROX vs UCON, 60 – 61
Estimación de la máxima verosimilitud incondicional UCON:
algoritmo de computadora, 62 – 65
ejemplo de computadora, 56 – 61
UCON vs. PROX, 60 – 61
Extendiendo una variable, 87 – 93

Factores de expansión X e Y, 21 – 22, 30, 40 – 44, 50, 62, 148
Factor de espaciamiento γ (ver Escala: Factor de espaciamiento)
Fiabilidad de la calibración (ver Error estándar de la calibración del ítem)
Fiabilidad de la medida (ver Error estándar de la medida de la persona)
Formularios de informes, 205 – 209

Grados de libertad, 23 – 24, 71, 74, 77, 79
ajuste burdo, 125
ajuste de persona, 23, 76 – 77, 79, 165 – 168
ajuste del ítem, 24, 77, 79
análisis de anclaje, 96

Habilidad β y b (ver Persona: medida de habilidad)

Igualación con ítems comunes, 108 – 109, 112 – 118
Igualación con personas comunes, 106 – 112
Igualación de formas de prueba (ver Enlace: formas de pruebas)
Información $I_{fi}$, 16 – 17, 73 – 75, 135, 161 – 164
Intensificando una variable, 87 – 94
Ítem:
ajuste, 52 – 55, 58 – 59, 77 – 79, 121 – 125
calibración de la dificultad δ y d, 17 – 22, 25, 30, 34 – 38, 40 – 42, 54 – 55, 61 – 65
curva característica, 12 – 14, 51 – 53, 58 – 59
discriminación, ix – x
índice de discriminación, 52 – 55
índice, 52 – 55
p valor, viii, xi – xiii, 25 – 26
puntaje $s_i$, 10, 18 – 22, 32 – 35
punto biserial, viii, x, 26

KCT (ver Prueba de los Cubos de Knox)

Linealidad, vii, 7 – 9, 15, 25, 27, 191 – 192
Líneas de control para gráficas de identidad, 94 – 95
Línea de identidad, 89, 92 – 95
LMD Least Measurement Difference (ver Mínima diferencia medible)
LOD Least Observable Difference (ver Mínima diferencia observable)
Logaritmo del momio (ver Lógito)
Logística:
distribución, ix – x
factor escalar 1.7 para la ojiva, 21 – 22
función, 15, 25, 27, 36
Lógito, 16 – 17, 25, 27, 30, 34, 36, 191 – 192
LSD Least significant difference (ver Mínima diferencia significativa)

Mapa (ver Variable: definición)
Matriz de datos, 10, 18,31, 33, 68, 107 – 109
MDM (ver Mínima diferencia medible)
MDO (ver Mínima diferencia observable)
MDS (ver Mínima diferencia significativa)
Media cuadrática residual v, 23 – 24, 26, 53, 71 – 74, 76 – 82, 165 – 170
Medida β y b (ver Persona: medida de la habilidad)
Medida objetivo (ver Objetivo: medición)
Medición de la persona independiente de la prueba, vii – xiii, 15, 20, 27, 141
Media cuadrática estandarizada t, 77 – 80, 165 – 169
Mejor diseño de prueba (ver Prueba: diseño)
Métodos de estimación, ix – x, 15 – 20, 44 – 45
PROX, 21 – 22, 28 – 45, 50 – 56, 60 – 62, 143, 149 – 150
UCON, 56 – 65, 142 – 143,148 – 150

UFORM, 143 – 151, 214, 216
Mínima diferencia medible MDM, 132, 135, 192 – 198
Mínima diferencia observable MDO, 132, 135, 194, 196
Mínima diferencia significativa MDS, 195 -196
Modelo de Rasch, 9 – 27

NIT (ver Escala: NIT)
No linealidad de los puntajes de la prueba, 7 – 9
Normas por el método corto, 126 – 128

Objetividad, viii xiii, 15, 141
Objetivo de medición, 129 – 131
dispersión S, 129 – 131, 133 – 134, 137 – 140
distribución D, 130 – 131, 134 – 139
ubicación M, 130 – 131, 137 – 139

