Evaluaciones nacionales del rendimiento académico

VOLUMEN 4

Análisis de los datos de una evaluación nacional del rendimiento académico

Gerry Shiel
Fernando Cartwright

GRUPO BANCO MUNDIAL

Análisis de los datos de una evaluación nacional del rendimiento académico

Evaluaciones nacionales del rendimiento académico

VOLUMEN 4

Análisis de los datos de una evaluación nacional del rendimiento académico

Gerry Shiel
Fernando Cartwright

Vincent Greaney y
Thomas Kellaghan, editores de la serie

GRUPO BANCO MUNDIAL

La presente obra fue publicada originalmente por el Banco Mundial en inglés en 2015, con el título *Analyzing Data from a National Assessment of Educational Achievement*. Vol. 4 of *National Assessments of Educational Achievement*. En caso de discrepancias, prevalecerá el idioma original.

ISBN (edición impresa): 978-1-4648-0749-7
ISBN (edición electrónica): 978-1-4648-0750-3; 978-0-8213-9583-7 (inglés)
DOI: 10.1596/978-1-4648-0749-7

Diseño de la portada: Naylor Design, Washington DC

ÍNDICE

RECUADRO

EJERCICIOS

FIGURAS DE LOS EJERCICIOS

TABLAS DE LOS EJERCICIOS

FIGURAS

TABLAS

PRÓLOGO

Medir los resultados del aprendizaje estudiantil es esencial para hacer un seguimiento del éxito de un sistema escolar y para mejorar la calidad de la educación. La información sobre el rendimiento estudiantil puede usarse para documentar una amplia variedad de programas y decisiones educacionales, entre ellos los relacionados con el diseño y la implementación de programas con el fin de mejorar la enseñanza y el aprendizaje en las aulas y la provisión de apoyo y formación convenientes allí donde más se necesite.

La serie de publicaciones *Evaluaciones nacionales del rendimiento académico*, de la cual este es el cuarto volumen, se centra en los procedimientos de más reciente desarrollo que deben seguirse para garantizar que los datos (tales como puntajes de pruebas e información contextual) generados por un ejercicio de evaluación a escala nacional sean de alta calidad técnica y respondan a las preocupaciones de los responsables de la formulación de políticas y la toma de decisiones, así como de otros actores del sistema educativo.

El volumen 1 de la serie describe los propósitos y características clave de las evaluaciones nacionales del rendimiento académico y se dirige principalmente a los responsables de la formulación de políticas y la toma de decisiones. El volumen 2 aborda el diseño de dos tipos de instrumentos de recopilación de datos para los ejercicios de evaluación nacional: las pruebas de rendimiento académico y los cuestionarios de contexto. El volumen 3 se centra en las tareas prácticas de la

implementación de un ejercicio de evaluación a gran escala, incluyendo instrucciones paso a paso sobre logística, muestreo y depuración y gestión de datos.

El cuarto volumen, *Análisis de los datos de una evaluación nacional del rendimiento académico*, trata de cómo generar información sobre los ítems y los puntajes de las pruebas y cómo relacionar los puntajes de las pruebas con los factores educacionales y sociales. Al igual que los volúmenes 2 y 3, este volumen está dirigido primordialmente a los equipos de economías en desarrollo y emergentes que tienen la responsabilidad de llevar a cabo evaluaciones nacionales.

Por último, el volumen 5 describe cómo redactar informes basados en las conclusiones de la evaluación nacional y cómo usar estos resultados para mejorar la calidad de la política educativa y la toma de decisiones. Es de particular relevancia para quienes tienen la responsabilidad de preparar informes de evaluación y de comunicar y usar los resultados.

A medida que los lectores progresen en la lectura de este cuarto volumen, quedarán patentes las complejidades y el potencial del análisis de los datos generados por una evaluación a escala nacional. A fin de explorar plenamente lo que estos datos nos pueden decir sobre calidad, equidad y otros aspectos del logro en un sistema educativo, el analista debe usar una variedad de técnicas descritas en la primera parte de este volumen. La segunda parte describe una técnica analítica clave, la Teoría de Respuesta al Ítem (TRI). El volumen viene acompañado de un programa de TRI de fácil uso específicamente diseñado que se denomina Análisis de Ítems y Pruebas (Item and Test Analysis, IATA). Es probable que los equipos de evaluación de cualquier lugar, con independencia de que estén aprendiendo acerca de la TRI o ya estén familiarizados con ella, consideren IATA una aportación muy útil a su colección de herramientas de análisis de datos.

Marguerite Clarke
Especialista Superior en Educación/Coordinadora de Evaluación del Aprendizaje
Agosto 2014

ACERCA DE LOS AUTORES Y EDITORES

AUTORES

Fernando Cartwright es psicómetra, investigador de ciencias sociales y desarrollador/arquitecto de software. Ha trabajado en numerosas evaluaciones nacionales e internacionales de habilidades y aprendizaje, incluido el Programa para la Evaluación Internacional de Alumnos (PISA) y la Encuesta Internacional de Alfabetización de Adultos y Habilidades para la Vida. Es el arquitecto de varios proyectos de medición social, entre ellos el Índice de Aprendizaje Combinado (Composite Learning Index), el Índice Europeo de Aprendizaje Permanente (European Lifelong Learning Index) y el índice Third Billion. Ha producido software relacionado con la medición educativa, entre otros el programa Item and Test Analysis (IATA) y aplicaciones basadas en la web para el desarrollo de pruebas, bancos de datos de ítems, administración de pruebas, y análisis y evaluación de datos. Reside en Ottawa.

Gerry Shiel es investigador en el Centro de Investigación Educativa del St. Patrick's College, Dublín. Ha dirigido el desarrollo de una serie de pruebas normalizadas de rendimiento académico en las áreas de comprensión lectora (tanto en inglés como en irlandés), matemáticas y ciencia. Ha trabajado extensamente en evaluaciones nacionales en el nivel de educación primaria y ha dirigido la implementación nacional

en Irlanda del Estudio Internacional sobre Docencia y Aprendizaje (TALIS) y el estudio PISA de la Organización para la Cooperación y el Desarrollo Económicos. Ha trabajado en temas relacionados con las evaluaciones en África, el Sudeste de Asia y Europa del Este.

EDITORES

Vincent Greaney ha sido jefe de especialistas en educación en el Banco Mundial. Ex docente, investigador en el Centro de Investigación Educativa del St. Patrick´s College, Dublín, y profesor visitante becario Fulbright en la Universidad del Oeste de Michigan, Kalamazoo; miembro de la Galería de Honor de la Lectura de la Asociación Internacional de Lectura. Sus áreas de interés incluyen la evaluación nacional, los exámenes públicos, la formación docente, la lectura y la promoción de la cohesión social mediante la reforma de los libros de texto. Ha trabajado en proyectos educativos principalmente en África, el Sudeste de Asia, Europa del Este y Oriente Medio.

Thomas Kellaghan fue director del Centro de Investigación Educativa del St. Patrick's College, Dublín, y miembro numerario de la Academia Internacional de Educación. Ha trabajado en la Universidad de Ibadán, en Nigeria, y en la Queen´s University de Belfast. Sus áreas de interés incluyen las evaluaciones, los exámenes públicos, la desventaja educativa, la formación docente y las relaciones entre el hogar y la escuela. Se desempeñó como presidente de la Asociación Internacional de Evaluación Educativa (IAEA) entre 1997 y 2001. Ha trabajado en temas relacionados con las evaluaciones en África, Europa del Este, el Sudeste de Asia, América Latina y Oriente Medio.

AGRADECIMIENTOS

Un equipo liderado por Vincent Greaney (consultor, Actividades Mundiales de Educación, Banco Mundial) y Thomas Kellaghan (Centro de Investigación Educativa, St. Patrick´s College, Dublín) ha preparado la serie de libros *Evaluaciones nacionales del rendimiento académico*, de la que este es el cuarto volumen. Han colaborado también en esta serie: Sylvia Acana (Junta Nacional de Exámenes de Uganda), Prue Anderson (Consejo Australiano de Investigación Educativa), Fernando Cartwright (Polymetrika, Canadá), Jean Dumais (Dirección General de Estadísticas de Canadá), Chris Freeman (Consejo Australiano de Investigación Educativa), J. Heward Gough (Dirección General de Estadísticas de Canadá), Sara J. Howie (Universidad de Pretoria), George Morgan (Consejo Australiano de Investigación Educativa), T. Scott Murray (Data Angel, Canadá), Kate O'Malley (Consejo Australiano de Investigación Educativa) y Gerry Shiel (Centro de Investigación Educativa, St. Patrick's College, Dublín).

El trabajo se llevó a cabo bajo la dirección general de Ruth Kagia, directora de educación; sus sucesores, Elizabeth King y Amit Dar; y Robin Horn y Harry Patrinos, directores, todos ellos del Banco Mundial. Robert Prouty inició el proyecto y lo dirigió hasta agosto de 2007. Marguerite Clarke lo ha dirigido desde entonces, ocupándose de las tareas de revisión y publicación.

Estamos muy agradecidos por las contribuciones realizadas por el panel de revisión: Eugenio González (Educational Testing Service), Pei-tseng Jenny Hsieh (Universidad de Oxford) y Laura Jane Lewis (Banco Mundial).

Diana Manevskaya (Banco Mundial) facilitó la preparación de este volumen. El Centro de Investigación Educativa, St. Patrick's College, Dublín, ofreció apoyo a través de Hilary Walshe, así como Peter Archer, John Coyle y Mary Rohan. La corrección fue llevada a cabo por Laura Glassman, Mary-Ann Moalli y Linda Stringer, de Publications Professionals LLC. El diseño, la edición y la producción del libro fueron coordinados por Janice Tuten y Paola Scalabrin, de la División de Publicaciones y Conocimiento del Banco Mundial; la impresión fue coordinada por Andrés Meneses.

El Consejo Australiano de Investigación Educativa, el Programa de Asociación Banco-Países Bajos, el Centro de Investigación Educativa, el Fondo Fiduciario de Irlanda para la Educación, la Dirección General de Estadísticas de Canadá y el Fondo Fiduciario de Rusia de Ayuda a la Educación para el Desarrollo (READ) brindaron su generoso apoyo para la preparación y publicación de esta serie.

SIGLAS

CCI	curva característica del ítem
CCP	curva característica de la prueba
EE	error estándar
FDI	funcionamiento diferencial del ítem
FRI	función de respuesta al ítem
IATA	Items and Test Analysis (Análisis de Ítems y Pruebas)
IC	intervalo de confianza
ID	identificación
IQR	rango de intercuartiles
JK	método jackknife
MLJ	modelización lineal jerárquica
NAEP	Evaluación Nacional del Progreso Educativo
PIRLS	Estudio sobre el Progreso Internacional de la Competencia en Lectura
PISA	Programa para la Evaluación Internacional de Alumnos
PR	probabilidad de respuesta
SPSS	Paquete SPSS (Paquete estadístico para ciencias sociales)
TCP	Teoría Clásica de las Pruebas
TIMSS	Estudio Internacional de Tendencias en Matemáticas y Ciencias
TRI	Teoría de Respuesta al Ítem
UPM	unidad primaria de muestreo

INTRODUCCIÓN

La economía del conocimiento en el mundo actual requiere que los gobiernos, los sistemas educativos y las escuelas hagan un seguimiento estrecho de una variedad de resultados educativos, entre ellos los rendimientos estudiantiles. Una evaluación nacional del rendimiento académico en áreas curriculares clave contribuye a este esfuerzo al abordar cuestiones relacionadas con

- *La calidad*—suministra información sobre el aprendizaje estudiantil con referencia al currículo, al logro de los niveles educativos previstos o a la preparación para el aprendizaje futuro.
- *La equidad*—determina si el sistema educativo está prestando atención insuficiente a grupos concretos de estudiantes, como ponen de manifiesto las diferencias en cuanto a logro académico relacionadas con el sexo, la ubicación, la pertenencia a grupos étnicos o lingüísticos, el grupo socioeconómico o la gestión escolar (pública-privada).
- *La dotación*—especifica factores relacionados con el aprendizaje estudiantil (por ejemplo, recursos escolares, implementación del plan de estudios, nivel de formación de los profesores, cualificación y experiencia, y circunstancias domésticas de los alumnos).

- *El cambio*—se refiere a los resultados educativos a lo largo del tiempo (Greaney y Kellaghan, 2008; Kellaghan y Greaney, 2001; Kellaghan, Greaney y Murray, 2009).

Los anteriores volúmenes de esta serie, *Evaluaciones nacionales del rendimiento académico*, describen los componentes de una evaluación nacional basada en muestras. Estos componentes comprenden la especificación del contenido de las pruebas y los cuestionarios; la definición de una población de interés y la selección de una muestra de probabilidad que represente a la población; la administración de instrumentos de evaluación y de otro tipo a los estudiantes y otros encuestados; la puntuación de las respuestas de los alumnos; y la depuración y gestión de los datos. El conjunto final de datos generado por estas actividades, en las que se ha creado y reunido ítems de prueba en un cuadernillo de prueba y se ha recopilado los datos de respuesta, proporciona la fuente para los análisis descritos en este volumen.

La primera parte del volumen se ha diseñado para ayudar a los equipos de evaluación nacional a llevar a cabo análisis de datos normalmente efectuados en una evaluación nacional. El capítulo 1 ofrece una descripción general de los conjuntos de datos usados en los ejemplos elaborados del CD que acompaña el volumen. Viene a continuación en el capítulo 2 un análisis exploratorio de los datos utilizando SPSS. Se definen conceptos tales como media, mediana, modo y desviación estándar, y se ejecuta una serie de análisis ilustrativos. El capítulo 3 presenta el concepto de error estándar de la estimación y describe procedimientos para calcular en qué grado puede preverse que difieran los datos de una muestra con respecto a los datos de la población. Se describe el modo en que WesVar computa los errores estándar para una muestra compleja, una característica importante de una evaluación nacional bien diseñada. El capítulo 4 describe maneras de abordar las cuestiones relacionadas con la equidad analizando las diferencias entre los puntajes promedio de categorías de alumnos para determinar si una diferencia obtenida es estadísticamente significativa. En el capítulo 5, la atención se traslada a las maneras en que puede describirse el desempeño de los alumnos con alto y bajo rendimiento académico. El capítulo 6 trata de las asociaciones entre las

variables (por ejemplo, las relaciones entre los recursos escolares y el aprendizaje de los alumnos), como se evidencia en la correlación, y ofrece una introducción al análisis de regresión. El capítulo 7 contiene ejemplos de cómo pueden presentarse los datos utilizando tablas y gráficas.

La segunda parte del volumen se centra en el desarrollo de escalas para describir el aprendizaje de los alumnos. Para abordar esta cuestión se utilizan dos populares marcos estadísticos (dentro de los cuales se han formulado varios modelos). El primero de ellos, la Teoría Clásica de las Pruebas (TCP) (véase Crocker y Algina, 2006; Haladyna, 2004; Lord y Novick, 1968), se ha utilizado durante la mayor parte del siglo XX y se ha usado asimismo para describir el desarrollo de pruebas en el volumen 2 de la presente serie (Anderson y Morgan, 2008). El segundo marco, que se describe en la segunda parte, es la Teoría de Respuesta al Ítem (TRI) (véase De Ayala, 2009; De Mars, 2010; Hambleton, Swaminathan y Rogers, 1991; Lord y Novick, 1968). Se originó a mediados del siglo XX y actualmente es ampliamente utilizado en las evaluaciones nacionales e internacionales de rendimiento estudiantil.

El programa Análisis de Ítems y Pruebas (IATA) descrito en este volumen emplea la TRI para analizar los datos de las pruebas. Se diseñó para ofrecer un modo fácil de abordar dos principales consideraciones estadísticas relacionadas con las evaluaciones nacionales: (a) aumentar la usabilidad y la interpretabilidad de los puntajes de las pruebas y (b) establecer escalas significativas y coherentes de acuerdo con las cuales comunicar los puntajes. Esta última requiere reducir el error de medición y proveer información que pueda generalizarse más allá de la muestra a partir de la cual se obtuvo los datos. La secuencia de análisis de la segunda parte se ha diseñado para imitar las fases de desarrollo e implementación de un programa de evaluación nacional, desde la realización de pruebas piloto hasta las pruebas definitivas y las pruebas de seguimiento en ciclos de evaluación posteriores. El capítulo 8 proporciona una descripción del menú principal de IATA, sus elementos interactivos y los resultados que produce. El capítulo 9 describe los pasos que se llevan a cabo al analizar los datos desde la administración de una prueba piloto, tras lo cual se describe en el capítulo 10 los pasos que contiene un análisis de los datos de

administración de una prueba final. Se describe análisis de rotación de los cuadernillos de prueba (capítulo 11) y de los ítems de crédito parcial (capítulo 12). La comparación de evaluaciones mediante vinculación y métodos especializados en IATA se examina en los capítulos 13 y 14, respectivamente. El volumen concluye con un anexo sobre TRI. Debe advertirse que IATA funciona sólo en Windows.

Las principales ventajas de la TRI son, a diferencia de la TCP, que produce estadísticas de ítems que son independientes de la distribución de capacidades de un conjunto de examinados y parámetros de caracterización de un examinado que son independientes del conjunto concreto de ítems de prueba a partir del cual se calibran. Sus ventajas se consideran especialmente convenientes en situaciones que requieren ajuste de los puntajes de las pruebas, identificación de sesgo de los ítems y diseño de pruebas adaptables computerizadas.

Una desventaja de la TRI es que requiere destrezas analíticas avanzadas y procedimientos de computación complejos, que puede que no estén disponibles en un equipo de evaluación nacional. Muchas evaluaciones nacionales de países en desarrollo continúan basando su desarrollo de pruebas en los índices de facilidad y discriminación de los ítems de la TCP. Debe reconocerse que estos formularios de datos proporcionan a los desarrolladores de las pruebas información útil con independencia del modelo de medición que se aplique en fases ulteriores del proceso de desarrollo de las pruebas. Además, la TCP y la TRI producen resultados muy similares en cuanto a la comparabilidad de las estadísticas de ítems y personas así como el grado de invarianza de las estadísticas de ítems en todas las muestras de examinados (Fan 1998).

Sea TCP o TRI la elección, dos cuestiones relacionadas con la práctica actual en las evaluaciones nacionales e internacionales, que siguen lo llevado a cabo habitualmente en el desarrollo de pruebas diseñadas para evaluar los logros de estudiantes individuales, merecen atención: (a) la suposición de que el rasgo o la capacidad que se está evaluando posee una sola dimensión y (b) la incidencia en maximizar las diferencias entre los rendimientos de los examinados. Ambas tienen implicaciones para la validez de las pruebas.

La suposición de unidimensionalidad subyacente al desarrollo de las pruebas tiene implicaciones importantes en una evaluación

nacional o internacional, no solo para la validez de los contenidos de las pruebas, sino también para determinar el sesgo de los ítems y la vinculación de las pruebas. No obstante, este supuesto queda cuestionado por la evidencia de que los alumnos varían en el ritmo al que adquieren competencia en diferentes áreas de rendimiento (lo que se ilustra, por ejemplo, cuando el logro en matemáticas se describe en términos de número, medición, forma y datos). Esta variación se debe, con toda probabilidad, a las diferencias en las experiencias educativas y culturales generales de los alumnos (Goldstein y Wood, 1989). Rechazar la inclusión de ítems en una prueba porque los datos estadísticos no admiten el supuesto de unidimensionalidad puede tener el efecto de excluir contenidos importantes, con el resultado consiguiente de una representación inadecuada de un constructo, lo que, por supuesto, afectaría a la validez de los contenidos de una prueba —un aspecto de la validez que se considera generalmente más importante que las inferencias basadas en datos estadísticos—.[1] La suposición de unidimensionalidad debe tenerse en cuenta especialmente en las evaluaciones internacionales, en las que se sabe que las experiencias de los alumnos, en la escuela y fuera de ella, varían en grado notable.

El objetivo de maximizar las diferencias entre los examinados, otra característica de los procedimientos diseñados para desarrollar pruebas para evaluar individualmente los logros de los estudiantes, interesa en una evaluación nacional (o internacional) porque el propósito de tal evaluación es describir los logros del sistema educativo, no diferenciar entre los logros de cada uno de los estudiantes. La implicación de esta situación es que deben tomarse en consideración otros factores distintos a la discriminación y la facilidad a la hora de decidir si se incluye ítems en una prueba. Por ejemplo, los ítems que todos los alumnos respondieron correctamente o los ítems que ningún alumno respondió correctamente normalmente no se incluirían en una prueba diseñada para estudiantes individuales porque no contribuirían a diferenciar entre los estudiantes. No obstante, en el caso de una evaluación nacional, podría ser importante saber que todos o ninguno de los alumnos han dominado determinadas áreas de rendimiento. Por consiguiente, los ítems que representan esas áreas se incluirían en la evaluación. Para garantizar que las pruebas usadas en una evaluación

nacional representen adecuadamente el constructo que se está evaluando y ofrezcan información exhaustiva sobre el abanico de logros obtenidos por los alumnos en el sistema educativo, es imperativo que los desarrolladores mantengan contacto regular con los especialistas en currículo y con los docentes a lo largo del proceso de desarrollo de las pruebas.

La introducción general al análisis estadístico precede a la parte dedicada a la TRI en este volumen porque presenta al lector muchos de los procedimientos analíticos usados en TRI. No obstante, en la situación real de una evaluación nacional, el escalado de los datos para describir el rendimiento estudiantil, tal como se describe en la segunda parte, debería completarse antes de llevar a cabo los análisis de la primera parte.

Se parte de que los usuarios de este volumen tienen un conocimiento básico del uso de carpetas y archivos, Excel y SPSS, y tienen la capacidad de navegar, sin dificultad, entre los componentes de SPSS.

NOTA

1. Cronbach (1970, 457) ha señalado que, incluso en el caso de las pruebas desarrolladas para evaluar individualmente a los alumnos, "no hay nada en la lógica de la validación de contenidos que exija que el contenido del universo o de la prueba sea homogéneo".

PARTE I

INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO DE LOS DATOS DE LA EVALUACIÓN NACIONAL

Gerry Shiel

CAPÍTULO 1

BASE DE DATOS PARA ANÁLISIS

Los datos de la evaluación nacional contienen un indicador del rendimiento estudiantil, que puede representarse de diversas maneras, tales como el número de ítems que el estudiante ha respondido correctamente (si bien este indicador no siempre es muy significativo); el porcentaje de ítems respondidos correctamente; y los puntajes escalados, en los que una distribución de puntajes con datos de media y desviación estándar se transforma en una distribución con una media y desviación estándar diferentes. La mayoría de las evaluaciones nacionales también recopilan datos adicionales. Estos pueden referirse a las escuelas (como tipo o tamaño); los docentes (como cualificaciones o experiencia); los alumnos (como edad o tiempo empleado en los deberes); y los padres y el entorno doméstico (como el nivel educativo de los padres o el número de libros en el hogar).

Los datos recolectados contendrán una serie de tipos de variables. Algunas de estas variables serán *categóricas* y supondrán la asignación de los individuos a categorías o grupos claramente definidos, como el nivel educativo o el sexo. Otras variables, descritas como *discretas*, consisten en mediciones numéricas o recuentos, como el número de niños en una familia. Se obtienen por recuento y poseen valores para los que no hay valores intermedios. Las variables *continuas*, por el contrario, describen mediciones numéricas que pueden ser cualquier

valor situado entre dos valores especificados, como la distancia desde la casa de un alumno a la escuela. El tipo de datos impone limitaciones sobre el tipo de análisis estadístico que puede llevarse a cabo, así como sobre el modo en que puede representarse los datos gráficamente.

Los análisis comenzarán generalmente con una exploración de datos numéricos simples, presentados como resúmenes estadísticos, en gráficos o diagramas, o de ambas maneras. La atención en esta fase, como se explica más detalladamente a lo largo del capítulo, se centra en la descripción, si bien lo aprendido puede generar hipótesis que se someterán a prueba en una fase posterior. La fase exploratoria del análisis de datos ofrece también la oportunidad de inspeccionar la calidad de los datos comprobando la existencia de valores faltantes, valores atípicos, lagunas y valores erróneos, si bien estos deberían haber sido identificados en la fase de depuración de datos (véase Freeman y O'Malley, 2012). Revela también la naturaleza de los datos, indicando si la distribución es simétrica, sesgada o agrupada. En esta fase temprana, una representación gráfica, en forma de gráfico de barras, histograma o diagrama de caja y bigotes, puede resultar muy informativa para identificar los patrones presentes en los datos.

Cuando se dispone de más de una observación sobre los individuos, es posible investigar las relaciones entre variables, como la relación entre los logros en lectoescritura y aritmética elemental de los estudiantes o entre el desempeño en matemáticas y los factores del entorno doméstico. Una asociación entre un par de variables se denomina *bivariante*. Puesto que muchas de las variables en una evaluación nacional estarán interrelacionadas, es necesario llevar a cabo análisis *multivariantes* que implican procedimientos para predecir el desempeño en una variable (por ejemplo, logros en lectura) a partir de los valores de un conjunto de otras variables (por ejemplo, sexo de los estudiantes o factores del entorno doméstico). Un primer paso en un análisis multivariante es mostrar y examinar correlaciones por pares entre las variables de una matriz de correlación. El presente volumen contiene una introducción al análisis multivariante (análisis de regresión; véase el capítulo 6). Sin embargo, no se ocupa de formas de análisis más complejas, como la modelación multinivel, en la que los análisis se han diseñado para reflejar la estructura hallada en los

sistemas educativos (estudiantes agrupados en clases, clases en escuelas, escuelas en regiones).

Los lectores pueden desarrollar sus habilidades analíticas trabajando en un conjunto de ejercicios con ayuda de la base de datos del CD que acompaña este libro. Esta base de datos, que es similar a la que se utiliza en la parte dedicada al muestreo de *Implementación de una evaluación nacional del rendimiento académico* (Dumais y Gough, 2012a), contiene pruebas de rendimiento académico y otros datos que se han modificado a partir de los datos recabados en una evaluación real de desempeño en matemáticas llevada a cabo en 4.º grado en un país pequeño y se presentan en esta serie como si procediesen de un país ficticio de nombre Sentz.

Los siguientes capítulos describen una serie de tareas analíticas que se suelen llevar a cabo con datos obtenidos en una evaluación nacional. Al efectuar estos análisis, los lectores podrán familiarizarse con un conjunto de técnicas estadísticas que pueden aplicar a sus propios datos. La mayoría de los análisis emplean el programa WesVar. A diferencia de muchos otros paquetes de software, WesVar tiene en cuenta la complejidad del diseño de la evaluación nacional cuando lleva a cabo análisis estadísticos, como calcular la varianza y el error de muestreo. Las partes 2 y 4 del volumen 3 de esta serie, *Implementación de una evaluación nacional del rendimiento académico*, describen en detalle el muestreo complejo (Dumais y Gough 2012a, 2012b).

GUARDAR LOS ARCHIVOS DEL CD EN UN DISCO DURO O SERVIDOR PROPIO

Es posible guardar los archivos del CD en un disco duro o un servidor. Copie o cree una carpeta llamada ***NAEA DATA ANALYSIS*** desde el CD en su escritorio. Debería tener siete subcarpetas dentro de la carpeta ***NAEA DATA ANALYSIS***: ***SPSS DATA***, ***EXERCISE SOLUTIONS***, ***WESVAR UNLABELED DATA***, ***MY WESVAR FILES***, ***WESVAR DATA & WORKBOOKS***, ***MY SPSS DATA*** y ***MY SOLUTIONS***. Para copiar la carpeta ***NAEA DATA ANALYSIS*** desde el CD a su escritorio, localice la carpeta en el CD y haga clic derecho en **Copy**. Luego abra **Desktop**, y haga clic derecho en **Paste**. Compruebe que la carpeta ***NAEA DATA ANALYSIS*** se haya

copiado efectivamente. A continuación se exponen con detalle las siete subcarpetas.

- ***SPSS DATA***. Los archivos de datos de SPSS (***NATASSESS.SAV*** y ***NATASSESS4.SAV***) usados para completar los ejercicios del capítulo 2 de este volumen, así como un archivo sobre las escuelas (***SCHOOLS.SAV***), pueden encontrarse en esta carpeta.
- ***EXERCISE SOLUTIONS***. Aquí podrá encontrar soluciones, principalmente en archivos de texto, para los ejercicios de los capítulos 2 a 7 de este volumen. Una vez completados los ejercicios, puede comprobar las soluciones que obtenga comparándolas con las de esta subcarpeta.
- ***WESVAR UNLABELED DATA***. Use esta fuente para los ejercicios de WesVar propuestos en el capítulo 3. Este archivo de datos (***NATASSESS4.VAR***) debería ser el mismo que el archivo de datos obtenido después de crear su propio archivo de datos de WesVar aplicando los pasos expuestos en el Anexo I.C. El archivo de datos de este directorio puede servir de copia de respaldo.
- ***MY WESVAR FILES***. Use esta subcarpeta para guardar los archivos de datos y libros de trabajo de WesVar que cree al completar los ejercicios de los capítulos 3 a 6. Cuando abra esta carpeta por primera vez, verá que está vacía, debido a que aún no ha guardado ningún archivo en ella. Se recomienda encarecidamente que cree sus propios archivos de datos y libros de trabajo de WesVar empleando los procedimientos reseñados más adelante así como en el Anexo I.C. Tenga en cuenta que todos los archivos de datos y libros de trabajo de WesVar que cree deberían guardarse en ***MY WESVAR FILES***.
- ***WESVAR DATA & WORKBOOKS***. Esta subcarpeta contiene el archivo de datos ***NATASSESS4.VAR***, su archivo de registro asociado ***NATASSESS4. LOG*** y cuatro libros de trabajo completos, ***CHAPTER3 WORKBOOK. WVB***, ***CHAPTER4 WORKBOOK. WVB***, ***CHAPTER5 WORKBOOK.WVB*** y ***CHAPTER6 WORKBOOK.WVB***. Puede acudir a estas fuentes para comprobar la exactitud de su trabajo en WesVar.
- ***MY SPSS DATA***. Use esta carpeta para guardar archivos de datos de SPSS nuevos o modificados, como aquellos que cree antes de transferir un archivo de datos de SPSS a WesVar (véase el Anexo I.C).

- ***MY SOLUTIONS***. Guarde sus soluciones a los ejercicios de los capítulos 2 a 7 en esta subcarpeta. Al igual que en los casos de ***MY WESVAR FILES*** y ***MY SPSS DATA***, estará vacía cuando la abra por primera vez.

El Anexo I.B contiene detalles del contenido de cada carpeta y archivo.

INSTRUMENTOS PARA LA ENCUESTA

Esta sección describe los principales instrumentos usados para recopilar los datos empleados en la base de datos.

Prueba de desempeño en matemáticas

La prueba constó de 125 ítems basados en el marco curricular nacional para 4.º grado. La Tabla 1.1 muestra la distribución de ítems en las principales áreas de contenido y procesos cognitivos (o comportamientos intelectuales) de matemáticas. La mayoría de los ítems evaluaron las áreas de contenido de una serie de indicadores, que reflejan las ponderaciones asignadas a dichas áreas en el currículo nacional y en los libros de texto. Más de la mitad de los ítems evaluaron dos procesos cognitivos: "implementar los procedimientos" (28 por ciento) y "aplicar y solucionar problemas" (32 por ciento).

TABLA 1.1

Prueba de matemáticas: distribución de ítems por área de contenido y proceso

Áreas de contenido			Procesos cognitivos		
	Número de ítems	Porcentaje de ítems		Número de ítems	Porcentaje de ítems
Números	46	36,8	Entender y recordar	16	12,8
Álgebra	6	4,8	Implementar procedimientos	35	28,0
Forma y espacio	18	14,4	Razonar	26	20,8
Medidas	44	35,2	Integrar y conectar	8	6,4
Datos y posibilidad	11	8,8	Aplicar y resolver problemas	40	32,0
Total	125	100,0	Total	125	100,0

Los ítems se agruparon en cinco bloques (A, B, C, D), cada uno de ellos consistente en 25 ítems. Cada cuadernillo de prueba contenía 75 ítems del total de 125. Cada bloque (excepto el bloque común B) aparecía una vez al principio y otra vez al final de cada cuadernillo de prueba.

Cada ítem empleó un formato o bien de opción múltiple o de respuesta corta. Los ítems de opción múltiple tenían cuatro respuestas posibles (A, B, C y D). Los alumnos debían marcar la respuesta que creyesen correcta. Respecto a los ítems de respuesta corta, los alumnos debían escribir respuestas a las preguntas o hacer dibujos (por ejemplo, dibujar líneas simétricas atravesando una forma bidimensional, como un rectángulo). Cada pregunta de respuesta múltiple tenía una única opción de respuesta correcta. Cada ítem de respuesta corta se puntuaba como correcta o como incorrecta según un sistema de puntaje proporcionado a los correctores de ítems.

Cuestionarios contextuales

La evaluación nacional incluyó cuestionarios diferenciados para directores, docentes, alumnos y padres (Tabla 1.2). Los docentes completaron también un formulario de calificación referente a cada alumno en la evaluación.

PONDERACIONES DE MUESTREO

Las ponderaciones de muestreo se computaron e incluyeron en el archivo; reflejan la probabilidad de selección para cada estudiante. La manera en que estas ponderaciones se computan y utilizan se describe en el volumen 3 de esta serie (Dumais y Gough, 2012b). Para cada estudiante se computó una ponderación de diseño que incluía los siguientes componentes:

- *Componente de selección de escuelas*. Las escuelas se seleccionaron con probabilidad proporcional al tamaño. Para la escuela *i*, en el estrato *h*, esta fue el recíproco del producto del número de escuelas seleccionadas multiplicado por el número de alumnos en el grado objetivo de la escuela (medida del tamaño), dividido por el número de

TABLA 1.2

Descripciones del cuestionario abreviadas

Cuestionario	Completado por	Los temas abordados incluyen:
Cuestionario de la escuela	Directores de escuela	Tamaño de la escuela, recursos escolares, personal docente, planificación de desarrollo de la escuela, calificaciones del director de la escuela
Cuestionario de docentes	Profesores de los alumnos participantes en 4.° grado	Cualificaciones de los docentes, años de experiencia docente, distancia recorrida hasta la escuela cada día, tamaño de las clases, tiempo dedicado a la enseñanza de matemáticas, frecuencia de evaluación del progreso de los alumnos, disponibilidad y uso de los recursos educativos en el aula
Cuestionario de alumnos	Alumnos	Edad, frecuencia de realización de deberes, interés por las matemáticas
Cuestionario de padres	Padres de los alumnos participantes	Nivel educativo alcanzado (propio y del cónyuge o la pareja), número de libros en el hogar, tamaño de la propiedad (tierra), disponibilidad de luz eléctrica en el hogar, apoyo y aliento de los padres
Formulario de calificación de los alumnos	Profesores de los alumnos participantes en 4.° grado con respecto a cada alumno	Asistencia a clase de los alumnos, competencia de los alumnos en el idioma de instrucción, calificación del rendimiento académico de los estudiantes por parte de los profesores, comportamiento, apoyo de los padres.

estudiantes en el estrato de la población. Por ejemplo, si había 5000 estudiantes en el estrato, y se había seleccionado a 10 escuelas del estrato, con 50 estudiantes en la escuela *i*, el componente de selección de la escuela para la escuela *i* (**Schwgt**) sería 5000/(10*50) = 10.

- *Componente de corrección por respuestas omitidas de las escuelas.* Puesto que todas las escuelas seleccionadas participaron, el factor de ajuste por respuestas omitidas de la escuela se fijó en 1.0. (**Schnrfac**).[1]
- *Componente de selección de los alumnos*. Puesto que todos los alumnos de 4.° grado de una escuela se seleccionaron, la probabilidad de que se sometiese a examen a un alumno de una escuela seleccionada era 1.0, y su recíproco era también 1.0 (**Studfac**).[2]
- *Componente de ajuste dentro de la escuela por respuestas omitidas de los alumnos*. Se efectuó una corrección de ajuste con relación a las respuestas omitidas de los alumnos dentro de las escuelas. Esta fue el inverso del número de cuadernillos de prueba válidos entregados

para los alumnos en la escuela sobre el número de alumnos de 4.º grado en la escuela menos los alumnos exentos (**Stunrfac**). Por ejemplo, si había 90 alumnos matriculados en 4.º grado en el momento del estudio, ninguno de los cuales tenía derecho a exención, y 80 tomaron parte, el factor de ajuste sería 90/80.

La ponderación referente a cada alumno se obtuvo calculando el producto de estos cuatro componentes (**Schwgt × Schnrfac × Studfac × Stunrfac**). Utilizando los ejemplos precedentes, para el alumno *x* en la escuela *i*, la ponderación habría sido 10 × 1 × 1 × 90/80. Este producto da como resultado el coeficiente de ponderación de diseño (**Wgtpop** en el archivo de datos). Cuando se pondera los datos del archivo de datos empleando el coeficiente de ponderación de diseño (población), el tamaño de población estimado es 51 713 (el número previsto de alumnos en 4.º grado en la población). Cada alumno de la muestra representaba en promedio 51 713/4747 = 10,89 alumnos.

Al llevar a cabo análisis sobre el archivo de datos de evaluación nacional de SPSS que acompaña este libro de trabajo, se recomienda aplicar la ponderación de población (**Wgtpop**). Esto pondera los datos para garantizar una representación proporcional de cada estrato.

El cálculo de las ponderaciones de la encuesta se describe en el capítulo 14 del volumen 3 de esta serie, *Implementación de una evaluación nacional del rendimiento académico* (Dumais y Gough, 2012b). Los pasos expuestos en el capítulo generarán automáticamente las ponderaciones requeridas para analizar los datos de la evaluación nacional. En los análisis presentes en este volumen, desde el capítulo 2 en adelante, se utiliza **Wgtpop** para ponderar los datos.

SPSS

Algunos de los archivos de datos del CD que acompaña este libro (como ***NATASSESS.SAV***)[3] se encuentran en formato SPSS. La versión concreta de SPSS usada para analizar los datos de este archivo fue SPSS, versión 18; los archivos de datos se han analizado también utilizando versiones más recientes de SPSS. A los efectos de los ejercicios presentados en este volumen, todos los archivos de datos (los datos de evaluación y los archivos basados en cada uno de los

cuestionarios) se fusionaron en un solo archivo SPSS consistente en datos sobre rendimiento estudiantil y otros datos en 4747 casos.[4]

A efectos del análisis, las variables en los niveles de escuela y profesores se desagregaron en el nivel de alumnos. En otras palabras, se asignó a cada alumno valores para estas variables correspondientes a valores asignados a su escuela y profesor. Por ejemplo, una de las variables en el cuestionario de docentes era el número de minutos asignados a la enseñanza de matemáticas cada semana. Cuando esta variable se desagregó, se asignó a cada alumno de una clase el mismo número de minutos de instrucción semanal impartidos por su profesor. En el curso de la evaluación nacional se recopiló datos referentes a varios cientos de variables. No obstante, el archivo de datos ***NATASSESS.SAV*** está limitado a un subconjunto de estas variables para mantener el tamaño y la estructura del archivo en un nivel manejable.

Abrir un archivo de datos de SPSS

Hay dos maneras de abrir un archivo de datos. Una es ir a **(My) Computer** en el menú **Windows (Start)** de su escritorio y hacer clic en la unidad y el directorio en que haya guardado su archivo de datos **SPSS**: por ejemplo, ***NAEA DATA ANALYSIS – SPSS DATA – NATASSESS.SAV***.

Puede también abrirse SPSS haciendo clic en **Start**, **All Programs – (IBM) SPSS Statistics**. Pulse en la versión específica de SPSS que aparece en su pantalla. Una vez lanzada, localice el archivo de datos SPSS requerido seleccionando **File – Open – Data**, y busque luego ***NAEA DATA ANALYSIS – SPSS DATA – NATASSESS.SAV***. Haga doble clic en ***NATASSESS.SAV*** para abrirlo.

Usar la barra de herramientas para llevar a cabo análisis preliminares

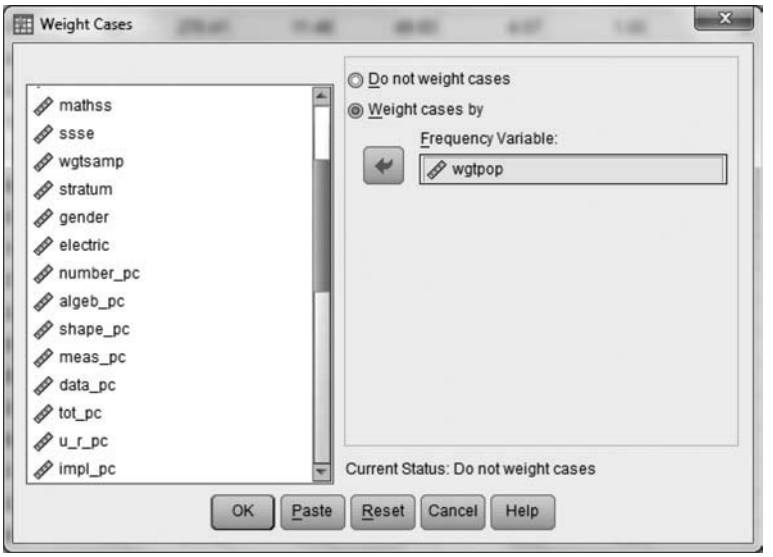
Es posible hacer análisis en SPSS de dos maneras principales, utilizando archivos de sintaxis o bien la barra de herramientas. En nuestro caso se usará la barra de herramientas. Puede hallarse en la parte superior del archivo de datos SPSS abierto. Simplemente haga clic sobre los procedimientos que desee ejecutar, tal como se muestra en el ejercicio 1.1.

EJERCICIO 1.1

Ejecutar estadísticas descriptivas en SPSS y guardar los archivos

1. Abra el archivo de datos de SPSS ***NATASSESS.SAV***, que puede encontrarse en ***NAEA DATA ANALYSIS\SPSS DATA***.
2. Compruebe que las ponderaciones estén activadas: Data – Weight Cases – Weight Cases by – Wgtpop (figura del ejercicio 1.1A) y haga clic en OK. Use Wgtpop para asegurarse de que las estadísticas calculadas representen a la población. Debería ver el mensaje Weight On en la parte inferior derecha de la pantalla.

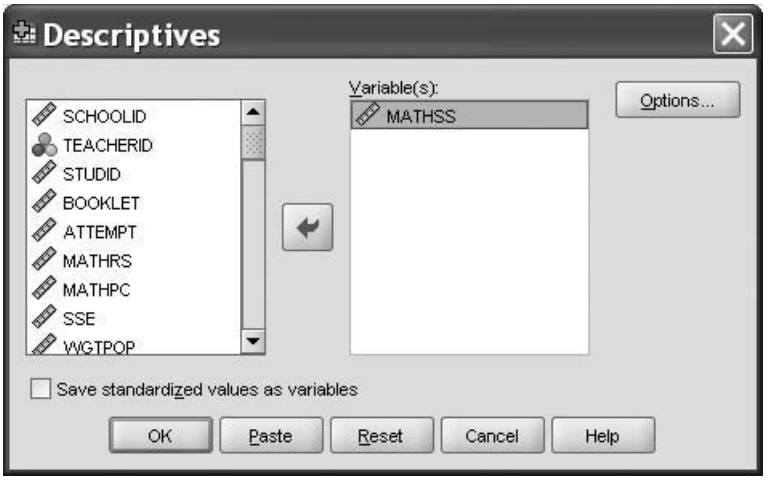
FIGURA DEL EJERCICIO 1.1.A Cuadro de diálogo Weight Cases (ponderar los casos)



3. Seleccione **Analyze – Descriptive Statistics – Descriptives.**[a]
4. En el cuadro de diálogo **Descriptives**, resalte la variable requerida en el panel izquierdo (en este caso, **Mathss**, los puntajes de escala en la prueba de desempeño en matemáticas). Haga clic en la flecha para moverla a **Variable(s)** (véase la figura del ejercicio 1.1.B).[b] Haga clic en **OK**. Su resultado debería mostrar una tabla con una media ponderada de 249,99 (que se redondea a 250) y una desviación estándar de 49,99780 (que se redondea a 50).[c]
5. Use **File – Save As** para asignar un nombre adecuado a su archivo de resultados de SPSS (por ejemplo, ***EXERCISE 1.1.SPV***), y guárdelo en ***NAEA DATA ANALYSIS\MY SOLUTIONS***. Luego seleccione **File – Close**.

EJERCICIO 1.1 *(continúa)*

FIGURA DEL EJERCICIO 1.1.B Cuadro de diálogo Descriptives de SPSS



6. Para guardar su archivo de datos SPSS, que debería estar en el modo de editor de datos, seleccione **File – Save As** …. Guárdelo en ***NAEA DATA ANALYSIS\MY SPSS DATA*** usando el nombre de archivo ***NATASSESS.SAV***. Luego seleccione **File – Exit**

a. Si ve texto en la pantalla en vez de datos, cambie del modo visor al modo editor de datos haciendo clic en **Window** y en (IBM) **SPSS Statistics Data Editor**.
b. Cuando abra un cuadro de diálogo, puede que vea la lista de etiquetas de variables (etiquetas asignadas a cada nombre de variable) en vez de los nombres de las variables. Asimismo, puede que vea que las variables están en orden alfabético en vez del orden en que aparecen en el archivo de datos. Para modificar estas preferencias, cierre el cuadro de diálogo y haga clic en **Edit – Options – General**. Luego seleccione las opciones que desee en el cuadro de la lista de variables.
c. Para reducir el número de posiciones decimales a una, haga doble clic y resalte los dígitos en la celda correspondiente (como por ejemplo 249,99) en la tabla. Haga clic con el botón derecho en **Cell Properties – Format Value – Number – Decimals – 1**.

WESVAR

WesVar es un paquete estadístico que se emplea a menudo conjuntamente con SPSS para analizar datos de evaluaciones nacionales. Además de presentar algunos ejercicios preliminares con SPSS, el capítulo 2 describe las razones para usar WesVar, mientras que los capítulos 3 a 6 describen diversos análisis que utilizan WesVar. El programa WesVar (que incluye un extenso menú de ayuda) puede descargarse del cibersitio Westat.[5]

NOTAS

1. Si en un estrato hubiese 20 escuelas, y de ellas participasen 18, el factor de corrección apropiado habría sido 20/18 o 1,11.

2. Si hubiera habido cinco clases de 4.º grado en la escuela, y se hubiera seleccionado a tres para participar, el componente de selección de alumnos habría sido 5/3. Por otro lado, si hubiera habido 100 alumnos de 4.º grado, y se hubiese seleccionado aleatoriamente a 35 para participar, el componente habría sido 100/35.

3. La extensión .SAV se usa cuando se guarda un archivo de datos de SPSS, mientras que se usa .SPV cuando se guarda un archivo de resultados de SPSS.

4. El capítulo 12 del volumen 3 de esta serie, *Implementación de una evaluación nacional del rendimiento académico*, contiene detalles de cómo fusionar archivos usando Access (Freeman y O'Malley 2012). Los archivos pueden también fusionarse en SPSS, mediante las opciones **Data** y **Merge Files** de la barra de herramientas (véase el Anexo I.C).

5. Puede bajarse la versión 5.1 de WesVar de forma gratuita en http://www.westat.com/our-work/information-systems/wesvar-support/download-wesvar.

CAPÍTULO 2

ANÁLISIS DE LOS DATOS DE UNA EVALUACIÓN NACIONAL UTILIZANDO SPSS

Este capítulo analiza una serie de datos de una evaluación nacional utilizando SPSS. Los ejercicios se han diseñado para permitir al analista comprender y calcular datos como, por ejemplo, la puntuación media global, las puntuaciones medias de los grupos constituyentes (como las regiones) y la variabilidad de las puntuaciones de las pruebas de grupo. Los análisis descritos en el capítulo se basan en datos ponderados.

La idea de una distribución de puntuaciones es un concepto central en el capítulo. Una distribución es un grupo de puntuaciones de una muestra sobre una variable única, como las puntuaciones de una prueba de rendimiento. Por ejemplo, si una prueba de matemáticas con una puntuación máxima de 10 puntos se administra a una muestra de 20 estudiantes, se puede obtener como resultado la siguiente distribución de puntuaciones: 0, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 10. En una evaluación nacional, en la que cientos o incluso miles de estudiantes pueden hacer la misma prueba, evidentemente el número de puntuaciones será mucho mayor. Este capítulo describe las medidas de (a) la tendencia central de las puntuaciones, (b) la dispersión de las puntuaciones, (c) la posición de las puntuaciones y (d) la forma de las distribuciones. Los ejemplos se basan en una distribución ponderada de puntuaciones de pruebas de matemáticas para

los 4747 casos para los cuales están disponibles los datos de la prueba de evaluación nacional.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de resumen más comunes que representan el *valor central* o típico de un conjunto de puntuaciones de pruebas son la media, la mediana y la moda.

Para calcular la *media* (sin ponderar) de un conjunto de datos, como las puntuaciones de las pruebas de rendimiento en matemáticas de los alumnos, hay que sumar cada uno de los valores. Luego se divide la suma por el número de puntos de los datos que contribuyeron a la suma (el número de estudiantes que realizó la prueba).

La *mediana* es el punto medio de un conjunto de números organizados por orden de magnitud.

La *moda* es el valor más frecuente en un conjunto de datos. Una distribución con dos modas se denomina *bimodal*.

A continuación presentamos un conjunto de nueve puntuaciones de pruebas para alumnos que hicieron una prueba de historia: 45, 52, 55, 55, 59, 60, 70, 71, y 73. La puntuación media es 60, la mediana de la puntuación es 59 y la puntuación modal es 55.

MEDIDAS DE DISPERSIÓN

La dispersión es un concepto central en estadística. Las medidas estadísticas de dispersión que se usan más comúnmente son la varianza, la desviación estándar y el rango.

La *varianza* es una medida para conocer cuántas puntuaciones de prueba varían o son dispersas. Para calcular la varianza de un conjunto de puntuaciones, se calcula la distancia (denominada una desviación) entre cada puntuación y la puntuación media. Las desviaciones se elevan al cuadrado y se suman, y luego se dividen por el número de casos menos uno. De este modo, la varianza es la diferencia media elevada al cuadrado entre cada uno de los puntos de la distribución y la media.

Un estadístico relacionado, la *desviación estándar*, es la raíz cuadrada de la varianza.

Existen también otras medidas de dispersión de las puntuaciones menos utilizadas. El *rango* de las puntuaciones de una distribución es la diferencia entre las puntuaciones superiores e inferiores. Si la puntuación inferior es 30 y la superior es 70, el rango es 40. El *rango intercuartil* (RIC) es la diferencia entre puntuaciones en los percentiles 25 (cuartil 1) y 75 (cuartil 3) de una distribución. (El percentil se describe en la próxima sección). El RIC es útil como valor de referencia para identificar valores atípicos (tales como valores mayores a 1,5 RIC por debajo del valor en el cuartil 1 o por encima del valor en el cuartil 3).

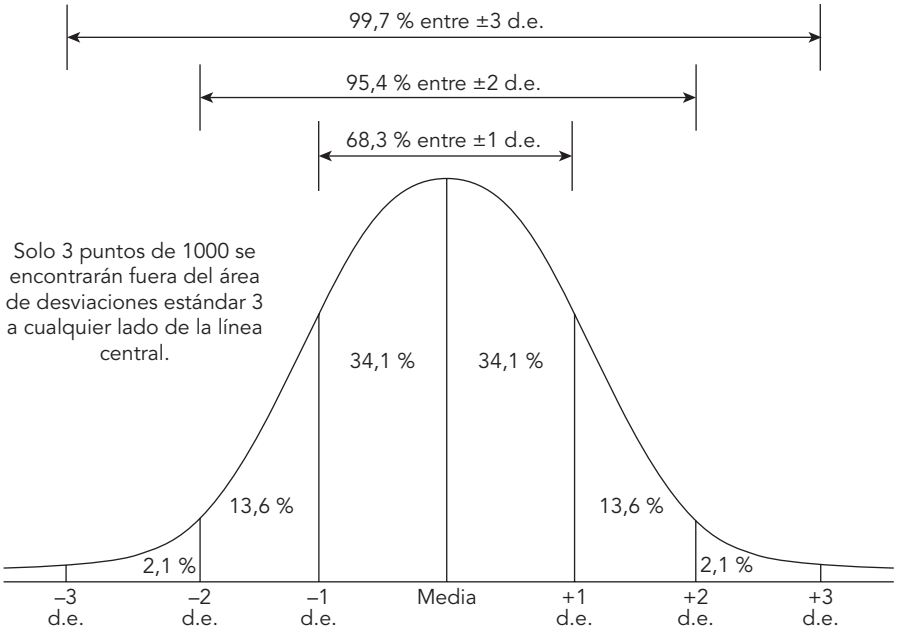
MEDIDAS DE POSICIÓN

La posición relativa de un miembro determinado de un conjunto, tal como la puntuación de un estudiante en comparación con las actuaciones de los demás alumnos que hicieron una prueba, puede identificarse de diversas maneras. Una es la clasificación de los percentiles de una puntuación o un valor específicos. Esto es el porcentaje de puntuaciones o valores que se sitúan dentro de una puntuación concreta. Por ejemplo, una puntuación con un rango percentil de 62 en una evaluación nacional significa que el 62 por ciento de los alumnos obtuvieron una puntuación inferior. Para calcular el rango percentil, las puntuaciones de las pruebas se clasifican de menor a mayor, a continuación se calcula el porcentaje de puntuaciones que son inferiores a una puntuación especificada. Algunas evaluaciones nacionales e internacionales ofrecen información sobre las puntuaciones de las pruebas junto con sus errores estándar (véase el capítulo 3) para percentiles determinados como el 10, 25, 50, 75 y 90. El rango percentil es fácil de comprender pero un análisis estadístico significativo es limitado debido a que la propiedad del intervalo del sistema de medidas se destruye durante la transformación de puntuaciones en percentiles.

Una puntuación o un valor pueden indicarse en términos del número de desviaciones estándar por las que se desvían de la media. En una distribución normal, aproximadamente el 68 por ciento de las puntuaciones se sitúan dentro de una desviación estándar de la media, el 95 por ciento caen dentro de dos desviaciones estándar, y casi el 100 por ciento dentro de tres desviaciones estándar. La Figura 2.1 lo ilustra gráficamente.

FIGURA 2.1

Distribución normal que muestra las unidades de desviación estándar



Nota: d.e. = desviación estándar.

Consideremos, por ejemplo, una distribución normal de puntuaciones con una media de 250 y una desviación estándar de 50. Dado que las puntuaciones se distribuyen normalmente, aproximadamente el 34 por ciento de los alumnos obtuvieron puntuaciones de entre 250 y 300; las puntuaciones conseguidas por otro 34 por ciento fueron de entre 200 y 250. Una puntuación de 325 correspondería a desviaciones estándar 1,5 (75 puntos) por encima de la media, en tanto que una puntuación de 125 correspondería a desviaciones estándar 2,5 (125 puntos) por debajo de la media.

MEDIDAS DE FORMA

Al examinar una distribución de puntuaciones de pruebas, es preciso considerar la forma de los datos, es decir, si la distribución está agrupada en una dirección o en la otra, porque una desviación significativa

de la normalidad puede vulnerar las presunciones para algunas técnicas estadísticas. En una distribución con un sesgo positivo, la mayoría de las puntuaciones están agrupadas en el extremo inferior, con unas pocas puntuaciones que se extienden hacia el extremo superior (Figura 2.2). Esto puede suceder cuando una prueba es particularmente difícil y la mayoría de los alumnos obtienen buenas calificaciones. En algunas evaluaciones nacionales existe este problema cuando la prueba es demasiado difícil para la población. En una distribución con sesgo positivo, la media es normalmente superior a la mediana.

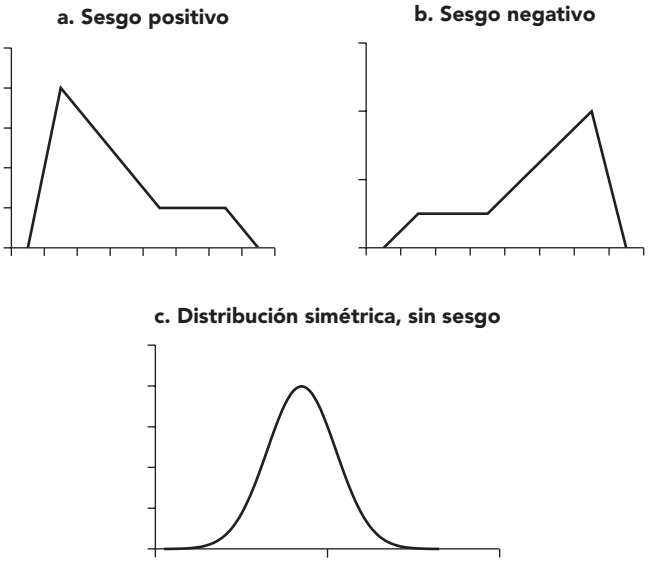
En una distribución con sesgo negativo, la mayoría de las puntuaciones están agrupadas en el extremo superior, y unas pocas se extienden hacia el extremo inferior (Figura 2.2). Esto puede suceder cuando una prueba es particularmente fácil y muchos alumnos obtienen calificaciones altas. En una distribución con sesgo negativo, la media es normalmente menor que la mediana.

En una distribución simétrica (Figura 2.2) la media, la mediana y la moda están próximas una a la otra y cercanas al centro.

La curtosis es una medida de los "valores máximos" de las puntuaciones en torno a la media. Indica si el gráfico de la distribución de las

FIGURA 2.2

Ejemplos de distribuciones con sesgos positivo, negativo, y sin sesgo



puntuaciones tiene más picos o es más uniforme que el de una distribución normal. Un conjunto de datos con curtosis alta (leptocúrtica) tiende a presentar un pico pronunciado en la media, mientras que un conjunto de datos con curtosis baja (platicúrtica) tiende a presentar un pico relativamente uniforme sobre la media, en comparación con la distribución normal. En una distribución normal, el valor del estadístico curtosis es cero o está muy próximo a cero.[1]

ANÁLISIS DE UN CONJUNTO DE DATOS UTILIZANDO SPSS

El comando **Explore** de SPSS ofrece una variedad de estadísticos y gráficos asociados que son muy útiles para analizar una distribución de puntuaciones. Además de los estadísticos descriptivos como la media, la mediana, la moda y la desviación estándar, **Explore** proporciona medidas de sesgo y curtosis, diagramas de tallo y hojas, histogramas, diagramas de caja y diagramas normales. Utilice **Explore** para analizar todos los casos en una distribución de los resultados de una evaluación nacional o para centrarse en los subgrupos, como por ejemplo, alumnos de género masculino y femenino, o estudiantes que asisten a centros escolares de diferentes regiones.

En el siguiente ejercicio, **Explore** se ejecuta utilizando el mismo conjunto de datos empleado en Implementación de una evaluación nacional del rendimiento académico (Greaney y Kellaghan, 2012) (***NATASSESS4.SAV***).[2] Se hace hincapié en los resultados para una variable, **Mathss** (puntuación en una escala de matemáticas). El ejercicio 2.1 presenta un número de enfoques alternativos para obtener estadísticos descriptivos. Antes de iniciar el ejercicio, vaya a la barra de herramientas de SPSS. Haga clic en **Edit – Options – General**, y compruebe que se ha seleccionado **Display Names** de entre las opciones de **Variable Lists**. Haga clic en OK.

También puede utilizarse **Explore** con el fin de realizar un análisis inicial para comparar los niveles de variables únicas como **Gender** (sexo) o **Region** (región). El ejercicio 2.2 describe como considerar los estadísticos de resumen para las cuatro regiones para las que se obtuvieron datos en la evaluación nacional.

EJERCICIO 2.1

Ejecución de Explore en SPSS, variable dependiente única (un nivel)

1. Abra el archivo de datos ***NAEA DATA ANALYSIS\MY SPSS DATA\NATASSESS.SAV***. (Tenga en cuenta que los datos del ejercicio 1.1 se han guardado en esta subcarpeta.)
2. Compruebe que se hayan activado las ponderaciones: haga clic en **Data – Weight Cases – Weight Cases by – Wgtpop – OK**.
3. En la barra de herramientas, seleccione **Analyze – Descriptive Statistics – Explore**.
4. Mueva **Mathss** (puntuación en una escala de matemáticas) a **Dependent List**. Mueva **Studid** a **Label Cases by**.[a]
5. Confirme que se haya marcado **Both** en **Display** (esto garantizará que se presenten en sus resultados tanto los diagramas como los estadísticos). Haga clic en **Statistics** (esquina superior derecha).

   Compruebe que aparezca una marca de verificación frente a **Descriptives**. Haga clic en **Continue – Plots**. Asegúrese de que se haya marcado **Stem-and-Leaf**. Haga clic en **Continue – OK**.
6. Haga clic en **Window** en **toolbar**, y luego en **Output1**. Guarde el resultado en ***NAEA DATA ANALYSIS\MY SOLUTIONS\EJERCICIO 2.1.SPV***.[b]

La tabla del ejercicio 2.1.A ofrece un resumen de procesamiento de casos. Como los datos se ponderaron para el tamaño de la población, los 4747 casos de la base de datos representan una población de 51 713. No falta ningún caso (porcentaje válido: 100).

TABLA DEL EJERCICIO 2.1.A Resumen de procesamiento de casos

	Casos					
	Válido		Faltan		Total	
Variable	*N*	Porcentaje	*N*	Porcentaje	*N*	Porcentaje
Matemáticas	51 713	100,0	0	0,0	51 713	100,0

Las siguientes explicaciones describen cómo interpretar los estadísticos en la tabla de ejercicios h 2.1.B:[c]

- ***Puntuación media*** (250,0) es la media aritmética ponderada. El error estándar de la media es 0,22. (Véase el capítulo 4 donde se incluye una descripción de los errores estándar.)
- ***El 95 % del intervalo de confianza para la media*** es el intervalo aproximado estimado de los valores que tienen una probabilidad del 95 por ciento de incluir la media desconocida de la población de la evaluación nacional. El intervalo de confianza se extiende desde 249,6 (límite inferior) hasta 250,4 (límite superior). (Esto se basa en la puntuación media ±1,96 multiplicado por el error estándar.)

(continúa)

EJERCICIO 2.1 *(continúa)*

TABLA DEL EJERCICIO 2.1.B Estadísticos descriptivos

Variable	Descripción		Estadístico	Error estándar
Matemáticas	Media		250,0	0,22
	95 % de intervalo de confianza para la media	Límite inferior	249,6	
		Límite superior	250,4	
	Media recortada al 5 %		251,1	
	Mediana		256,3	
	Varianza		2499,8	
	Desviación estándar		50,0	
	Mínimo		88,4	
	Máximo		400,0	
	Intervalo		311,6	
	Rango intercuartil		67,1	
	Sesgo		−0,380	0,011
	Curtosis		−0,101	0,022

- ***Media recortada al 5 %*** es la media aritmética calculada cuando se ha descartado el 5 por ciento superior y el 5 por ciento inferior de las puntuaciones (casos). Ofrece una mejor medida de la tendencia central si los datos no son simétricos. La media recortada al 5 por ciento es 251,1.
- ***La mediana*** es el valor por debajo del cual se sitúa el 50 por ciento de los casos. Es también el percentil 50º. La mediana se calculó en 256,3.
- ***La varianza*** es una medida del alcance de la dispersión o extensión de la distribución de las puntuaciones de la prueba. La varianza es 2499,8.
- ***La desviación estándar*** (igual a la raíz cuadrada de la varianza) es 50,0.
- ***Mínimo y máximo*** son los valores inferior (88,4) y superior (400,0) de la distribución.
- ***El rango*** de puntuaciones es la diferencia (311,6) entre los valores de puntuación inferior y superior de la distribución.
- ***El rango intercuartil*** es la distancia que hay entre los valores del tercer cuartil (percentil 75) y del primer cuartil (percentil 25) y ofrece una medida de la dispersión de los datos. El RIC para las puntuaciones es 67,1.
- ***El sesgo*** proporciona una medida de la asimetría de una distribución. Una distribución normal es simétrica y tiene un valor de sesgo en torno a cero. El sesgo es ligeramente negativo (−0,38). Un valor de sesgo entre −1 y +1 se considera muy adecuado para la mayoría de los usos psicométricos, pero normalmente un valor que esté entre −2 y +2 es aceptable.
- ***La curtosis*** estudia en qué medida las observaciones se agrupan en torno a un punto central (los valores máximos de la distribución de la probabilidad). En una distribución normal, el valor de la curtosis es cero, o está muy próximo a cero. Una curtosis positiva

EJERCICIO 2.1 *(continúa)*

excesiva indica que las observaciones (puntuaciones) se agrupan más en torno al eje central y tienen colas más planas (distribución leptocúrtica) que las de una distribución normal. Una curtosis negativa excesiva indica que las observaciones se agrupan menos y tienen colas más altas (distribución platicúrtica). Tal como sucede con el sesgo, un valor de curtosis entre −1 y +1 se considera muy bueno, pero un valor entre −2 y +2 también es aceptable normalmente. Nuestro valor obtenido de −0,101 se sitúa dentro de ambos límites.

El análisis produce un diagrama de tallo y hojas (figura de ejercicio 2.1.A), que presenta la forma y densidad relativa de los datos. Es un método de presentar las frecuencias de las puntuaciones. Cada valor observado (clasificado en orden ascendente) se divide en dos componentes: los primeros dígitos (tallo) y los dígitos finales (hoja). El tallo representa los dígitos que marcan la decena (o superior) de una puntuación, y la hoja contiene los últimos dígitos. Por ejemplo, el tallo 15 muestra que 821 (ponderado) alumnos consiguieron puntuaciones de entre 150 y 159 (inclusive). Los datos indican también que los valores inferiores o iguales a 117, e iguales o superiores a 386, se consideran "extremos" por razones que explicaremos más adelante.

FIGURA DEL EJERCICIO 2.1.A Diagrama de tallo y hojas para puntuaciones por escala matemáticas

Diagrama de tallo y hojas para puntuaciones en una escala de matemáticas

```
 345,60 Extremas    (=<117)
  101,22    11 .  79&
  314,96    12 .  03799&&
  420,52    13 .  1135788&&
  838,18    14 .  0133455556678888999&
  821,09    15 .  000234566778899&
 1031,31    16 .  01112234556667788889
 1151,17    17 .  00123444555566677888899
 1453,47    18 .  000112222223333345555566677899
 2092,02    19 .  00011111223333444445555566667778888899999
 2451,99    20 .  000011112222222333334444455555566666777777888999
 2716,84    21 .  0000011122222333344444455555556666667777777888899999999
 2557,21    22 .  000001111222233334444455556666667777777888888999999
 3687,82    23 .  00000000011112222222333333334444444555555555566666667777777888888899999999
 3404,89    24 .  0000000001111111222222233333333344444445555556666667777777888888889999999
 4134,60    25 .  00000000111112222222333333444444444555555555666666667777777777888888888999999999
 4588,34    26 .  000000000111111111111111112222222222222333333333444444444455555555666666677778888888899999999
 4204,67    27 .  00000000011111111112222222223333333334444445555555566666666667777777888888999999999999
 4387,03    28 .  0000000001111111112222222222223333333333344444444445555556666666667777777888888899999999
 3105,82    29 .  000001111122222223333333344444445555566666666677777888888899999999
 3052,00    30 .  00001111111122222333333333444444445555566666667777788899999
 1795,08    31 .  000000011222223333444455555667788889
 1085,84    32 .  011112233445566678899
  817,82    33 .  0000113445666778&
  493,64    34 .  01234778998&
  271,30    35 .  012268
  296,97    36 .  01779&&
   56,81    37 .  &
   23,51    38 .  &
   11,27 Extremas    (>=386)

Ancho del tallo:  10,00
Cada hoja:      50 caso(s)

& denota hojas de fracciones
```

Cierre el archivo de datos de SPSS ***NATASSESS.SAV*** seleccionando **File-Exit** en la barra de herramientas. En la carpeta **Exercise Solutions** (Soluciones para ejercicios) hay un archivo para este ejercicio.

(continúa)

EJERCICIO 2.1 *(continúa)*

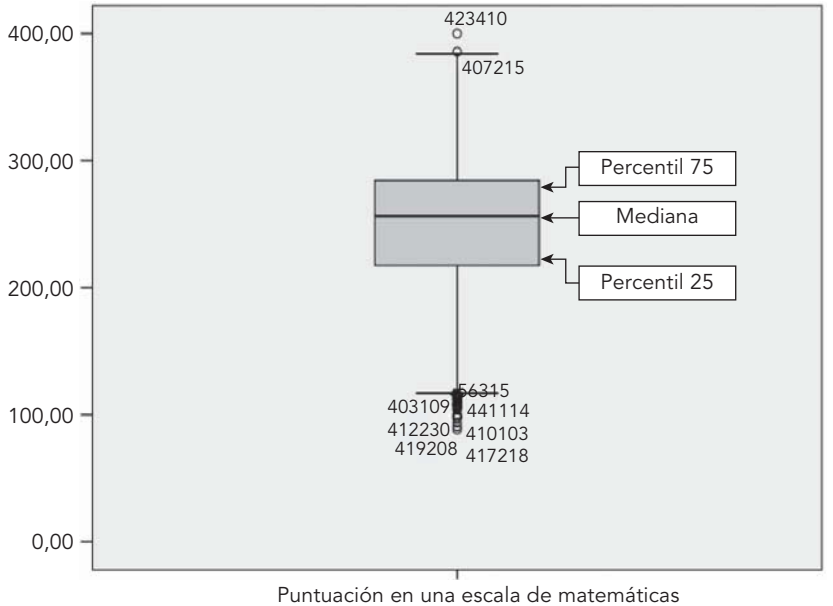
El comando **Explore** también genera un diagrama de caja (o diagrama de caja y bigotes) (figura del ejercicio 2.1.B). Esta es una representación gráfica de la distribución de las puntuaciones de pruebas que incluye la mediana (percentil 50) y los percentiles 25 y 75. La distancia entre la parte superior e inferior de la caja (entre los percentiles 25 y 75) es el RIC, o la distancia entre las puntuaciones mínima y máxima en el 50 por ciento de las puntuaciones en la distribución.

El diagrama de caja muestra también valores atípicos y valores extremos. Los bigotes (las líneas que se extienden entre la parte superior e inferior de la caja) representan los valores máximo y mínimo que no son atípicos ni extremos. Los valores atípicos (valores que se sitúan entre 1,5 y 3 veces el RIC) y los valores extremos (valores que son más de 3 veces el RIC) se representan mediante pequeños círculos detrás de los bigotes. Los números incluidos en la lista (tal como 423410) corresponden a los números de identificación de los alumnos (**Studid**) que han registrado puntuaciones atípicas o extremas.

El diagrama de caja (figura del ejercicio 2.1.B) puede ofrecer información útil en un formato visual. Representa las siguientes características de los datos:

- La mediana (la línea que cruza el centro de la caja) es el punto medio de la distribución y, al igual que la puntuación media, es una medida de tendencia central.

FIGURA DEL EJERCICIO 2.1.B Diagrama de caja para puntuaciones en una escala de matemáticas



EJERCICIO 2.1 *(continúa)*

- La altura de la caja (el RIC) muestra en qué medida varían los valores de puntuación de la prueba en la distribución.
- Una mediana situada en la mitad inferior de la caja sugiere un sesgo positivo, pero si se encontrara en la mitad superior de la caja indicaría un sesgo negativo. En la figura del ejercicio 2.1.B, la mediana se encuentra hacia la parte media de la caja, indicando un sesgo relativamente poco importante.

Las puntuaciones de valores atípicos, definidas como puntuaciones entre 1,5 y 3 longitudes de caja desde los valores de caja superior (percentil 75) e inferior (percentil 25) y los valores extremos, definidos como puntuaciones que corresponden a más de 3 longitudes de caja, se deben examinar para determinar si son puntuaciones incorrectas o realmente válidas.

a. Quizá desee ejecutar algunos estadísticos opcionales o generar gráficos opcionales. Haciendo clic en **Statistics** (Estadísticos) después del paso 3, puede seleccionar alumnos con valores extremos seleccionando **Outliers** (Valores atípicos) (donde hay una lista de las cinco puntuaciones más altas y las cinco más bajas de la distribución). De un modo similar, al seleccionar **Percentiles** (Percentiles) puede obtener puntuaciones en los percentiles 5, 10, 25, 50, 75, 90 y 95, además de **Descriptives** (Descriptivos) por defecto. Al hacer clic en **Plots** (Diagramas), puede seleccionar **Histogram** (Histograma) además de **Stem-and-Leaf** plot (Diagrama de tallo y hojas) por defecto. Ambos se pueden copiar en un documento Word.
b. Nótese que también existe una copia disponible del resultado para el ejercicio 2.1 en ***NAEA Data Analysis\Exercise Solutions\Exercise2.1.SPV***.
c. Nótese que todas las estimaciones, menos las estimaciones finales de la Tabla 2.1.B se han redondeado—al primer decimal en el caso de las estimaciones, y a la segunda cifra decimal en el caso de los errores estándar. Esto se realizó destacando los valores de la tabla (excepto los de sesgo y curtosis), haciendo clic con el botón derecho del ratón, seleccionando **Cell – Properties – Format – Value – Number**, y estableciendo el número de decimales en 1 (o 2).

EJERCICIO 2.2

Ejecución de Explore en SPSS, variable dependiente única (más de un nivel)

1. Abra el archivo de datos de **SPSS: *NAEA DATA ANALYSIS\MY SPSS DATA\ NATASSESS.SAV.***
2. Seleccione **Analyze – Descriptive Statistics – Explore**.

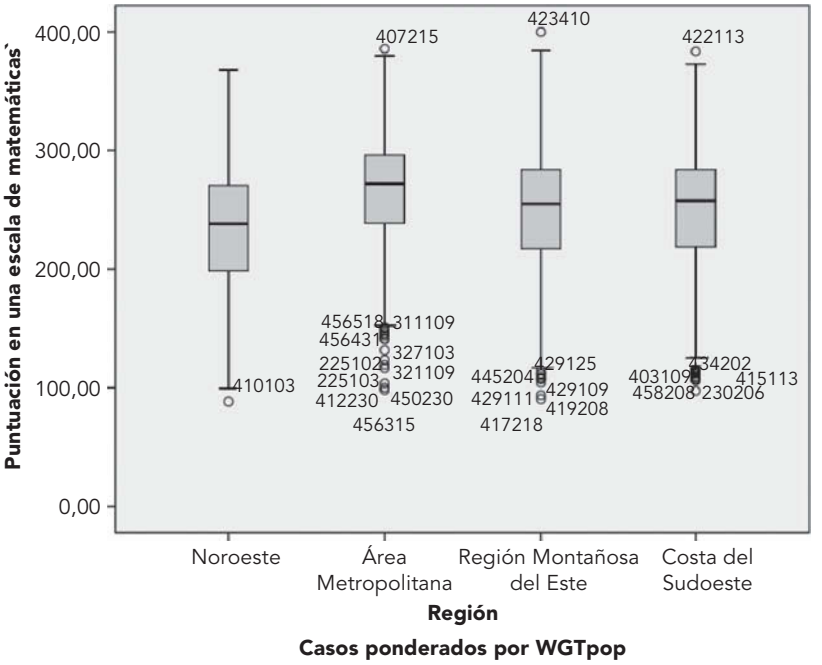
   Mueva **Mathss** a **Dependent List**. Mueva **Region** a **Factor List**. Mueva **Studid** a **Label Cases by**. Compruebe que se haya marcado **Both** en **Display** (esto garantizará que en su resultado se visualicen tanto los diagramas como los estadísticos). Haga clic en **Statistics** (esquina superior derecha). Compruebe que se haya marcado **Descriptives**. Haga clic en **Continue – Plots**. Compruebe que se haya marcado **Stem-and-Leaf**. Haga clic en **Continue – OK**.

(continúa)

EJERCICIO 2.2 *(continúa)*

En el archivo de salida, desplácese hacia abajo para ver los estadísticos descriptivos para cada una de las cuatro regiones. Nótese, por ejemplo, que la puntuación media para la región Noroeste es 233,3. Los datos correspondientes al Área Metropolitana, la Región Montañosa del Este y la Costa del Sudoeste son 265,7; 249,1; y 251,2, respectivamente. Desplácese hacia abajo hasta el extremo final para ver los diagramas de caja para cada una de las cuatro regiones (figura del ejercicio 2.2.A).

FIGURA DEL EJERCICIO 2.2.A Diagramas de caja para puntuaciones en una escala de matemáticas por región



3. El diagrama de caja (figura del ejercicio 2.2.A) muestra la mediana de las puntuaciones para las cuatro regiones (Noroeste, Área Metropolitana, Región Montañosa del Este y Costa del Sudoeste). El analista observará el número relativamente grande de puntuaciones "extremas" del Área Metropolitana, que es una función de la puntuación relativamente alta en el percentil 25 de esa región, comparado con, por ejemplo, el noreste.
4. Guarde el resultado en ***NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 2.2.SPV***. Guarde el archivo de datos de SPSS y salga de SPSS: **File – Save** y **File – Exit**.

a. Los errores estándar pertinentes se calculan en el ejercicio 3.3.

NOTAS

1. Por lo general, las evaluaciones nacionales no ofrecen información sobre sesgo y curtosis. No obstante, pueden tener valor diagnóstico para identificar curvas de distribución que pueden ser problemáticas.

2. A diferencia de ***NATASSESS4.SAV, NATASSESS.SAV*** no contiene zonas *jackknife* ni replicaciones *jackknife* (indicadores). Véase Anexo 1.C.

CAPÍTULO 3

UNA INTRODUCCIÓN A WESVAR

Este capítulo describe los procedimientos para calcular las estimaciones de la población, como las puntuaciones medias, los rangos percentiles y los errores estándar asociados a ellos, utilizando los datos ponderados de una evaluación nacional. Los análisis presentados en este capítulo y en los siguientes se han realizado utilizando el paquete estadístico WesVar que considera la complejidad de una muestra seleccionada en múltiples pasos.

CONFIGURAR UN ARCHIVO DE DATOS EN WESVAR

En primer lugar, verifique si Wesvar está instalado en su ordenador. Wesvar se puede descargar de la página web de Wesvar.[1] Es preciso modificar el archivo de datos de SPSS NATASSESS.SAV antes de guardar el archivo de datos de Wesvar. Para crear el archivo de datos en WesVar, su archivo de datos de SPSS debe incluir variables como las siguientes (aunque los nombres reales de las variables son arbitrarios):

- **Studid**: número de identificación del alumno, un solo número asignado a cada estudiante

- **Wgtpop**: el coeficiente de la población que pondera los datos para ofrecer una estimación acertada de una característica de la población
- **Jkzone**: zona jackknife;[2] cada par de escuelas es asignada a una zona jackknife diferente
- **Jkindic**: indicador jackknife; dentro de cada zona jackknife,[3] se asigna aleatoriamente un valor de 0 a una escuela y a la otra escuela un valor de 1.

El archivo de SPSS, ***NATASSESS4.SAV***, al que se puede acceder en ***NAEA DATA ANALYSIS\SPSS DATA***, contiene los datos de cada una de estas variables clave. Para traer el archivo de datos de SPSS que contiene los datos del cuestionario y de la prueba a WesVar y crear ponderaciones replicadas, siga las instrucciones ofrecidas en el Anexo I.C de este volumen y guarde el nuevo archivo de datos WesVar, ***NATASSESS4.VAR***, en ***NAEA DATA ANALYSIS\MY WESVAR FILES***. Como alternativa, para realizar los siguientes ejercicios, se puede utilizar el archivo de datos de WesVar listo para usar pero sin etiquetar ***NATASSESS4.VAR*** que está en la subcarpeta ***WESVAR UNLABELED DATA***.

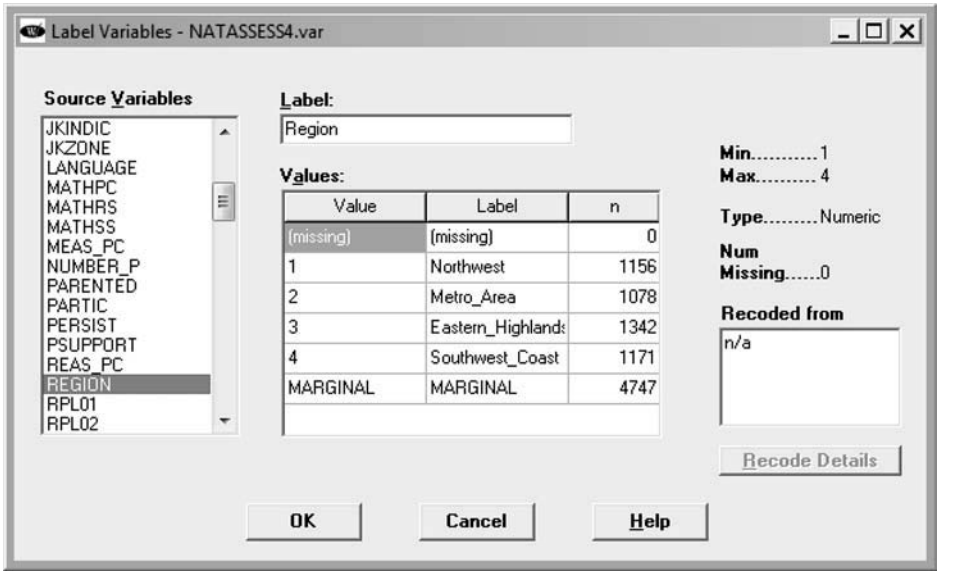
AÑADIR ETIQUETAS A LAS VARIABLES

Siga las siguientes instrucciones:[4]

1. Inicie **WesVar–OpenWesVarDataFile–*NAEADATAANALYSIS\WESVAR UNLABELED DATA*** y abra ***NATASSESS4.VAR***.
2. En la barra de herramientas, seleccione **Format – Label**.
3. Seleccione **Region** en **Source Variables** (a la izquierda).
4. En la casilla después de 1, escriba **Northwest** en la columna de etiquetas; en la casilla después de 2, escriba **Metro_Area**; en la casilla después de 3, escriba **Eastern_Highlands**; y en la casilla después de 4, escriba **Southwest_Coast** (Figura 3.1).
5. Haga clic en **OK**. Se obtendrá el siguiente mensaje: *This operation will create a new VAR file (*Esta acción creará un nuevo archivo). Haga clic en **OK**. Guarde como ***NATASSESS4.VAR*** en ***NAEA DATA ANALYSIS\MY WESVAR FILES***.[5] Esta acción

FIGURA 3.1

Añadir etiquetas a las variables en WesVar



sobrescribirá cualquier archivo de datos existente que tenga el mismo nombre.

6. Seleccione **File – Close**.

CALCULAR ESTADÍSTICOS DESCRIPTIVOS EN WESVAR

Los estadísticos descriptivos se pueden generar de diversas maneras en WesVar. Aquí, se utiliza el comando del menú **Descriptive Stats** para generar algunos estadísticos descriptivos, **Mathss** (puntuación en una escala de matemáticas).

Al abrir WesVar se presentan cuatro accesos separados:

1. **Nuevo archivo de datos WesVar**: Utilizar este acceso para crear un archivo de datos Wesvar a partir de otro formato de archivos como SPSS. (El Anexo I.C describe el proceso para crear un nuevo archivo de datos Wesvar.)
2. **Abrir el archivo de datos Wesvar**: Utilizar este acceso para abrir y modificar un archivo de datos Wesvar, por ejemplo para etiquetar

o recodificar variables. En los siguientes ejercicios se utilizará el archivo de datos creado en Wesvar ***NATASSESS4.VAR***.

3. **New WesVar Workbook**: WesVar requiere que todos los análisis se realicen en un libro de trabajo (véase el ejercicio 3.1). El libro de trabajo debe estar asociado a un archivo de datos Wesvar. Deberá crear un nuevo libro de trabajo para el ejercicio 3.1.
4. **Abrir el libro de trabajo de WesVar**: Este acceso implica abrir un libro de trabajo previamente guardado para realizar nuevos análisis o modificar los existentes. Puede abrir el libro de trabajo WesVar que creó para el ejercicio 3.1 y utilizarlo para realizar otros análisis dentro del mismo capítulo (como los ejercicios 3.2 y 3.3).

Seguir los pasos del ejercicio 3.1.

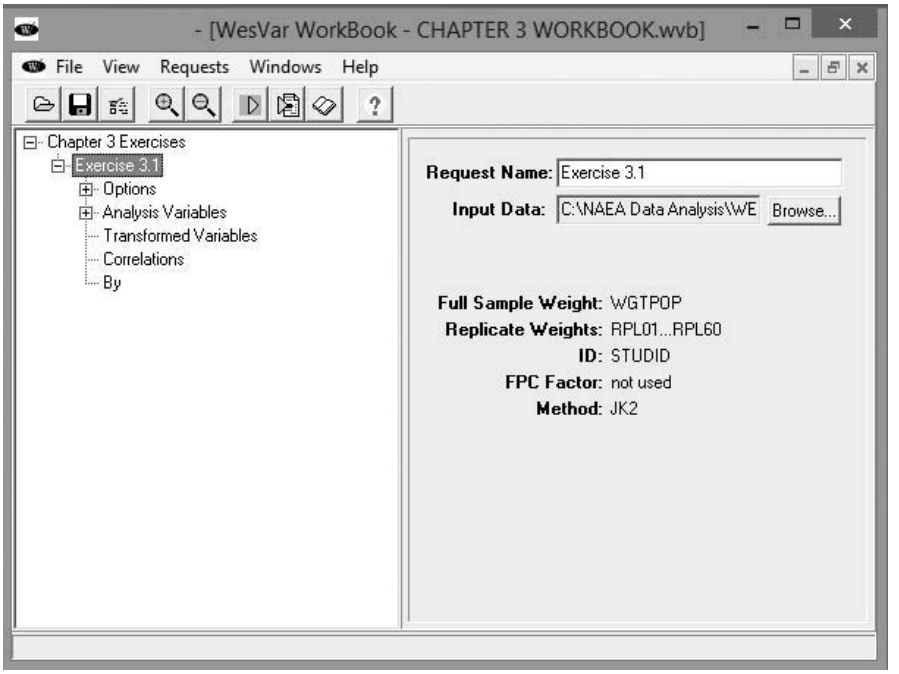
EJERCICIO 3.1

Generar estadísticos descriptivos en WesVar

1. Inicie WesVar, y haga clic en **New WesVar Workbook**. El libro de trabajo le permitirá indicar a WesVar cuáles análisis desea realizar. Aparecerá en pantalla la siguiente advertencia: *Before creating a new Workbook, you will be asked to specify a data file that will be used as the default data file for new Workbook requests* (Antes de crear un nuevo libro de trabajo debe especificar un archivo de datos que se empleará como archivo de datos predeterminado para el nuevo libro de trabajo). Haga clic en **OK**.
2. Aparece en pantalla una ventana con el título **Open WesVar Data File for Workbook**. Localice el archivo de datos ***NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR***.[a]
3. Seleccione **Open** – ***NATASSESS4.VAR***. Haga clic en **Descriptive Stats** (parte inferior derecha de la pantalla). Destaque **WorkBook Title 1** en el panel izquierdo. Cambie este título por otro más específico, destacando el texto en **Title** en el panel derecho e ingresando luego las palabras ***Chapter 3 Exercises***. Cambie también **Request Name** en **Descriptive Request One** ingresando un nuevo nombre, ***Exercise 3.1*** (véase la figura del ejercicio 3.1.A). Guarde el libro de trabajo seleccionando **File – Save As** y localizando el directorio ***NAEA DATA ANALYSIS\MY WESVAR FILES***. Guarde el archivo como ***CHAPTER 3 WORKBOOK***.

EJERCICIO 3.1 ***(continúa)***

FIGURA DEL EJERCICIO 3.1.A Libro de trabajo NEW WESVAR

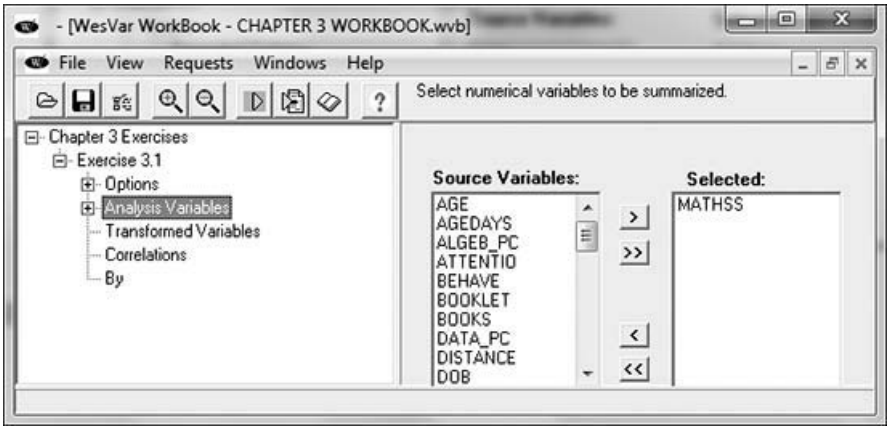


4. Seleccione **Options – Output Control** en el panel izquierdo. Esto permite controlar el número de decimales del resultado. Muchas evaluaciones nacionales utilizan una cifra decimal para las estimaciones (como por ejemplo, puntuaciones medias o rangos percentiles) y dos para los errores estándar. Escriba estas cifras en las dos últimas casillas del panel derecho. (Puede establecer que estas especificaciones sean permanentes para el libro de trabajo actual desde **File – Preferences – General**. Especifique los números significativos de las cifras decimales y haga clic en **Save**).
5. Seleccione **Analysis Variables** en el panel izquierdo. Mueva **Mathss** de **Source Variables** a **Selected** (figura del ejercicio 3.1.B). Las variables adicionales se pueden pasar a la columna **Selected**.
6. Haga clic en el primer ícono **flecha verde** de la barra de herramientas (o seleccione **Requests – Run Workbook Requests** en la barra del menú) para proceder con el análisis. Dé tiempo para calcular el análisis. Haga clic en el ícono **libro abierto** de la barra de herramientas.
7. Para visualizar el resultado, amplíe la pantalla haciendo clic en los signos **+** frente a **Exercise 3.1**, **Variables**, y **Mathss**. Seleccione **Statistics** para ver los resultados (véase la figura del ejercicio 3.1.C).

(continúa)

EJERCICIO 3.1 *(continúa)*

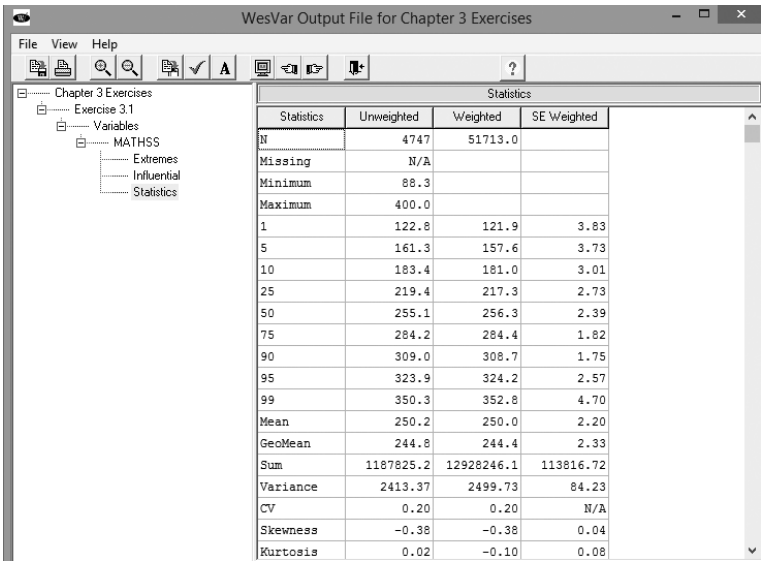
FIGURA DEL EJERCICIO 3.1.B Especificar las variables que se desea analizar en Descriptives de WesVar



Descriptives en WesVar genera los siguientes estadísticos:

- *N* es el número de casos. La columna con el título "Weighted" (Ponderado) indica que cuando se aplica **Wgtpop**, la estimación de la población es de 51 713 casos. El tamaño de la muestra sin ponderar es 4747.