PCK (ver Prueba de los Cubos de Knox)
Persona:
ajuste , 2 – 4, 76 – 77, 121 – 125,165 – 180, 205 – 209
curva característica, 12 – 14
Medida de la habilidad β y b, 17 – 22, 134 – 136, 142 – 151
PROX, 37 – 39, 43 – 44, 51, 56, 61 – 62, 143, 148 – 149
puntaje $r_v$,4 – 10,18 – 22
convertir a medida $b_r$, 21 – 22, 27, 37 – 40, 43 – 44, 61 – 65, 142 – 151
dependencia de la prueba, 4 – 6
no linealidad, 7 – 9
puntaje relativo $f_r$, 132, 134, 140, 144 – 146, 149, 193 – 194
respuesta, $x_{vi}$, 9 – 14, 68 – 77, 165 – 180
UCON, 57, 61 – 65, 142 – 143, 147 – 149
UFORM, 143 – 147,149 – 151, 212
Precisión de la medida (ver Error estándar: medida de la persona)
PROX (ver Estimación aproximadamente normal)
Prueba de los Cubos de Knox PCK, 28 – 29
banco de ítems de la PCKB, 106 – 118
definición de la variable, 83 – 91,119 – 120, 206 – 207
KCTB, 106 – 121
matriz de respuesta, 31 – 33, 66
PCKB, 106 – 121
referencia a criterio, 118 – 121, 206 – 207
referencia a norma, 120, 126 128, 200, 204, 206 – 207
Prueba, diseño 131 – 140
altura H y h, 132 – 133, 137 – 140
ancho (o rango) W y w, 132 – 133, 136 – 140
curva operativa, 132 – 133, 138
distribución de ítems o personas, 130 – 139
forma, 132 – 140
longitud L, 132 – 133, 136 – 140
Pruebas individualizadas (ver Diseño o confección a la medida)
Puntaje r de la prueba, 2 – 10, 18 – 20, 27
convertir a medida $b_r$, 142 – 164
PROX, 21 – 22, 27, 37 – 40, 43 – 44, 61 – 62, 143
UCON, 62 – 65, 142 – 143
UFORM, 143 – 151, 212
Puntaje (ver Prueba: puntaje)

Red (armadura), 102 – 106 (ver Construcción del banco de ítems)
completa, 103 – 104
incompleta, 104 – 106
Red, 101 – 103 (ver Construcción del banco de ítems)
Referencia a criterio, 118 – 121, 199 – 202, 204, 206 – 207
Referencia a un desempeño (ver Escala CHIP)
Referencia a norma, 120 – 121, 126 – 128, 151, 198 – 200, 204, 206 – 207
Residuo (ver Residuo estandarizado)
Residuo estandarizado z y $z^2$, 23 – 24, 70 – 80, 121 – 125, 165 – 180, 205 – 209
Respuesta:
ajuste, 69 – 75, 121 – 125, 165 – 180
curva, 9 – 14
improbable (no plausible), 71 – 74
matriz de la PCK, 31, 33, 66 – 69
modelo, 9 – 14
patrón, 2 – 4, 170 – 180
adivinador, 171, 174 – 177, 181, 185 – 187
cauteloso, 171, 176, 178 – 180, 188
descuidado, 171, 176, 178 – 180, 188 – 190
dormilón, 171 – 177, 181 – 184

SEM (ver Error estándar de medida)
SIT (ver Escala: SIT)

UCON (ver Estimación de la máxima verosimilitud incondicional)
UFORM (ver Estimación de aproximación uniforme)
Unidad de probabilidad de Chicago (ver Escala: CHIP)
Unidad de escalamiento normativo (ver Escala: NIT)
Unidades de escalamiento sustantivo (ver Escala: SIT)
Unidad de probabilidad (ver Escala: CHIP)
Unidad de escala de probabilidad de respuesta (ver Escala CHIP)

Valor de escala aditivo α (ver Escala: Valor aditivo α)
Validez de la calibración (ver Ajuste: ítem)
Validez de la medida (ver Ajuste: persona)
Variable definición de, 1 – 4, 98 – 106
PCK, 83 – 91, 119 – 120, 206 – 207