FIGURA DEL EJERCICIO 3.1.C Resultado desde Descriptives de WesVar



Statistics

Statistics	Unweighted	Weighted	SE Weighted
N	4747	51713.0	
Missing	N/A		
Minimum	88.3		
Maximum	400.0		
1	122.8	121.9	3.83
5	161.3	157.6	3.73
10	183.4	181.0	3.01
25	219.4	217.3	2.73
50	255.1	256.3	2.39
75	284.2	284.4	1.82
90	309.0	308.7	1.75
95	323.9	324.2	2.57
99	350.3	352.8	4.70
Mean	250.2	250.0	2.20
GeoMean	244.8	244.4	2.33
Sum	1187825.2	12928246.1	113816.72
Variance	2413.37	2499.73	84.23
CV	0.20	0.20	N/A
Skewness	-0.38	-0.38	0.04
Kurtosis	0.02	-0.10	0.08

EJERCICIO 3.1 *(continúa)*

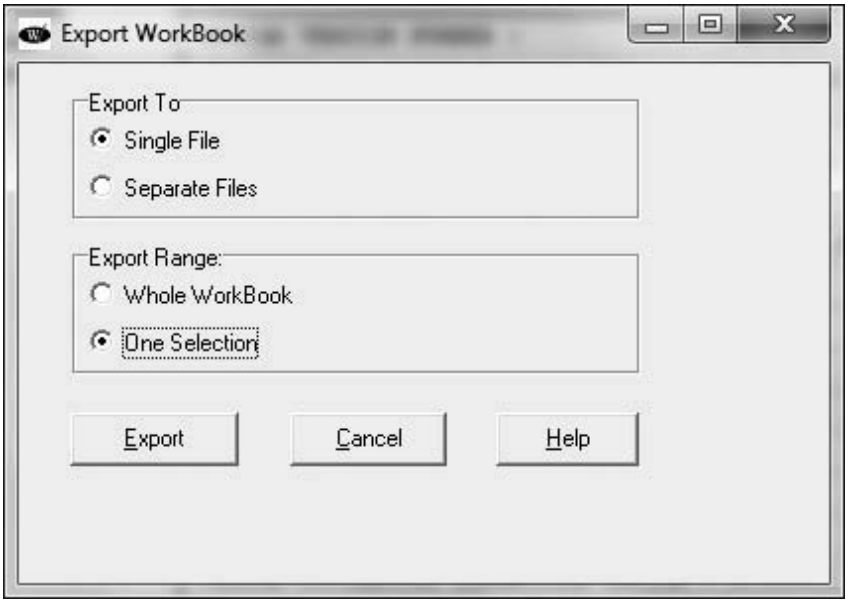
- *Mínimo* y *Máximo*. Los valores obtenidos son 88,3 y 400,0, respectivamente.
- *Rangos percentiles*. Se presentan valores sin ponderar y ponderados para los cuantiles 1, 5, 10, 25, 50, 75, 90, 95 y 99 (rangos percentiles). Normalmente, se reportan los valores ponderados. Por ejemplo, la puntuación ponderada en el percentil 25 es 217,3. Este es el valor estimado por debajo del cual se sitúa la puntuación del 25 por ciento de la población del grado 4. El error estándar (2,73) correspondiente a esta estimación se puede hallar en la última columna, "SE Weighted." La siguiente acción en este capítulo contiene información adicional sobre el significado del error estándar y cómo computarlo.
- *Media*. La media sin ponderar es 250,2. La media ponderada es 250,0, y el error estándar correspondiente es 2,20. No deberíamos sorprendernos si obtenemos una media ponderada de 250 porque la media se estableció en este valor al hacer las escalas. Si lo desea, puede establecer un intervalo de confianza del 95 por ciento en torno a la media, utilizando el error estándar de 2,20 (véase más abajo).
- *Media geométrica*. Se calcula tomando el producto de los valores de una distribución y obteniendo su raíz enésima, donde *n* es el cómputo de los números de la distribución. En general, no se ofrece información sobre la media geométrica en las evaluaciones nacionales.
- *Suma*. Es la suma de los valores de una distribución. No se ofrece información sobre ella en las evaluaciones nacionales.
- *Varianza*. La desviación estándar es la raíz cuadrada de la varianza. La varianza ponderada es 2499,73 (prácticamente 2500). La raíz cuadrada de este valor es 50, que es la desviación estándar del conjunto **Mathss** al hacer las escalas.
- *CV* (coeficiente de variación). Se calcula dividiendo la raíz cuadrada de la varianza por la puntuación media. Los valores sin ponderar y ponderados para el CV son iguales (0,20).
- *Sesgo*. La distribución de **Mathss** tiene un sesgo ligeramente negativo (−0,38). Se debe tener en cuenta que un valor de sesgo de rango ±1 se considera satisfactorio (el capítulo 2).
- *Curtosis*. El valor ponderado para la curtosis es −0,10. Una vez más, este valor se encuentra dentro del rango ±1 recomendado, lo que indica que no existen serios problemas de curtosis.

8. Guarde el resultado de WesVar como un archivo de texto[b] haciendo clic en **File** y **Export** de la barra del menú del archivo de salida. Seleccione **Single File** y **One Selection**, y haga clic en **Export** (figura del ejercicio 3.1.D). Guarde en ***NAEA DATA ANALYSIS\MY SOLUTIONS*** utilizando un nombre de archivo adecuado (por ejemplo, ***EXERCISE 3.1.TXT***).[c]

(continúa)

EJERCICIO 3.1 *(continúa)*

FIGURA DEL EJERCICIO 3.1.D Exportar un archivo de WesVar



9. Haga clic en el ícono **Exit Door** (el último ícono de la barra de herramientas) para volver al libro de trabajo de WesVar. Guarde el libro de trabajo seleccionando **File** – **Save** en la barra de menú. El archivo debería guardarse en ***NAEA DATA ANALYSIS\MY WESVAR FILES*** como ***CHAPTER 3 WORKBOOK***.[d] Es posible utilizar nuevamente este libro de trabajo para comprobar sus respuestas o realizar otros análisis que se describen en este capítulo. Haga clic en **File** y luego en **Exit** para concluir la sesión en WesVar.

Para abrir nuevamente este libro de trabajo en la próxima sesión de WesVar, inicie WesVar y seleccione **Open WesVar Workbook** (panel derecho). Luego localice ***CHAPTER 3 WORKBOOK*** en ***MY WESVAR FILES***.

a. Si todavía no ha guardado ***NATASSESS4.VAR*** en ***NAEA DATA ANALYSIS\MY WESVAR FILES***, cópielo desde ***NAEA DATA ANALYSIS\Wesvardata & WORKBOOKS*** y guárdelo en ***NAEA DATA ANALYSIS\MY WESVAR FILES***. Si se está utilizando **Copy** y **Paste**, es preciso copiar también el archivo de registro asociado ***NATASSESS4.LOG*** en la misma subcarpeta.
b. Los ficheros de texto pueden copiarse en Excel, donde se pueden reformatear para incluirlos en el informe de una evaluación nacional. El resultado también puede copiarse directamente desde el archivo de salida de WesVar en un archivo Excel.
c. Para abrir en una próxima ocasión el resultado guardado, haga clic en **(My) Computer**, localice el archivo (como en ***NAEA DATA ANALYSIS\MY SOLUTIONS***) y luego haga doble clic sobre él. Los archivos .TXT también se pueden pegar en Excel. Destaque y copie los datos relevantes en el archivo de texto y luego utilice **Paste Special** (**Unicode Option**) para conservar el formato.
d. En ***NAEA DATA ANALYSIS\WESVAR DATA & WORKBOOKS*** se puede observar que WesVar produce archivos de listados de salida con extensiones como .001 o .002 cada vez que utiliza un libro de trabajo. No es necesario acceder a estos archivos para utilizar WesVar ni al resultado de WesVar. Para acceder a todos los archivos de WesVar, excepto los archivos de texto que haya creado, es preciso iniciar primero el programa WesVar.

CALCULAR LA PUNTUACIÓN MEDIA Y SU ERROR ESTÁNDAR

El *error estándar* de un estadístico es la estimación de la desviación estándar de ese estadístico si se extrajese muestras infinitas de la población, tal como la que nos ocupa (por ejemplo, todos los alumnos del grado 4). El error estándar de la media es un estadístico importante porque se utiliza en las pruebas para obtener significación estadística. Siempre se debe ofrecer dicha información con los resultados de una evaluación nacional. Los errores estándar pueden calcularse también para otros estadísticos, como los rangos percentiles. El ejercicio 3.2 describe la forma de calcular la puntuación media, su error estándar y su intervalo de confianza en WesVar.

EJERCICIO 3.2

Calcular una puntuación media y su error estándar en WesVar

1. Inicie WesVar, y haga clic en **Open WesVar Workbook.** Localice el libro de trabajo utilizado en el ejercicio 3.1 (***NAEA DATA ANALYSIS\MY WESVAR FILES\ CHAPTER 3 WORKBOOK***).
2. Seleccione **Chapter 3 Ejercicios** (panel izquierdo). Luego haga clic en **Table** (panel derecho). Destaque **Table Request One** (panel izquierdo). Haga clic en **Add Table Set Single** (panel derecho). Haga clic en **Table Request One** (panel izquierdo). Cambie **Request Name** (panel derecho) a ***Exercise 3.2***.
3. Seleccione **Options – Generated Statistics** (panel izquierdo), y compruebe que **Estimate**, **Error estándar**, y **Confidence Interval (Standard)** estén marcados. Elimine las marcas de las otras casillas. En **Ejercicio 3.2** (panel izquierdo), haga clic en **Computed Statistics**. Destaque **Mathss** en el menú **Source Variables** (panel derecho), y haga clic en **Block Mean** (también en el panel derecho) (figura del ejercicio 3.2.A). Haga clic en el ícono **flecha verde** de la barra de herramientas (**Run Workbook Request**). Haga clic en el ícono **Open Book** de la barra de herramientas para ver el resultado.
4. En **Output**, seleccione **Exercise 3.2**, **Table Set #1**, y luego **Overall.** Haga clic en el ícono **+** (plus) para ampliar los nodos según lo necesario. El resultado se muestra en la figura del ejercicio 3.2.B.

   El resultado de WesVar que se muestra en la figura del ejercicio 3.2.B ofrece los siguientes datos:

   - *M_Mathss* es la puntuación media ponderada de 250,0.

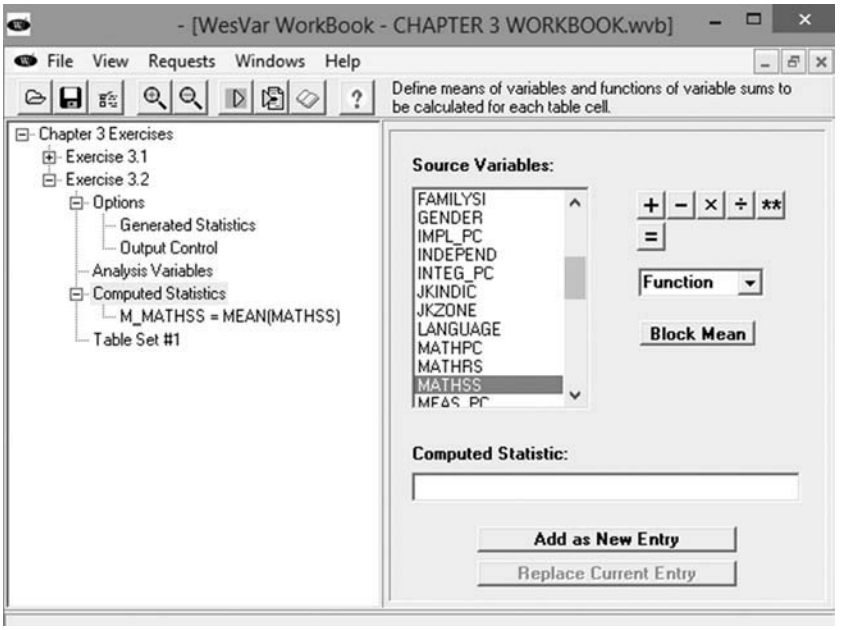
(continúa)

EJERCICIO 3.2 *(continúa)*

- *StdError (Error estándar)* es 2,20.
- El *95 % inferior* y el *95 % superior* forman el intervalo de confianza del 95 por ciento en torno a la puntuación media de 250. Abarca desde 245,6 (límite inferior) hasta 254,4 (límite superior). Es posible encontrar la verdadera puntuación media de la población en este intervalo del 95 por ciento del tiempo.[a] Este intervalo también se podría haber calculado con los datos de la Figura 3.1.C, donde la media y el error estándar se establecieron en 250,0 y 2,20, respectivamente.

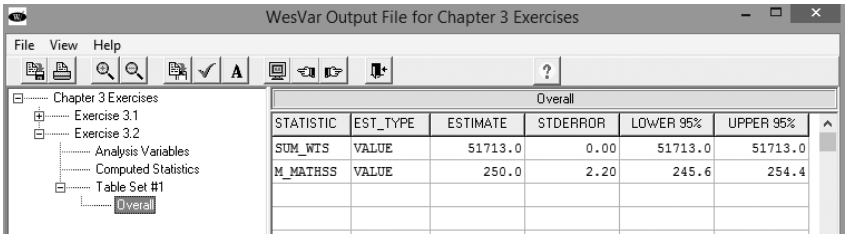
5. Guarde el resultado haciendo clic en **File y Export**. Seleccione las opciones **Single File One Selection**. Exporte el resultado como un archivo de texto en ***NAEA DATA ANALYSIS\MY SOLUTIONS***, usando un nombre de archivo adecuado (como por ejemplo ***Exercise 3.2.Txt***).
6. Salga del resultado mediante el ícono **Exit Door** de la barra de herramientas y guarde el libro de trabajo de WesVar seleccionando **File – Save** y luego **File – Close**. Esta acción debería guardar su libro de trabajo en ***NAEA DATA ANALYSIS\MY WESVAR FILES***.

FIGURA DEL EJERCICIO 3.2.A Especificar un estadístico calculado en una tabla de WesVar



EJERCICIO 3.2 *(continúa)*

FIGURA DEL EJERCICIO 3.2.B Resultado para las tablas de WesVar: Calcular la puntuación media



a. Tenga en cuenta que WesVar multiplica el error estándar por 2,00 en vez de multiplicarlo por el valor más convencional de 1,96 cuando calcula los intervalos de confianza del 95 por ciento. Esto resulta en un intervalo de confianza ligeramente mayor.

CALCULAR LAS PUNTUACIONES MEDIAS Y LOS ERRORES ESTÁNDAR PARA LOS SUBGRUPOS DE LA POBLACIÓN

Es posible que los responsables de las políticas, los investigadores y otras personas deseen observar los estadísticos descriptivos para diferentes niveles de una variable. Por ejemplo, pueden interesarse por las puntuaciones medias del rendimiento en matemáticas a nivel nacional para alumnos de género masculino y femenino, o en las puntuaciones medias de alumnos que asisten a centros educativos de diferentes regiones de un país.

Una simple adición al libro de trabajo de WesVar nos permite calcular la puntuación media y el error estándar para cada una de las cuatro regiones de la evaluación nacional de Sentz (Noroeste, Área Metropolitana, Región Montañosa del Este, Costa del Sudoeste) (Ejercicio 3.3).

EJERCICIO 3.3

Calcular las puntuaciones medias y los errores estándar en WesVar, cuatro regiones

1. Inicie WesVar, y haga clic en **Open WesVar Workbook.** Localice el libro de trabajo utilizado en el ejercicio 3.1 (***NAEA DATA ANALYSIS\MY WESVAR FILES\ CHAPTER 3 WORKBOOK***).

1. Inicie WesVar. Seleccione **Open WesVar Workbook**. Abra el libro de trabajo de WesVar guardado tras completar el ejercicio 3.2. Debería ser ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 3 WORKBOOK.***

2. Destaque el nodo **Exercise 3.2** en el panel izquierdo de la pantalla de su libro de trabajo, y seleccione **Clone** (haciendo clic con el botón derecho). Esta acción crea una copia de la solicitud de la tabla **Compute Mean Score**. Haga clic en **Table Request Two** y etiquétela en el panel derecho como ***Exercise 3.3.*** Amplíe **Exercise 3.3** en el panel izquierdo. En **Options – Output Control,** establezca **Estimate** en una cifra decimal y **Std. Error** en dos. Compruebe que se hayan marcado **Variable Label** y **Value Label**. Elimine la selección de otras opciones.

3. Haga clic en **Table Set #1**. Luego, en el panel derecho, mueva **Region** de **Source Variables** a la casilla **Selected**. Haga clic en **Add as New Entry**.

4. Aplique etiquetas para **Region** si aún no se han aplicado al archivo de datos de WesVar (***Natassess4.var***). En el panel izquierdo, haga clic en **Region** – **Cells**. En **Cell Definition** (panel derecho), haga clic en 1 (en **Region)**, e introduzca ***Northwest*** en el panel **Label**. Pulse **Return** en el teclado, o haga clic en **Add as New Entry**. Continúe este proceso asignando la etiqueta ***Metro_Area*** a 2, ***Eastern_Highlands*** a 3, y ***Southwest_Coast*** a 4. (WesVar no permite dejar espacios en blanco entre palabras de modo que se debe utilizar un subrayado.)

5. Seleccione **Region** (panel izquierdo) y marque **Percent** en **Sum of Weights** en el panel derecho; elimine la selección de otras casillas en **Sum of Weights**. La opción **Percent** ofrecerá el porcentaje de alumnos de cada región, mientras que la puntuación media en matemáticas se generará por separado para cada región porque **Mathss** ya se especificó en **Computed Statistics** (véase la figura del ejercicio 3.3.A).

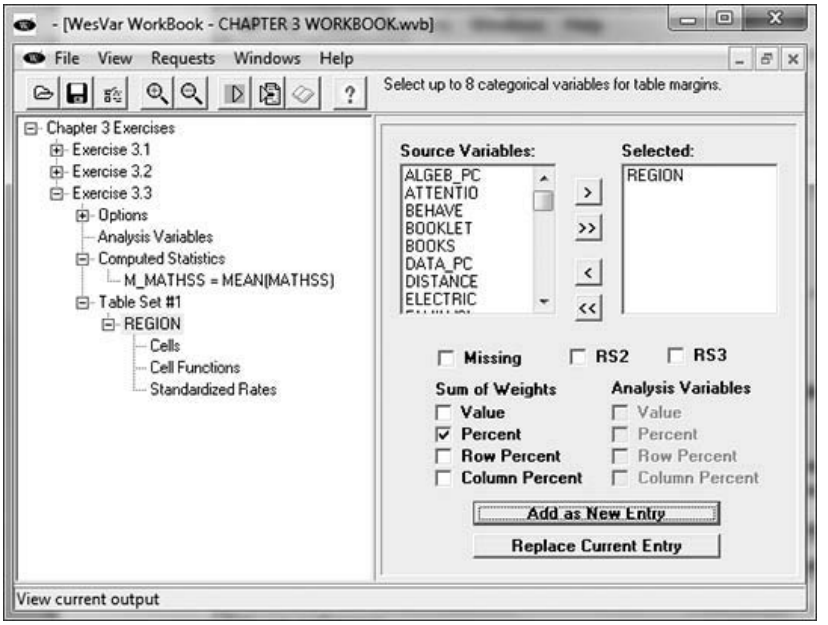
6. Haga clic en el ícono **flecha verde** de la barra de herramientas para ejecutar el análisis y luego en el ícono **Open Book** para visualizar el resultado. El resultado se muestra en la figura del ejercicio 3.3.B. Puede ser necesario hacer clic en el ícono + (más) para ampliar **Exercise 3.3**, **Table Set #1,** y luego en **Region** para visualizar el resultado.

7. Guarde el resultado haciendo clic en **File** y luego en **Export**. Seleccione las opciones **Single File** y **One Selection**. Exporte el resultado como un archivo de texto a ***NAEA DATA ANALYSIS\ MY SOLUTIONS*** usando un nombre de archivo adecuado (por ejemplo ***EXERCISE 3.3.TXT***).

8. Salga del resultado mediante el ícono **Exit Door**, y guarde el libro de trabajo de WesVar seleccionando **File – Save** y luego **File – Close.** Como alternativa, use **File – Save As**, y sobrescriba el libro existente del capítulo 3 en ***NAEA DATA ANALYSIS\MY WESVAR FILES\ CHAPTER 3 WORKBOOK.***

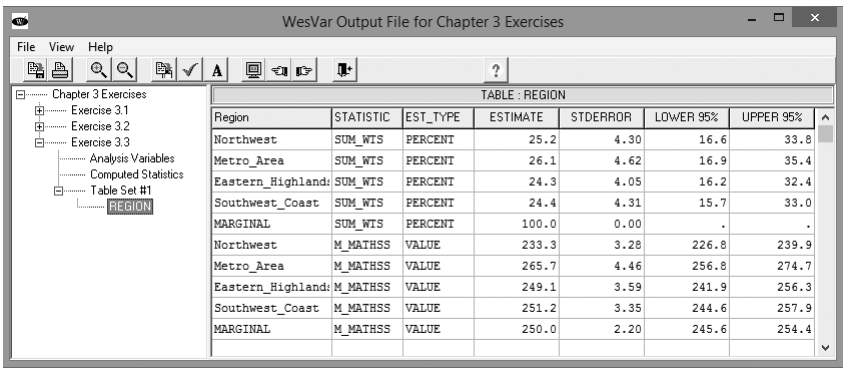
EJERCICIO 3.3 *(continúa)*

FIGURA DEL EJERCICIO 3.3.A Libro de trabajo de WesVar antes de calcular las puntuaciones medias por región



El resultado presentado en la figura del ejercicio 3.3.B comprende las puntuaciones medias, los errores estándar y los intervalos de confianza del 95 por ciento para cada región. Por ejemplo, la puntuación media para la región Noroeste es 233,3 y su error estándar es 3,28. El intervalo de confianza del 95 por ciento correspondiente es de 226,8 a 239,9. La figura muestra también el porcentaje de la población total del grado 4 de cada región (por ejemplo, el 26,1 por ciento corresponde al Área Metropolitana).

FIGURA DEL EJERCICIO 3.3.B Resultado de WesVar para el cálculo de las puntuaciones medias por región



Region	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%
Northwest	SUM_WTS	PERCENT	25.2	4.30	16.6	33.8
Metro_Area	SUM_WTS	PERCENT	26.1	4.62	16.9	35.4
Eastern_Highland:	SUM_WTS	PERCENT	24.3	4.05	16.2	32.4
Southwest_Coast	SUM_WTS	PERCENT	24.4	4.31	15.7	33.0
MARGINAL	SUM_WTS	PERCENT	100.0	0.00	.	.
Northwest	M_MATHSS	VALUE	233.3	3.28	226.8	239.9
Metro_Area	M_MATHSS	VALUE	265.7	4.46	256.8	274.7
Eastern_Highland:	M_MATHSS	VALUE	249.1	3.59	241.9	256.3
Southwest_Coast	M_MATHSS	VALUE	251.2	3.35	244.6	257.9
MARGINAL	M_MATHSS	VALUE	250.0	2.20	245.6	254.4

NOTAS

1. WesVar puede descargarse de forma gratuita desde http://www.westat.com/our-work/information-systems/wesvar-support/download-wesvar.

2. Esta variable se puede denominar también **Jkpair** o "código de emparejamiento de conglomerados para la estimación de la varianza." Es preciso establecer zonas Jackknife (pares JK) para crear las ponderaciones apropiadas (replicaciones).

3. Esta variable se puede denominar también **Jkrep** o "dentro del código de replicación de un par dentro de un conglomerado."

4. Las instrucciones para el etiquetado requieren abrir un archivo de datos en ***WESVAR UNLABELED DATA***. No obstante, si ya se ha creado un archivo de datos siguiendo las instrucciones del Anexo 1.C, se puede utilizar el archivo de datos guardado en ***NAEA DATA ANALYSIS\MY WESVAR FILES*** en vez del archivo de datos de ***WESVAR UNLABELED DATA***.

5. Si se copia ***NATASSESS4.VAR*** de ***WESVAR UNLABELED DATA*** en ***MY WESVAR FILES*** usando **Copy** y **Paste**, el archivo no se ejecutará a menos que se copie también el archivo de registro asociado ***NATASSESS4.LOG*** en la misma subcarpeta.

CAPÍTULO 4

COMPARAR LOS RENDIMIENTOS DE DOS O MÁS GRUPOS

Es posible que los responsables políticos quieran saber si los niveles de rendimiento de subpoblaciones de estudiantes que participaron en una evaluación nacional (por ejemplo, niños y niñas, alumnos de escuelas de diferentes regiones) difieren significativamente entre sí. Este capítulo describe procedimientos que permiten responder a dichas preguntas.

EXAMINAR LA DIFERENCIA ENTRE DOS PUNTAJES PROMEDIO

Para evaluar si la diferencia en el rendimiento entre dos grupos es estadísticamente significativa, se necesita una variable categórica con dos o más niveles y una variable continua o una de escala de intervalo. A continuación se presentan ejemplos de preguntas de interés que implican variables categóricas y continuas:

- *Género* (femenino/masculino) (variable categórica) y *rendimiento en matemáticas* (variable continua). ¿Difiere significativamente el rendimiento en matemáticas de los estudiantes dependiendo de su sexo?

- *Disponibilidad de luz eléctrica para estudiar* (sí/no) (variable categórica) y *rendimiento en lectura* (variable continua). ¿Difiere significativamente el rendimiento en lectura de aquellos estudiantes que no tienen luz eléctrica en su hogar en comparación con el rendimiento de los que sí tienen luz en su hogar?
- *Acceso a ayuda con las tareas para el hogar* (sí/no) (variable categórica) y *rendimiento en ciencias* (variable continua). ¿Los estudiantes que disponen de ayuda con sus tareas para casa tienen un rendimiento significativamente diferente en ciencias al de aquellos estudiantes que no disponen de ayuda?
- *Uso del idioma de instrucción en el hogar* (sí/no) (variable categórica) y *conocimiento cívico* (variable continua). ¿Los estudiantes que utilizan el idioma de instrucción de la escuela en su entorno familiar difieren significativamente en su rendimiento en una evaluación de conocimiento cívico de aquellos estudiantes que hablan un idioma distinto en su entorno familiar?
- *Acceso a su propio libro de texto de lectura en la escuela* (sí/no) (variable categórica) y *rendimiento en lectura* (variable continua). ¿Los estudiantes que tienen sus propios libros de texto de lectura en la escuela obtienen un puntaje de lectura promedio que difiere significativamente del puntaje promedio de los estudiantes que deben compartir un libro de texto?

Una prueba que compara las puntuaciones medias de dos grupos responde a la siguiente pregunta: ¿Dos poblaciones representadas por las dos muestras difieren significativamente en su puntaje promedio en alguna variable? Los resultados se presentan en forma de niveles de significación, llamados valores *p*. El término "estadísticamente significativo" se utiliza para indicar que es improbable que cierta diferencia observada se haya dado por azar. Si, por ejemplo, los resultados muestran que la diferencia media a favor de las niñas es significativa en un nivel de 0,05 (valor *p*), esto quiere decir que hay menos de un 5 % de probabilidad de que la diferencia haya sido causada azarosamente. Un valor *p* de 0,01 significa que la probabilidad de que se produzca el resultado observado si los grupos no difieren en la variable de interés es de uno entre cien.

El error estándar de la diferencia (entre promedios) es un concepto importante al analizar la significación estadística de las diferencias de

puntaje promedio. Si la diferencia de puntaje promedio es lo suficientemente grande como para hallarse fuera de un intervalo de confianza próximo a la diferencia, que se basa en el error estándar de esta, se puede concluir que la diferencia es estadísticamente significativa. Si la diferencia se encuentra dentro del intervalo de confianza, se puede concluir que no es estadísticamente significativa.

En el siguiente ejemplo (ejercicio 4.1), el objetivo es determinar si la diferencia media en el rendimiento en matemáticas entre los estudiantes que tienen luz eléctrica en el hogar y los que no es estadísticamente significativa.

Antes de realizar el ejercicio 4.1, etiquete la variable como **Electric** en su archivo de datos. Para hacer esto, ejecute WesVar y seleccione **Open WesVar Data File**. Abra el archivo de datos ***NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR.*** Luego siga los pasos para etiquetar una variable que se explicaron en el capítulo 3 (páginas 34-35). Etiquete el 1 como ***Yes*** para indicar la disponibilidad de electricidad en el hogar. Etiquete el 2 como ***No*** para indicar la falta de electricidad. Guarde su archivo de datos WesVar en ***NAEA DATA ANALYSIS\MY WESVAR FILES***, sobrescribiendo el archivo ***NATASSESS4.VAR*** existente de ser necesario.

EJERCICIO 4.1

Evaluar la diferencia entre dos puntajes promedio

1. Ejecute WesVar y haga clic en **New WesVar Workbook**. Se mostrará la siguiente advertencia: *Before creating a new Workbook, you will be asked to specify a Data file that will be used as the default Data file for new Workbook requests. (Antes de crear un nuevo libro de trabajo, se le solicitará que especifique un archivo de datos que se utilizará como el archivo de datos predeterminado para las nuevas solicitudes de libro de trabajo).*

   Haga clic en **OK**.

2. Se abrirá una ventana llamada **Open WesVar Data File for Workbook.** Seleccione el archivo de datos ***NATASSESS4.VAR*** en ***NAEA DATA ANALYSIS\MY WESVAR FILES***. Haga clic en **Open**.

(continúa)

EJERCICIO 4.1 *(continúa)*

3. Guarde su nuevo libro de trabajo como ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 4 WORKBOOK.WVB***.
4. Resalte **Workbook Title 1** (panel izquierdo) y escriba ***Chapter 4 Exercises*** en la casilla **Title** (panel derecho). En la casilla **New Request** (panel derecho, mitad inferior), haga clic en **Table**. Resalte **Table Request One** (panel izquierdo) y cambie **Request Name** (panel derecho) a ***Exercise 4.1.*** Haga clic en **Add Table Set Single** (panel derecho).
5. Vaya a **Options** – **Generated Statistics** en el panel izquierdo. Asegúrese de que **Estimate**, **Standard Error** y **Confidence Interval (Standard)** estén seleccionados. Deseleccione las otras casillas.
6. Vaya a **Options** – **Output Control** en el panel izquierdo. Asegúrese de que **Estimates** esté configurado con un decimal y **Std. Error** con dos decimales. Compruebe que las etiquetas variable y valor estén seleccionadas.
7. Seleccione **Computed Statistics**, resalte **Maths** en **Source Variables** (panel derecho) y haga clic en **Block Mean.**
8. Haga clic en **Table Set #1** (panel izquierdo). Mueva **Electric** del cuadro **Source Variables** al cuadro marcado como **Selected** en el panel derecho. Compruebe que la opción **Percent** esté seleccionada bajo **Sum of Weights.** Si está seleccionada, deseleccione **Value**, **Row Percent** y **Column Percent**. Haga clic en el botón **Add as New Entry.**
9. Resalte **Electric** (panel izquierdo) y haga clic en **Cells** debajo. Haga clic en **1** (panel derecho) y escriba ***Yes*** en el cuadro de la etiqueta, y presione **Return** (o haga clic en **Add as New Entry**). Haga clic en **2** y escriba ***No*** en el cuadro de la etiqueta y pulse **Return**.
10. A continuación, seleccione **Cell Functions** (panel izquierdo). Escriba lo siguiente en la casilla llamada
    Function Statistic: ***MeanDiff = Yes – No***. Haga clic en **Add As New Entry.**
11. Debajo de **MeanDiff = Yes – No** (panel izquierdo), Haga clic en **For**. Asegúrese de que **M_Mathss** se encuentre en el cuadro llamado **Selected** (figura del ejercicio 4.1.A). De ser necesario, mueva **Sum_Wts** al cuadro llamado **Source Variables**.
12. Haga clic en la **Flecha verde** en la barra de herramientas para ejecutar el análisis. Espere que se ejecute el programa. Para ver el resultado, haga clic en el ícono **libro abierto** en la barra de herramientas.
13. Para ver los puntajes promedio relacionados con tener o no luz eléctrica en el hogar, expanda **Exercise 4.1** – **Table Set #1** y haga clic en **Electric**. Los datos de los estudiantes en hogares con y sin electricidad se pueden ver en la figura del ejercicio 4.1.B. Tenga en cuenta que los estudiantes de hogares con electricidad obtuvieron un puntaje promedio de matemáticas de 254,3 y un error estándar de 2,30.

EJERCICIO 4.1 *(continúa)*

FIGURA DEL EJERCICIO 4.1.A **WesVar Workbook antes de evaluar la diferencia entre dos puntajes promedio**

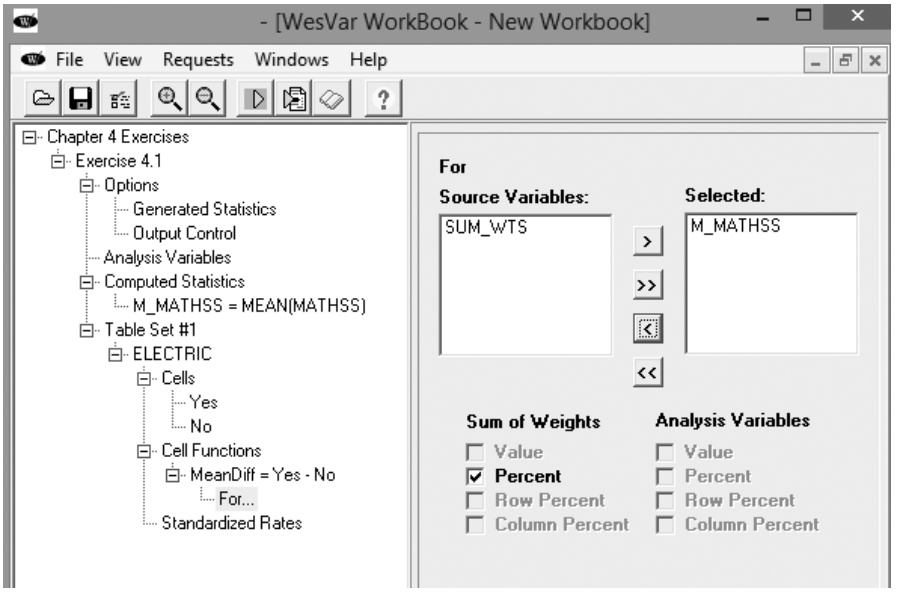
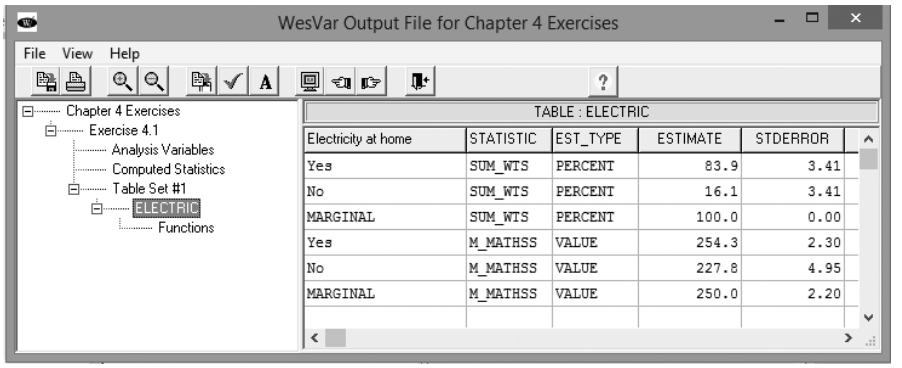


FIGURA DEL EJERCICIO 4.1.B **Resultado WesVar: Puntajes promedio en matemáticas de estudiantes con y sin electricidad en el hogar**



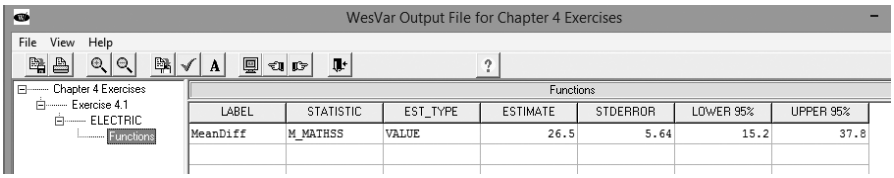
14. Para guardar el resultado, seleccione **File – Export – Single File – One Selection** y haga clic en **Export**. Guarde en ***NAEA DATA ANALYSIS\MY SOLUTIONS*** con un nombre de archivo adecuado (como ***EXERCISE 4.1A.TXT***).

(continúa)

EJERCICIO 4.1 *(continúa)*

15. Para ver el cálculo de la diferencia en los puntajes de matemáticas entre los estudiantes de hogares con y sin electricidad (26,5), haga clic en **Functions (**debajo de **Electric)** (figura del ejercicio 4.1.C).

FIGURA DEL EJERCICIO 4.1.C Resultado WesVar: Diferencia de puntaje promedio en matemáticas de estudiantes con y sin electricidad en el hogar



En la figura del ejercicio 4.1.B se pueden observar los puntajes promedio de matemáticas de los estudiantes con electricidad en el hogar (254,3) y de los que no tienen electricidad (227,8). Los errores estándar correspondientes son 2,30 y 4,95. La figura del ejercicio 4.1.C muestra la diferencia de puntaje promedio (26,5). Esta es la diferencia en los puntajes promedio de matemáticas entre aquellos con y sin electricidad en el hogar (Yes − No). El error estándar de la diferencia es 5,64. El intervalo de confianza del 95 % (alrededor de la diferencia de puntaje promedio) se extiende de 15,2 (límite inferior) a 37,8 (límite superior). El intervalo de confianza puede ayudar a determinar rápidamente si existe una diferencia significativa entre dos promedios. Si el intervalo de confianza incluye el valor cero (como por ejemplo, de −4,5 a +7,9), se puede afirmar que la diferencia promedio no es significativamente diferente de cero en el nivel 0,05. En el caso de los datos actuales, dado que el intervalo de confianza del 95 % (15,2 a 37,8) no incluye el cero, se puede concluir que la diferencia promedio de 26,5 es considerablemente diferente de cero (p <0,05). La diferencia en el rendimiento promedio en matemáticas entre estudiantes con y sin electricidad en el hogar es estadísticamente significativa.

La información obtenida en este análisis se puede resumir en una tabla (tabla del ejercicio 4.1.A). Una tabla como esta se puede incluir en un informe de una evaluación nacional.

TABLA DEL EJERCICIO 4.1.A Comparación de puntajes promedio en matemáticas de estudiantes con y sin electricidad en el hogar

Estado	Porcentaje de estudiantes (EE)	Puntaje promedio (EE)	
Electricidad en el hogar	83,9 (3,41)	254,3 (2,30)	
Sin electricidad en el hogar	16,1 (3,41)	227,8 (4,95)	
Comparación		Diferencia (EED)	IC (95 %)
Electricidad - Sin electricidad en el hogar		26,5 (5,64)	**15,2 a 37,8**

Nota: IC (95 %)= intervalo de confianza del 95 %; EE= error estándar de la estimación; EED= error estándar de la diferencia. Los intervalos de confianza asociados a las diferencias estadísticamente significativas están en negrita.

EJERCICIO 4.1 *(continúa)*

16. Vuelva a **Functions** y guarde la comparación de los puntajes promedio como ***EXERCISE 4.1B.TXT*** en ***NAEA DATA ANALYSIS\MY SOLUTIONS***.
17. Vuelva al libro de trabajo **Chapter 4 Exercises** haciendo clic en el ícono **puerta abierta**. Haga clic en **Save** en la barra de herramientas para guardar los cambios. Seleccione **File – Close** en la barra del menú para cerrar el libro de trabajo o continúe con el ejercicio 4.2.

EXAMINAR LAS DIFERENCIAS ENTRE TRES O MÁS PUNTAJES PROMEDIO

En esta sección examinamos las diferencias entre tres o más categorías de estudiantes (por ejemplo, estudiantes de escuelas en distintas regiones de un país). Comparamos el rendimiento de los estudiantes en una región de Sentz (como el Área Metropolitana) con el rendimiento de los estudiantes en cada una de las otras tres regiones del país. En el análisis, el Área Metropolitana se denomina *grupo de referencia* y el rendimiento en cada una de las demás regiones se compara con este.

Al realizar múltiples comparaciones (por ejemplo, comparar el puntaje promedio de los estudiantes de una región con los puntajes promedio de los de las otras tres regiones), se debe ajustar el nivel alfa o de significación. Si no se ajusta, se corre el riesgo de reseñar que una diferencia es estadísticamente significativa cuando no lo es. El valor estándar de alfa (0,05) que se utiliza cuando se compara un par de puntajes promedio (con un intervalo de confianza del 95 % alrededor de la diferencia de puntaje promedio) se debe ajustar hacia abajo (es decir, dividir por la cantidad de comparaciones a realizar) si se hace más de una comparación. Por ejemplo, si se realizan tres comparaciones, el nivel del valor alfa de 0,05 debe dividirse por 3, lo que resulta en un valor ajustado de 0,0167 (0,05/3).

En el ejemplo del ejercicio 4.2, el puntaje promedio de los estudiantes del Área Metropolitana (el grupo de referencia) se compara

con el puntaje promedio de los estudiantes de cada una de las otras regiones. En consecuencia, se realizan tres comparaciones:

- Área Metropolitana – Noroeste
- Área Metropolitana – Región Montañosa del Este
- Área Metropolitana – Costa del Sudoeste

EJERCICIO 4.2

Examinar las diferencias entre tres o más puntajes promedio

1. Ejecute WesVar. Seleccione **Open WesVar Workbook**. Abra el libro de trabajo WesVar que guardó al finalizar el ejercicio 4.1. Este es ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 4 WORKBOOK***.
2. Resalte el nodo **Exercise 4.1**. Haga clic derecho y seleccione **Clone**. De esta manera, se hace una copia de Exercise 4.1 (Tabla de solicitud de diferencias entre dos promedios). Resalte **Table Request Two** en el panel derecho y etiquételo como ***Exercise 4.2***.
3. Expanda **Exercise 4.2** (panel izquierdo). Seleccione **Options** (panel izquierdo) y en el panel derecho, cambie el nivel **Alpha** a 0,0167 (porque se realizarán tres comparaciones) (figura del ejercicio 4.2.A).
4. Expanda el menú **Options** en el panel izquierdo. En **Generated Statistics,** asegúrese de que **Estimate**, **Standard Error** y **Confidence Interval (Standard)** estén seleccionados. Dentro de **Options – Output Control**, asegúrese de que **Variable Label** y **Value Label** estén seleccionados. Deseleccione las otras casillas. Expanda **Table Set #1** en el panel izquierdo y resalte **Electric**, deselecciónelo en el panel derecho y seleccione **Region**. Haga clic en **Replace Current Entry**. Si recibe el mensaje: *Table structure has changed… Do you want to make this change? (Se ha modificado la estructura de la tabla... ¿Desea realizar este cambio?)*, seleccione **Yes**. Asegúrese de que **Percent** esté seleccionado. Expanda **Region** (panel izquierdo). Seleccione **Cells** y defina las celdas de la siguiente manera: 1= ***Northwest***; 2= ***Metro_Area***; 3= ***Eastern_Highlands***; 4= ***Southwest_Coast***. (Dado que WesVar no permite dejar espacios en blanco entre palabras, debe utilizar un guión bajo). Luego de ingresar cada etiqueta, haga clic en **Add as New Entry** o presione **Return** en su teclado (figura del ejercicio 4.2.B).
5. Seleccione **Cell Functions** (inmediatamente debajo de **Cells** en el panel izquierdo). Ingrese lo siguiente en el cuadro **Function Statistic** y haga clic en **Add as New Entry** después de cada una.
   - **MeanDiffMetro_NW = Metro_Area – Northwest** (figura del ejercicio 4.2.C).

     Haga clic en **For** debajo de cada función (panel izquierdo) y asegúrese de que aparezca **Mathss** debajo de **Selected** cada vez. Puede ser necesario que tenga que mover **Sum_Wts** a **Source Variables**. Repita el proceso para

EJERCICIO 4.2 *(continúa)*

FIGURA DEL EJERCICIO 4.2.A Libro de trabajo de WesVar que muestra el ajuste del nivel alfa

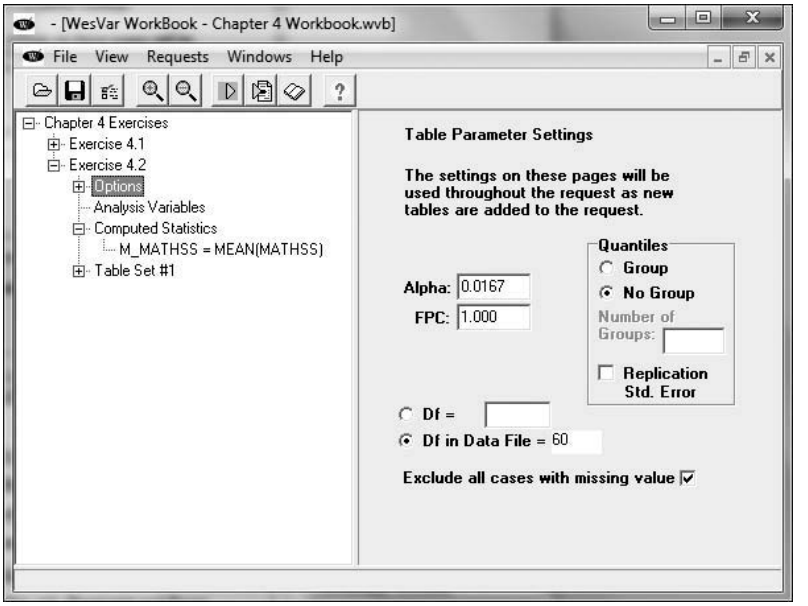
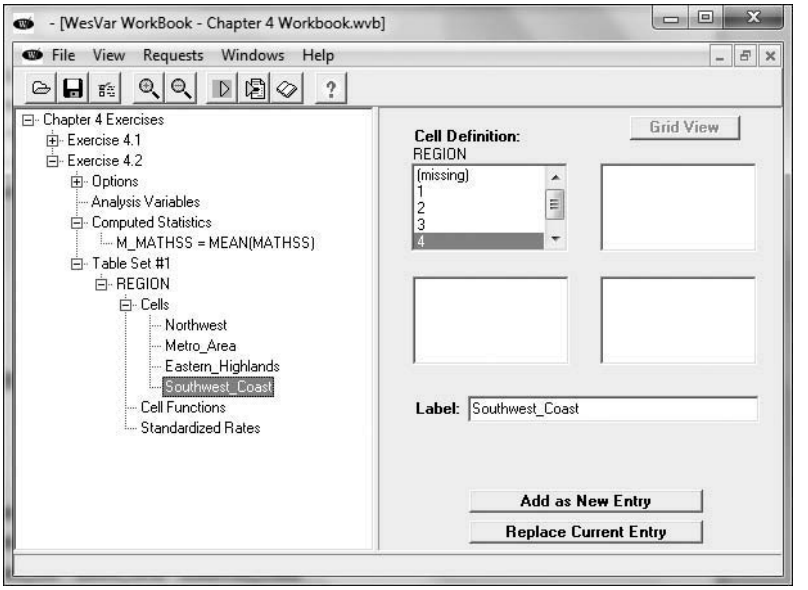


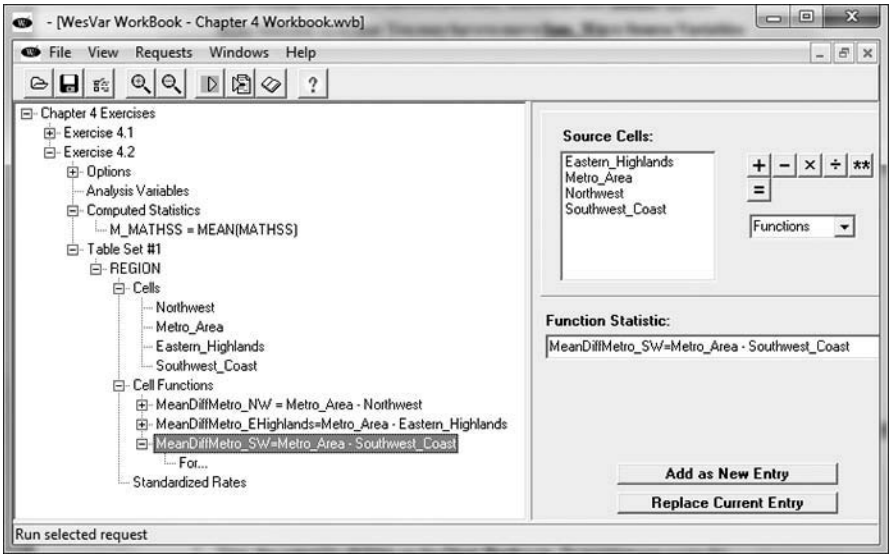
FIGURA DEL EJERCICIO 4.2.B Completar la definición de celdas en WesVar



(continúa)

EJERCICIO 4.2 *(continúa)*

FIGURA DEL EJERCICIO 4.2.C Funciones de visualización de celdas del libro de trabajo de WesVar



- **MeanDiffMetro_EHighlands = Metro_Area – Eastern_Highlands**
- **MeanDiffMetro_SW = Metro_Area – Southwest_Coast**

6. Haga clic en la **flecha verde** en la barra de herramientas para ejecutar el análisis.
7. Para ver el resultado, haga clic en el ícono **libro abierto** en la barra de herramientas. Para acceder a los puntajes promedio de cada región, seleccione **Exercise 4.2** – **Table Set #1** y **Region** (figura del ejercicio 4.2.D). Para acceder a los datos de las diferencias entre los puntajes promedio, seleccione **Functions** (el nodo inmediatamente debajo de **Region**) (figura del ejercicio 4.2.E).
8. Use **File** y **Export** (**Single File**, **One Selection**) para guardar su resultado como archivos de texto. Primero guarde los puntajes promedio (Regions) y luego las diferencias de puntajes promedio (Functions) en ***NAEA DATA ANALYSIS\MY SOLUTIONS,*** con los nombres de archivo ***EXERCISE 4.2A.TXT*** y ***EXERCISE 4.2B.TXT***, respectivamente.
9. Vuelva a ***CHAPTER 4 WORKBOOK*** haciendo clic en el ícono **puerta abierta** en la barra de herramientas. Seleccione **Save – Close** en la barra del menú (o seleccione el ícono **Save** en la barra de herramientas).

La figura del ejercicio 4.2.E proporciona las tres comparaciones solicitadas. La diferencia en los puntajes promedio entre el Área Metropolitana y el Noroeste es de 32,4 puntos en la escala de puntaje y el error estándar es de 5,74. El intervalo de confianza alrededor de la diferencia es de 18,2 a 46,5.

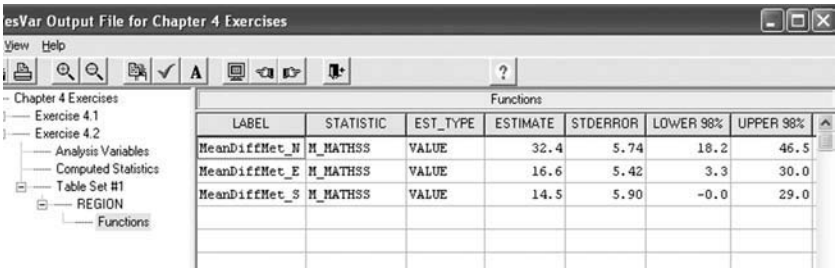
EJERCICIO 4.2 *(continúa)*

FIGURA DEL EJERCICIO 4.2.D Resultado WesVar: Puntajes promedio de matemáticas por región

TABLE : REGION

Region	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	LOWER 98%	UPPER 98%
Northwest	SUM_WTS	PERCENT	25.2	4.30	14.6	35.8
Metro_Area	SUM_WTS	PERCENT	26.1	4.62	14.8	37.5
Eastern_Highland:	SUM_WTS	PERCENT	24.3	4.05	14.3	34.3
Southwest_Coast	SUM_WTS	PERCENT	24.4	4.31	13.8	35.0
MARGINAL	SUM_WTS	PERCENT	100.0	.	.	.
Northwest	M_MATHSS	VALUE	233.3	3.28	225.3	241.4
Metro_Area	M_MATHSS	VALUE	265.7	4.46	254.8	276.7
Eastern_Highland:	M_MATHSS	VALUE	249.1	3.59	240.3	257.9
Southwest_Coast	M_MATHSS	VALUE	251.2	3.35	243.0	259.5
MARGINAL	M_MATHSS	VALUE	250.0	2.20	244.6	255.4

FIGURA DEL EJERCICIO 4.2.E Resultado WesVar: Diferencias entre puntajes promedio de matemáticas por región



Dado que el intervalo no incluye el cero, la diferencia entre el puntaje promedio del Área Metropolitana y el puntaje promedio de la región Noroeste resulta estadísticamente significativa. De manera similar, la diferencia entre los puntajes promedio del Área Metropolitana y de las Región Montañosa del Este (16,6 puntos) es estadísticamente significativa porque el intervalo alrededor de la diferencia (3,3 a 30,0) no incluye cero. Finalmente, la diferencia entre el Área Metropolitana y la Costa del Sudoeste (14,5 puntos) no resulta estadísticamente significativa ya que el intervalo en torno a la diferencia (-0,0 a 29,0) incluye cero.

La tabla del ejercicio 4.2.A presenta el resultado de este análisis como se presentaría en el informe de una evaluación nacional. La tabla muestra cada puntaje promedio regional, el puntaje promedio nacional y los errores estándar asociados. La mitad inferior de la tabla indica, en negrita, las diferencias regionales estadísticamente significativas.

(continúa)

EJERCICIO 4.2 *(continúa)*

TABLA DEL EJERCICIO 4.2.A Comparación de puntajes promedio en matemáticas de estudiantes con y sin electricidad en el hogar por región

Región	Puntaje promedio (EE)	
Área Metropolitana	233,3 (3,28)	
Noroeste	265,7 (4,46)	
Región Montañosa del Este	249,1 (3,59)	
Costa del Sudoeste	251,2 (3,35)	
Nacional	250,0 (2,20)	
Comparación	Diferencia (EED)	ICA (95 %)
Área Metropolitana – Noroeste	32,4 (5,74)	**18,2 a 46,5**
Área Metropolitana – Región Montañosa del Este	16,6 (5,42)	**3,3 a 30,3**
Área Metropolitana – Costa del Sudoeste	14,5 (5,90)	−0,0 a 29,0

Nota: ICA (95 %)= intervalo de confianza del 95 % ajustado; EE= error estándar de la estimación; EED= error estándar de la diferencia Los intervalos de confianza asociados a las diferencias estadísticamente significativas están en negrita.

Otras comparaciones que pueden interesar a los responsables normativos (según las variables en la base de datos) son las siguientes:

- *Grupo étnico* y *rendimiento en matemáticas:* ¿Existen diferencias significativas entre los grupos étnicos y sus rendimientos promedio en matemáticas? ¿Qué grupo tiene el puntaje promedio más alto?
- *Nivel de educación parental* y *rendimiento en lectura:* ¿Existen diferencias significativas entre los estudiantes con padres que han completado estudios universitarios y los estudiantes en cada uno de los otros grupos (padres sin educación formal, padres cuyo nivel más alto de educación se encuentra entre los grados 1 a 3, 4 a 6, 7 a 9, o 10 a 12)?
- *Acceso a los medios de comunicación* y *rendimiento en lenguaje:* ¿Existen diferencias significativas en el rendimiento promedio en lenguaje entre los estudiantes que viven en hogares con (a) radio y televisión, en comparación con aquellos que viven en hogares con (b) solo una radio, (c) solo un televisor o (d) sin radio ni televisión?

CAPÍTULO 5

IDENTIFICACIÓN DE LOS ALUMNOS CON ALTO Y BAJO RENDIMIENTO

Además de considerar las diferencias entre los niveles medios de rendimiento en las subpoblaciones de alumnos de una evaluación nacional (capítulo 4), los responsables de las políticas y quienes utilizan los resultados de la evaluación podrían también estar interesados en identificar los factores asociados a la distribución de dicho rendimiento. Pueden, entonces, surgir preguntas como las siguientes:

- ¿La proporción de alumnos en riesgo (con bajo rendimiento) es mayor en una región del país (provincia) que en otra?
- ¿Los niños y las niñas están igualmente representados entre los alumnos con alto rendimiento en matemáticas?
- ¿En qué tipo de escuela (rural, urbana, pública o privada) se encuentra la mayor proporción de alumnos con bajo rendimiento?

En el presente capítulo, responderemos a este tipo de preguntas identificando las proporciones de alumnos que se encuentran por encima o por debajo de los valores clave, o valores de referencia, tales como los percentiles 10 o 90.

ESTIMACIÓN DE LOS PUNTAJES CORRESPONDIENTES A LOS RANGOS DE PERCENTILES NACIONALES

La presente sección describe una manera de estimar las puntuaciones de los alumnos correspondientes a rangos de percentiles clave en una evaluación nacional. En el capítulo 3, se utilizó el método **Descriptive Stats** en **WesVar** para hallar los puntajes ponderados y no ponderados correspondientes a diferentes rangos de percentiles y sus errores estándar (véase la figura del ejercicio 3.1. C). En el ejercicio 5.1, se utiliza un método alternativo (**WesVar Tables**) para estimar los puntajes correspondientes a los percentiles 10, 25, 50, 75 y 90, y sus errores estándar para cada región de Sentz. Este método también

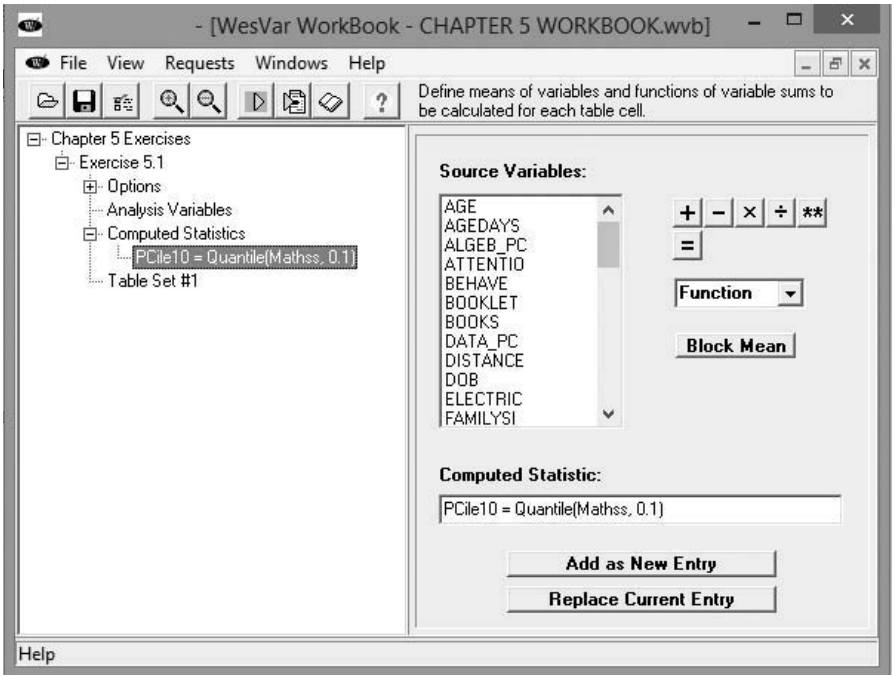
EJERCICIO 5.1

Cálculo de los puntajes correspondientes a los percentiles nacionales

1. Inicie WesVar y haga clic sobre **New WesVar Workbook.** Es posible que reciba la siguiente advertencia: *Antes de crear un nuevo libro de trabajo, se le solicitará que especifique un archivo de datos que será utilizado como el archivo de datos por defecto para el nuevo libro de trabajo.* Haga clic en **OK.**
2. Aparecerá una ventana con el título **Open WesVar Data File for Workbook.** Seleccione el archivo de datos ***NAEA DATA ANALYSIS\MY WESVAR FILES\ NATASSESS4.VAR***.
3. Guarde su nuevo libro en ***NAEA DATA ANALYSIS\MY WESVAR FILES*** con el nombre ***CHAPTER 5 WORKBOOK.WVB***.
4. Cambie **Workbook Title One** en el panel derecho por ***Chapter 5 Exercises***. En **New Request,** resalte **Table.** Haga clic sobre **Table Request One** (en el panel izquierdo). Cambie **Request Name** por ***Exercise 5.1*** (en el panel derecho). Haga clic sobre **Add Table Set Single** en el panel derecho.
5. Seleccione **Options – Generated Statistics** y asegúrese de que **Estimate**, **Standard Error** y **Confidence Interval (Standard)** estén marcados. Deseleccione los otros casilleros. En **Output Control** configure **Estimate** con un decimal y **Std. Error** con dos decimales. Asegúrese de que **Variable Label** y **Value Label** estén marcados.
6. Seleccione **Computed Statistics** en el panel izquierdo. En el panel derecho, escriba ***Pcile10 = Quantile(Mathss, 0.1)*** y haga clic sobre **Add as New Entry** (figura del ejercicio 5.1.A.). Esto le ordena a WesVar que calcule el puntaje correspondiente al percentil 10.

EJERCICIO 5.1 *(continúa)*

FIGURA DEL EJERCICIO 5.1.A Libro de trabajo WesVar: cálculo de los puntajes correspondientes a los percentiles



7. Realice el mismo procedimiento para los percentiles 25, 50, 75 y 90, haciendo clic sobre **Computed Statistics** primero, cada vez que lo haga. Por ejemplo, la fórmula para el percentil 25 es **PCile25 = Quantile(Mathss, 0.25)**. Recuerde hacer clic sobre **Add as New Entry** luego de ingresar cada fórmula.[a]
8. Comience el análisis haciendo clic sobre la **flecha verde** en la barra de herramientas.
9. Observe los resultados (figura del ejercicio 5.1. B) haciendo clic sobre el ícono del **libro abierto** en la barra de herramientas. Expanda **Exercise 5.1 – Table Set #1** en el panel de la izquierda. Haga clic sobre **Overall** para ver los puntajes correspondientes a los percentiles.

El resultado en la figura del ejercicio 5.1. B muestra los puntajes correspondientes a cada rango de percentil seleccionado, junto con los errores estándar. Por ejemplo, el puntaje en el percentil 10 es 181,0 y el error estándar es 3,01. El intervalo de confianza de 95 por ciento para 181,0 es de 174,5 a 186,5. Por lo tanto, hay un 95 % de probabilidad de que el puntaje en el percentil 10 sea entre 174,5 y 186,5. Los puntajes y los errores estándar pueden ser tabulados y publicados en el reporte de la evaluación nacional (véase la tabla del ejercicio 5.1. A).

(continúa)

EJERCICIO 5.1 *(continúa)*

FIGURA DEL EJERCICIO 5.1.B Resultado WesVar: Cálculo de los puntajes correspondientes a los percentiles

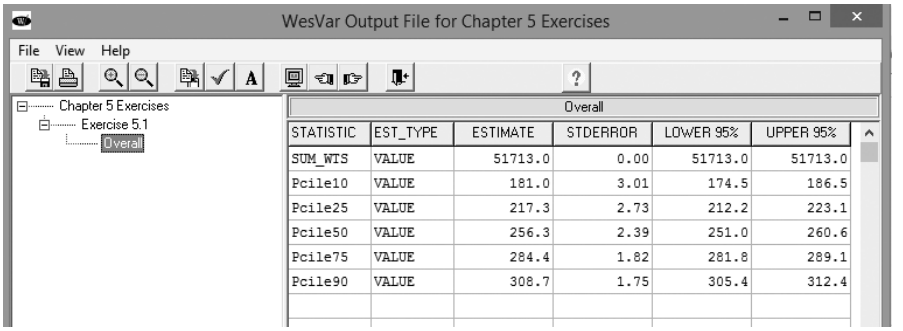


TABLA DEL EJERCICIO 5.1.A Puntajes de matemáticas a nivel nacional (y errores estándar) en diferentes rangos de percentiles

Percentil	Puntaje	Error estándar
10	181,0	3,01
25	217,3	2,73
50	256,3	2,39
75	284,4	1,82
90	308,7	1,75

10. Seleccione **File** y haga clic sobre **Export** (**Single File – One Selection**) para guardar su resultado como un archivo de texto. Guárdelo en ***NAEA DATA ANALYSIS\MY SOLUTIONS*** utilizando un nombre de archivo apropiado, como por ejemplo ***EXERCISE 5.1.TXT.***

11. Haga clic sobre el ícono de la **puerta abierta** en la barra de herramientas para volver a ***CHAPTER 5 WORKBOOK***. Haga clic sobre **File – Save** y luego **File – Close.**

a. *De manera alternativa, puede hacer clic sobre* **Pcile10 = Quantile(Mathss, 0.10)** *en el panel izquierdo y hacer clic derecho sobre* **Clone** *cuatro veces, modificando cada uno según sea necesario y utilizando la función* **Replace Current Entry.**

puede utilizarse para identificar los puntajes correspondientes a otros rangos de percentiles.

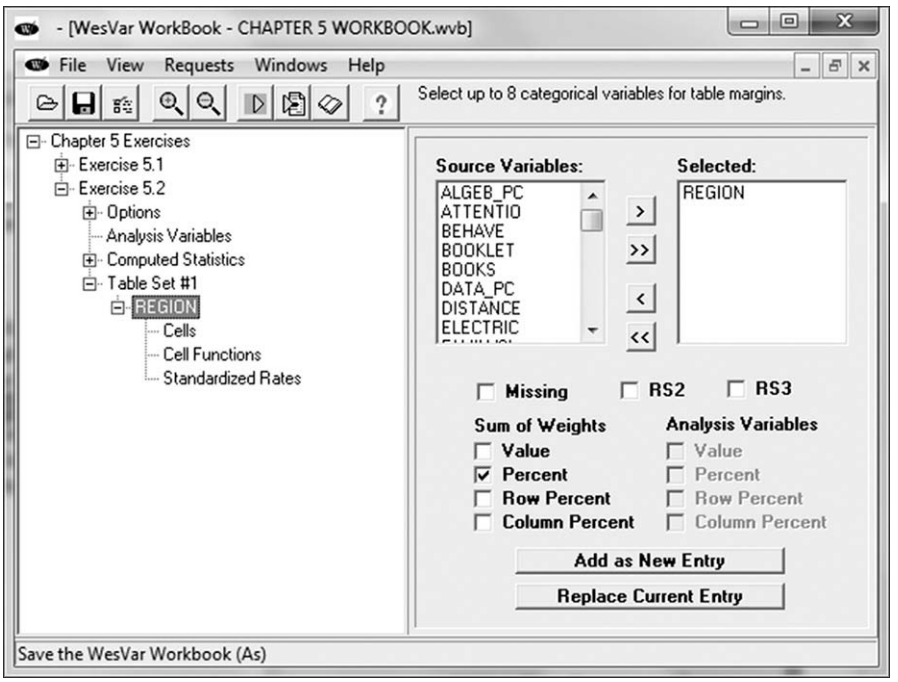
Una vez obtenidos los puntajes correspondientes al rango de percentiles a nivel nacional, pueden calcularse los puntajes correspondientes a cada región de Sentz (ejercicio 5.2).

EJERCICIO 5.2

Cálculo de los puntajes correspondientes a los percentiles según la región

1. Inicie WesVar. Seleccione **Open WesVar Workbook**. Abra el libro de trabajo WesVar que guardó mientras completaba el ejercicio 5.1: ***NAEA DATA ANALYSIS\My WesVar FILES\Chapter 5 Workbook***.[a]
2. Si es necesario, minimice **Exercise 5.1** (panel izquierdo), haciendo clic sobre el signo – (menos).
3. Haga clic derecho sobre **Exercise 5.1** en el panel izquierdo, y seleccione **Clone**. Cambie el nombre de **Table Request Two** en el panel de la derecha por ***Exercise 5.2***. Todos los pasos subsiguientes del ejercicio 5.2 se completarán en este nodo.
4. Si es necesario, expanda **Exercise 5.2**. Seleccione **Table Set #1**. Mueva la opción **Region** de **Source Variables** a **Selected** en el panel derecho.
5. Haga clic sobre **Add as New Entry** en el panel derecho.
6. Seleccione **Region** en el panel izquierdo, y marque el casillero **Percent** debajo de **Sum of Weights** en el panel derecho. Desmarque **Value**, **Row Percent**, y **Column Percent** si es necesario (véase figura del ejercicio 5.2. A).

FIGURA DEL EJERCICIO 5.2.A Libro de trabajo WesVar antes de calcular los puntajes correspondientes al percentil según la región



(continúa)

EJERCICIO 5.2 *(continúa)*

7. Haga clic sobre la **flecha verde** en la barra de herramientas para comenzar el análisis.
8. Haga clic sobre el ícono del **libro abierto** en la barra de herramientas. Si fuera necesario, expanda **Exercise 5.2**. Seleccione **Table Set #1** – **REGION** para ver el percentil estimado para cada región.

La figura del ejercicio 5.2. B muestra los puntajes en el percentil 10 de cada región, junto con los errores estándar asociados y los intervalos de confianza de 95 %. Los datos de los puntajes correspondientes a los otros percentiles pueden verse deslizando hacia abajo el archivo de resultados. La tabla del ejercicio 5.2. A presenta los datos en un formato que podría utilizarse para el reporte de una evaluación nacional.

FIGURA DEL EJERCICIO 5.2.B Resultado parcial de WesVar: Cálculo de los puntajes correspondientes al percentil 10 según la región

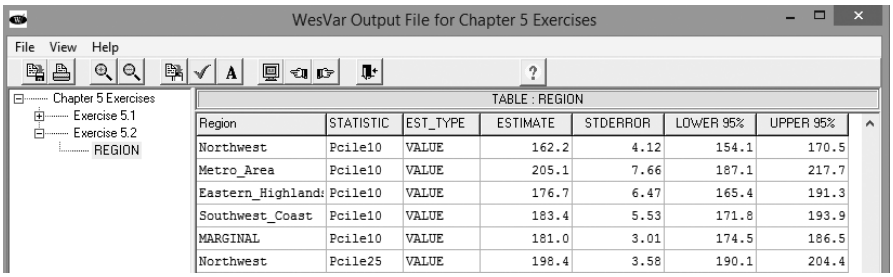


TABLA DEL EJERCICIO 5.2.A Puntajes en matemáticas (y errores estándar) en diferentes niveles de percentil según la región

Percentil	Noroeste	Región metropolitana	Región Montañosa del Este	Costa sudoeste	Nacional
10	162,2 *(4,12)*	205,1 *(7,66)*	176,7 *(6,47)*	183,5 *(5,53)*	181,0 *(3,01)*
25	198,4 *(3,58)*	238,7 *(5,95)*	217,1 *(5,68)*	218,6 *(5,08)*	217,3 *(2,73)*
50	238,0 *(4,62)*	271,6 *(5,16)*	255,2 *(3,41)*	257,6 *(4,09)*	256,3 *(2,39)*
75	270,2 *(4,11)*	296,3 *(3,27)*	283,9 *(3,49)*	283,9 *(3,58)*	284,4 *(1,82)*
90	293,8 *(4,42)*	313,7 *(3,02)*	309,1 *(2,87)*	310,3 *(3,33)*	308,7 *(1,75)*

La tabla del ejercicio 5.2. A muestra que los puntajes correspondientes a los rangos de percentiles cambian según la región. Por ejemplo, los puntajes correspondientes al percentil 10 varían desde 162,2 en el Noroeste hasta 205,1 en el Área Metropolitana. El examen de los puntajes regionales sugiere que los alumnos con bajo rendimiento (en el percentil 10) en el Noroeste, la Región Montañosa del Este y la Costa del Sudoeste tienen un desempeño peor que los alumnos con bajo rendimiento en el Área Metropolitana. Los puntajes de los alumnos correspondientes al percentil 90 del Noroeste son menores que los de los alumnos de otras regiones.

EJERCICIO 5.2 *(continúa)*

9. Guarde el resultado utilizando **File – Export – Single File – One Selection**. Seleccione **Export**. Guárdelo como un archivo de texto en ***NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 5.2***. Salga del archivo de resultado haciendo clic en el ícono de la **puerta abierta** en la barra de herramientas.
10. Guarde los cambios del libro seleccionando **File – Save** y luego **File – Close.**

a. Si no encuentra el archivo, seleccione ***NAEA DATA ANALYSIS\WESVAR DATA FILES & WORKBOOKS\CHAPTER 5 WORKBOOK***. *Luego de completar el ejercicio, guarde el libro en* ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 5 WORKBOOK***.

ESTIMACIÓN DE LOS PORCENTAJES DE ALUMNOS EN LOS SUBGRUPOS MEDIANTE EL USO DE LOS RANGOS DE PERCENTILES NACIONALES

Es posible obtener información sobre el porcentaje de alumnos en los subgrupos de la población (región, raza y sexo) con puntajes inferiores al valor de referencia designado, tales como los que se encuentran en el percentil nacional 10. Para ello es necesario identificar primero los puntajes correspondientes a los rangos de percentiles nacionales en la base de datos de la evaluación nacional en WesVar (véase la tabla del ejercicio 5.1. A). El ejercicio 5.3 muestra cómo crear nuevas variables correspondientes a los percentiles 25, 50 y 75 en el archivo de datos de WesVar. Cree una nueva variable correspondiente al valor de referencia de cada percentil y designe con un "1" a los alumnos que obtuvieron puntajes inferiores al valor de referencia y con un "2" a los alumnos que obtuvieron un puntaje igual o superior a ese valor.

Luego se calculan los porcentajes de los alumnos que obtuvieron puntajes inferiores a los valores de referencia de los percentiles nacionales de cada región (ejercicio 5.4).

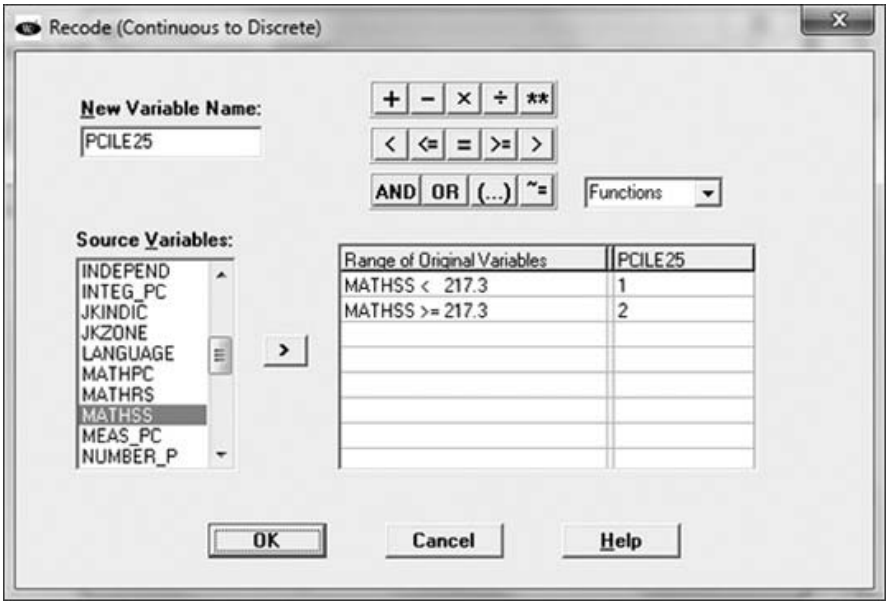
El resultado puede ser utilizado para determinar el porcentaje de alumnos en cada región cuyo puntaje es inferior al percentil nacional 25 (figura del ejercicio 5.4.B). Se estima que en la región Noroeste el 36,5 por ciento de los alumnos tienen puntajes inferiores al valor de referencia de este percentil. El error estándar de la estimación es 2,74, y el intervalo de confianza del 95 por ciento oscila entre el 31,0 % y

EJERCICIO 5.3

Registro de una variable en las categorías de percentiles mediante el uso de WesVar

1. Inicie WesVar. Seleccione **Open WesVar Data File**, y seleccione ***NAEA DATA ANALYSIS\ MY WESVAR FILES\NATASSESS4.VAR***.
2. Haga clic sobre **Format** en la barra de menú, y seleccione **Recode**. En **Pending Records**, haga clic en **New Continuous (to Discrete)**.
3. Escriba un nuevo nombre en el casillero **New Variable Name**. Para este ejercicio, asigne el nombre de la variable del grupo ***Pcile25*** que dividirá a los alumnos en dos categorías: aquellos con puntajes inferiores al percentil nacional 25 y aquellos con puntajes iguales o superiores.
4. En la columna denominada **Range of Original Variables**, escriba ***Mathss < 217.3*** (el valor del percentil nacional 25; véase la Tabla 5.2). En la misma fila, en la columna denominada **Pcile25**, escriba el número ***1***. En la siguiente fila, debajo de **Range of Original Variables**, escriba ***Mathss*** ≥ ***217.3***, y en la columna **Pcile25**, escriba el número ***2*** (figura del ejercicio 5.3. A). Haga clic en **OK**.

FIGURA DEL EJERCICIO 5.3.A Libro de trabajo WesVar: Registro de Mathss con una variable discreta



5. Haga clic sobre **New Continuous to Discrete** nuevamente, y repita el mismo procedimiento para crear las siguientes variables:

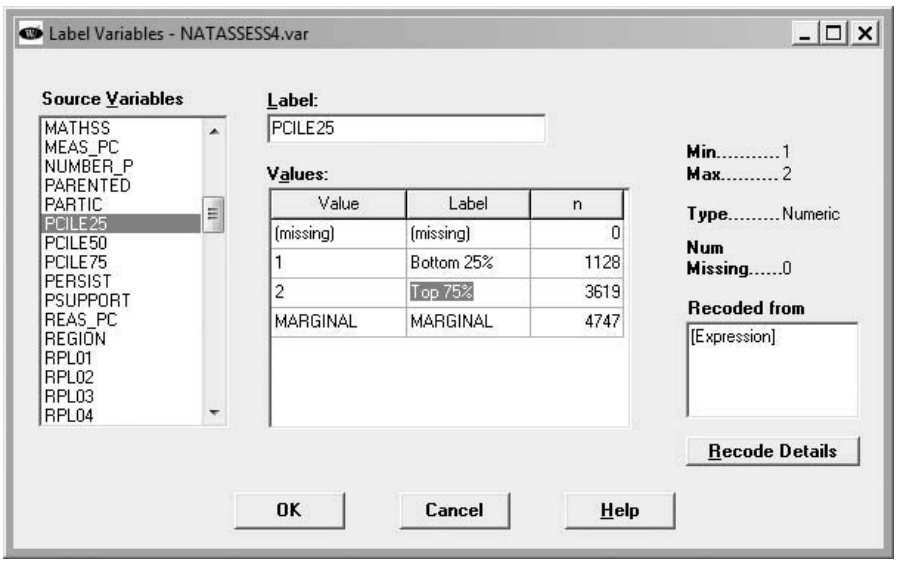
Pcile50: Mathss < 256,3 → 1 **Mathss** ≥ 256,3 → 2 **(OK)**

Pcile75: Mathss < 284,4 → 1 **Mathss** ≥ 284,4 → 2 **(OK)**

EJERCICIO 5.3 *(continúa)*

6. Luego de ingresar la última variable (**Pcile75**), haga clic en **OK**. Luego haga clic en **OK** en la pantalla **Pending Recodes**. Verá un mensaje que dice *Esta operación creará un nuevo archivo VAR. Deberá asignarle un nombre al archivo*. Guárdelo como ***NAEA DATA ANAYSIS\MY WESVAR FILES\NATASSESS4.VAR***, sobrescribiendo la versión guardada anteriormente.
7. Seleccione **Format** y **Label** en la barra de menú. En **Source Variables**, haga clic sobre **Pcile25**. En **Values**, en la columna denominada **Label**, reemplace el número **1** con ***Bottom 25 %*** y reemplace el **2** con ***Top 75 %*** (figura del ejercicio 5.3. B). Luego haga un clic en **OK**. Verá un mensaje que dice *Esta operación creará un nuevo archivo VAR. Deberá asignarle un nombre al archivo*. Haga clic en **OK**. Guárdelo como ***NAEA DATA ANALYSIS\MY WESVAR FILES\ NATASSESS4.VAR*** para sobrescribir la versión guardada anteriormente.

FIGURA DEL EJERCICIO 5.3.B Denominación de las categorías de percentiles en WesVar



8. Repita el proceso para **Pcile50** y **Pcile75.**
9. Seleccione **File** y **Exit.**[a]

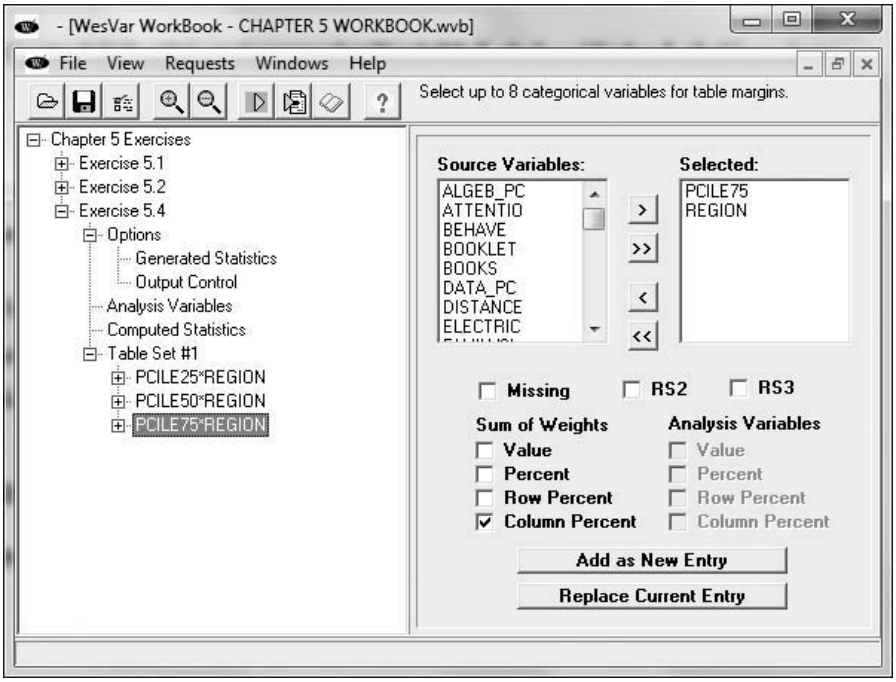
a. Observe que este ejercicio no tiene un archivo de resultado para guardar. En su lugar, usted guardó el archivo de datos WesVar modificado (***NATASSESS4.VAR***), que utilizará en el próximo ejercicio.

EJERCICIO 5.4

Cálculo de los porcentajes de alumnos con puntajes inferiores a los valores de referencia de los percentiles nacionales y errores estándar en cada región

1. Inicie WesVar. Seleccione **Open WesVar Workbook**. Abra ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 5 WORKBOOK***.
2. Seleccione **Chapter 5 Exercises** en el panel izquierdo. Seleccione **New Request – Table** en el panel derecho. Haga clic en **Table Request Three** en el panel izquierdo, y reescriba el nombre ***Exercise 5.4***. Seleccione **Exercise 5.4** en el panel izquierdo y luego **Add Table Set – Single** en el panel derecho.
3. Seleccione **Table Set #1** en el panel izquierdo, haga clic sobre **Pcile25** en el casillero **Source Variables** en el panel derecho, y muévalo al casillero **Selected.** Luego mueva **Region** de **Source Variables** a **Selected**. Asegúrese de que el casillero **Column Percent** en el panel derecho, bajo **Sum of Weights**, esté marcado, y de que las otras opciones (**Value, Percent** y **Row Percent**) estén desmarcadas. Haga clic en **Add as New Entry** en el panel derecho. Vuelva a mover **Pcile25** y **Region** a **Source Variables**.
4. Repita el paso 3 para **Pcile50** y **Region** y para **Pcile75** y **Region**, asegurándose de escribir **Pcile** antes que **Region**. Haga clic en **Add as New Entry** luego de cada modificación (figura del ejercicio 5.4. A).

FIGURA DEL EJERCICIO 5.4.A Captura de pantalla del libro de trabajo WesVar antes de calcular los porcentajes de puntajes inferiores a los valores de referencia nacionales en cada región



EJERCICIO 5.4 *(continúa)*

5. Seleccione **Output Control** en el panel izquierdo y asegúrese de que **Variable Label** y **Value Label** estén marcados.
6. Seleccione **Generated Statistics** y asegúrese de que solo **Estimate**, **Standard Error** y **Confidence Interval (Standard)** estén marcados.
7. Ejecute el análisis haciendo clic en la **flecha verde** en la barra de herramientas. Observe el resultado seleccionando el ícono del **libro abierto** en la barra de herramientas. Expanda el lado izquierdo abriendo **Exercise 5.4**, **Table Set #1** y el rango **Pcile25*Region** para ver el primer bloque de resultados (figura del ejercicio 5.4. B).

FIGURA DEL EJERCICIO 5.4.B Resultado parcial: Porcentajes de alumnos con puntajes inferiores a los valores de referencia nacionales en cada región

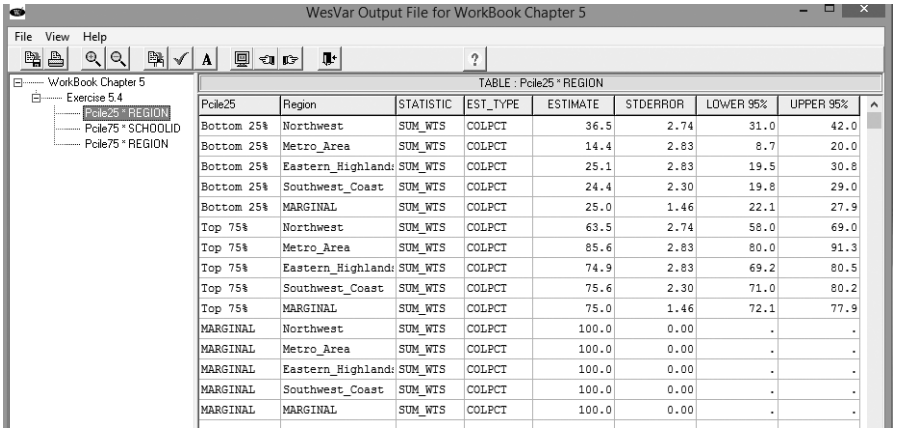


TABLE : Pcile25 * REGION

Pcile25	Region	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%
Bottom 25%	Northwest	SUM_WTS	COLPCT	36.5	2.74	31.0	42.0
Bottom 25%	Metro_Area	SUM_WTS	COLPCT	14.4	2.83	8.7	20.0
Bottom 25%	Eastern_Highlands	SUM_WTS	COLPCT	25.1	2.83	19.5	30.8
Bottom 25%	Southwest_Coast	SUM_WTS	COLPCT	24.4	2.30	19.8	29.0
Bottom 25%	MARGINAL	SUM_WTS	COLPCT	25.0	1.46	22.1	27.9
Top 75%	Northwest	SUM_WTS	COLPCT	63.5	2.74	58.0	69.0
Top 75%	Metro_Area	SUM_WTS	COLPCT	85.6	2.83	80.0	91.3
Top 75%	Eastern_Highlands	SUM_WTS	COLPCT	74.9	2.83	69.2	80.5
Top 75%	Southwest_Coast	SUM_WTS	COLPCT	75.6	2.30	71.0	80.2
Top 75%	MARGINAL	SUM_WTS	COLPCT	75.0	1.46	72.1	77.9
MARGINAL	Northwest	SUM_WTS	COLPCT	100.0	0.00	.	.
MARGINAL	Metro_Area	SUM_WTS	COLPCT	100.0	0.00	.	.
MARGINAL	Eastern_Highlands	SUM_WTS	COLPCT	100.0	0.00	.	.
MARGINAL	Southwest_Coast	SUM_WTS	COLPCT	100.0	0.00	.	.
MARGINAL	MARGINAL	SUM_WTS	COLPCT	100.0	0.00	.	.

8. Guarde el resultado seleccionando **File – Export – Single File – One Selection** en la barra de menú. Haga clic en **Export**. Guárdelo como un archivo de texto en ***NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 5.4 – 25th*** (para los resultados correspondientes al percentil 25), ***EXERCISE 5.4 – 50th*** (para los resultados correspondientes al percentil 50) y así sucesivamente. Salga del archivo de resultados a través del ícono de la **puerta abierta** en la barra de herramientas.
9. Guarde el libro de WesVar haciendo clic en el ícono **Save** en la barra de herramientas (o seleccionando **File** – **Save** en la barra de menú). Haga clic en **File – Close.**

el 42,0 %. La Tabla 5.1 presenta los datos en forma tabular. Se puede observar que el 36,5 por ciento de los alumnos del Noroeste, en comparación con el 14,4 por ciento del Área Metropolitana, obtuvieron puntajes inferiores al valor de referencia del percentil nacional 25. En las otras dos regiones, la Región Montañosa del Este y la Costa del

Sudoeste, los porcentajes de puntuaciones en este nivel son similares al porcentaje nacional (25 por ciento).

De manera similar, se pueden organizar en forma de tabla los porcentajes de alumnos con puntajes inferiores (o iguales o superiores) a otros valores de referencia nacional, tales como el percentil 50 **(Pcile50)** y el percentil 75 **(Pcile75).** La Tabla 5.2 muestra el porcentaje de alumnos con alto rendimiento en cada región (aquellos alumnos que obtuvieron puntajes iguales o superiores al percentil nacional 75). Los datos de esta tabla se basan en el resultado del ejercicio 5.4 y muestran que el 15,1 por ciento de los alumnos del Noroeste obtuvieron resultados iguales o superiores al valor de referencia del percentil nacional 75, en comparación con el 35,3 por ciento de alumnos del Área Metropolitana. En las dos regiones restantes, los porcentajes de alumnos que lograron un porcentaje igual o superior al valor de referencia del percentil 75 (24,7 por ciento de los casos) son similares al porcentaje nacional (25 por ciento).

TABLA 5.1

Porcentajes de alumnos con puntajes inferiores al valor de referencia del percentil nacional 25 según la región

	Alumnos con puntajes inferiores al valor de referencia del percentil nacional 25	
Región	Porcentaje	Error estándar
Noroeste	36,5	2,74
Región Metropolitana	14,4	2,83
Región Montañosa del Este	25,1	2,83
Costa del Sudoeste	24,4	2,30
Nacional	25,0	1,46

TABLA 5.2

Porcentajes de alumnos con puntajes superiores al valor de referencia del percentil nacional 75 según la región

	Alumnos con puntajes superiores al valor de referencia del percentil nacional 75	
Región	Porcentaje	Error estándar
Noroeste	15,1	2,24
Región Metropolitana	35,3	3,88
Región Montañosa del Este	24,7	2,53
Costa del Sudoeste	24,7	2,48
Nacional	25,0	1,61

CAPÍTULO 6

ASOCIACIÓN ENTRE VARIABLES: CORRELACIÓN Y REGRESIÓN

CORRELACIÓN

Los responsables de las políticas pueden estar interesados en identificar la medida en que el aprendizaje de un estudiante se relaciona con una serie de factores. El coeficiente de correlación (r), que es una medida de la *asociación lineal* entre dos variables, proporciona esta información. Los siguientes son ejemplos de preguntas a las que las correlaciones pueden responder:

- ¿Hay alguna relación entre la frecuencia de asistencia a la escuela y el rendimiento de los estudiantes en matemáticas?
- ¿Hay alguna relación entre el nivel educativo alcanzado por los padres y el rendimiento de los estudiantes en lectura?
- ¿Hay alguna relación entre la distancia que un estudiante recorre para asistir a la escuela y el rendimiento en matemáticas?
- ¿La experiencia de los docentes (número de años ejerciendo la docencia) se relaciona con el rendimiento de los estudiantes en áreas clave del currículo?