## NOTACIÓN

| Para las **Personas** v = 1, N | | Para los **Ítems** i = 1, L | |
|---|---|---|---|
| $\beta_v$ | Parámetro de habilidad de la persona v | $\delta_i$ | Parámetro de dificultad del ítem i |
| $b_v$ | Estadístico estimador de $\beta_v$ | $d_i$ | Estadístico para estimar $\delta_i$ |
| $SE(b_v)$ | Error estándar del estadístico $b_v$ | $SE(d_i)$ | Error estándar del estadístico $d_i$ |
| $r_v$ | Puntaje en la prueba observado para la persona v | $s_i$ | Puntaje observado en la muestra para el ítem i |
| $b_r$ | Habilidad estimada para el puntaje r | $p_i$ | Valor p para el ítem i en la muestra |
| $n_r$ | Número de personas con puntaje r | | |
| $y_v$ | Puntaje en la prueba en lógitos para la persona v | $X_i$ | Puntaje en lógitos para el ítem i en la muestra |
| $y_r$ | Lógito del puntaje en la prueba r | x. | Media de los ítems de la prueba en lógitos |
| y. | Media de la muestra de personas en lógitos | U | Varianza de los ítems de la prueba en lógitos |
| V | Varianza de la muestra de personas en lógitos | Y | Factor de expansión de los ítems en lógitos |
| X | Factor de expansión en lógitos de las personas<br>Para ajustar por ancho de la prueba | | para ajustar por la dispersión en la muestra |

| | |
|---|---|
| $x_{vi}$ | Respuesta de la persona v al ítem i |
| $p\{x_{vi} \mid \beta_v, \delta_i\}$ | Probabilidad de respuesta $x_{vi}$ dados $\beta_v$ y $\delta_i$ |
| $\pi_{vi}$ | Probabilidad de respuesta correcta, es decir: $x_{vj} = 1$ |
| $p_{vi}$ | Estimado de $\pi_{vi}$ basado en $b_v$ y $d_i$ |
| $p_{ri}$ | Estimado de $\pi_{vi}$ para el puntaje r basado en $b_r$ y $d_i$ |
| $l_{vi}$ | Información en $x_{vi}$ sobre la persona v y el ítem i |
| $z_{vi}$ | Residuo estandarizado de $x_{vj}$ a partir de la expectativa estimada |

| | | | |
|---|---|---|---|
| $v_v$ | Residuo medio cuadrático para la persona v | $v_i$ | Residuo medio cuadrático para el ítem i |
| $f_v$ | Grados de libertad en $v_v$ | $f_j$ | Grados de libertad en $v_i$ |
| $t_v$ | Media cuadrática estandarizada $v_v$ | $t_i$ | Media cuadrática estandarizada $v_i$ |

Puede haber excepciones a esta notación cuando es conveniente localmente, en particular con “s”.

## NOTACIÓN
## (Continuación)

| | **Para la <u>muestra</u> de N personas** | | **Para una <u>prueba</u> de L ítems** |
|---|---|---|---|
| M | Habilidad media de las personas | H | Media de dificultad de los ítems |
| m | Estimador de M | h | Estimador de H |
| $\sigma$ | Desviación estándar de la habilidad de las personas | $\omega$ | Desviación estándar de la dificultad de los ítems |
| s | Estimador de $\sigma$ | W | Rango de dificultades de los ítems |
| | | w | Estimador de W |

| | |
|---|---|
| e | Logaritmo neperiano o natural en base e = 2.71828... |
| $e^{(\beta v - \delta i)}$<br>$\exp(\beta_v - \delta_i)$ | base e elevada al exponente $(\beta_v - \delta_i)$ |
| $\sum_j^M (y_j)$ | Suma de valores $v_j$ para j = 1, M |
| $\prod_j^M (y_j)$ | Producto de valores $y_j$ para j = 1, M |
| $\ell n(y)$ | Logaritmo natural de y |
| E {y} | Valor esperado de y |
| V {y} | Varianza de y |
| 1.7<br>$2.89 = 1.7^2$<br>$8.35 = 2.89^2 = 1.7^4$ | Coeficiente que aproxima la ojiva logística de distribución acumulada en 0.01 a la ojiva de la distribución normal acumulada |

### MODELO DE RASCH

**Para la respuesta <u>correcta</u>**

$$x_{vi} = 1$$

$$P\{x_{vi}=1|\beta_v, \delta_i\} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)]$$

**Para la respuesta <u>incorrecta</u>**

$$x_{vi} = 0$$

$$P\{x_{vi}=0|\beta_v, \delta_i\} = 1/[1 + \exp(\beta_v - \delta_i)]$$

**Para <u>cualquier</u> respuesta**

$$x_{vi} = 1 \text{ o } 0$$

$$P\{x_{vi}|\beta_v, \delta_i\} = \exp[x_{vi}(\beta_v - \delta_i)]/[1 + \exp(\beta_v - \delta_i)]$$