El coeficiente de correlación nos puede decir la dirección de la relación entre dos variables y la solidez o magnitud de la relación entre ellas.

Dirección de la relación

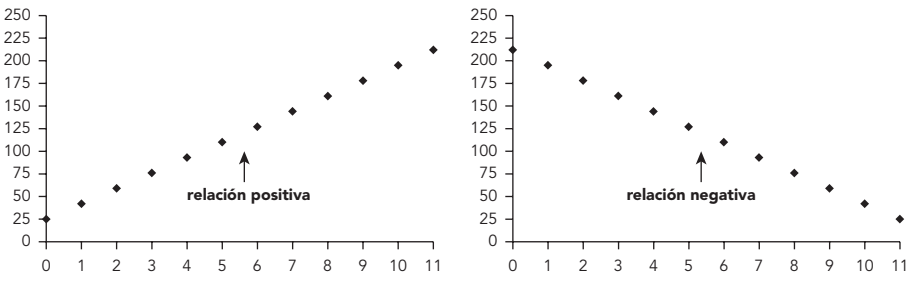
Los coeficientes de correlación pueden ser positivos, negativos, o cero. Una correlación positiva indica que los valores de las dos variables tienden a moverse en la misma dirección; a medida que los puntajes de una variable aumentan, también aumentan los puntajes de la otra, en promedio. Una correlación positiva (por ejemplo, 0,60) entre desempeño en lectura y escritura indicaría que, en promedio, a medida que aumenta el desempeño en lectura, también aumenta el desempeño en escritura, y viceversa. Por el contrario, una correlación negativa indicaría que a medida que el valor de una variable aumenta, el valor de la otra variable tiende a decrecer. Por ejemplo, una correlación negativa (-0,28) entre ansiedad respecto de las matemáticas y desempeño en una evaluación de matemáticas indicaría que, en general, a medida que la ansiedad del estudiante aumenta, el desempeño decrece (y viceversa). Una correlación igual a cero indica que no existe evidencia de una relación entre dos variables (tales como la altura de los estudiantes y el rendimiento). La Figura 6.1 ilustra gráficamente relaciones positivas y negativas entre variables. En el primer diagrama la relación es positiva, en el segundo es negativa.

Fuerza o magnitud de la relación

Los coeficientes de correlación cercanos a –1,0 o a +1,0 indican una fuerte relación. Los valores entre estos extremos indican relaciones bastante más débiles.

FIGURA 6.1

Correlaciones positivas y negativas



En general, las correlaciones entre mediciones de estatus socioeconómico y rendimiento tienden a variar entre 0,20 y 0,30. Las correlaciones entre los puntajes de los estudiantes en las evaluaciones de lectura y ciencias a menudo se ubican en el rango entre 0,80 y 0,90 (véase, por ejemplo, OCDE 2007), lo que sugiere que el rendimiento en una evaluación (por ejemplo, lectura) se apoya en las mismas habilidades que el rendimiento en la otra (ciencias). En las evaluaciones nacionales o en las investigaciones en educación en general, raramente se encuentran correlaciones perfectas o casi perfectas.

Elaboración de un diagrama de dispersión

Antes de calcular un coeficiente de correlación y evaluar su significado estadístico, puede ser útil elaborar un diagrama de dispersión que ilustra la relación entre dos variables en forma de gráfico. Si la relación es lineal, los puntos tenderán a ubicarse alrededor de una línea recta que atraviesa los datos. Cuanto más cerca de una línea recta se ubiquen los puntos, más fuerte será la relación entre las variables. En el ejercicio 6.1 se usa SPSS para describir la solidez de la relación entre dos subescalas de puntuación en matemáticas en la base de datos ***NATASSESS***, **Impl_pc** (porcentaje de respuestas correctas en los ítems que suponen la implementación de procedimientos matemáticos) y **Solve_pc** (porcentaje de respuestas correctas en análisis y resolución de problemas matemáticos).[1]

EJERCICIO 6.1

Elaboración de un diagrama de dispersión en SPSS

1. Abra el archivo de datos ***NAEA DATA ANALYSIS\SPSS DATA\NATASSESS4.SAV.***
2. Seleccione **Data – Weight Cases – Weight Cases by** …, mueva **Wgtpop** al cuadro etiquetado **Frequency Variable**, y haga clic en **OK**.
3. En la barra de herramientas, seleccione **Graphs – Legacy Dialogs – Scatter/Dot – Simple Scatter – Define**.

(continúa)

EJERCICIO 6.1 *(continúa)*

4. Asigne **Impl_pc** a **Y Axis** y **Solve_pc** a **X Axis** (figura del ejercicio 6.1.A). Haga clic en **OK**. Dele un tiempo para que procese. El resultado se muestra en la figura del ejercicio 6.1.B.

FIGURA DEL EJERCICIO 6.1.A Cuadro de diálogo parcial de SPSS antes de diseñar el diagrama de dispersión

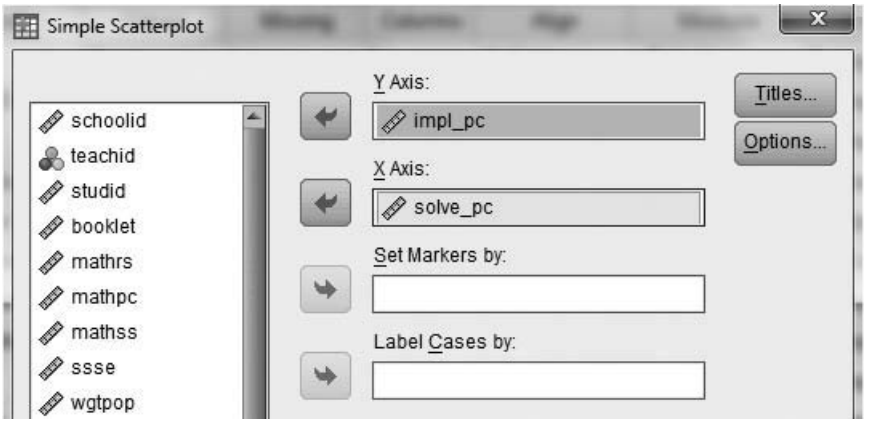
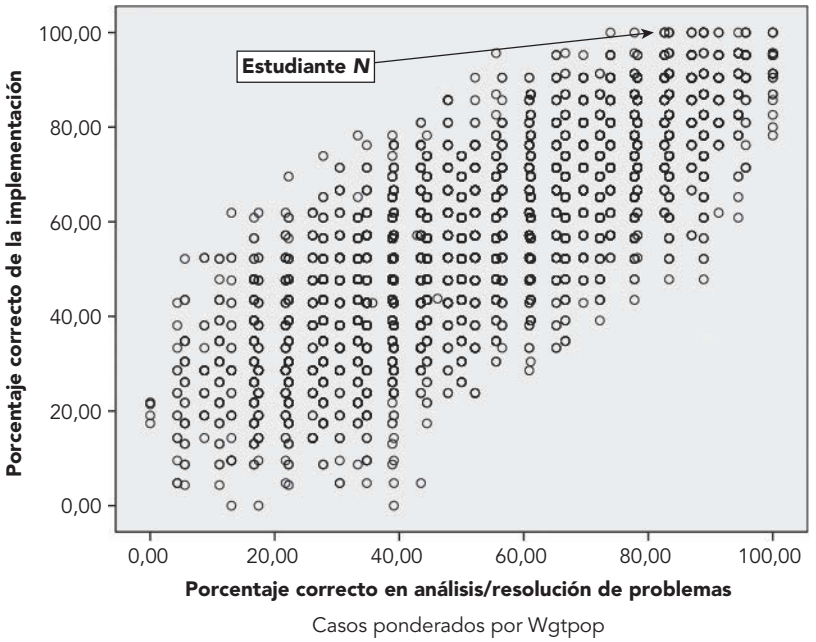


FIGURA DEL EJERCICIO 6.1.B Diagrama de dispersión de las relaciones entre la implementación de procedimientos y la resolución de problemas en matemáticas

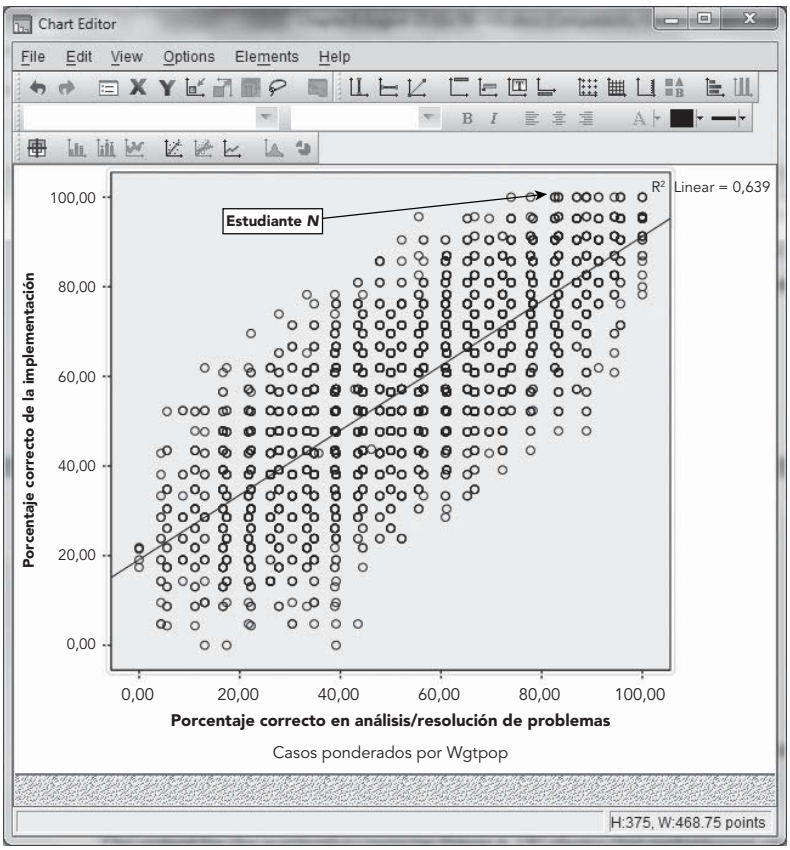


EJERCICIO 6.1 *(continúa)*

Podemos añadir una *línea de ajuste óptimo* al diagrama de dispersión. Esta es la línea recta que mejor resume los datos del diagrama. Para trazar la línea de ajuste óptimo en un diagrama de dispersión en SPSS, haga doble clic en el diagrama de dispersión para acceder a **Chart Editor** y, desde la barra de menú, seleccione **Elements – Fit Line at Total – Linear – Confidence Intervals – None**. El resultado se muestra en la figura del ejercicio 6.1.C.

En un diagrama de dispersión se representa a uno o más individuos mediante un punto, que es la intersección de sus puntuaciones en dos variables. Por ejemplo, el estudiante *N* (figura del ejercicio 6.1.B) obtuvo el 100 % de las respuestas correctas en implementación de matemáticas (eje *y*) y el 83 % de respuestas correctas en resolución de problemas matemáticos (eje *x*). Nótese que los puntos se agrupan en una banda, que va desde el borde inferior izquierdo hasta el borde superior derecho, un indicador de correlación positiva entre las dos variables.

FIGURA DEL EJERCICIO 6.1.C Diagrama de flujo que muestra la línea de ajuste óptimo



(continúa)

EJERCICIO 6.1 *(continúa)*

El resultado del diagrama de dispersión (figura de los ejercicios 6.1.B y 6.1.C) muestra que el desempeño en resolución de problemas matemáticos tiende a aumentar a medida que aumenta el desempeño en implementación de procedimientos matemáticos (y viceversa).

5. En el resultado en SPSS, haga clic fuera del área del gráfico. Guarde el resultado seleccionando **File – Save As: *NAEA DATA ANALYSIS\MY SOLUTIONS***, y nombre el archivo como ***EXERCISE 6.1.SPV***.

Cálculo del coeficiente de correlación y evaluación de su relevancia estadística

Este apartado muestra cómo calcular un coeficiente de correlación utilizando WesVar (ejercicio 6.2). El propósito es determinar la magnitud de la relación entre desempeño en implementación de procedimientos matemáticos (**Impl_pc**) y desempeño en resolución de problemas matemáticos (**Prob_pc**). También se necesita una medida del error en relación al coeficiente de correlación obtenido para permitir evaluar si el coeficiente de correlación difiere significativamente de cero.

EJERCICIO 6.2

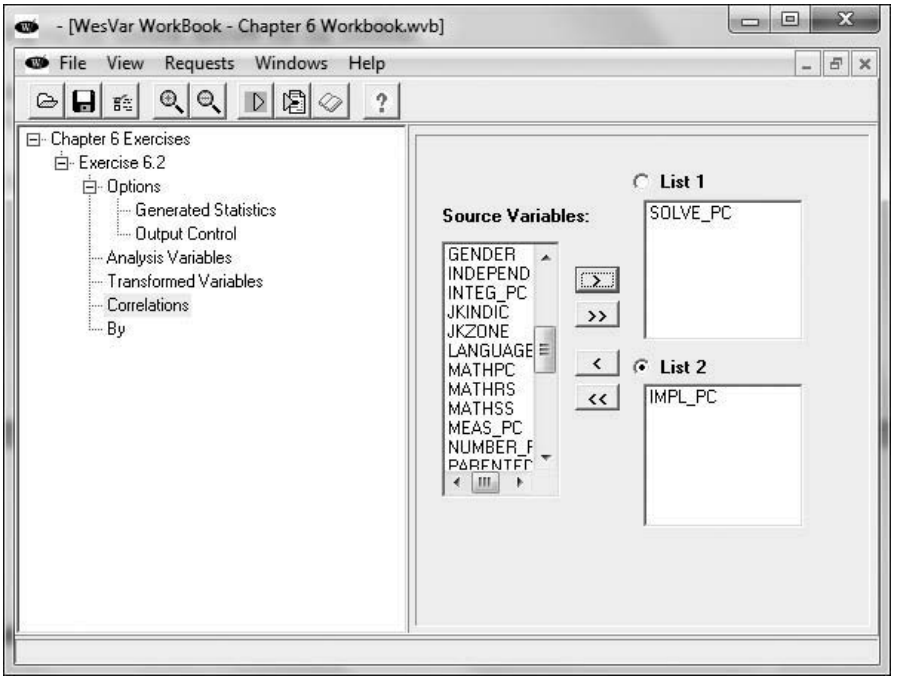
Cálculo del coeficiente de correlación, nivel nacional

1. Inicie WesVar, y haga clic en **New WesVar Workbook**. Puede recibir la siguiente advertencia: *Before creating a new Workbook, you will be asked to specify a Data file that will be used as the default Data file for new Workbook requests (Antes de crear un nuevo libro de trabajo, se le pedirá que especifique el archivo de datos que será utilizado como archivo de datos por defecto para las peticiones de nuevo libro de trabajo.*) Si esto sucede, haga clic en **OK**.
2. Aparecerá una ventana llamada **Open WesVar Data File for Workbook**. Seleccione el archivo de datos ***NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR.***
3. Guarde su nuevo libro de trabajo en ***NAEA DATA ANALYSIS\MY WESVAR FILES\ CHAPTER 6 WORKBOOK.WVB.*** Haga clic en **Workbook Title 1** (panel de la izquierda) y escriba ***Chapter 6 Exercises*** (panel de la derecha).

EJERCICIO 6.2 *(continúa)*

4. Haga clic en **Descriptive Stats**. Seleccione **Descriptive Request One** (panel de la izquierda) y escriba ***Exercise 6.2*** (panel de la derecha). Seleccione **Options – Output Control** en el panel izquierdo. Asegúrese de que el número de posiciones decimales para **Estimates** y **Std. Error** esté establecido en tres. Además, asegúrese de que **Variable Name** y **Variable Label** hayan sido seleccionados.
5. Seleccione **Correlations** en el panel izquierdo. Seleccione **List 1** (panel de la derecha), y mueva **Solve_pc** desde **Source Variables** a **List 1**. Seleccione **List 2**, y mueva **Impl_pc** desde **Source Variables** a **List 2** (figura del ejercicio 6.2.A). Puede agregar variables adicionales a cada lista si así lo desea. Cada variable de **List 1** se correlacionará con una de **List 2**.

FIGURA DEL EJERCICIO 6.2.A Libro de trabajo de WesVar antes de ejecutar un análisis de correlación

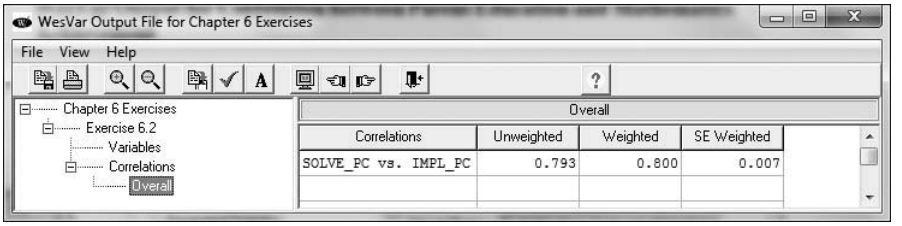


6. Ejecute el análisis haciendo clic en la **flecha verde** en la barra de herramientas.
7. Vea el resultado haciendo clic en el ícono **libro abierto** en la barra de herramientas. Haga clic en el signo **+** (más) para expandir **Correlations.** Seleccione **Overall** (figura del ejercicio 6.2.B).

(continúa)

EJERCICIO 6.2 *(continúa)*

FIGURA DEL EJERCICIO 6.2.B Resultado en WesVar: Correlación entre la resolución de problemas y la implementación de procedimientos matemáticos



8. Seleccione **File – Export – Single File – One Selection – Export** en la barra de menú, salve su resultado en ***NAEA DATA ANALYSIS\My Solutions*** como ***EXERCISE 6.2.TXT***.
9. Salga del resultado de WesVar, y guarde el libro de trabajo WesVar seleccionando **File** – **Save** (véase el paso 3 más arriba).

La figura del ejercicio 6.2.B muestra los coeficientes de correlación no ponderados y ponderados entre **Solve_pc** y **Impl_pc**. En una evaluación nacional se presenta el resultado ponderado. Para determinar si la correlación de 0,800 es estadísticamente relevante, calcule una estadística llamada *t* dividiendo la correlación por su error estándar. En este caso, *t* es igual a 0,800/0,007 o 114,3. Una tabla[a] de valores *t* revela que, para 60 grados de libertad (el número de *jackknife* se replica en el análisis WesVar) en una prueba bilateral, se requiere que el valor *t* sea igual o mayor que 2,0 para que tenga relevancia significativa ($p < 0,05$). Puesto que 114,3 excede este valor, se puede concluir con gran seguridad que no es probable que el valor *r* sea igual a cero. Puesto que el coeficiente de correlación es positivo, se puede decir que, en general, a medida que aumenta el desempeño de los estudiantes en la implementación de procedimientos matemáticos, también aumenta su desempeño en la resolución de problemas (y viceversa).

a. Véase la tabla de valores t en un libro de texto sobre estadísticas estándar. Como alternativa, también puede consultar http://surfstat.anu.edu.au/surfstat-home/tables/t.php para ver tablas de valores t en línea. Introduzca los grados de libertad (60) y probabilidad (0,05) para las pruebas bilaterales t (tabla final). Haga clic en la flecha inversa para calcular el valor t requerido para que se verifique la significación para un nivel de 0,05. En este ejemplo, este valor es 2.

REGRESIÓN

La regresión difiere de la correlación en varios modos. En primer lugar, el modelo de correlación no especifica la naturaleza de la relación entre variables. Por el contrario, la regresión ejemplifica la dependencia de una variable respecto de una o más variables distintas. Sobre la

base de la teoría de investigación, se considera que los puntajes en una variable (la variable dependiente, en los gráficos generalmente representada en el eje vertical [y]) dependen de o están supeditados a los puntajes de otra variable (la variable independiente, generalmente representada en el eje horizontal [x]). Por ejemplo, se espera que los puntajes obtenidos en una prueba de lectura dependan de la cantidad de tiempo que un estudiante dedique a la lectura por placer.[2]

En segundo lugar, en la regresión, la relación funcional entre variables independientes y dependientes se puede establecer formalmente como una ecuación con valores asociados que describen en qué medida la ecuación se ajusta a los datos. La información acerca del desempeño de un grupo de individuos se utiliza para especificar una ecuación (conocida como *ecuación de regresión*) asumiendo que la relación es lineal (esto es, que la variación en el valor de una variable será similar para todos los valores de la variable).

Los analistas pueden necesitar ir más allá de esta forma de regresión cuando analizan los datos de una evaluación nacional. Enfoques más sofisticados, tales como el modelo lineal jerárquico (MLJ), generalmente son más apropiados para tomar en consideración la estructura jerárquica o multinivel de los datos obtenidos en estos estudios (véase Raudenbush y Bryk, 2002; Snijders y Bosker, 1999). El MLJ puede distinguir los efectos de las variables de nivel de la escuela y del alumno. Por ejemplo, si tiene datos acerca del estatus de la escuela y del nivel socioeconómico de los estudiantes, ambos pueden ser incluidos en el modelo y se pueden determinar los efectos de cada uno en el rendimiento de los estudiantes. Un modelo de dos niveles también permite identificar la proporción de varianza entre escuelas y la proporción al interior de las escuelas que explican las variables incluidas en el modelo. De manera similar, un modelo de tres niveles (escuelas, grados y estudiantes) ofrece una estimación de la proporción de varianza explicada por las variables a nivel de las escuelas (tales como localización y tamaño), a nivel del grado (tales como características del docente y disponibilidad de recursos pedagógicos) y a nivel del estudiante (tales como edad y ansiedad frente a las matemáticas). Los modelos multinivel son especialmente útiles cuando la varianza entre escuelas en la variable dependiente es amplia (por ejemplo, cuando excede el 5 % de la varianza total).

La modelización multinivel está más allá de la finalidad del presente volumen. De todos modos, la forma del análisis de regresión que se describe proporciona una introducción a algunos de los conceptos que subyacen a la modelización multinivel y puede ser utilizada cuando un modelo multinivel no resulta apropiado.

La ecuación de la regresión en los casos con una variable dependiente [y] y una variable independiente [x], es como sigue:

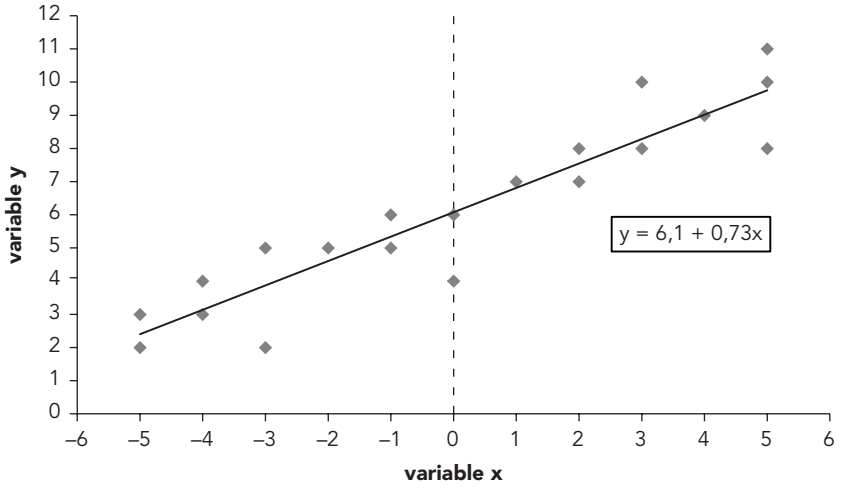
$$\hat{y} = \alpha + bX + \varepsilon,$$

donde
α = intercepción (el punto en el eje y cuando x es cero)
b = gradiente o pendiente de la línea de regresión (el coeficiente de regresión)
X = puntuación en la variable independiente
ε = término de error (reflejado en los residuales o diferencias entre los valores esperados y los observados).[3]

La Figura 6.2 muestra la ecuación de regresión y la línea de regresión en un diagrama de dispersión de dos variables: x (independiente) e y (dependiente).

FIGURA 6.2

Línea de regresión y ecuación de regresión en un diagrama de dispersión



RECUADRO 6.1

Variables en regresión estándar

Variable dependiente única (resultado), tales como el rendimiento en lectura o matemáticas

Una o más variables independientes (explicativas), tales como el tamaño de la clase, la región geográfica, las cualificaciones del docente, la educación de los padres, el género del estudiante

La ecuación de regresión (a) indica si existe una tendencia al incremento o decremento en las puntuaciones de *y* (predicción) a medida que cambian los valores de *x* (indicador); (b) puede ser utilizada para estimar o predecir valores de *y* a partir de valores conocidos de *x*; y (c) estima el valor de *y* cuando el valor de *x* es cero (véase ejercicio 6.3).

La regresión describe la relación entre dos o más variables en forma de ecuación. Esto hace posible predecir, por ejemplo, la puntuación de un estudiante basándose en una prueba de rendimiento a partir de lo que se sabe del contexto familiar del estudiante u otras variables.

En una evaluación nacional típica, es probable que muchas variables se correlacionen significativamente con los puntajes en las pruebas de matemáticas, lengua o ciencia. En esta situación, se puede utilizar la regresión múltiple para cuantificar la asociación entre múltiples variables independientes y una variable dependiente. En el recuadro 6.1 se presentan ejemplos de variables dependientes e independientes que aparecen frecuentemente en evaluaciones nacionales.

Implementación del análisis de regresión con una variable dependiente y una variable independiente

Las secciones que siguen proporcionan ejemplos acerca de cómo llevar a cabo el análisis utilizando la regresión en WesVar y acerca de cómo interpretar el resultado. WesVar se utiliza porque toma en cuenta la naturaleza compleja de la muestra en la evaluación nacional (véase el capítulo 3) al momento de evaluar los niveles de significación (por ejemplo, las diferencias originadas en la agrupación de estudiantes en escuelas y clases). El primer ejemplo se ocupa del tipo más

simple de regresión. Dirige su atención hacia la asociación entre una variable dependiente (**Mathss**) —con un intervalo que abarca de 88 a 400— y una variable independiente —número de libros en el hogar (**Books**)— con un intervalo que va desde "no hay libros" (0) hasta 120. El programa de regresión lineal en WesVar le pedirá que seleccione variables independientes de dos listas: **Class Variables** y **Source Variables**. La lista **Class Variables** comprende variables categóricas que tienen 255 categorías de respuesta o menos, incluyendo valores omitidos (una característica de WesVar). Las variables **Source Variables** son variables continuas. Algunas variables aparecen tanto en la lista de **Class** como en la de **Source Variables**. Un ejemplo es **Books**. Aquí, de todos modos, tratamos **Books** como una variable **Source**. Véase el ejemplo en el ejercicio 6.3.

El análisis de regresión puede utilizarse con variables categóricas así como con variables continuas. En el ejemplo del ejercicio 6.4, la variable independiente es **Region**. Recuerde que en el capítulo 4 se

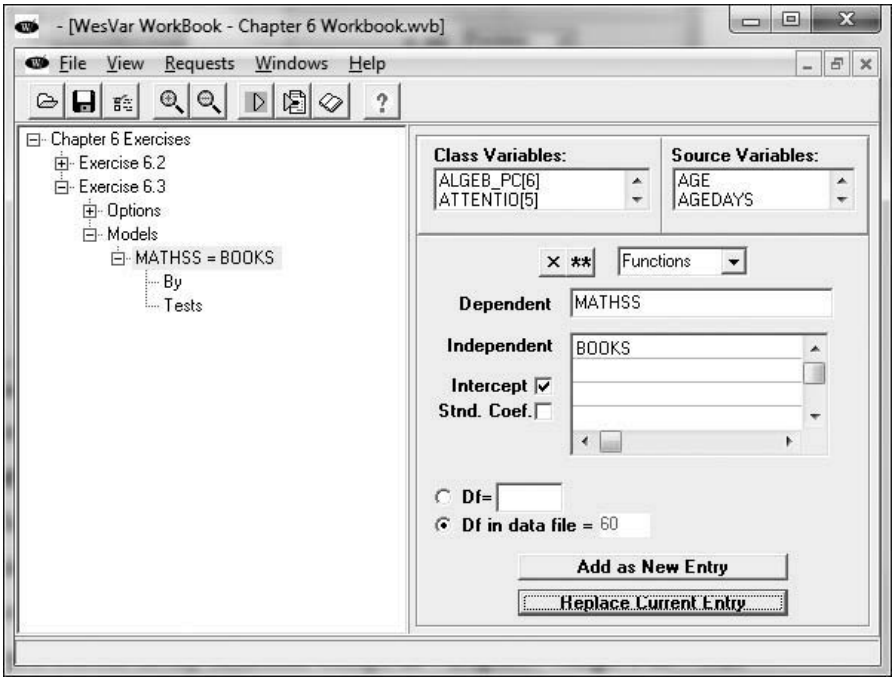
EJERCICIO 6.3

Ejecución del análisis de regresión en WesVar con una variable independiente (continua)

1. Ejecute WesVar, y abra el libro de trabajo utilizado en el ejercicio 6.2, ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK.***
2. Seleccione **Chapter 6 Exercises** en el panel izquierdo, y haga clic en **Regression** en el panel derecho. Seleccione **Regression Request One** (panel izquierdo). Escriba **Exercise 6.3** (panel derecho).
3. Seleccione **Options** en el panel izquierdo y asegúrese de que esté seleccionado **Linear**. En **Options**, seleccione **Generated Statistics – Confidence Interval**. En **Output Control**, establezca tres posiciones decimales para **Estimates** y **Std. Error.**
4. Seleccione **Models** en el panel izquierdo. Arrastre **Mathss** desde la lista de **Source Variables** hasta el cuadro con la etiqueta **Dependent**. Esta es la variable dependiente para el análisis de regresión. (Si solo puede ver una o dos variables en los cuadros del panel derecho, mueva el cursor hasta el borde inferior y arrástrelo hacia abajo para mostrar más variables).
5. Vaya al cuadro **Source Variables** y arrastre **Books** hasta el primer espacio con la etiqueta **Independent**. Asegúrese de que el cuadro **Intercept** haya sido seleccionado. Haga clic en **Add as New Entry** (figura del ejercicio 6.3.A).

EJERCICIO 6.3 *(continúa)*

FIGURA DEL EJERCICIO 6.3.A Libro de trabajo WesVar antes de ejecutar el análisis de regresión con una variable independiente



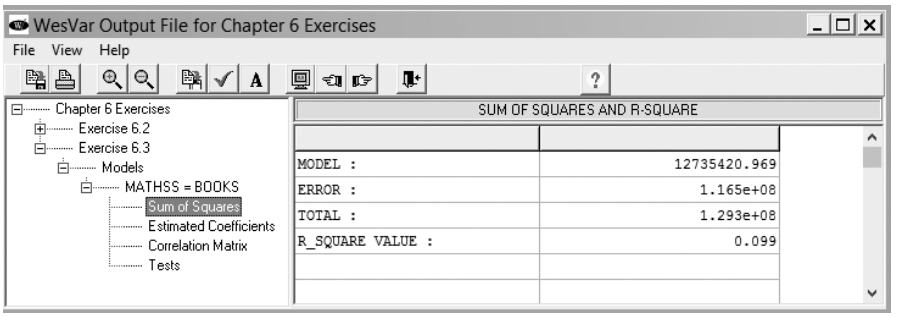
6. Ejecute el análisis de regresión (haga clic en la **flecha verde** en la barra de herramientas), y abra el resultado (haga clic en el ícono **libro abierto** en la barra de herramientas). Expanda **Exercise 6.3**, **Models**, y **Mathss = Books**. Haga clic en **Sum of Squares** (figura del ejercicio 6.3.B). El valor de *R elevado al* cuadrado **(R_Square Value)** es 0,099. Esto indica que **Books** explicó casi el 10 % de la varianza en los puntajes obtenidos en las pruebas de rendimiento en matemáticas. El valor de *R al cuadrado* se obtiene dividiendo la suma de valores al cuadrado del modelo (explicado) por la suma total de los valores al cuadrado.
7. Guarde este resultado como un archivo de texto en ***MY SOLUTIONS*** utilizando ***File – Export* – Single File– One Selection**. Dé al archivo el nombre ***EXERCISE 6.3 SUM OF SQUARES***.
8. Seleccione **Estimated Coefficients** en el archivo de resultados (figura del ejercicio 6.3.C). Aquí se muestra la estimación del parámetro o la variación esperada en la puntuación en relación con la cantidad de libros en el hogar del estudiante. Aplique la fórmula descrita anteriormente, $\hat{y} = \alpha + bX$, para calcular la relación o

(continúa)

EJERCICIO 6.3 *(continúa)*

asociación entre la cantidad de libros y el rendimiento en matemáticas. No utilice ε en estos cálculos[a]. Nótese que *X* representa la cantidad de libros y *b* es la pendiente de la línea de regresión. El valor de la estimación para la variable libros (0,515) indica que un aumento en un libro se asocia con un incremento de 0,515 puntos en el rendimiento en matemáticas. La puntuación esperada en matemáticas de un estudiante que no tiene libros es el valor del punto de intersección[b] 225,196 o, utilizando la fórmula, 225,196 + 0 * 0,515. La puntuación esperada de un estudiante que tiene 10 libros es 230,346 (225,196 + 10 * 0,515). Por consiguiente, tener 10 libros en el hogar se asocia a un incremento de cinco puntos en el rendimiento en matemáticas. La puntuación media para Books es 48,2[c]. Por consiguiente, el puntaje esperado en matemáticas para un estudiante con una cantidad promedio de libros en su hogar es 225,196 + 0,515 * 48,2, o 250,019, lo cual es un puntaje muy próximo al puntaje medio general registrado de 250,0.

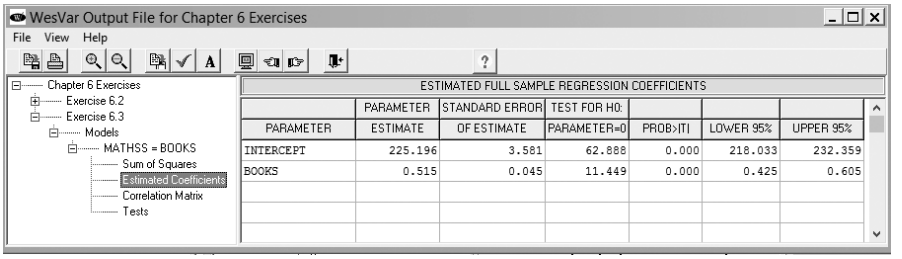
FIGURA DEL EJERCICIO 6.3.B Resultado del análisis de regresión en WesVar con una variable independiente: la suma de valores al cuadrado y el valor de R al cuadrado



Los valores *t* de la figura del ejercicio 6.3.C se calculan dividiendo cada valor estimado por su error estándar. El valor *t* es una medida de significación estadística y evalúa la probabilidad de que el valor efectivo del parámetro no sea cero. Para **Books**, el valor *t* es 11,449 y el valor de la probabilidad (*p*) (Prob>|T|) es o está cerca de cero (0,000). Esto indica que existe una probabilidad ínfima de que el valor efectivo del parámetro **Books** sea cero. El intervalo de confianza del 95 % de una variable se estima de forma aproximada sumando dos veces su error estándar al parámetro y restando dos veces su error estándar. De este modo, luego de redondear, el intervalo de confianza del 95 % para **Books** es de 0,425 a 0,605 (0,515 ± 2 * 0,045). Hay un 95 % de certeza de que la estimación para el valor de **Books** en la población se halle entre 0,425 y 0,59. (Estos valores son casi idénticos a los representados en la figura del ejercicio 6.3.C.)

EJERCICIO 6.3 *(continúa)*

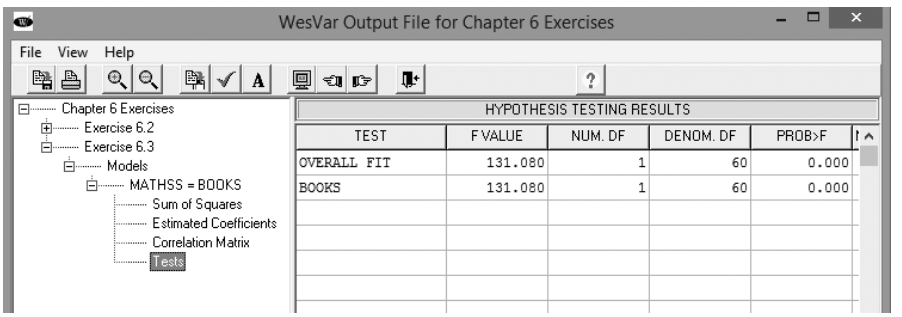
FIGURA DEL EJERCICIO 6.3.C Resultado del análisis de regresión en WesVar con una variable independiente: coeficientes estimados



9. Guarde el resultado en ***MY SOLUTIONS*** utilizando **File – Export – Single File – One Selection**. Dé al archivo el nombre ***EXERCISE 6.3 ESTIMATES.TXT***.

10. A continuación resultará de interés conocer lo bien que se ajusta el modelo estadístico que estima el valor de **Mathss** sobre la base de los datos de una variable independiente, **Books**. Seleccione **Tests** (figura del ejercicio 6.3.D). Nótese que el ajuste global del modelo de regresión es estadísticamente significativo (primera fila); la probabilidad de obtener un valor F^{d} de 131,080 se aproxima a cero (0,000). Esto significa que el modelo de regresión obtenido que contiene **Books** es estadísticamente diferente del modelo que no incluye esta variable. La siguiente fila en la figura del ejercicio 6.3.D presenta idénticos datos. Esto confirma que un modelo que contiene la variable **Books** es diferente en términos estadísticos comparado con un modelo que no contiene variables independientes (modelo nulo). Desde la perspectiva de las políticas, este hallazgo indica que la cantidad de libros en el hogar se relaciona con el puntaje del rendimiento en matemáticas de los estudiantes.

FIGURA DEL EJERCICIO 6.3.D Resultado para el análisis de regresión en WesVar con una variable independiente



(continúa)

EJERCICIO 6.3 *(continúa)*

11. Guarde el resultado en ***MY SOLUTIONS*** utilizando **File – Export – Single File – One Selection**. Nombre el archivo como ***EXERCISE 6.3 TESTS.TXT***.
12. Regrese a su libro de trabajo (utilizando el ícono **puerta abierta** en la barra de herramientas). Seleccione **File** – **Save** y luego haga clic en **File** – **Close**. Su libro de trabajo debería guardarse en ***NAEA DATA ANALYSIS\MY WESVAR FILES\ CHAPTER 6 WORKBOOK***.

a. Esto sucede porque los errores positivos y negativos se contrarrestan.
b. El punto de intersección α estima el valor promedio de *y* (matemáticas, en este caso), donde *X* (cantidad de libros en el hogar) = 0. El valor de α es el punto donde la línea de regresión interseca el eje *y*.
c. Para calcular esto, seleccione **Descriptives** en WesVar, y mueva **Books** desde **Source Variables** a **Selected**, tal como se indica en el ejercicio 3.1.
d. La estadistica *F*, que debe utilizarse cuando se comparan más de dos variables, evalúa la significación o las diferencias entre medias.

EJERCICIO 6.4

Ejecución del análisis de regresión en WesVar con una sola variable independiente (categórica)

1. Ejecute WesVar y abra el libro de trabajo utilizado en el ejercicio 6.3, ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK***.
2. Seleccione **Chapter 6 Exercises** (panel de la izquierda) y haga clic en **Regression** (panel de la derecha). Seleccione **Regression Request Two**[a] (panel izquierdo). Escriba ***Exercise 6.4*** (panel derecho). Bajo **Options – Generated Statistics,** seleccione **Confidence Interval**. Bajo **Options – Output Control**, establezca el número de puntos decimales para el error estimado y el estándar en tres posiciones decimales.
3. Seleccione **Models** en el panel izquierdo.
4. Arrastre **Mathss** desde la lista **Source Variables** hasta el cuadro etiquetado **Dependent** en el panel derecho. Esta es la variable dependiente para su análisis de regresión.
5. Vaya al cuadro **Class Variables** y arrastre hacia abajo **Region[4]** hasta la fila para la primera variable independiente. Asegúrese de que se haya seleccionado el cuadro **Intercept**. Haga clic en **Add as New Entry**. Aquí se toma **Region[4]** de la lista **Class Variables** porque es una variable categórica (ya que cada estudiante fue asignado a una de las cuatro regiones).
6. Ejecute el análisis de regresión (haga clic en la **flecha verde** en la barra de herramientas), y abra el resultado (haga clic en el ícono **libro abierto** en la barra de herramientas). Expanda **Exercise 6.4**, **Models**, y **Mathss = region[4]**. Seleccione

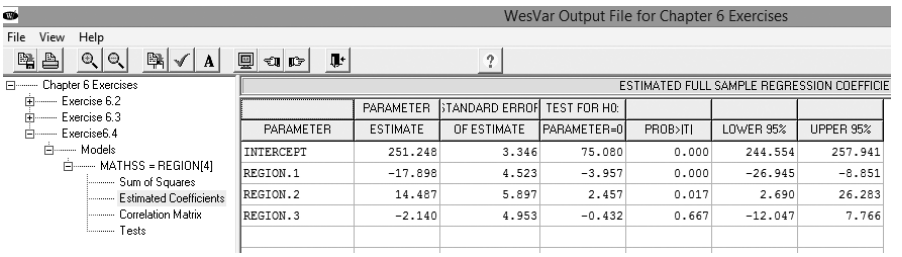
EJERCICIO 6.4 *(continúa)*

Sum of Squares en el panel izquierdo. El valor de *R al cuadrado* es 0,054. Esto indica que **Region[4]** explica o justifica el 5 % de la varianza en los puntajes de las pruebas de rendimiento en matemáticas.

7. Haga clic en **Estimated Coefficients** en el panel izquierdo.

   Se brindan las estimaciones de parámetros para tres de las cuatro regiones (véase la figura del ejercicio 6.4.A). La estimación del parámetro para la región de referencia es la intersección (251,248). Esto equivale a la puntuación media para la región de la Costa del Sudoeste indicada en el ejercicio 4.2. La estimación del parámetro para **Region.1** (Noroeste) es −17,898 (esto es, 17,898 puntos por debajo de la categoría de referencia). Por lo tanto, la puntuación esperada de un estudiante de rendimiento promedio en la región Noroeste es de 233,350 (251,248 − 17,898). Esto equivale a la puntuación media estimada para la región Noroeste en el ejercicio 4.2. La puntuación estimada de un estudiante con desempeño promedio en el Área Metropolitana (**Region.2**) es de 265,735 (esto es, 14,487 puntos por encima de la de un estudiante de la región de referencia). Por último, la puntuación estimada para un estudiante que tiene un desempeño promedio en la Región Montañosa del Este (**Region.3**) es de 249,108, que es 2,140 puntos por debajo de la región de referencia. El valor *t* no significativo asociado a la estimación del parámetro para la Región Montañosa del Este (ProbITI = 0,667) indica que −2,14 no es significativamente diferente de cero y, por consiguiente, el rendimiento de un estudiante promedio de esta región no es significativamente diferente —desde un punto de vista estadístico— del de un estudiante promedio de la región de referencia (Costa del Sudoeste).

FIGURA DEL EJERCICIO 6.4.A Resultado del análisis de regresión en WesVar: variable independiente categórica



PARAMETER	PARAMETER ESTIMATE	STANDARD ERROR OF ESTIMATE	TEST FOR H0: PARAMETER=0	PROB>|T|	LOWER 95%	UPPER 95%
INTERCEPT	251.248	3.346	75.080	0.000	244.554	257.941
REGION.1	-17.898	4.523	-3.957	0.000	-26.945	-8.851
REGION.2	14.487	5.897	2.457	0.017	2.690	26.283
REGION.3	-2.140	4.953	-0.432	0.667	-12.047	7.766

8. Guarde este resultado como un archivo de texto en ***MY SOLUTIONS*** utilizando **File – Export – Single File – One Selection**. Nombre el archivo como ***EXERCISE 6.4 ESTIMATES.TXT***.

a. Este número puede variar. Por ejemplo, si usted ya ejecutó un análisis de regresión y lo eliminó, el número será más alto.

estableció que los estudiantes del Área Metropolitana tuvieron un rendimiento significativamente mejor que los estudiantes de las otras tres regiones de Sentz (ejercicio 4.2). Nótese que, cuando se selecciona una variable como **Region** como variable independiente en un análisis de regresión, hay que crear una serie de variables para indicar la región en la cual se localiza la escuela de un estudiante. WesVar crea una serie de variables ficticias, cada una de las cuales corresponde a una única región, que están codificadas mediante 1 o 0, según si el estudiante pertenece a la región o no. Por ejemplo, cuando WesVar crea la variable ficticia **Northwest**, todos los estudiantes que asisten a la escuela en dicha región están codificados como 1 en la variable ficticia, y los estudiantes de cada una de las otras tres regiones están codificados como 0. De manera análoga, los estudiantes que asisten a escuelas en el Área Metropolitana deberían ser codificados como 1 en la variable ficticia **Metro**, y los estudiantes de cada una de las otras tres regiones deberían ser codificados como 0. Lo mismo aplica para los estudiantes de la Región Montañosa del Este. Se incluye en el análisis la región o categoría final, que es conocida como *categoría de referencia* y que no está codificada por separado. En el ejemplo dado en el ejercicio 6.4, donde **Region** es la variable categórica (variable de clase), el rendimiento de los estudiantes en cada una de las primeras tres regiones (Noroeste, Área Metropolitana y Región Montañosa del Este) se compara con el de los estudiantes de la cuarta región (Costa del Sudoeste).

Implementación del análisis de regresión múltiple con una variable dependiente y dos o más variables independientes

En esta sección se describe el efecto del incremento del número de variables independientes en la explicación o justificación de los resultados de las pruebas de matemáticas. Se toma en consideración tres variables independientes:

- **Books**, la cantidad de libros en el hogar del estudiante, una variable discreta
- **Distance**, la distancia en kilómetros entre el hogar del estudiante y la escuela, una variable continua

- **Parented [6]**, el nivel de educación más alto alcanzado por cada progenitor, una variable categórica con seis categorías: 1 = sin educación formal; 2 = grados 1–3; 3 = grados 4–6; 4 = grados 7–9; 5 = escuela media superior; y 6 = diploma.

La interrelación entre variables independientes debería revisarse antes de ejecutar un análisis de regresión. Debería ser de particular interés la *multicolinealidad*, que surge cuando dos o más variables independientes están fuertemente correlacionadas. Cuando esto ocurre, los errores estándar de los análisis de regresión aumentan, haciendo más difícil evaluar el rol particular de cada variable independiente para explicar el rendimiento[4]. El ejercicio 6.5 muestra cómo pueden estimarse los coeficientes de correlación en WesVar.

EJERCICIO 6.5

Estimación de los coeficientes de correlación

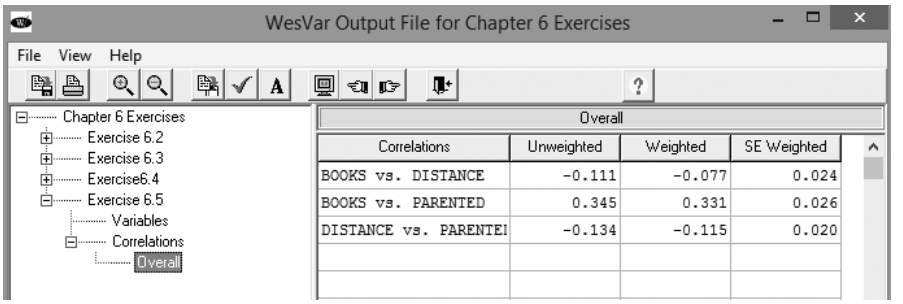
1. Ejecute WesVar y abra el libro de trabajo utilizado en el ejercicio 6.4, ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK***.
2. Seleccione **Chapter 6 Exercises** (panel izquierdo), y haga clic en **Descriptive Stats** (panel derecho). Seleccione **Descriptive Request Two** en el panel izquierdo, y escriba ***Exercise 6.5*** en el espacio previsto en el panel derecho.
3. Seleccione **Options – Output Control** y establezca en tres el número de posiciones decimales para **Estimates** y **Std. Error**.
4. Seleccione **Correlations** (panel izquierdo). Seleccione **List 1** (panel derecho). Mueva las tres variables: **Books, Distance** y **Parented** a **List 1**.
5. Ejecute correlaciones (haga clic en la **flecha verde** en la barra de herramientas). Abra el resultado (haga clic en el ícono **libro abierto** en la barra de herramientas). Seleccione y expanda **Exercise 6.5** en el panel izquierdo. Seleccione **Correlations** – **Overall**.

   Los datos de resultados que figuran en la columna **Weighted** (figura del ejercicio 6.5.A) muestran que no hay evidencia de multicolinealidad ya que ninguna de las correlaciones se aproxima a 0,80 (véase Hutcheson y Sofroniou, 1999). La correlación negativa entre **Books** y **Distance** (−0,077) indica que a medida que la distancia entre el hogar y la escuela tiende a crecer, la cantidad de libros en el hogar de los estudiantes tiende a decrecer. La correlación entre **Books** y **Parented** (tratada aquí como una variable continua) es de 0,331. Esta indica que niveles educativos más altos de los padres se asocian con una mayor cantidad de libros en sus hogares.

(continúa)

EJERCICIO 6.5 *(continúa)*

FIGURA DEL EJERCICIO 6.5.A Resultado de correlaciones entre variables independientes



6. Guarde este resultado como un archivo de texto en **My Solutions** utilizando **File – Export – Single File – One Selection**. Nombre el archivo como ***EXERCISE 6.5 CORRELATIONS***. Cierre su libro de trabajo en WesVar utilizando **File – Save** and **File – Close**.

Puesto que las tres variables no están altamente interrelacionadas, puede ejecutar un análisis de regresión con una variable dependiente (**Mathss**) y tres variables independientes (**Books**, **Distance** y **Parented [6]**) (ejercicio 6.6).

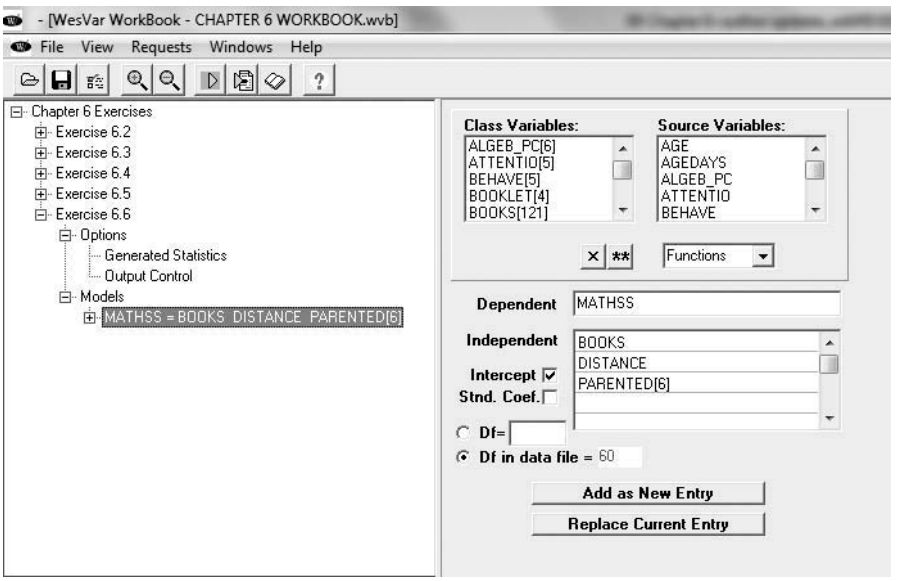
En síntesis, los datos muestran que un modelo con tres variables independientes (**Books**, **Distance** y **Parented [6]**) explica el 24 % de la variación en el rendimiento en matemáticas. El modelo sugiere una asociación positiva entre la cantidad de libros en el hogar de un estudiante y su desempeño en matemáticas, incluso luego de haber tomado en consideración las otras dos variables. También sugiere que el nivel de educación parental se asocia con el desempeño en matemáticas; los estudiantes cuyos padres tienen un mayor nivel de educación tienden a tener puntuaciones esperadas más altas que los alumnos cuyos padres tienen un menor nivel, tomando en consideración las otras dos variables (cantidad de libros y distancia). Por último, el modelo indica una asociación negativa entre la distancia respecto de la escuela y el rendimiento en matemáticas, tomando en consideración la cantidad de libros y la educación de los padres; los estudiantes que viven más lejos de la escuela tienden a tener un desempeño más bajo que los que viven cerca de la misma.

EJERCICIO 6.6

Ejecución de un análisis de regresión en WesVar con más de una variable independiente

1. Abra el libro de trabajo de WesVar ***CHAPTER 6 WORKBOOK*** que guardó en ***MY WESVAR FILES*** (el último utilizado en el ejercicio 6.5).
2. Seleccione **Chapter 6 Exercises** (panel izquierdo) y haga clic en **Regression** (panel derecho). Seleccione **Regression Request Three** (panel izquierdo). Escriba ***Exercise 6.6*** (panel derecho).
3. Bajo **Options – Output Control**, establezca en tres las posiciones decimales para **Estimates** y **Std. Error**. Haga clic en **Models** (panel izquierdo).
4. Arrastre **Mathss** desde la lista **Source Variables** hasta el cuadro etiquetado **Dependent**.
5. Vaya a S**ource Variables** y arrastre **Books** y **Distance** hasta la lista de variables independientes de manera que cada variable esté en una fila. Luego, vaya al cuadro **Class Variables** y arrastre **Parented [6]** hasta la fila para la tercera variable independiente.[a]
6. Asegúrese de que esté seleccionado el cuadro **Intercept** (panel derecho). Haga clic en **Add as New Entry** (panel derecho) (figura del ejercicio 6.6.A). Ejecute el análisis de regresión (haga clic en la **flecha verde** en la barra de herramientas).

FIGURA DEL EJERCICIO 6.6.A Pantalla de WesVar antes de ejecutar un análisis de regresión con más de una variable independiente

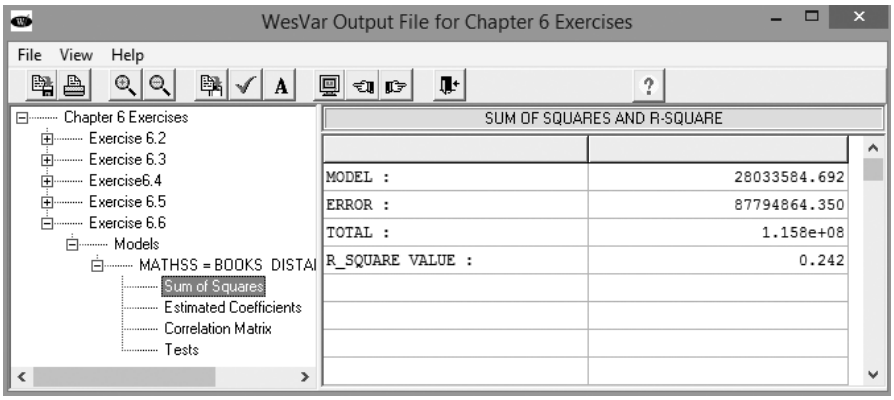


(continúa)

EJERCICIO 6.6 *(continúa)*

7. Vea el resultado (haga clic en el ícono **libro abierto** en la barra de herramientas). Expanda **Exercise 6.6** (panel izquierdo) hasta que vea **Sum of Squares**. Esto muestra que el nuevo modelo de tres variables justifica el 24 % de la varianza en el rendimiento en matemáticas (R^2 = 0.242). Esto es una mejora respecto del modelo anterior en el que **Books**, como una única variable independiente, justificaba en menos del 10 % la varianza en el rendimiento en matemáticas (véase la figura del ejercicio 6.3.B).

FIGURA DEL EJERCICIO 6.6.B Resultado de análisis de regresión en WesVar con más de una variable independiente: suma de cuadrados



8. Guarde el resultado en la figura del ejercicio 6.6.B seleccionando **File – Export – Single File One Selection – Export** como ***My Solutions\EXERCISE 6.6 SUM OF SQUARES***.

9. Seleccione **Estimated Coefficients** en el archivo de resultado (panel izquierdo, expandiéndolo todo lo necesario). El resultado ofrece estimaciones de parámetro para **Intercept**, **Books** y **Distance**, y cinco o seis niveles de **Parented**. Nótese que todos los parámetros en el modelo son estadísticamente significativos; el valor de Prob>|T|) es o está cerca de cero. Esto indica una probabilidad muy baja de que cualquiera de los parámetros sea cero.

10. La estimación de parámetro para **Intercept** es 278,909. Esto se corresponde con la puntuación esperada de un estudiante que no tenga libros (**Books**) en su hogar, que viva a una distancia (**Distance)** nula de la escuela (cero kilómetros) y que tenga por lo menos uno de los padres con el nivel de educación parental más alto (**Parented.6**, la categoría de referencia para **Parented [6]**). La estimación del parámetro para **Books** es 0,309. Este es el incremento en rendimiento asociado con un libro adicional en el hogar. Por consiguiente, un estudiante con una cantidad promedio de libros en el hogar (**Books** 48,191), que vive a menos de un kilómetro de la escuela, y uno de cuyos padres al menos posee un diploma, tendría una puntuación estimada de

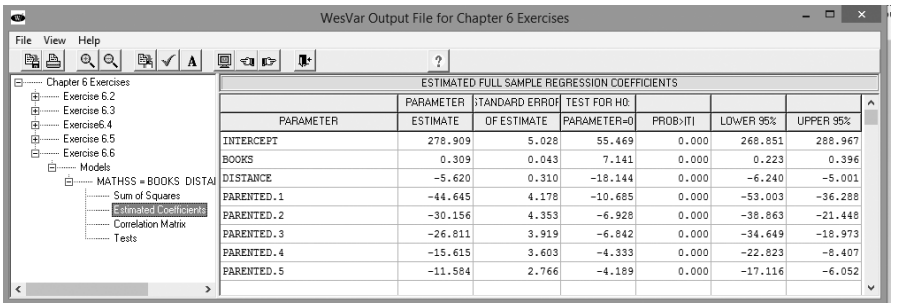
EJERCICIO 6.6 *(continúa)*

278,909 + (0,309 * 48,191) + (−5,620 * 0), o 293,800. (Nótese que **Parented [6]** tiene una ponderación cero en este cálculo porque es la categoría de referencia).

11. La estimación del parámetro para **Distance** es −5,620. El signo negativo significa que los puntajes esperados de los estudiantes en matemáticas decrecen a medida que aumenta la distancia entre la escuela y el hogar en el que viven. La distancia promedio a la escuela es de 4,257 kilómetros[b]. Por lo tanto, la puntuación esperada en matemáticas para un estudiante que vive a 4,257 kilómetros de la escuela, que tiene cero libros en su hogar y que tiene por lo menos uno de sus progenitores en la categoría de referencia **Parented [6]** es 278,909 + (−5,620 * 4,257) + (0,309 * 0), o 254,985.

12. Se proporciona estimaciones de parámetros para cinco niveles de **Parented [6]**. Tal como se indicó antes, **Parented.6** (por lo menos uno de los padres posee un diploma), la categoría de referencia, es el nivel más alto de educación de los padres. Estimaciones de parámetros negativas se asocian con niveles más bajos de educación de los padres. Por ejemplo, la estimación de parámetro para **Parented.2** (grados 1–3 completos) es −30,156 (véase la figura del ejercicio 6.6.C). Por lo tanto, la puntuación estimada de un estudiante que no tiene libros en su hogar, que vive junto a la escuela o muy cerca de ella, y de cuyos padres por lo menos uno completó el 3.° grado (**Parented.2)** es 278,909 − 30,156, o 248,753.

FIGURA DEL EJERCICIO 6.6.C Resultado del análisis de regresión en WesVar con más de una variable independiente: coeficientes estimados



ESTIMATED FULL SAMPLE REGRESSION COEFFICIENTS

PARAMETER	PARAMETER ESTIMATE	STANDARD ERROR OF ESTIMATE	TEST FOR H0: PARAMETER=0	PROB>|T|	LOWER 95%	UPPER 95%
INTERCEPT	278.909	5.028	55.469	0.000	268.851	288.967
BOOKS	0.309	0.043	7.141	0.000	0.223	0.396
DISTANCE	-5.620	0.310	-18.144	0.000	-6.240	-5.001
PARENTED.1	-44.645	4.178	-10.685	0.000	-53.003	-36.288
PARENTED.2	-30.156	4.353	-6.928	0.000	-38.863	-21.448
PARENTED.3	-26.811	3.919	-6.842	0.000	-34.649	-18.973
PARENTED.4	-15.615	3.603	-4.333	0.000	-22.823	-8.407
PARENTED.5	-11.584	2.766	-4.189	0.000	-17.116	-6.052

13. De manera similar, se puede generar puntuaciones esperadas para los estudiantes contando con cantidades variables de libros en el hogar, distancias variables entre hogar y escuela, y con progenitores que tengan niveles educativos variables. Por lo tanto, la puntuación esperada para un estudiante que tiene en su hogar 50 libros (**Books)**, que vive a 5 kilómetros de distancia (**Distance)** de la escuela y cuyo nivel de educación parental más alto son los grados 7–9 (**Parented.4**) es igual a 278,909 + (0,309 * 50) + (−5,620 * 5) + (−15,615), o 250,644.

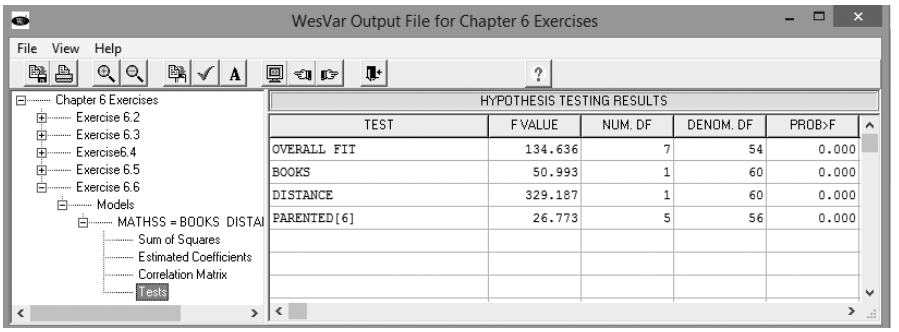
(continúa)

EJERCICIO 6.6 *(continúa)*

14. Guarde el resultado en ***MY SOLUTIONS*** utilizando **File – Export – Single File – One Selection**. Nombre el archivo como ***EXERCISE 6.6 ESTIMATES.TXT***.

15. Seleccione **Tests** en el resultado. Aquí puede ver (primera fila de la figura del ejercicio 6.6.D) que el ajuste general del modelo de regresión es estadísticamente significativo, con la probabilidad de obtener un valor para *F* de 134,636, que se aproxima a cero (0,000). Esto significa que por lo menos una de las variables, **Books**, **Distance** o **Parented [6]** es estadísticamente significativa. Los valores *p* para **Books**, **Distance** y **Parented [6]** muestran que para cada una de estas variables, el coeficiente de regresión difiere significativamente de cero, luego de haber controlado las otras. Por ejemplo, **Distance** es estadísticamente significativa aun después de haber tomado en cuenta los efectos de **Books** y **Parented [6]**.

FIGURA DEL EJERCICIO 6.6.D Resultado de análisis de regresión en WesVar con más de una variable independiente: prueba de ajuste del modelo



16. Guarde este resultado como un archivo de texto en **My Solutions** utilizando **File – Export – Single File – One Selection**. Nombre el archivo como ***EXERCISE 6.6 TESTS.TXT***.

17. Regrese a su libro de trabajo (vía el ícono **Exit Door** en la barra de herramientas). Seleccione **File** – **Save** y luego **File** – **Close**. Su libro de trabajo debería guardarse en ***NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK***.

a. Nótese que **Parented** también aparece en la lista de **Source Variables**. En este ejemplo, de todos modos, debe seleccionar **Parented [6]** en la lista de **Class Variables** porque en un análisis de regresión se la trata como una variable categórica.
b. Para el cálculo, seleccione **Descriptives** en WesVar y mueva **Distance** desde **Source Variables** a **Selected**, de acuerdo con el procedimiento indicado en el ejercicio 3.1.

CORRELACIÓN Y CAUSALIDAD

En la mayoría de las evaluaciones nacionales se obtiene información acerca de un abanico de variables personales y situacionales a través de cuestionarios completados por los estudiantes, los docentes y (a veces) los padres. Las variables generalmente se seleccionan al considerarse que están relacionadas con el rendimiento de los estudiantes (tanto si ello se basa en una investigación, como si no). Los responsables de las políticas también pueden incluir variables para guiar la elección de las intervenciones diseñadas con el fin de mejorar el rendimiento de los estudiantes.

La asociación entre la variable de contexto y el rendimiento del estudiante puede representarse mediante una correlación. Sin embargo, el descubrimiento de que las variables están correlacionadas (incluso fuertemente) en una evaluación nacional, no significa que una variable sea la causa de otra. Hay muchas razones para afirmar esto: En primer lugar, la mayoría de las evaluaciones nacionales son por naturaleza transversales. Los datos relativos a los factores contextuales y al rendimiento se recopilan al mismo tiempo. Por consiguiente, la secuencia temporal entre eventos, en la cual la causa precede al efecto, que es normalmente necesaria para una inferencia de causalidad, no está presente. Este problema puede evitarse cuando algunas de las variables contextuales describen eventos pasados (por ejemplo, la cantidad de tiempo de instrucción empleado en la enseñanza de un área curricular). Este problema también es abordado en los raros casos en los cuales los datos de los estudiantes se recogen en momentos diferentes. Tomar en consideración las características de los estudiantes en un estadio precedente de sus carreras educativas (incluyendo datos de rendimientos previos y de contexto) fortalecerá las inferencias que se puedan hacer respecto de los efectos de sus experiencias escolares (Kellaghan, Greaney y Murray 2009).

Un segundo problema para identificar relaciones causales en una evaluación nacional es que los factores que afectan al rendimiento de un estudiante son complejos y están interrelacionados. En numerosas referencias bibliográficas se señala un amplio abanico de factores asociados con el rendimiento, incluyendo características personales de los estudiantes, factores familiares y comunitarios, así como prácticas

instructivas particulares de los maestros. Los análisis estadísticos en los cuales una variable individual se relaciona con el rendimiento, no solo serán incapaces de reconocer esta complejidad sino que también pueden conducir a una conclusión errada. Un simple ejemplo servirá para ilustrar este punto. El descubrimiento de que los estudiantes de escuelas privadas presentan niveles más altos de rendimiento que los alumnos que asisten a escuelas públicas, puede ser interpretado como que la educación proporcionada en las escuelas privadas es superior a la proporcionada en las escuelas públicas. Sin embargo, tal conclusión debería ser modificada si se incluyeran en los análisis las mediciones de otros factores, tales como el nivel de rendimiento cuando ingresaron a la escuela o sus circunstancias familiares (por ejemplo, el nivel socioeconómico de los padres).

La complejidad de los factores que afectan al rendimiento de los estudiantes se aborda con análisis de regresión múltiple cuando los "efectos" de una variable se aíslan por la eliminación o el ajuste sistemático de los "efectos" de las otras variables. Una inferencia relativa a una causalidad se fortalece si se encuentra que existen relaciones significativas luego de los ajustes.

Una variedad de métodos de análisis de regresión más elaborados (cuya descripción está más allá de la finalidad del presente trabajo), en los cuales se examinan los patrones de correlación entre variables predictoras y los resultados para identificar sendas causales, es útil para reforzar inferencias acerca de la causalidad. Estos métodos están diseñados para identificar variables que pueden ser consideradas como *moderadoras*, que afectan a la dirección y solidez de la relación entre variables de criterio y variables independientes (por ejemplo, género, edad y estatus socioeconómico), y *mediadoras*, que explican cómo o por qué se verifican los efectos (Bullock, Green y Ha 2010).

Sin embargo, incluso los más sofisticados métodos de análisis no abordan un ulterior problema que surge al inferir la causalidad sobre la base de los datos obtenidos en un estudio transversal (como lo es una evaluación nacional). Por ejemplo, puede que no haya datos sobre información importante, a menudo incontrolable y oscura, que probablemente afecte al rendimiento de los estudiantes. Este problema puede abordarse solo a través de diseños de investigación aleatorios y controlados en los cuales la influencia directa de una condición sobre

otra se estudia cuando se han eliminado todas las otras posibles causas de variación.

Aunque es problemático hacer inferencias basadas en datos de las evaluaciones nacionales, la información relacionada con el rendimiento que se obtiene en una evaluación mantiene probablemente un valor pragmático. Los datos recogidos en una evaluación que se refieren, por ejemplo, a género, localización o pertenencia a un grupo étnico, pueden actuar como una señal importante, identificando áreas problemáticas en el sistema educativo que merecen atención al momento de formular políticas, e indicando —posiblemente— medidas de intervención o correctivas. No obstante, la determinación de la naturaleza de tales medidas de intervención o correctivas requiere que se tengan en cuenta las condiciones locales y los recursos, exista un compromiso por parte de los sectores interesados, y que los resultados de las investigaciones se consideren relevantes (véase Kellaghan, Greaney y Murray, 2009).

NOTAS

1. Si una etiqueta de variable (**Implement – Percent Correct**) aparece en la columna de la izquierda en **Simple Scatterplot** en lugar del nombre de la variable (**Impl_pc**), seleccione **Cancel** y luego seleccione **Edit – Options – General – Variable Lists – Display Names – Apply – OK** (al mensaje) **– OK**.

2. El intercambio entre las dos arrojaría diferentes líneas de ajuste optimizado. Esto quiere decir que la línea que mejor predice *y* a partir de *x* no es la misma que la línea que mejor predice *x* a partir de *y*.

3. El término de error o "ruido" refleja que los factores añadidos a los ya incluidos en el modelo de análisis de regresión influencian la variable dependiente.

4. Al preparar la ejecución de un análisis de regresión, el analista debería buscar puntuaciones atípicas o extremas en una variable (véase el capítulo 2) porque pueden distorsionar los resultados. Es también importante determinar si alguna de las variables en la ecuación de regresión está altamente sesgada. Las puntuaciones atípicas o extremas pueden contribuir a distorsionar las estimaciones de parámetros y las estadísticas.

CAPÍTULO 7

PRESENTACIÓN DE DATOS A TRAVÉS DE GRÁFICOS Y DIAGRAMAS

"Una imagen vale más que mil palabras". Es posible que muchos lectores de los informes de la evaluación nacional entiendan más rápidamente los datos si se presentan en forma de gráfico o diagrama que si se presentan con formato de tabla. Los primeros capítulos contenían diagramas de dispersión, gráficos de barra y diagramas de caja producidos por programas informáticos durante varios análisis. (Por ejemplo, el ejercicio 6.1 demostró cómo se construye un diagrama de dispersión con SPSS). Este capítulo proporciona ejemplos sobre un número de propuestas gráficas para la preparación y presentación de los resultados de la evaluación nacional. El énfasis se ha puesto sobre los gráficos y diagramas de línea que se pueden producir en Excel y se pueden construir y pegar en un informe de la evaluación nacional. A medida que se familiarice con estas técnicas, se encontrará otros programas que se pueden utilizar para presentar datos de manera gráfica. No se intenta reemplazar las tablas por los gráficos y los diagramas. Estos exponen los datos de manera diferente. Si están incluidas en un informe, las tablas correspondientes (las tablas de datos en las que se basan los gráficos y diagramas) también deben incluirse, ya sea en el cuerpo del texto o en un apéndice (por ejemplo, en un apéndice electrónico). Los gráficos de la sección siguiente son bidimensionales (2-D).

GRÁFICOS

Un gráfico de columnas tiene barras rectangulares verticales de longitudes generalmente proporcionales a las frecuencias que representan. Por ejemplo, se puede utilizar un gráfico de columnas para comparar la población de cuatro regiones o los puntajes promedio de los estudiantes en distintos tipos de escuela en un área curricular específica. En el ejercicio 7.1, los datos sobre los niveles de competencia en matemáticas de los estudiantes están presentados en forma de gráfico de columnas, en el que la altura de la barra indica el porcentaje de estudiantes en cada nivel. Los datos están basados en una evaluación nacional que notificó resultados usando cinco niveles de competencia (niveles del 1 al 4 y debajo del nivel 1). El procedimiento para crear un gráfico de columnas en Excel también se describe en el ejercicio 7.1.[1]

También se puede informar el porcentaje de estudiantes en cada nivel de competencia para cada región de un país. Esta vez, los datos regionales sobre niveles de competencia estarán representados por un gráfico de barras (véase el ejercicio 7.2), que es muy parecido a un gráfico de columnas, excepto porque representa los datos de manera horizontal.

EJERCICIO 7.1

Dibujar un gráfico de columnas para mostrar el rendimiento por nivel de competencia, datos nacionales

1. Disponga los datos en una hoja de Excel (figura del ejercicio 7.1.A). Copie los datos de una tabla existente, o simplemente teclee los datos en una hoja de Excel.
2. Resalte los datos (figura del ejercicio 7.1.A). En la barra de menú, seleccione **Insert** y **Column** (figura del ejercicio 7.1.B).[a]
3. Seleccione **2-D Column** y **Clustered Column**: la primera columna del gráfico. Para etiquetar el eje vertical (*y*), coloque el cursor sobre la izquierda donde vea los valores (0 a 25) y haga clic derecho. En la barra de herramientas, seleccione **Chart Tools** (parte superior de la pantalla) **– Layout – Axis Titles — Primary Vertical Axis Title – Rotated Title**. Haga clic sobre el cuadro que se encuentra al lado del eje vertical, resalte el texto en el cuadro y escriba ***Porcentaje de estudiantes*** (figura del ejercicio 7.1.C).

EJERCICIO 7.1 ***(continúa)***

FIGURA DEL EJERCICIO 7.1.A Porcentajes de estudiantes en cada franja de rendimiento

Nivel	Porcentaje de estudiantes
Debajo del nivel 1	19,8
Nivel 1	20,4
Nivel 2	23,6
Nivel 3	19,2
Nivel 4	17,1

FIGURA DEL EJERCICIO 7.1.B Insertar opciones de gráfico en Excel

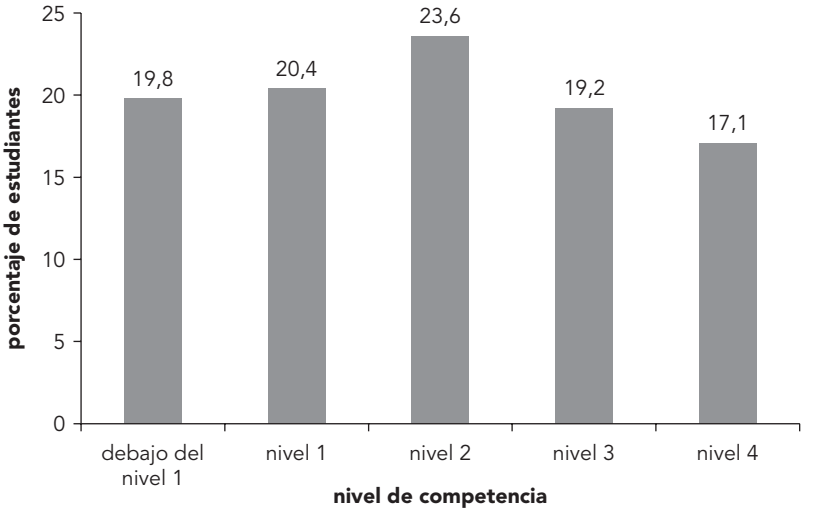
4. Para etiquetar el eje horizontal (*x*), coloque el cursor en la parte inferior del gráfico y haga clic derecho. (Se resaltarán los diferentes niveles de competencia). En la barra de herramientas, seleccione **Chart Tools** (parte superior de la pantalla) **– Layout – Primary Horizontal – Axis Title – Title Below Axis**. Resalte el texto (**Axis Title**) en el cuadro y escriba ***Niveles de competencia*** (figura del ejercicio 7.1.C).
5. Si Excel ha asignado un título de gráfico, puede cambiarlo resaltando el título y modificando lo escrito. Para asignar un título, si no se hubiera proporcionado ninguno, coloque el cursor sobre el gráfico y haga clic para resaltar el gráfico. En la barra de herramientas, seleccione **Chart Tools** (parte superior de la pantalla) **– Layout – Chart Title – Above Chart**. Seleccione el texto y escriba ***Percentage of Students at Each Mathematics Proficiency Level (Porcentaje de estudiantes en cada nivel de competencia en matemáticas)***. El título se ha colocado sobre el gráfico en la figura del ejercicio 7.1.C.
6. Puede modificar varias características del gráfico (como el tamaño y estilo de fuente) haciendo clic en el área apropiada del gráfico y utilizando la herramienta de edición (clic derecho) o utilizando **Design** y **Layout** (debajo de **Chart Tools**). Si quiere mostrar

(continúa)

EJERCICIO 7.1 *(continúa)*

los valores (porcentajes) en cada columna, haga clic derecho en una de las columnas (esta acción debería resaltarlas a todas). Luego seleccione **Chart Tools – Layout – Data Labels – Outside End**. Esto muestra el porcentaje de estudiantes en cada nivel de competencia.

FIGURA DEL EJERCICIO 7.1.C Porcentaje de estudiantes en cada nivel de competencia en matemáticas



7. Resalte el área del gráfico en Excel haciendo clic en el perímetro y seleccione **Copy**. Pegue el gráfico en su informe.
8. Guarde su hoja de trabajo de Excel en ***MY SOLUTIONS***, utilizando el nombre de archivo ***EXERCISE 7.1.XLS***.

a. Las versiones anteriores de Excel pueden llevarlo al Asistente para gráficos con el que también podrá seleccionar una columna del gráfico.

EJERCICIO 7.2

Dibujar un gráfico de barras para mostrar el porcentaje de cada nivel de competencia por región

1. Copie los datos de la figura del ejercicio 7.2.A a un archivo de Excel.
2. Resalte los datos, como se muestra en la figura del ejercicio 7.2.A. Haga clic en **Insert** (o seleccione **Insert – Charts**) **– Bar – 2-D Bar – 100% Stacked Bar** (tercera opción)

EJERCICIO 7.2 *(continúa)*

(figura del ejercicio 7.2.B). Este gráfico le permite comparar los porcentajes de estudiantes a través de los niveles de competencia y las regiones. Puede modificar varias características de este gráfico para hacerlo más presentable.

FIGURA DEL EJERCICIO 7.2.A Porcentaje de estudiantes en cada nivel de competencia en matemáticas por región

	Debajo del 1	Nivel 1	Nivel 2	Nivel 3	Nivel 4
Noroeste	27,8	23,8	22,9	16,8	8,7
Área Metropolitana	13,5	13,9	16,9	23,8	31,9
Región Montañosa	17,3	21,6	23,1	25,9	12,1
Costa del Sudoeste	18	16,9	24,7	30,5	9,9

FIGURA DEL EJERCICIO 7.2.B Opciones de gráfico de barras 2-D en Excel

- Si el resultado que obtiene (véase la figura del ejercicio 7.2.C) presenta las regiones en el orden incorrecto (por ejemplo, Costa del Sudoeste en la parte superior en lugar de Noroeste), haga clic en el gráfico y luego clic en las etiquetas del eje vertical (regiones). Seleccione **Chart Tools – Format – Current Selection – Format Selection – Axis Options.** Marque el cuadro **Categories in Reverse Order – Close.**
- Si el nivel aparece en el eje *y* y las regiones en el eje *x* (esto es, al revés de como se muestra en la figura del ejercicio 7.2.C), vuelva a Excel. Haga clic derecho en el área del gráfico. Seleccione **Data**. Haga clic en **Switch Row/Column** (figura del ejercicio 7.2.D).

  Se puede, si así lo desea, indicar el porcentaje de estudiantes en cada nivel de competencia en cada región. Haga clic en la barra Below Level 1 [Debajo del nivel 1] en la primera región. Esto resaltará las barras de Debajo del nivel 1 en todas las regiones. Haga clic derecho y seleccione **Add Data Labels** (o seleccione **Layout – Labels – Data Labels – Center**). Repita el proceso para los otros niveles (figura del ejercicio 7.2.E).

(continúa)

EJERCICIO 7.2 *(continúa)*

FIGURA DEL EJERCICIO 7.2.C Porcentaje de estudiantes en cada nivel de competencia en matemáticas por región

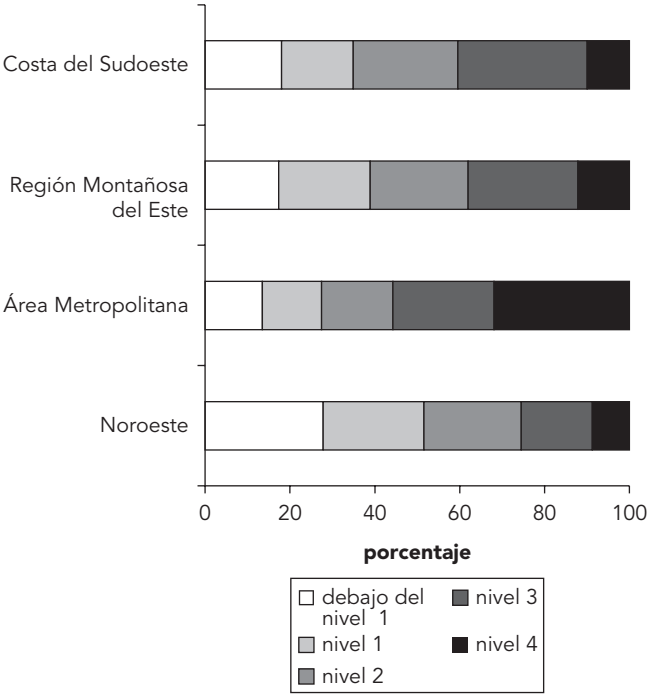
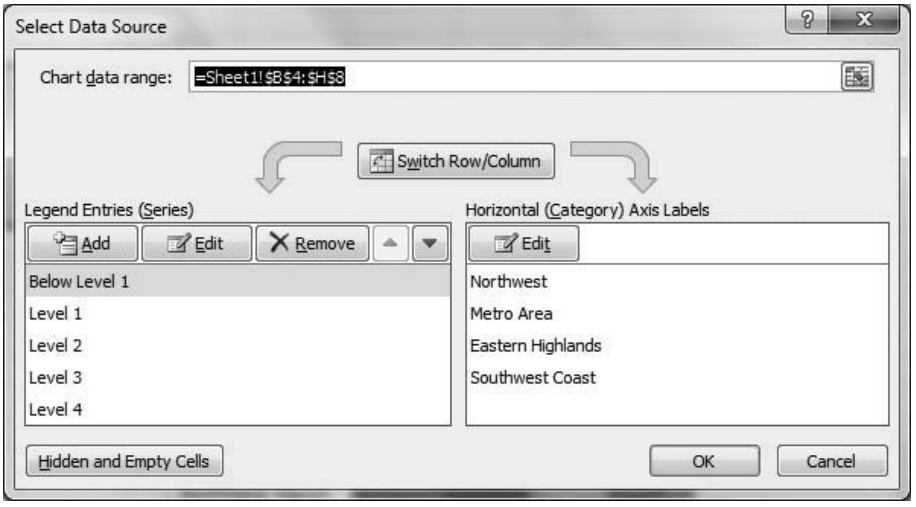
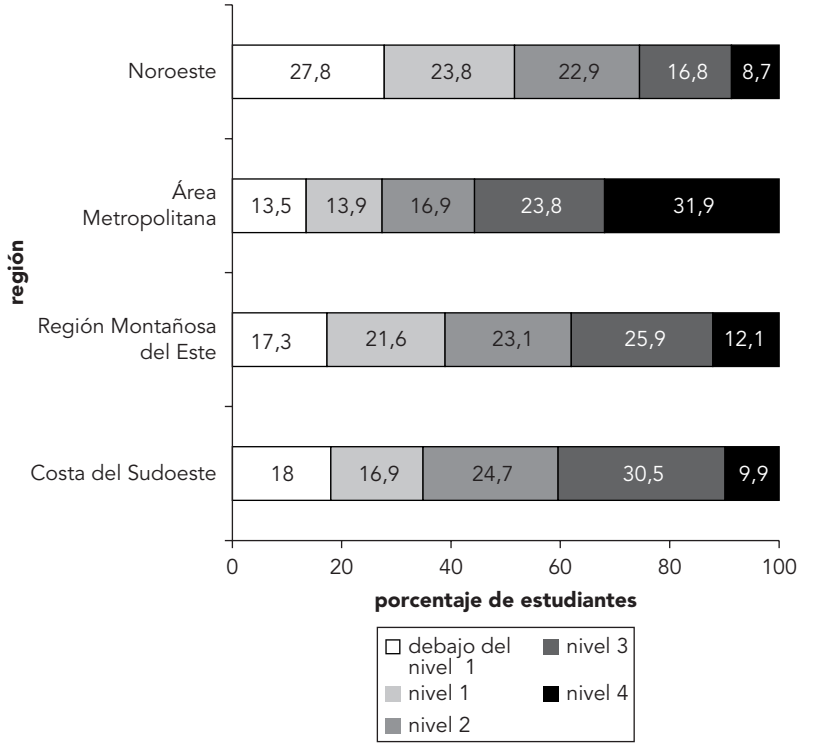


FIGURA DEL EJERCICIO 7.2.D Opción Switch Row/Column (Intercambiar filas/columnas) en Chart Tools/Design (Herramientas para gráficos/Diseño) de Excel



EJERCICIO 7.2 *(continúa)*

FIGURA DEL EJERCICIO 7.2.E Porcentaje de estudiantes en cada nivel de competencia en matemáticas por región



- Si el eje horizontal (con las figuras de porcentaje) aparece en la parte superior del gráfico en lugar de en la parte inferior, seleccione el eje horizontal (el eje con los porcentajes) haciendo clic derecho en cualquier etiqueta. Luego haga clic derecho en las etiquetas Hi o Lo del eje (según la posición actual de las figuras de porcentaje) y seleccione **Format Axis – Horizontal Axis Crosses – Automatic**.
- Si las regiones están en el orden inverso, seleccione el eje vertical haciendo clic izquierdo en cualquiera de las etiquetas. Luego haga clic derecho en **Format Axis** y marque **Categories in Reverse Order**.
- Si desea agregar un título a los ejes *x* o *y*, resalte el área del gráfico en la que aparecen los porcentajes. Luego seleccione **Insert – Layout – Axis Titles – Primary Horizontal Axis Title – Title Below Axis**. Resalte **Axis Title** y escriba el nuevo título (por ejemplo, ***Porcentaje de estudiantes***) en la barra de fórmula o directamente en el cuadro de texto recientemente creado.[a] Siga el mismo procedimiento para agregar un título al otro eje. Para agregar un título al gráfico, siga las instrucciones en el paso 5 del ejercicio 7.1.

(continúa)

EJERCICIO 7.2 *(continúa)*

3. Resalte el área del gráfico en Excel haciendo clic en el perímetro; seleccione **Home – Copy**. Luego pegue el gráfico en su informe y otórguele un título adecuado.
4. Guarde su hoja de trabajo de Excel (***NAEA DATA ANALYSIS\MY SOLUTIONS\ EJERCICIO 7.2.XLS***).
5. Seleccione **File – Save** y **File – Close.**

a. Algunas versiones recientes de Excel implementan **Chart Layout** seguido de **Axis Title** y **Chart Title** para el etiquetado.

GRÁFICOS DE LÍNEAS CON INTERVALOS DE CONFIANZA

Un gráfico que muestra una serie de puntajes promedio, junto con sus intervalos de confianza del 95 por ciento, permite que el lector no técnico sepa cuáles de las regiones tuvieron un buen rendimiento, en comparación con las otras, y también si un puntaje promedio y su intervalo de confianza de una región se superpone al puntaje promedio e intervalo de confianza de otra región.

EJERCICIO 7.3

Dibujar intervalos de confianza del 95 por ciento para una serie de puntajes promedio

1. Abra Excel. Ingrese los datos manualmente en la hoja de Excel; distribúyalos como se muestra en la figura del ejercicio 7.3.A[a] Otra posibilidad consiste en transferir estos datos al archivo de Excel desde ***NAEA DATA ANALYSIS\MY SOLUTIONS\ EXERCISE 3.3.TXT***. Esto requerirá que elimine algunos datos no necesarios (como errores típicos y los que brindan datos de porcentaje) de la hoja de Excel.
2. Resalte los datos, como se indica en la figura del ejercicio 7.3.A. Seleccione **Insert – Other Charts** (o **Insert Chart**) **– Stock – High-Low-Close** (primera opción).
3. Resalte los números del eje vertical. Haga clic derecho, seleccione **Format – Axis**. Luego, bajo **Axis Options**, marque **Fixed** para **Minimum** (o **Scale**) e inserte 220. De la misma manera, establezca **Maximum** a 280 (figura del ejercicio 7.3.B). Haga **clic** (u **OK**). Tenga en cuenta que estos valores están justo por debajo y por encima de los valores más altos en la figura del ejercicio 7.3.A.

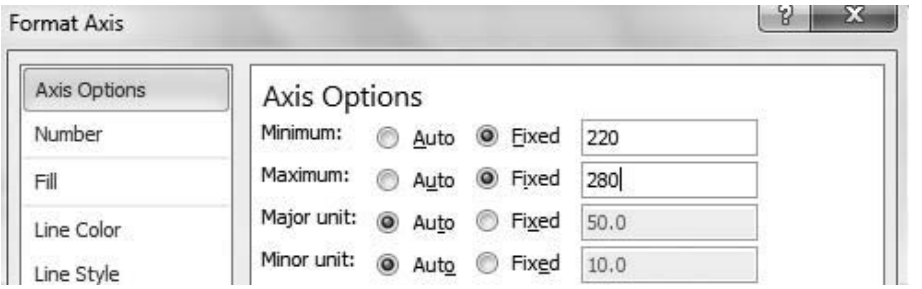
EJERCICIO 7.3 *(continúa)*

FIGURA DEL EJERCICIO 7.3.A Puntajes promedio de matemáticas y puntajes en los intervalos de confianza superiores e inferiores por región

Región	IC superior al 95%	Pro	IC inferior al 95%
Noroeste	239,9	233,3	226,8
Área Metropolitana	274,7	265,7	256,8
Región Montañosa del Este	256,3	249,1	241,9
Southwest coast	257,9	251,2	244,6

Nota: IC = Intervalo de confianza

FIGURA DEL EJERCICIO 7.3.B Opciones de formato de eje en Excel



4. En el gráfico de Excel, haga clic izquierdo en el límite inferior de la línea de la primera región (Noroeste). Esta acción debería resaltar el límite inferior de las cuatro regiones. En la barra de menú, seleccione Format Selection – Marker Options (**o** Marker Style) – Built In – Type. Seleccione una opción adecuada (como "**–**") de la selección debajo de **Type** (figura del ejercicio 7.3.C) (o use la ventana del menú desplegable **Style**). Aumente el tamaño con **Size** a 10 o mayor, asegurándose de que el tamaño sea lo suficientemente grande para destacar dentro del gráfico. Haga clic en **Close**. Repita el proceso para los límites superiores.
5. Resalte el punto central de cada línea (el puntaje promedio) haciendo clic en la línea de la primera región y repita el mismo procedimiento que en el paso 4 para seleccionar un marcador adecuado. Utilice el tamaño 20 para diferenciar el marcador de puntaje promedio del resto de los marcadores. Tenga en cuenta que la leyenda en el lado izquierdo del gráfico se produce automáticamente, basándose en su hoja de trabajo inicial. Tiene que etiquetar los ejes y los cuadros de series en el lado derecho además de decidir un título para el gráfico.

   Copie el gráfico (figura del ejercicio 7.3.D) en su informe, usando **Copy** y **Paste**.

(continúa)

EJERCICIO 7.3 *(continúa)*

FIGURA DEL EJERCICIO 7.3.C Opciones de formato de series de datos en Excel

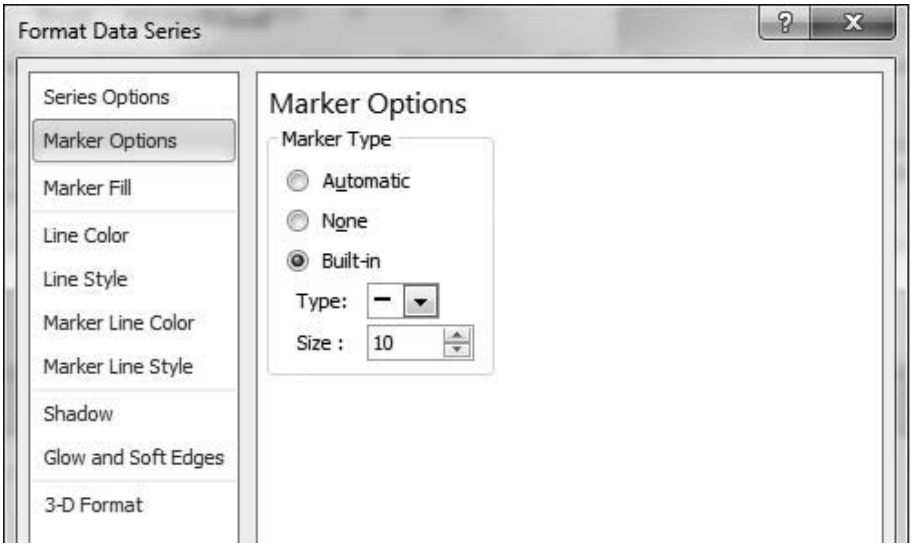
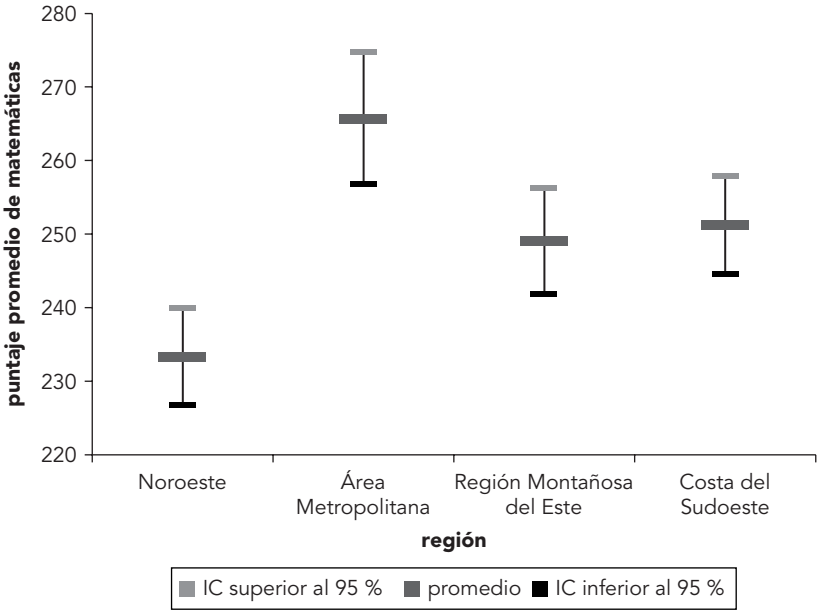


FIGURA DEL EJERCICIO 7.3.D Gráfico de línea para puntajes promedio de matemáticas e intervalos de confianza del 95 por ciento por región



Nota: IC = Intervalo de confianza

EJERCICIO 7.3 *(continúa)*

6. Guarde la hoja de Excel en ***NAEA DATA ANALYSIS\MY SOLUTIONS*** utilizando un nombre adecuado (como ***EXERCISE 7.3.XLS***).

El gráfico en la figura del ejercicio 7.3.D describe las diferencias en el rendimiento promedio entre regiones. Si no existen superposiciones entre los intervalos de confianza, se puede decir que los puntajes promedio son significativamente diferentes. Así, no existe superposición entre el intervalo de confianza del 95 por ciento ni para el Noroeste ni para ninguna de las otras regiones. Los puntajes promedio de matemáticas de las Región Montañosa del Este y la Costa del Sudoeste no difieren significativamente uno del otro (porque existe una superposición importante entre sus líneas respectivas).

a. Como se informa en el capítulo 3, WesVar utiliza 2,0 en lugar de 1,96 para computar los intervalos de confianza del 95 por ciento.

GRÁFICOS DE LÍNEAS PARA REPRESENTAR DATOS SOBRE TENDENCIAS

El ejercicio 7.4 ilustra cómo un gráfico de líneas se puede utilizar para presentar datos sobre tendencias. El gráfico resume el rendimiento de los niños y las niñas para cada uno de los cuatro años en los que se administró una evaluación nacional (2004, 2007, 2010 y 2013).

EJERCICIO 7.4

Mostrar datos de tendencias con un gráfico de líneas

1. Abra Excel. Disponga los datos en una hoja de Excel (figura del ejercicio 7.4.A).[a]
2. Resalte los datos y etiquetas para incluirlos en el gráfico (figura del ejercicio 7.4.A). En la barra de herramientas, seleccione **Insert – Line – Line with Markers**[b] (primera opción, segunda fila) para crear el gráfico de líneas que se muestra en la figura del ejercicio 7.4.B. Puede modificar el estilo de las líneas haciendo clic sobre alguna de ellas. Seleccione **Format – Data Series** (o debajo de **Current Selection – Format Selection**) **– Line Style.** Utilice la flecha hacia abajo **Dash Type** para seleccionar el estilo de línea que desee. Haga clic en **Close.**
3. Puede agregar títulos a los ejes *x* e *y*- y al gráfico en su conjunto siguiendo los pasos del ejercicio 7.1. Para agregar un título al eje *y*-, haga clic en la escala adyacente (210 a 255). Seleccione **Chart Tools – Layout** (o **Chart Layout**) **– Axis Titles** (o **Labels – Axis Titles**) **– Primary Vertical Axis Title – Rotated Title**. Haga clic en el espacio para el título de eje y escriba ***Puntaje promedio de matemáticas***.

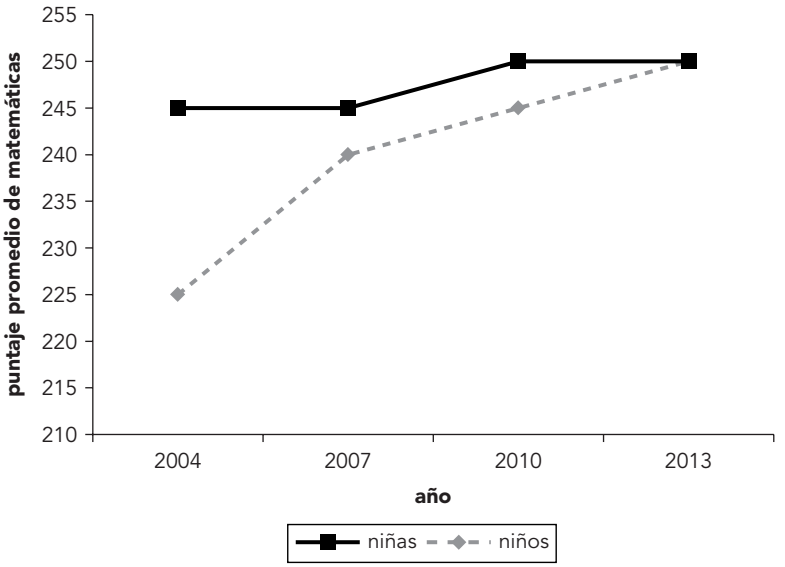
(continúa)

EJERCICIO 7.4 *(continúa)*

FIGURA DEL EJERCICIO 7.4.A Hoja de trabajo de Excel con puntajes promedio de matemáticas por género, 2004–13

	Niños	Niñas
2004	225	245
2007	240	245
2010	245	250
2013	250	250

FIGURA DEL EJERCICIO 7.4.B Puntajes promedio de matemáticas por género, 2004–13



4. De manera similar, etiquete el eje horizontal *x*- ***Year***.
5. Utilice el ícono **Chart Title** para crear un título para la etiqueta (como ***Puntajes promedio de matemáticas por género, 2004–13***).
6. Copie el gráfico (figura del ejercicio 7.4.B) en su informe. Luego guarde la hoja de Excel utilizando un nombre adecuado (***NAEA DATA ANALYSIS\MY SOLUTIONS\ EXERCISE 7.4.XLS***).
7. En **Excel**, seleccione **File – Close**.

a. Tenga en cuenta que los datos de la figura del ejercicio 7.4.A no están tomados del archivo de datos ***NATASSESS4.VAR.***
b. Esto es **Marked Line** en algunas versiones de Excel.
c. Algunas versiones de Excel pueden requerir: **Format – Data Series – Line – Weights & Arrows – Dashed.** Utilice la flecha descendente para seleccionar el estilo de línea deseado. Haga clic en **OK**.

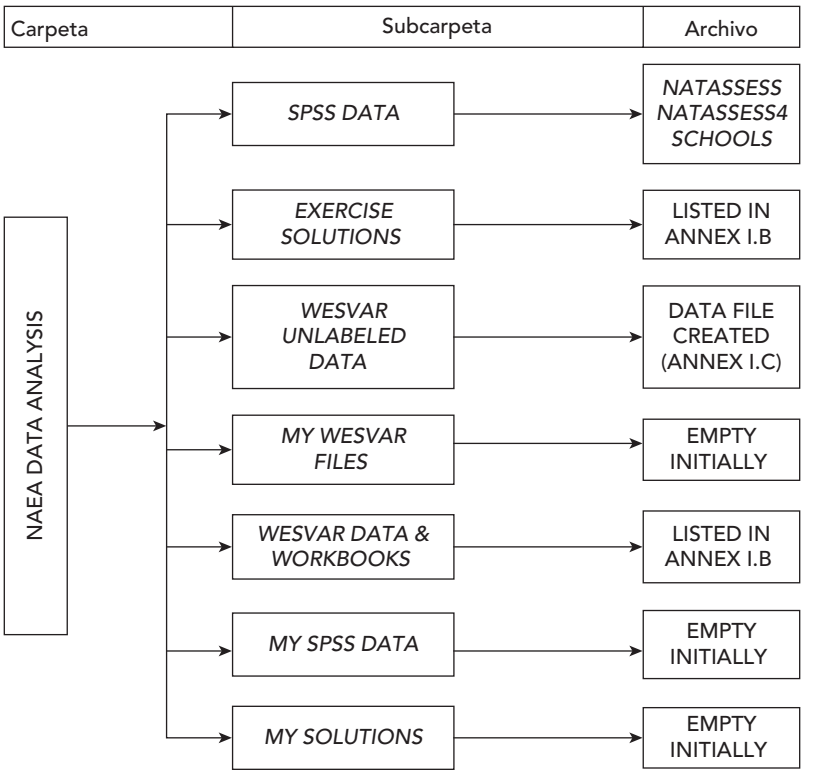
NOTA

1. Las instrucciones pueden variar ligeramente según la versión de Excel que se utilice. Excel para Mac y Excel para Windows tienen pequeñas diferencias de funciones, según la plataforma y la versión que se utilicen.

ANEXO

I.A

ANÁLISIS DE DATOS DE LA EVALUACIÓN NACIONAL DEL RENDIMIENTO ACADÉMICO (NAEA): ESTRUCTURA DEL DIRECTORIO DE ARCHIVOS



ANEXO

I.B

ANÁLISIS DE DATOS DE LA EVALUACIÓN NACIONAL DEL RENDIMIENTO ACADÉMICO: SUBCARPETAS Y ARCHIVOS

Subcarpetas	Archivos	Descripción
SPSS DATA [datos en SPSS]	***NATASSESS.SAV***	El archivo incluye los identificadores, las variables contextuales, los puntajes, la participación y el coeficiente de ponderación de diseño de cada alumno seleccionado. También, la clase, la escuela y los identificadores regionales.
	NATASSESS4.SAV	El archivo incluye los identificadores, las variables contextuales, los puntajes, la participación y el coeficiente de ponderación de diseño de cada alumno seleccionado. También, la clase, la escuela y los identificadores regionales. Asimismo, se incluyen estratos de *jackknife* (JK) (zonas) y réplicas de JK (indicadores).
	SCHOOLS.SAV	El archivo incluye los números de identificación de la escuela, estratos de JK (zonas) y réplicas de JK (indicadores). Se funde con ***NATASSESS.SAV*** para crear ***NATASSESS4.SAV***.
EXERCISE SOLUTIONS [soluciones de los ejercicios]	***EXERCISE 1.1.SPV***	Estadísticas descriptivas en SPSS para **Mathss** (media, desviación estándar) con Descriptives.

(continúa)

Subcarpetas	Archivos	Descripción
	EXERCISE 2.1.SPV	Estadísticas descriptivas en SPSS para **Mathss** (varias) con Explore (nacional).
	EXERCISE 2.2.SPV	Estadísticas descriptivas en SPSS para **Mathss** (varias) con Explore (por región).
	EXERCISE 3.1.TXT	Estadísticas descriptivas WesVar de **Mathss**
	EXERCISE 3.2.TXT	Puntaje medio, desviación estándar e intervalo de confianza de **Mathss** (nacionales).
	EXERCISE 3.3.TXT	**Puntaje medio, desviación estándar e intervalo de confianza de Mathss (por región).**
	EXERCISE 4.1A.TXT	**Puntaje medio de los alumnos que tienen y que no tienen electricidad en sus hogares de Mathss.**
	EXERCISE 4.1B.TXT	**Puntaje medio e intervalo de confianza de los alumnos que tienen y que no tienen electricidad en sus hogares (dos grupos de comparación) de Mathss.**
	EXERCISE 4.2A.TXT	**Puntaje medio de Mathss (por región).**
	EXERCISE 4.2B.TXT	**Valor estadístico de las diferencias entre los puntajes medios regionales de Mathss (comparación de más de dos grupos)**
	EXERCISE 5.1.TXT	**Puntajes correspondientes al percentil de Mathss (nacional).**
	EXERCISE 5.2.TXT	**Puntajes correspondientes al percentil de Mathss (por región).**
	EXERCISE 5.4-25TH.TXT	**Porcentaje de alumnos por debajo del percentil 25 de Mathss (errores estándar [ES] e intervalos de confianza [IC]) (por región).**
	EXERCISE 5.4-50TH.TXT	**Porcentaje de alumnos por debajo del percentil 50 de Mathss (ES e IC) (por región).**
	EXERCISE 5.4-75TH.TXT	**Porcentaje de alumnos por debajo del percentil 75 de Mathss (ES e IC) (por región).**
	EXERCISE 6.1.SPV	**Diagrama de dispersión de los puntajes de implementación y análisis/resolución de problemas en matemáticas.**
	EXERCISE 6.2.TXT	**Correlación entre implementación y análisis/resolución de problemas en matemáticas.**

Subcarpetas	Archivos	Descripción
	EXERCISE 6.3 SUM OF SQUARES.TXT	**Suma de cuadrados de la regresión y valor de R cuadrado, una variable independiente (Books)**
	EXERCISE 6.3 ESTIMATES.TXT	**Coeficiente de regresión estimado, una variable independiente (Books).**
	EXERCISE 6.3 TESTS.TXT	**Prueba de significancia de la regresión, una variable independiente (Books).**
	EXERCISE 6.4 ESTIMATES.TXT	**Coeficientes de regresión estimados, una variable independiente (Books).**
	EXERCISE 6.5 CORRELATIONS.TXT	**Correlaciones entre tres variables (Books, Parented, Distance).**
	EXERCISE 6.6 SUM OF SQUARES .TXT	**Suma de cuadrados de la regresión y valor de R cuadrado (para tres variables independientes).**
	EXERCISE 6.6 ESTIMATES.TXT	**Coeficientes de regresión estimados (para tres variables independientes).**
	EXERCISE 6.6 TESTS.TXT	**Pruebas de significancia de la regresión (para tres variables independientes).**
	EXERCISE 7.1.XLS	**Cuadro: Porcentaje de alumnos en cada nivel de competencia en matemáticas (nacional).**
	EXERCISE 7.2.XLS	**Cuadro: Porcentaje de alumnos en cada nivel de competencia en matemáticas (por región).**
	EXERCISE 7.3.XLS	**Gráfico de líneas: Puntajes medios e intervalos de confianza de 95 % de Mathss (por región).**
	EXERCISE 7.4.XLS	**Gráfico de líneas: Puntajes medios de Mathss (por género).**
WESVAR UNLABELED DATA [datos de WesVar sin nombre]	***NATASSESS4.VAR (NATASSESS4.LOG)***	**El archivo incluye los identificadores, las variables contextuales, los puntajes, la participación y la ponderación de diseño de cada alumno seleccionado. También la clase, la escuela y los identificadores regionales. Asimismo, incluye estratos de JK (zonas) y réplicas de JK (indicadores). No incluye los nombres de las variables y las variables creadas.**

(continúa)

Subcarpetas	Archivos	Descripción
MY WESVAR FILES [mis archivos de WesVar]	***La subcarpeta inicialmente está vacía. El usuario guarda en esta subcarpeta los libros y los archivos de datos de WesVar que crea cuando completa los ejercicios en los capítulos 3–6.***	
WESVAR DATA & WORKBOOKS [datos y libros de trabajo de WesVar]	***NATASSESS4.VAR (NATASSESS4.LOG)***	**El archivo de datos incluye los identificadores, las variables contextuales, los puntajes, la participación y la ponderación de diseño de cada alumno seleccionado. También la clase, la escuela y los identificadores regionales. Asimismo, incluye estratos de JK (zonas) y réplicas de JK (indicadores). También contiene los nombres de las variables y las variables creadas.**
	CHAPTER 3 WORKBOOK.WVB	**Los archivos son libros de trabajo completos para ejecutar los ejercicios de los capítulos 3-6.**
	CHAPTER 4 WORKBOOK.WVB	
	CHAPTER 5 WORKBOOK.WVB	
	CHAPTER 6 WORKBOOK.WVB	
MY SPSS DATA [mis datos de SPSS]	***La subcarpeta inicialmente está vacía. El usuario guarda archivos de datos de SPSS nuevos o modificados en esta subcarpeta.***	
MY SOLUTIONS [mis soluciones]	***La subcarpeta inicialmente está vacía. El usuario guarda todas las soluciones de los ejercicios en esta subcarpeta.***	

ANEXO

I.C

ABRIR UN ARCHIVO DE SPSS EN WESVAR

Una vez que se hayan preparado los datos para la evaluación nacional en SPSS con la prueba, los datos del cuestionario y las ponderaciones de la encuesta calculadas, deberán realizarse algunas modificaciones de los datos antes de transferir el archivo a WesVar.

Las siguientes instrucciones lo guiarán durante la creación de las ponderaciones con el método *jackknife* para el diseño de una encuesta de dos etapas. Podrá encontrar más información sobre el remuestreo con el método *jackknife* en el volumen 3 de esta serie (Dumais y Gough 2012b, anexos IV.C y IV.D).

SPSS se utiliza para crear la información de muestreo que necesita WesVar para generar las ponderaciones para el análisis de los datos de la evaluación nacional. Dado que las ponderaciones replicadas se generan para escuelas, se debe crear un registro para cada escuela que participe. Este registro debe incluir el ***Schoolid*** real de cada escuela. Si no cuenta con un archivo que contenga un número de identificación para cada escuela, debe utilizar el siguiente procedimiento para crear uno.

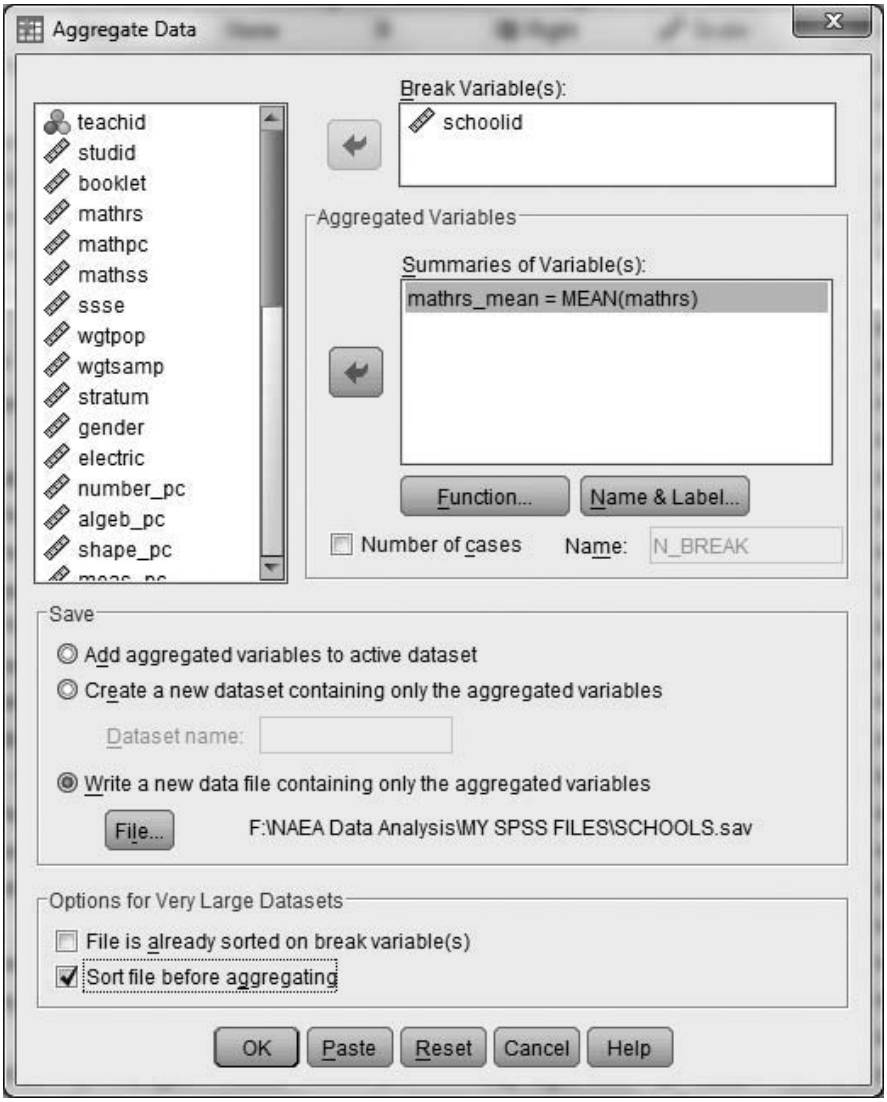
Si aún no ha transferido el archivo ***NATASSESS.SAV*** a ***MY SPSS DATA***, debe copiarlo de ***SPSS DATA*** a ***MY SPSS DATA***.

(Es importante destacar que se trata del archivo ***NATASSESS.SAV*** y no del archivo ***NATASSESS4.SAV***.)

1. Busque ***NAEADATAANALYSIS\MYSPSSDATA*** y abra ***NATASSESS.SAV*** haciendo doble clic sobre el nombre del archivo. Si el archivo no se encuentra en ***MY SPSS DATA***, ábralo en ***NAEA DATA ANALYSIS\SPSS DATA*** y guárdelo en ***NAEA DATA ANALYSIS\MYSPSS DATA*** utilizando **File – Save As**.
2. En la barra de herramientas de SPSS, seleccione **Data – Aggregate**. Mueva ***Schoolid*** a **Break Variable(s).** Mueva **MATHRS** al casillero **Aggregated Variables** (Figura I.C.1). En **Save**, seleccione **Write a new data file using only the aggregated variables**. Seleccione **File** (justo debajo). Seleccione ***NAEA DATA ANALYSIS\MY SPSS DATA*** como el directorio donde desea guardar el nuevo archivo, utilizando el nombre ***SCHOOLS***.
3. En **Options for Very Large Datasets**, marque **Sort file before aggregating**. Haga clic en **OK**.
4. Busque ***NAEA DATA ANALYSIS\MY SPSS DATA*** y abra **Schools**. Abra **Variable View** en la parte inferior y elimine la fila **Mathrs_Mean** con **Edit – Clear**. Seleccione **Data – Sort Cases**, y mueva **Schoolid** al casillero **Sort by**; seleccione **Ascending** y haga clic en **OK**. (Si necesita partir desde la pantalla de resultados en SPSS, seleccione Window en la barra de herramientas y luego los datos correspondientes).
5. En este paso, se asignan los valores 1 o 2 a cada escuela participante.[1] Seleccione **Transform – Compute Variable**, y escriba ***Jkzone*** en **Target Variable**. Luego escriba ***RND($Casenum/2)*** en **Numeric Expression**, y haga clic en **OK**. Luego, seleccione **Transform – Compute Variable**. Seleccione **Reset**. Luego escriba ***Randompick*** en **Target Variable** y ***RV.Uniform(0,1)*** en **Numeric Expression**. Haga clic en **OK**. Ahora, en **Data View,** puede observar 120 escuelas en 60 pares numerados del 1 al 60, y cada escuela debe tener asignado un número aleatorio entre 0 y 1. Si los números aleatorios se muestran como ceros y unos, aumente el número de decimales desde la celda **Variable View**. Ahora ya puede comenzar a crear las réplicas **JK**.
6. Los pasos 5 y 6 asignan uno de dos valores a cada escuela dentro de su par (zona). Seleccione **Data – Sort Cases**, luego mueva

FIGURA I.C.1

Agregar datos en SPSS



Jkzone y **Randompick** a **Sort by**. Haga clic en **Ascending** y en **OK**. Ahora, seleccione **Data – Identify Duplicate Cases**, y mueva **Jkzone** a **Define matching cases by**. En **Variables to Create**, marque **Indicator of Primary Cases** y **Last case in each group is primary (PrimaryLast)**. Haga clic en **OK**.

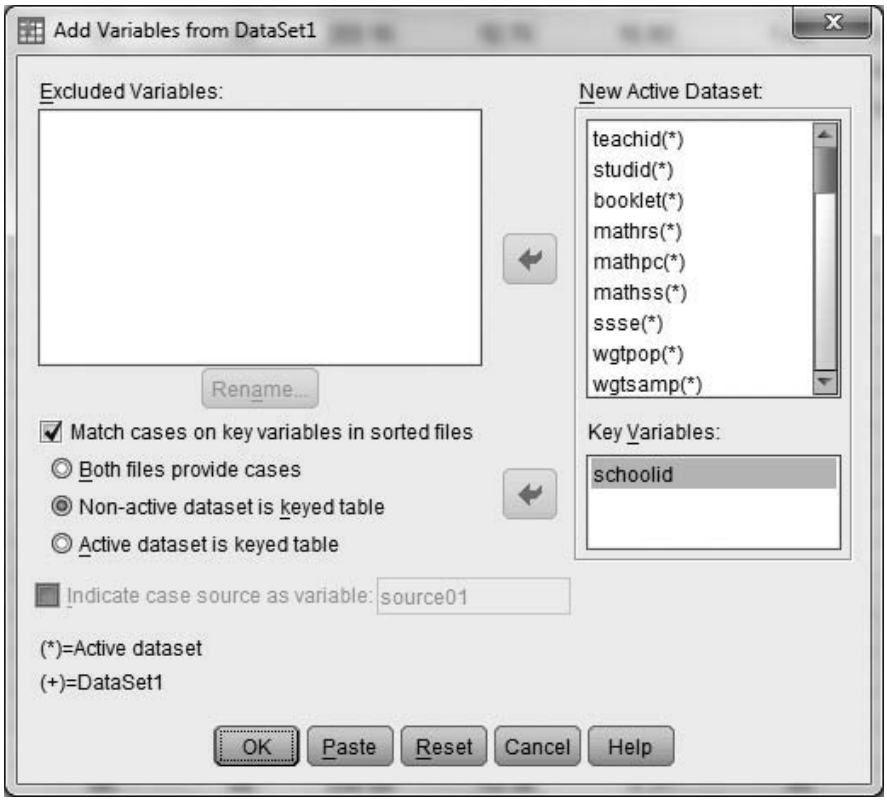
7. Dado que en WesVar las réplicas deben numerarse a partir del 1, no del 0, los códigos de las réplicas necesitan ser modificados con **Transform – Recode into Different Variables**. Mueva **PrimaryLast** a **Input Variable,** y escriba ***Jkindic*** en **Output Variable Name**.[2] Si lo desea, puede ingresar un nombre. Haga clic en **Change** y luego en **Old and New Values**. En **Old Value**, haga clic en **Value** y escriba ***0***. En **New Value** (a la derecha), escriba ***1*** y haga clic en **Add**. En **Old Value**, haga clic en **All other values**. Ahora, en **New Value**, escriba ***2***. Haga clic en **Add**, **Continue** y **OK**. Todos los valores 0 de **PrimaryLast** han sido transformados a valores ***JKINDIC 1***, y todos los valores ***1*** se han convertido en ***2***. (Es importante señalar que el hecho de que los valores 0 y 1 para la variable **PrimaryLast** aparezcan o no en la pantalla depende de su configuración de datos en SPSS). Seleccione **Data – Sort Cases** en el menú. Seleccione **Reset.** Mueva **Schoolid** al casillero **Sort by**; haga clic en **Ascending** y en **OK**.
8. Abra **Variable View** y elimine las variables **RANDOM PICK** y **PrimaryLast**. Seleccione **File – Save** para volver a guardar el archivo de datos ***SCHOOLS***. No cierre el archivo. Ahora, los números (replicados) correspondientes a la zona JK y al indicador JK han sido asignados a las escuelas participantes. Esta información debe ser enviada ahora al archivo ***NATASSESS.SPV***, donde se han guardado los datos y las ponderaciones de la encuesta.
9. Abra el archivo ***NAEA DATA ANALYSIS\MY SPSS DATA\ NATASSESS.SAV***. Seleccione **Data, Sort Cases**, y mueva **Schoolid** al casillero **Sort by**. Seleccione **Ascending** y **OK**.
10. Luego, las variables **Jkzone** y **Jkindic** (que actualmente se encuentran en el archivo ***SCHOOLS.SAV***) se funden con el archivo de datos ***NATASSESS.SAV***. Aún dentro de ***NATASSESS.SAV***, seleccione **Data – Merge files – Add variables**. Seleccione ***SCHOOLS. SAV*** de **Open Dataset**, y haga clic en **Continue**. Marque el casillero próximo a **Match cases on key variables in sorted files**. Seleccione ***SCHOOLID*** y muévalo a **Key Variables** (Figura I.C. 2). Seleccione **Non-active dataset is keyed table**.
11. Luego, las variables **Jkzone** y **Jkindic** (que actualmente se encuentran en el archivo ***SCHOOLS.SAV***) se funden con el archivo de datos ***NATASSESS.SAV***. Aún dentro de ***NATASSESS.SAV***,

seleccione **Data – Merge files – Add variables**. Seleccione ***SCHOOLS.SAV*** de **Open Dataset**, y haga clic en **Continue**. Marque el casillero próximo a **Match cases on key variables in sorted files**. Seleccione ***SCHOOLID*** y muévalo a **Key Variables** (Figura I.C. 2). Seleccione **Non-active dataset is keyed table**.

12. Cierre los archivos abiertos ***SCHOOLS.SAV*** y ***NATASSESS4.SAV***.
13. Ahora ***NATASSESS4.SAV*** se ha incorporado a WesVar y se ha transformado en un archivo de datos de WesVar. Abra WesVar. Haga clic en **New WesVar Data File**. En **Input Database**, seleccione ***NAEA DATA ANALYSIS\MY SPSS DATA\NATASSESS4.SAV***. Haga clic en **Open**. Aparecerá el mensaje: *Create extra formatted variables.* Haga clic en **Done**.

FIGURA I.C.2

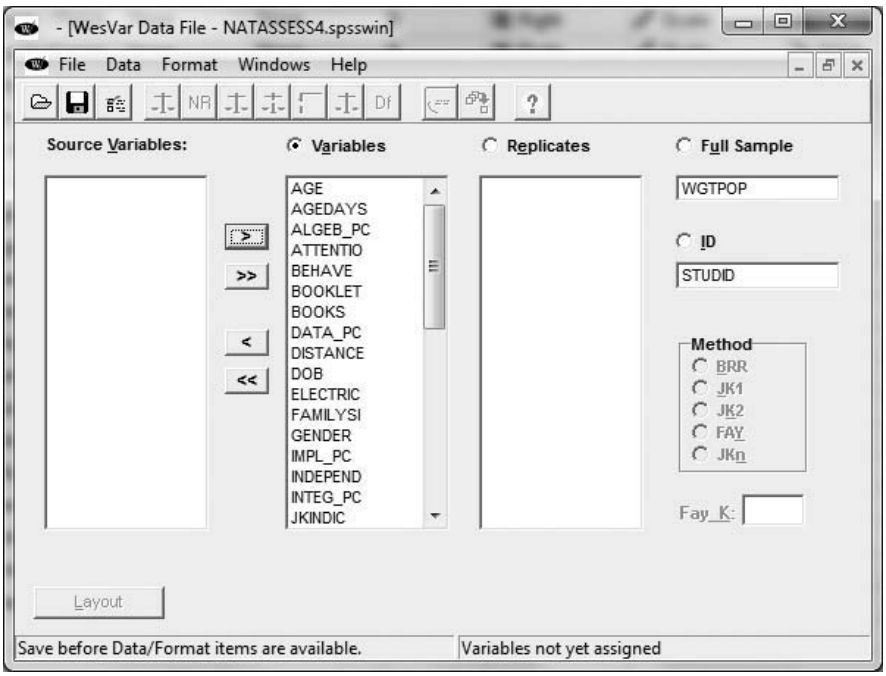
Agregar variables a un archivo de SPSS



14. En **Source Variables**, seleccione **Wgtpop** (la variable que pondera los datos según sus representaciones en la población). Seleccione **Full Sample**. Mueva **Wgtpop** a **Full Sample** con la flecha **>**. Luego seleccione **Studid** en **Source Variables**. Haga clic en **ID** (a la derecha) y mueva **Studid** a **ID**, nuevamente con la flecha **>**. Ahora, seleccione las variables restantes en **Source Variables**. Seleccione **Variables**, y haga clic sobre las flechas **>>** para mover las variables restantes al casillero **Variables** (Figura I.C.3).
15. Guarde su archivo de datos (**File – Save As**) en ***NAEA DATA ANALYSIS\MYWESVARFILES*** con el nombre de archivo ***NATASSESS4.VAR***. (Puede utilizar el mismo nombre porque el formato y la extensión son específicos de WesVar; no se confundirán con los originales de SPSS.)
16. Antes de realizar cálculos con las tablas, WesVar debe crear réplicas de las ponderaciones para estimar los errores de muestreo.

FIGURA I.C.3

Lista de variables en el archivo de datos de WesVar



En la misma pantalla, haga clic sobre el ícono de la balanza o seleccione **Data – Create weights**. Desde **Source Variables**, mueva **Jkzone** a **VarStrat** y mueva **Jkindic** a **VarUnit**. Haga clic en **JK2** en **Method** (Figura I.C.4). Haga clic en **OK**. WesVar muestra un mensaje, *This operation will create a new VAR file. You will be asked to specify a file name (Esta operación creará un nuevo archivo VAR. Deberá asignar un nombre al archivo)*. Haga clic en **OK**. Guarde el archivo en ***NAEA DATA ANALYSIS\MYWESVARFILES*** con el nombre de archivo ***NATASSESS4.VAR***. Puede solicitársele que confirme esta operación: *Confirm Save As*. Haga clic en **Yes**. WesVar ha agregado las réplicas de las ponderaciones para poder realizar la estimación de los errores de muestreo, y el archivo se ve ahora como la Figura I.C.5.

17. Para cerrar el archivo de datos de WesVar, seleccione **File – Close**. Ahora está en condiciones de ejecutar los análisis del capítulo 3

FIGURA I.C.4

Crear ponderaciones en WesVar

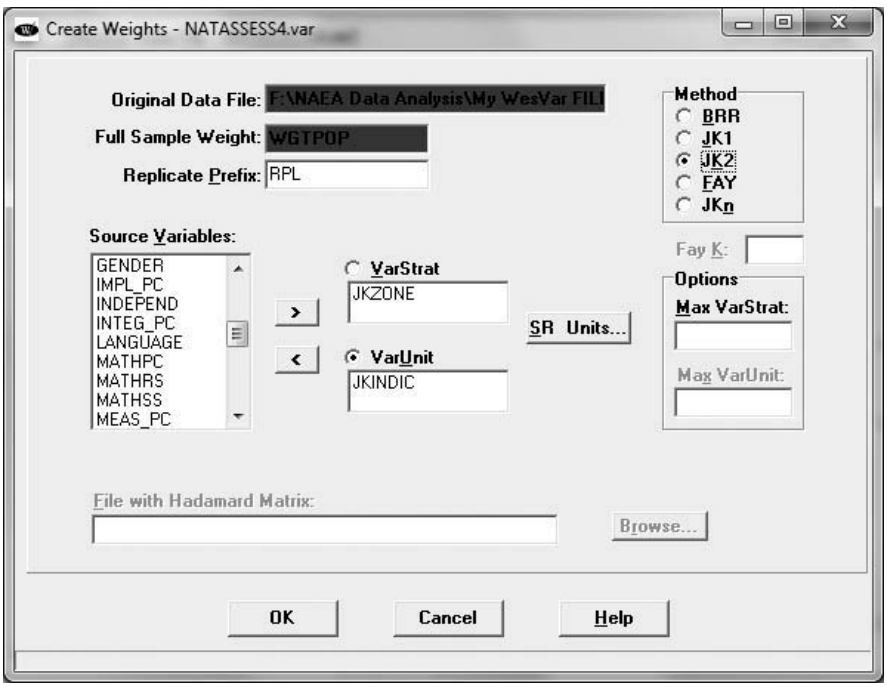
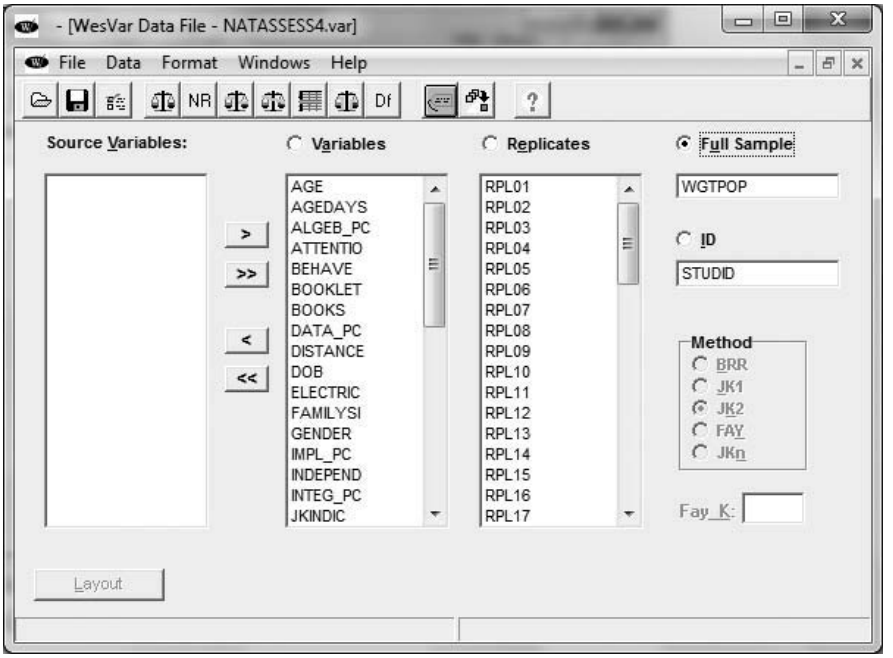


FIGURA I.C.5

Réplicas de las ponderaciones creadas por WesVar



con el archivo de datos de WesVar ***NATASSESS4.VAR***. Es importante mencionar que este archivo es idéntico al archivo ***NATASSESS4.VAR*** que se encuentra en ***NAEA DATA ANALYSIS\ WESVAR DATA & WORKBOOKS***.

NOTAS

1. En los datos que se utilizan aquí, el número de unidades primarias de muestreo (UPM), en este caso escuelas, es par. En el caso de que hubiera un número impar de escuelas, la zona final de *jackknife* tendrá tres escuelas (1, 2 y 3). En tal caso, será necesario seleccionar JK como método de análisis (paso 15) dado que JK2 requiere un número par de UPM.

2. **Jkindic** es el nombre de la variable de los indicadores *Jackknife*.

PARTE II

ANÁLISIS DE LOS ÍTEMS Y DE LAS PRUEBAS

Fernando Cartwright

CAPÍTULO 8

INTRODUCCIÓN A IATA

Este capítulo presenta una descripción de los requisitos para datos en el programa IATA (Item and Test Analysis) y un resumen de los tipos de información que brinda. Se incluye una introducción a la interfaz de IATA, por ejemplo, interfaces de tareas, menú principal y navegación durante el flujo de trabajo.

INSTALACIÓN DE IATA

Para ejecutar este programa, la computadora debe contar con los requisitos mínimos recomendados para un sistema Windows XP (SP3) o un sistema operativo posterior basado en Windows, por ejemplo, Windows Vista, Windows 7 o Windows 8. Si hay una versión anterior de IATA instalada en la computadora, se debe desinstalar, y la carpeta de datos de muestra de IATA se debe eliminar del escritorio antes de comenzar la instalación nueva. Durante el proceso de instalación, puede que se requiera actualizar el sistema operativo con la última versión de .NET Framework de Microsoft (.NET 4 o posterior). .NET Framework es un componente regular del sistema operativo Windows y normalmente se actualiza automáticamente. Sin embargo, si el acceso a internet es esporádico o si la función de

actualización automática está deshabilitada, el sistema operativo puede estar desactualizado. Si la instalación de IATA se realiza desde un CD o una memoria USB, el archivo de instalación para ***.NetFx40_Full_x86_x64.exe*** y el archivo ***IATASETUP.exe*** se encontrarán en el mismo directorio en el CD. También se puede descargar y ejecutar la actualización desde la siguiente URL: http://www.microsoft.com/en-us/download/details.aspx?id=17718.

Para instalar IATA, debe registrarse como administrador del sistema o como usuario con permiso para instalar el software. Abra el CD, abra la carpeta IATA, luego haga doble clic en el archivo ***IATASETUP.exe.*** Copie este archivo y péguelo en el escritorio de Windows, al que se puede acceder haciendo clic en el ícono de escritorio que se encuentra en el ángulo inferior derecho de la barra de tareas de Windows, o abriendo Explorer y haciendo doble clic en el ícono de escritorio. Es posible que el programa solicite su permiso para que ***IATASETUP*** pueda acceder o realizar cambios en su equipo. Debe hacer clic en **Allow** o **Yes** para continuar. Luego aparecerá la pantalla de bienvenida que confirma que IATA se está instalando. En cualquier momento, puede hacer clic en **Cancel** para salir del proceso de instalación; en ese caso no se producirá ningún cambio en su equipo. Haga clic en **Next>>** para continuar. Lea atentamente el acuerdo y seleccione **I accept the agreement** para confirmar que respetará las condiciones. Continúe con el resto del proceso de instalación haciendo clic en **Next>>**. Puede dejar los parámetros de instalación predeterminados sin modificar o cambiarlos si desea guardar sus archivos o menús en otros directorios. Antes de la instalación, aparecerá un resumen de las especificaciones. Haga clic en **Install**. Una vez completada la instalación, el programa presentará la pantalla de confirmación. Si desea ejecutar IATA de inmediato, marque la casilla de verificación **Launch IATA**. Haga clic en **Finish** para salir del asistente de instalación.

DATOS DE LA EVALUACIÓN

En el análisis de las evaluaciones se generan y se utilizan dos tipos principales de datos: los datos de respuesta y los datos del ítem.

Cada alumno genera los datos de respuesta a medida que responde las preguntas de la *prueba* (conjunto específico de preguntas que evalúa un dominio común de competencia o conocimiento). Cada pregunta de una prueba se denomina ítem, y puede tratarse de tareas con opción múltiple, con respuesta corta, con respuesta abierta, o de desempeño. Los datos del ítem se generan analizando y revisando los ítems y registrando sus propiedades estadísticas o cognitivas. Cada fila de un archivo de datos de respuesta describe las características de un alumno, mientras que cada fila de un archivo de datos del ítem describe las características de un ítem de la prueba.

IATA puede leer y escribir en muchos formatos comunes de tabla de datos (por ejemplo, archivos de texto delimitados de Access, Excel, SPSS [Statistical Package for Social Sciences]) si están formateados correctamente. Si el formato no tiene la estructura correcta, IATA no podrá realizar los análisis. Los formatos compatibles con bases de datos, como Access o SPSS, se encargan de la mayoría de las cuestiones de formateo de datos. Pero si los datos se almacenan con un formato menos restrictivo, por ejemplo, en Excel o en un archivo de texto, se deben tener en cuenta las siguientes convenciones:

- Los nombres de las variables deben aparecer en la celda superior de cada columna (conocida como *encabezado*). Cada columna con datos debe tener un encabezado. Cada variable debe tener un nombre distintivo respecto de otras variables en un archivo de datos. El nombre de una variable debe comenzar con una letra y no debe contener ningún espacio.
- El *rango de datos* es el rectángulo formado por celdas que contienen datos, que comienza con el nombre de la primera variable que aparece en el archivo de datos y termina con el valor de la última variable de la fila que está más abajo. El rectángulo de celdas que forma el rango de datos no debe tener filas o columnas vacías.
- Dicho rango debe comenzar en la primera celda de la hoja de cálculo o del archivo. En Excel, es la celda A1. En archivos de texto, corresponde a la posición del cursor en el extremo superior izquierdo del archivo.

La Figura 8.1 ilustra dos ejemplos de formato, uno correcto y el otro incorrecto. En el ejemplo de formato incorrecto (a), hay una fila en blanco arriba del rectángulo de datos y una columna en blanco a su izquierda. También hay filas y columnas en blanco dentro del rectángulo, y una de las columnas con datos no tiene encabezado. En el formato correcto (b), todos los datos están reunidos dentro de un solo rectángulo en la parte superior izquierda de la hoja de cálculo sin filas ni columnas en blanco.

FIGURA 8.1

Ejemplos de formato de datos correcto e incorrecto

a. Formato incorrecto: Filas y columnas vacías dentro del área de datos y alrededor de ella

	A	B	C	D	E	F	G
1							
2		Var_A		Var_B	Var_C		Var_E
3							
4		1		1	1	1	1
5		2		2	2	2	2
6		3		3	3	3	3
7		4		4	4	4	4
8							
9		5		5	5	5	5

b. Formato correcto: Rango de datos en la esquina superior sin filas ni columnas vacías

	A	B	C	D	E	F	G
1	Var_A	Var_B	Var_C	Var_D	Var_E		
2	1	1	1	1	1		
3	2	2	2	2	2		
4	3	3	3	3	3		
5	4	4	4	4	4		
6	5	5	5	5	5		
7							
8							
9							

Datos de respuesta

Estos datos incluyen la respuesta de cada alumno a cada ítem de la prueba. Los resultados de la prueba importados en el archivo de datos de respuesta deben permitir la calificación automática, es decir, que esos datos deben incluir códigos que representen cómo los alumnos respondieron realmente. Por ejemplo, si los datos de respuesta se obtienen de una prueba con opción múltiple, esos datos deben registrar los códigos que representen las opciones que cada alumno eligió (por ejemplo, A, B, C, D). IATA convertirá las respuestas codificadas en puntajes usando la clave de las respuestas que se ingresó manualmente o que se encuentra en un archivo de clave de respuestas.

En un archivo de datos de respuesta se puede guardar información adicional que resultará útil cuando se analicen los resultados de la prueba. Por ejemplo, información demográfica según variables de edad, grado, sexo, escuela y región. Otras informaciones útiles se pueden recopilar de cuestionarios (por ejemplo, cuestionarios alumno-docente) u obtenerse de los registros administrativos. Si se empleó una muestra estratificada de alumnos, en este archivo se debe incluir la ponderación de la muestra para cada alumno.

Se debe asignar un identificador único para cada alumno; IATA los generará automáticamente según el orden del registro, en caso de que el identificador único no esté especificado. Sin embargo, si los resultados se han de vincular con otras fuentes de datos, como encuestas de seguimiento o registros administrativos, el empleo de un identificador predefinido, por ejemplo, el nombre o el número del alumno, es práctico para facilitar un posterior enlace entre conjuntos de datos.

Se deben asignar códigos a todas las respuestas. En los ítems de opción múltiple, este procedimiento es directo ya que cada opción de la respuesta ya está codificada como correcta o incorrecta. Para los ítems de respuesta abierta, se necesita una rúbrica de puntaje para calificar las respuestas usando un marco de codificación común. Estos ítems pueden calificarse como correcto/incorrecto o asignándose un crédito parcial a las respuestas diferentes. Un ítem con crédito parcial tiene más de un puntaje mayor a cero. Las respuestas a las preguntas de respuesta abierta deben codificarse previamente durante el proceso de preparación de los datos de respuesta. En los volúmenes

2 y 3 de esta serie se describen los procedimientos de codificación para los ítems de la prueba (Anderson y Morgan, 2008; Greaney y Kellaghan, 2012). Para la puntuación de los datos de respuesta, en la mayoría de los análisis, se debe cargar una clave de respuestas en IATA. Esta clave es una lista de códigos de respuesta que indican la o las respuestas correctas para cada ítem de la prueba. La clave de respuestas se puede importar como un archivo de datos o ingresar manualmente. Si el análisis utiliza parámetros de ítem anclados (por ejemplo, para vincular versiones distintas de una prueba) deben incluirse en el archivo de clave de respuestas y no se podrán ingresar manualmente (véase más abajo "Datos de ítems").

Manejo de datos faltantes u omitidos

La falta de datos se produce cuando un alumno no responde un ítem de la prueba. Cuando esto sucede, en lugar de dejar el campo de datos en blanco, se emplea un código de valor faltante para registrar por qué falta la respuesta. Existen dos tipos de respuestas faltantes: *faltante* y *omitida*.

El código de datos *faltante* se asigna a las variables cuando los alumnos podrían haber respondido un ítem pero no lo hicieron, dejando la respuesta en blanco. Esos datos faltantes se califican como incorrectos. Por el contrario, el código de datos *omitido* se usa cuando no se administró un ítem a los alumnos, por ejemplo, cuando se aplicó una rotación de cuadernillos en una evaluación nacional.

Según sean las circunstancias de administración de la prueba o del procesamiento de los datos, se debe decidir si se procesarán como datos faltantes o como omitidos los códigos que se aplican a las respuestas de alumnos que no se pudieron leer o que se contestaron de forma inadecuada, por ejemplo, si seleccionaron dos opciones en un ítem de opción múltiple. En general, si estos errores en los datos son el resultado de un error del alumno, los códigos se deben tratar como faltantes y se deben calificar como incorrectos. Pero si los errores resultan de las limitaciones del procesamiento de datos, como la falta de precisión en el escaneo de la ficha de puntuación que no se verificó, los códigos deben considerarse como omitidos.

Los códigos de datos omitidos también se usan cuando se emplea un diseño equilibrado de rotación de cuadernillos de la prueba.

Estos diseños equilibrados implican asignar muestras de ítems equivalentes aleatoriamente a distintos alumnos para que no respondan todos los mismos ítems de la prueba (véase Anderson y Morgan 2008). Además, permiten una amplia cobertura de la materia y al mismo tiempo limitan el tiempo requerido para la prueba. En un diseño de rotación de cuadernillos, los códigos omitidos se deben asignar a todos los ítems de un alumno, excepto los ítems contenidos en el cuadernillo de prueba que se presentó al alumno. Normalmente, los códigos omitidos no se asignan a los ítems en los casos en que todos los alumnos deben responder a todos los ítems.

Las convenciones usuales emplean algunos valores específicos para los distintos tipos de datos de respuestas omitidas (véase Freeman y O´Malley 2012 para obtener información sobre códigos de respuestas). Los siguientes son los valores comúnmente empleados:

- 9 para una respuesta faltante, donde los alumnos no han respondido a un ítem
- 8 para una respuesta que no se puede calificar; normalmente sucede cuando los alumnos indican varias respuestas en una prueba de opción múltiple y cuando las respuestas a los ítems de respuesta abierta son ilegibles.
- 7 para ítems omitidos o ausentes, que podría usarse en un diseño de rotación de cuadernillos

Independientemente de los códigos específicos utilizados, se debe especificar el modo en que IATA debe tratar cada código de la respuesta omitida, si como faltante u omitido.

Nomenclatura del ítem

En un programa de evaluación nacional, es importante asignar un nombre único a cada ítem (véase Anderson y Morgan, 2008; Freeman y O'Malley, 2012). Todos los análisis estadísticos que se realizan sobre un ítem de la prueba deben estar claramente vinculados con el nombre o la etiqueta de un ítem. Si el ítem se repite en varios ciclos de una evaluación nacional, debe conservar el mismo nombre en los archivos de datos para cada ciclo. Por ejemplo, un ítem de matemáticas utilizado por primera vez en 2009 podría llamarse **M003**, para indicar que fue el tercer ítem que aparecía en la prueba de ese año. Si se usa ese

mismo ítem en una prueba de 2010, deberá conservar el nombre **M003**, sin importar en qué lugar de la prueba aparezca. Nombrar ítems según la posición que ocupan en la prueba puede causar confusión si se repiten. Por esa razón, es más útil asignar nombres permanentes al elaborar el ítem la primera vez, en lugar de cuando se utilizan por primera vez en una evaluación.

La coherencia en la asignación de nombres también facilita la vinculación de los resultados de diferentes pruebas. Cuando IATA realiza estimaciones sobre los vínculos estadísticos entre pruebas, establece una correspondencia de nombres de los ítems durante el procedimiento de vinculación. Si el nombre de un ítem hace referencia a ítems diferentes de dos pruebas vinculadas, la vinculación arrojará resultados inexactos. Si bien se puede cambiar el nombre de los ítems para facilitar el proceso mencionado, es más simple mantener un nombre único desde el comienzo y es menos probable que se carguen errores en el sistema.

Variables reservadas por IATA

Durante el análisis de los datos de respuestas, IATA calcula una variedad de variables de salida o de trabajo. Los nombres de estas variables son de uso restringido y no deben emplearse como nombres de ítems de las pruebas ni de variables de los cuestionarios. Estas variables, que IATA agrega al archivo de datos con puntaje de la prueba, aparecen enumeradas en la Tabla 8.1.

Además de estos nombres específicos, evite usar nombres que contengan el símbolo @, que está reservado para procesar los *ítems de crédito parcial*, que son ítems con más de un valor posible de puntaje mayor a cero.

Datos de ítems

IATA genera y utiliza archivos de datos de ítems con un formato específico. Un archivo de datos de ítems contiene toda la información necesaria para realizar el análisis estadístico de los ítems y puede incluir los parámetros utilizados para describir las propiedades estadísticas de estos ítems. Un banco de ítems generado o empleado por IATA debe incluir las variables enumeradas en la Tabla 8.2.

TABLA 8.1

Variables generadas o utilizadas por IATA para describir la competencia del alumno y su desempeño en la prueba

Nombre del puntaje	Descripción
XWeight	Es la ponderación del diseño del caso empleado durante el análisis (si no está especificado, el valor es igual a 1 para todos los alumnos).
Missing	Esta variable describe la cantidad de ítems omitidos para un alumno.
PercentScore	El puntaje porcentual es la cantidad de ítems contestados correctamente por el alumno, expresado como porcentaje de la cantidad total de ítems administrados al alumno (excluidos los datos de respuesta omitidos).
PercentError	Es el error de medición para el puntaje porcentual. Es una estimación específica para cada alumno; su valor depende del puntaje porcentual y de la cantidad de ítems que respondió el alumno.
Percentile	El rango de percentil (un número entre 0 y 100) describe, para cada alumno, el porcentaje de otros alumnos con puntajes inferiores.
RawZScore	RawZScore es el puntaje porcentual transformado para que tenga un valor medio de 0 y una variación estándar de 1 dentro de la muestra.
ZScore	Este puntaje representa el equivalente de distribución normal del puntaje de percentil. También se conoce como *puntaje con curva de campana.* Mientras que la distribución de la variable RawZScore depende de la distribución de puntajes porcentuales correctos, la distribución de esta variable suele tener una forma de campana más perfecta.
IRTscore	Este puntaje (Teoría de Respuesta al Ítem; IRT, por sus siglas en inglés) representa la estimación de la competencia del alumno. Normalmente, este puntaje tiene una variación media y estándar de 0 y 1, respectivamente. Facilita la generalización más allá de una muestra específica de ítems porque su estimación considera las propiedades estadísticas de los ítems de la prueba.[a]
IRTerror	Esta variable es el error en la medición del puntaje IRT.
IRTskew	La asimetría en la estimación de la competencia indica si la prueba resulta mejor para medir el límite inferior o el superior de competencia del alumno. (Por ejemplo, una prueba fácil puede describir con exactitud si los alumnos alcanzaron un nivel mínimo de competencia pero no brinda una estimación más exacta sobre los niveles superiores de competencia).
IRTkurt	La curtosis de la estimación de competencia indica cuán precisa es la estimación para un nivel de error dado. (Por ejemplo, en dos puntajes con el mismo error de medición, el que tenga una medida de curtosis mayor será el más preciso).
TrueScore	Esta puntuación es la estimación de un puntaje porcentual calculado a partir de la variable IRT score. Es preferible a la puntuación porcentual bruta porque corrige las diferencias de error de medición entre ítems. Se calcula como el promedio de probabilidad de respuesta correcta a cada ítem, dados la variable IRT score del alumno y los parámetros de los ítems de la prueba.
Level	Esta variable es una estimación del nivel de competencia para un alumno asignada según los procedimientos de determinación de estándares. (Si no se han realizado estos procedimientos, por defecto se asignará a todos los estudiantes un valor de 1).

Nota: TRI = Teoría de Respuesta al Ítem.

a. Las funciones avanzadas de IATA incluyen opciones de escalado adicionales; consulte la guía de instalación en el CD de acompañamiento.

TABLA 8.2

Variables en un archivo de datos de ítems

Variable	Descripción
Name	(OBLIGATORIO) nombre unívoco de cada ítem de la prueba
Key	(OBLIGATORIO) información empleada para asignar una puntuación numérica a cada respuesta al ítem, que es, o bien el código único correspondiente a la respuesta correcta, o un conjunto delimitado de valores que define una variedad de respuestas aceptables y sus respectivas puntuaciones numéricas
a	(OPCIONAL) el primero de los tres parámetros que describe cómo el rendimiento de un ítem de la prueba se relaciona con la competencia en el sector del rendimiento; se lo conoce como *parámetro de discriminación* o *de pendiente*
b	(OPCIONAL) segundo parámetro del ítem, conocido como *parámetro de dificultad* o *de ubicación*
c	(OPCIONAL) tercer parámetro del ítem, conocido como el *parámetro de pseudoazar*[a]
Level	(OPCIONAL) un nivel de competencia previamente asignado en función de la especificación inicial del ítem y de la revisión de un experto (los valores deben ser números naturales, comenzando con 1)
Content	(OPCIONAL) un código o descripción empleado para describir el subsector del currículo, también conocido como *eje* o *hilo*, con el cual cada ítem está fuertemente alineado

a. Utilizar el parámetro **c** para describir ítems puede provocar que no funcionen correctamente determinadas funciones, como la de equivalencia. Para la mayoría de los propósitos, los ítems son más útiles si el valor del parámetro **c** es igual a cero o se establece en cero. El modelo de los tres parámetros debe ser utilizado únicamente por usuarios expertos que sean conscientes de sus limitaciones. La estimación y el empleo del parámetro **c** son propiedades de las funciones avanzadas de IATA. El registro en IATA es gratuita y permite acceder a estas características avanzadas. Para obtener indicaciones sobre el registro, consulte la guía de instalación en el CD adjunto.

La Tabla 8.3 presenta ejemplos de un archivo de datos del banco de ítems con información de cinco ítems de ciencia denominados **C1Sci31**, **C1Sci32**, **C1Sci33**, **C1Sci34** y **C1Sci35**. Observe que el ítem **C1Sci35** no incluye datos en las columnas **a, b, c** y ***Level***. Como se indica en la tabla, los únicos campos de datos obligatorios son **Name** y **Key**. Si faltan los parámetros **a**, **b** o **c**, serán calculados durante el análisis. En muchas situaciones puede que en IATA se necesite la captura de un archivo de datos de ítems que no tenga estos parámetros. El escenario más común se presenta cuando nunca se analizaron los datos de la respuesta al ítem; en ese caso, el archivo se emplea simplemente como una clave de respuestas.

TABLA 8.3

Ejemplo de sección de un archivo de datos de ítems

Name	a	b	c	Key	Level	Content
C1Sci31	0,34	0,83	0,01	3	3	Razonamiento científico
C1Sci32	0,46	0,40	0,12	4	2	Física
C1Sci33	0,32	0,31	0,06	3	2	Física
C1Sci34	0,18	0,75	0,16	1	3	Biología
C1Sci35				5		Entorno

Otro escenario ocurre cuando algunos ítems incluyen parámetros estimados en un análisis de datos anterior y se desea fijar los valores de estos ítems, en lugar de que IATA los vuelva a estimar; en este caso, se podrían dejar vacíos los valores **a**, **b** y **c** de los ítems para los que se quiera estimar nuevos parámetros. Los valores para **Level** y **Content** se pueden ingresar desde la interfaz de IATA en forma manual o se pueden dejar vacíos.

Un archivo de datos de ítems también puede incluir variables adicionales. Por ejemplo, normalmente junto con datos de ítems se puede guardar la pregunta enunciada en el banco de ítems, las estadísticas que describen el número de veces que el ítem se ha utilizado y una lista de las pruebas en las que aparece cada ítem. No obstante, IATA no utilizará otras variables que no sean los siete campos de datos necesarios enumerados en la Tabla 8.3.

El equipo de evaluación nacional puede emplear información de cualquier fuente si tiene los datos requeridos en el formato que se presenta en la Tabla 8.2. Por ejemplo, las evaluaciones nacionales pueden obtener permiso para utilizar los ítems de una evaluación a gran escala, como puede ser una administrada por la Asociación Internacional para la Evaluación del Rendimiento Educativo, que incluye el TIMSS (Estudio Internacional de Tendencias en Matemáticas y Ciencias) y el PIRLS (Estudio sobre el Progreso Internacional de la Competencia en Lectura) (http://timss.bc.edu/). Si los ítems existentes en las evaluaciones a gran escala se incluyen en una evaluación nacional, los parámetros del ítem de esas evaluaciones se pueden emplear para crear un archivo de datos de ítems que IATA pueda importar.

Formatos de clave de respuesta

En la columna con el encabezado **Key** de un archivo de datos banco de ítems es necesario indicarle a IATA la información que puede utilizar para calificar cada ítem. En el caso más sencillo, para los ítems de opción múltiple con una única opción correcta, el valor de cada columna debe ser el carácter alfanumérico correspondiente a la opción correcta. El valor distingue entre mayúsculas y minúsculas, lo que significa que, por ejemplo, si la respuesta correcta está codifica con A mayúscula, la letra A mayúscula debe estar incluida en la clave de respuesta; si la clave indica un valor *a* minúscula, entonces las respuestas con un valor A mayúscula se calificarán como incorrectas.

En raras ocasiones, durante el proceso de revisión de la prueba, se puede establecer que un ítem tenga más de una respuesta correcta. Para asignar más de un valor a una clave de respuesta, se debe ingresar una lista de valores correctos, separados por comas sin espacios entre los valores ni después de las comas. Por ejemplo, si en un ítem de la prueba, las respuestas A y C se consideran correctas, el valor de la clave del ítem debe ser A,C.[1]

Formatos de datos del ítem para ítems de crédito parcial

Los ítems de crédito parcial (o respuesta graduada) son ítems de la prueba que tienen más de un valor de calificación. Por ejemplo, en lugar de tener un puntaje de 0 o 1, un ítem con distintos niveles de respuesta correcta puede tener una puntuación de 0, 1 o 2, donde 0 representa un intento de respuesta; 1, una respuesta parcialmente correcta y 2, una respuesta perfecta. Con el fin de dar cabida a diferentes valores de puntaje, se debe ingresar la clave de respuesta para cada valor del puntaje mayor que 0. Si todos los puntajes de los ítems de crédito parcial del sistema de calificación son mayores que 0, no debe ingresarse información clave de respuesta para el valor más bajo. Por ejemplo, si las posibles calificaciones para los ítems son 1, 2 y 3, la clave de respuesta debe ofrecer información solo para los puntajes 2 y 3. El formato de una clave de respuesta de crédito parcial es <score1>:<value list 1>; <score 2>:<value list 2>; … <score *n*>:<value list *n*>. Por ejemplo,

para un ítem de crédito parcial con tres puntajes, codificados como A, B y C, con valores 1, 2, y 3, respectivamente, la clave de respuesta debe ingresarse como 1:A;2:B;3:C.

Si un ítem de crédito parcial ya ha sido analizado, tendrá un mayor número de parámetros que un ítem de la prueba regular. Cada valor de calificación tendrá un valor distinto para el parámetro **b**, aunque el parámetro **a** tendrá el mismo valor en todas las calificaciones. Estos datos del ítem se deben ingresar en un formato especial. Además de la entrada principal con la clave de respuesta completa, se debe agregar una nueva entrada para cada valor de la calificación (excepto para el valor más bajo) como si cada puntaje del ítem fuera un ítem independiente de la prueba. Los campos de los parámetros de la entrada del ítem principal deben quedar en blanco. Por ejemplo, si un ítem tiene puntuaciones de 0, 1 y 2, se requerirán en total tres filas en el archivo de datos del ítem: una para el ítem general, que tendrá solo la clave de respuesta y el nombre del ítem; y dos entradas específicas de puntajes para 1 y 2 con nombre, clave de respuesta e información del parámetro.

El valor del campo nombre de cada nueva entrada específica para puntaje es el nombre del ítem original seguido de @<score value>. Por ejemplo, si el nombre del ítem original es **TestItem**, el nombre para un puntaje 1 es **TestItem@1**. IATA emplea un modelo de respuesta al ítem que requiere que los valores de los parámetros **b** tengan el mismo orden que los puntajes. Por lo tanto, en caso de haber dos entradas de puntaje, 1 y 2, el valor del parámetro **b** para el puntaje 2 debe ser mayor que el del parámetro **b** para el puntaje 1 (consúltese la Tabla 8.4).

TABLA 8.4

Ejemplo de sección de un archivo de datos de ítems para un ítem de crédito parcial

Name	a	b	c	Key	Level	Content
PCItem001				1:1;2:2;3:3		Partes de la oración
PCItem001@1	0,61	−0,43	0	1,2,3	1	Partes de la oración
PCItem001@2	0,61	0,22	0	2,3	1	Partes de la oración
PCItem001@3	0,61	0,74	0	3	2	Partes de la oración

Cuando se agrega una nueva fila para cada puntaje de ítem de crédito parcial, los valores del campo clave de respuesta también se deben especificar de manera diferente. El análisis de un ítem de crédito parcial adopta el supuesto de que un alumno con un puntaje determinado también ha logrado dominar el nivel de habilidad asociado con el puntaje más bajo en ese ítem. En otras palabras, si cada puntaje se considera un ítem independiente de la prueba, se supone que un alumno con un puntaje de crédito parcial alto se ha desempeñado correctamente en las puntuaciones de crédito parcial más bajas. Para gestionar esta relación en IATA, la clave de respuesta para cada valor del puntaje debe enumerar su propio valor o los valores clave, así como los valores de los puntajes más altos.

En la Tabla 8.4 se muestra un ejemplo de formato de datos del ítem de crédito parcial adecuado para un ítem con puntajes de 0, 1, 2 y 3. Observe que no se proporciona información para el puntaje más bajo (0). La entrada del ítem principal no tiene valores de parámetro ni un valor para **Level**. Debido a que cada valor de puntaje podría corresponder a un estándar de desempeño diferente, no tiene sentido especificar el nivel de todo el ítem. Si bien las respuestas ya están calificadas, la información de puntaje se debe especificar con el formato de clave de respuesta adecuado. Para que IATA asigne correctamente el puntaje a las respuestas, la clave de respuesta debe proporcionar los valores encontrados en los datos y el puntaje asignado a cada valor.

DATOS GENERADOS POR IATA

IATA genera varias tablas de datos que incluyen las especificaciones del análisis actual y los resultados del análisis. En general, todos los resultados deben ser archivados como referencia a futuro. La Tabla 8.5 resume la lista de tablas de datos generada por IATA. Las tablas se pueden guardar individualmente o en conjunto directamente desde IATA en una variedad de formatos comunes, como Excel (*.xls/*.xlsx) y SPSS (*.sav), delimitadas por comas (*.csv) o tabulaciones (*.txt).

Además de las tablas de datos que se describen en la Tabla 8.5, IATA genera varios gráficos, resúmenes de texto y tablas de resultados

TABLA 8.5

Tablas de datos generadas por IATA

Tablas de datos	Descripción
Responses	Datos de respuesta originales (incluidos los datos que no forman parte de la prueba) importados en IATA
Values	Códigos de respuesta únicos para todos los ítems de la prueba e indicación de cómo se codificó el valor de respuesta, si como faltante válido (omitido válido) o faltante
Scored	Datos de respuesta que se calificaron como correctos (1) o incorrectos (0) empleando la clave de respuesta especificada, así como todas las puntuaciones resumidas y sus errores estándar
Items1[a]	Claves de respuesta al ítem y estadísticas relacionadas con el análisis actual y los parámetros del ítem
Items2	Parámetros y claves de respuesta al ítem del archivo de parámetros del ítem de referencia empleados para la vinculación
MergedItems	Correlación ítem por ítem, entre el nuevo archivo de parámetros y el de referencia empleados por el proceso de vinculación
Eigenvalues	Proporción de la varianza que se explica en cada una de las dimensiones inherentes a las respuestas al ítem
PatternMatrix	Proporción de la varianza de cada ítem que se explica en cada una de las dimensiones inherentes a las respuestas al ítem
Levels	Umbrales empleados para definir los niveles de competencia
LinkingConstants	Constantes de transformación a escala empleadas para ajustar la escala de rasgos latente entre las poblaciones o las muestras
BookmarkData	Enumeración ordenada de los ítems que se puede utilizar para facilitar la determinación de estándares o la creación de definiciones para los niveles de rendimiento
DIF_<specifications>	Resultados de un análisis de funcionamiento diferencial del ítem, donde la parte del nombre de la tabla <specifications> resume la variable y los grupos comparados en el análisis
CustomTest<name>	Grupo de ítems elegidos para minimizar el error de medición en un rango de competencia específico; <name> es un valor específico del usuario

a. La tabla de datos Item1 producida por IATA tras un análisis de los datos de respuesta no solo sirve de fichero de datos del banco de ítems sino que también genera varias estadísticas adicionales. Dichas estadísticas, incluidas en secciones posteriores acerca del análisis de los datos de respuesta, describen el comportamiento de los ítems en una muestra específica y son útiles como guía para los análisis y la elaboración de las pruebas pero no se requiere mantenerlas en un archivo de banco de ítems usado por IATA.

que solo se muestran en la interfaz de este programa. Los resultados se pueden copiar directamente desde IATA y se pueden pegar en otros documentos para su posterior consulta. La forma de copiarlos depende del tipo de resultado.

En gráficos, al hacer clic con el botón derecho sobre el cuerpo de la imagen aparecerá un menú emergente con las opciones para (a) copiar la imagen en el portapapeles, (b) guardar la imagen gráfica directamente en un archivo o (c) imprimir la imagen. Para mostrar los resultados en forma de tabla, primero debe seleccionar las celdas, filas o columnas, luego copiar los datos, ya sea seleccionando **Copy** con el botón derecho en el menú emergente o con **Ctrl+c**. Los datos copiados se pueden pegar en un archivo de texto o directamente en un programa con hojas de cálculo, como Excel o SPSS.

CÓMO INTERPRETAR LOS RESULTADOS DE IATA

Siempre que IATA genera resultados de análisis para ítems individuales, también presenta indicadores sintéticos "símbolo de tráfico", que proporcionan una idea general de cómo interpretar los resultados. IATA utiliza tres símbolos (consúltese la Tabla 8.6). Los símbolos

TABLA 8.6

Símbolos de tráfico en IATA y su significado

Símbolo	Significado
	Un círculo indica que no hay ningún problema grave.
	Un diamante indica que los resultados están por debajo del nivel óptimo. Este indicador se emplea para sugerir que se podrían requerir modificaciones en las especificaciones de análisis o en los ítems mismos. No obstante, el ítem no presenta ningún error significativo en los resultados del análisis.
	Un triángulo aparece al lado de un ítem potencialmente problemático. Se utiliza para indicar los ítems que no se pudieron incluir en el análisis debido a problemas con los datos o con las especificaciones, o para recomendar una revisión más exhaustiva de las especificaciones o de los ítems de la prueba y los datos inherentes. Cuando aparece este indicador, no significa necesariamente que exista un problema, pero sí sugiere que los resultados generales de los análisis podrían ser más exactos si se suprimiera el ítem de la prueba indicado o si se volviera a especificar el análisis.

(círculo, diamante, triángulo) aparecen en color en la pantalla de la computadora.

En el caso de análisis en los que se consideran múltiples elementos de información al interpretar los resultados de un ítem específico, como el análisis de ítems y los resultados de las pruebas de dimensionalidad, IATA genera estados de situación interpretativos que pretenden resumir las diferentes estadísticas. Estos estados pretender ser una sugerencia práctica sobre cómo proceder. No obstante, en caso de que IATA recomiende una modificación en cualquiera de las especificaciones del análisis o en los ítems de la prueba, deberá comprobar que la recomendación sea apropiada mediante la revisión de los resultados estadísticos, los cuadernillos de prueba reales, o ambos, o mediante el análisis de los ítems con un especialista en currículos.

DATOS DE LA MUESTRA

Al instalar IATA en una computadora, se crea una carpeta en el escritorio con el nombre IATA. Esta carpeta contiene datos de la muestra que se requieren para los ejemplos guiados que se presentan en este libro. La carpeta con los datos de la muestra contiene siete archivos: seis conjuntos de datos de respuestas, cada uno en formato Excel, y un archivo Excel que contiene claves de respuestas para cada uno de los conjuntos de datos de respuestas. Los archivos están en formato *.xls para compatibilidad con programas más antiguos y de acceso público (dependiendo de la configuración de la computadora, la extensión de archivo .xls puede que no sea visible). Los nombres y el contenido de los archivos son los siguientes:

- ***PILOT1:*** conjunto de datos de respuesta que corresponden a una prueba piloto que contiene ítems de opción múltiple
- ***CYCLE1:*** conjunto de datos de respuesta que corresponden a la administración de la evaluación nacional
- ***PILOT2:*** conjunto de datos de respuesta que corresponden a una prueba piloto que contiene ítems de opción múltiple en un diseño equilibrado de rotación de cuadernillos
- ***PILOT2PartialCredit:*** conjunto de datos de respuesta que corresponden a una prueba piloto que contiene ítems de opción

múltiple y de crédito parcial en un diseño equilibrado de rotación de cuadernillos
- ***CYCLE2:*** conjunto de datos de respuesta que corresponden a la administración de la evaluación nacional con ítems comunes a una administración anterior
- ***CYCLE3:*** conjunto de datos de respuesta que corresponden a una administración de evaluación nacional con elementos comunes con una administración anterior
- ***ItemDataAllTests:*** archivo Excel con varias hojas que incluye claves de respuestas e información acerca de los ítems de cada uno de los archivos de datos de respuesta

Los datos de esta muestra son conjuntos ficticios desarrollados con la única intención de proporcionar aplicaciones y ejemplos concretos para este programa. Si bien reflejan patrones típicos de respuesta de alumnos, y las relaciones en los datos son similares a las de la mayoría de las evaluaciones a gran escala, los resultados y los debates acerca de las conclusiones de los análisis no representan una evaluación nacional real.

Si se eliminaran algunos de los archivos de datos de la muestra, se pueden recuperar ejecutando el programa IATASetup.exe en el CD o desde el sitio web de IATA (http://polymetrika.com/home/IATA).

ANÁLISIS DE LOS FLUJOS DE TRABAJO Y DE LAS INTERFACES DE IATA

IATA se diferencia de los programas de análisis estadísticos que proporcionan una variedad de funciones de análisis a las que se puede acceder de forma individual. Por el contrario, en IATA se puede acceder a todas las funciones de análisis a través de flujos de trabajo, en los que los resultados de cada paso se pueden emplear para informar las especificaciones o para la interpretación de los resultados en pasos posteriores. Hay cinco flujos de trabajo disponibles en el programa:

- Response data analysis, para análisis de datos de respuesta
- Response data analysis with linking, para análisis de datos de respuesta con vinculación

- Linking item data, vinculación de datos del ítem
- Selecting optimal test items, selección de ítems de la prueba óptimos
- Developing and assigning performance standards, desarrollo y asignación de estándares de rendimiento

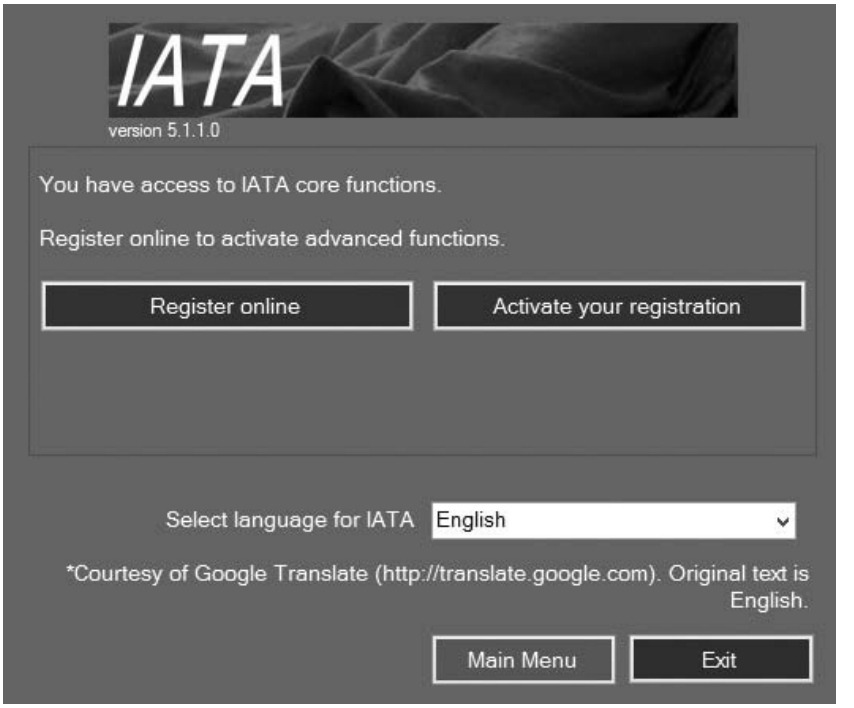
Cada flujo de trabajo refleja las necesidades de los objetivos específicos que pueden surgir en el contexto de una evaluación nacional. Las siguientes directrices se refieren a algunas situaciones asociadas con flujos de trabajo:

- Si se ha realizado una prueba piloto y se necesita información detallada sobre el comportamiento del ítem para determinar el contenido de la prueba final, se debe emplear el flujo de trabajo **Response data analysis**.
- Si se ha completado la recolección de datos de la primera evaluación nacional en una serie de evaluaciones planificada, se debe emplear el flujo de trabajo **Response data analysis**.
- Si se asigna una nueva escala de puntaje a una muestra de alumnos que ya rindieron la misma prueba en una evaluación nacional anterior, se debe emplear el flujo de trabajo **Response data analysis**.
- Si se ha realizado una evaluación nacional que comparte ítems con una evaluación anterior y se desea realizar una comparación de los resultados entre las dos, se debe emplear el flujo de trabajo **Response data analysis with linking** o el flujo de trabajo **Linking item data**.
- Si se necesita modificar una prueba y debe identificar los mejores ítems que se deseen conservar en la nueva prueba con el fin de mantener su comparabilidad con la prueba anterior, se debe utilizar el flujo de trabajo **selecting optimal test item.**
- Si ya se realizó la evaluación nacional y desea interpretar los resultados conforme a las expectativas del currículo, en lugar de simplemente comparar a los alumnos entre sí, se debe utilizar el flujo de trabajo **Developing and assigning performance standards**.

Para realizar un análisis con IATA, debe seleccionar uno de estos flujos en el menú principal. Para acceder al menú principal, haga clic en **Main Menu** en el extremo inferior derecho de la pantalla de

FIGURA 8.2

Selección inicial del idioma y registro opcional en IATA



selección de idioma y registro que aparece en IATA (Figura 8.2). Para ejecutar IATA, no es necesario ingresar información en esta pantalla. El idioma predeterminado es el inglés, y el registro es opcional. Al registrarse tiene acceso a las funciones avanzadas de análisis y a recibir notificaciones de las actualizaciones de IATA.

La Figura 8.3 muestra el menú principal de IATA.

Cada flujo de trabajo se compone de un conjunto de tareas que se completan en orden. La mayor parte de los flujos de trabajo comparten muchas tareas. IATA realiza 10 tareas, cada una con su propia interfaz, que generalmente aparecen en el orden indicado en la Tabla 8.7. No todas las tareas se muestran en todos los flujos de trabajo. Los flujos de trabajo están diseñados para que solo deba realizar las tareas relevantes para sus objetivos de análisis. La Tabla 8.7 combina las tareas con los flujos de trabajo.

FIGURA 8.3

Menú principal de IATA

TABLA 8.7

Tareas en IATA y flujos de trabajo que se emplean para realizarlas

Tareas	Flujos de trabajo				
	Response data analysis A	Response data analysis with linking B	Linking item data C	Selecting optimal test items D	Developing/ assigning performance standards E
Cargar datos	•	•	•	•	•
Determinar especificaciones de análisis	•	•			
Analizar ítems de la prueba	•	•			
Analizar la dimensionalidad de la prueba	•	•			
Analizar el funcionamiento diferencial del ítem	•	•			
Vincular		•	•		

(continúa)

TABLA 8.7

Tareas en IATA y flujos de trabajo que se emplean para realizarlas *(continúa)*

Tareas	Flujos de trabajo				
	Response data analysis A	Response data analysis with linking B	Linking item data C	Selecting optimal test items D	Developing/ assigning performance standards E
Establecer escala de resultados de la prueba	•	•			
Seleccionar ítems de la prueba óptimos	•	•		•	
Informar el desarrollo de estándares de rendimiento	•	•			•
Guardar los resultados	•	•	•	•	•

Los dos primeros flujos de trabajo (A, B) son muy similares en cuanto a sus tareas. Por el contrario, los últimos tres (C, D, E) analizan únicamente los datos del ítem. Todos los flujos de trabajo requieren que los datos se carguen en IATA y permiten guardar los resultados.

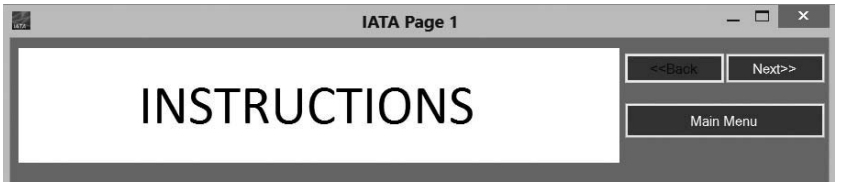
CÓMO AVANZAR POR LOS FLUJOS DE TRABAJO DE IATA

Después de seleccionar un flujo de trabajo en el menú principal, IATA lo guiará a través del conjunto de tareas que componen el flujo de trabajo. Cada tarea tiene su propia interfaz que permite especificar de qué forma IATA debe realizar la tarea y, en caso de corresponder, ver los resultados.

El recuadro de instrucciones de IATA y los botones de navegación se muestran en la Figura 8.4. El recuadro de instrucciones ofrece un resumen de las especificaciones que se requieren para cada tarea y las sugerencias interpretativas. Haga clic en el botón **<<Back** para revisar una tarea anterior; haga clic en **Next>>** para ir a la tarea siguiente. Tenga en cuenta que si bien IATA no impide que el usuario se desplace hacia atrás y hacia adelante en el flujo de trabajo, puede que las tareas posteriores en el flujo de trabajo no brinden resultados significativos a menos que se hayan completado correctamente las tareas anteriores.

FIGURA 8.4

Recuadro de instrucciones de la interfaz de tareas de IATA y botones de navegación



Independientemente del flujo de trabajo en el que aparezcan, las especificaciones generales para cada tarea siguen siendo las mismas. Las interfaces de las tareas se describen en detalle en las guías de ejemplo en los capítulos 9 al 14 de este volumen.

Los ejemplos en los capítulos siguientes muestran cómo utilizar IATA para realizar los análisis de datos de la prueba y del ítem necesarios para una evaluación nacional aplicando los datos de la carpeta de IATA en su escritorio. Para analizar los datos de su propia evaluación nacional, cree una carpeta nueva y asígnele un nombre adecuado, como por ejemplo ***NATIONAL_ASSESSMENT_YEAR_1***. (Cuando nombre archivos y carpetas evite el empleo de espacios o caracteres especiales, excepto _). Debe guardar los datos de respuesta de sus alumnos en esta carpeta y también en esta misma carpeta, todos los resultados producidos por sus análisis. Si bien IATA puede analizar datos en distintos formatos, como Excel y SPSS, los archivos de datos que emplee deben respetar las estructuras y convenciones de nombres que se describen en las tablas 8.2 y 8.3. Los nombres de los archivos de datos deben identificar claramente el origen de los datos.[2]

NOTAS

1. Este requerimiento de formato significa que nunca debe emplear comas como valores de clave de respuesta.

2. Podrá encontrar información adicional sobre IATA en http://polymetrika.com/home/IATA.

CAPÍTULO 9

ANÁLISIS DE LOS DATOS DE LA ADMINISTRACIÓN DE UNA PRUEBA PILOTO

Utilice el conjunto de datos de muestra ***PILOT1*** para llevar a cabo este ejercicio. Las claves de respuesta para esta prueba están en el libro de Excel ***ItemDataAllTests,*** en la hoja ***PILOT1.***

Este capítulo presenta el análisis de los datos de la prueba piloto utilizando el programa Item and Test Analysis (IATA). Se utilizará el flujo de trabajo **Response data analysis** para analizar los datos de respuesta con un archivo de claves de respuesta. Las etapas en el flujo de trabajo incluyen la carga de los datos, la especificación del análisis, el análisis de ítems, el análisis dimensional, el análisis de funcionamiento diferencial de los ítems y la selección de los ítems. No se calculan puntuaciones escaladas ni estándares de rendimiento porque es poco probable que la distribución de competencia en la muestra piloto sea representativa de la población.

Imagine el siguiente caso: un equipo de evaluación nacional y sus expertos curriculares han creado un conjunto de ítems nuevos de opción múltiple diseñados para evaluar las habilidades matemáticas de los estudiantes del 5.° grado. Los nuevos ítems de prueba, que se consideraron adecuados para representar el currículo nacional, se habían creado para reflejar las principales categorías de contenidos (conocimiento numérico, forma y espacio, relaciones, resolución de

problemas y la incertidumbre) determinadas por un comité directivo nacional. La versión final de la prueba, para la que se han previsto 50 ítems, está destinada a ser administrada a los estudiantes del 5.º grado de todos los niveles de competencia.

Como primer paso, el equipo de evaluación nacional administró una prueba de 80 ítems a un total de 262 estudiantes, seleccionados en siete escuelas de cada una de las tres regiones. El equipo utilizó más ítems de los que se iban a incluir en la prueba final previendo, como es habitual, que muchos de los temas propuestos para la prueba no funcionarán bien por diversos motivos. (Por ejemplo, pueden resultar demasiado fáciles o demasiado difíciles o algunas instrucciones pueden ser confusas). De hecho, algunos ítems pueden haber sido rechazados por comisiones de revisión antes de la prueba previa. En previsión de que algunos ítems puedan presentar problemas adicionales, se deben probar previamente al menos un 50 % más de ítems además de los que se requieren para la prueba final. Tenga en cuenta también que una prueba piloto tiene por objeto tanto poner a prueba los protocolos operativos de una encuesta como determinar la composición de los ítems en la prueba final.

El archivo de datos de respuesta de los estudiantes contiene las respuestas de opción múltiple de cada estudiante para cada uno de los 80 ítems, así como algunas variables relacionadas con la escuela (identificación de la región, identificación de la escuela, tipo de escuela, tamaño de la escuela) e información sobre los estudiantes (sexo de los estudiantes, idioma hablado en el hogar).

Haga clic en la primera opción del menú principal, **Response data analysis,** para introducir el flujo de trabajo del análisis (Figura 9.1).

FIGURA 9.1

El flujo de trabajo Response Data Analysis

Si en cualquier etapa del flujo de trabajo se detecta un error o resultados distintos de los esperados, vuelva al paso anterior o comience el análisis de nuevo desde el menú principal.

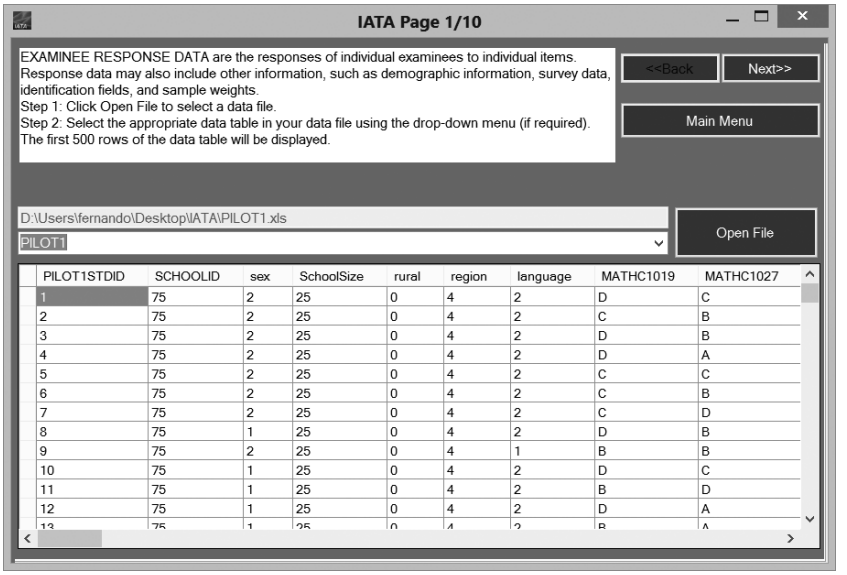
PASO 1: CARGAR LOS DATOS DE RESPUESTA

Independientemente de la vía de análisis elegido, debe cargarse en IATA los datos previamente recogidos o producidos (por ejemplo, los datos de la prueba piloto de evaluación nacional o un archivo de datos del ítem). IATA es flexible y tiene procedimientos y botones sencillos para cargar datos de respuesta, datos de ítems o ambas cosas. Independientemente de la vía de análisis o el tipo de datos, debe indicarse al programa IATA cuál es el archivo de datos que debe importar y qué datos del archivo debe usar. IATA puede importar datos en SPSS (Paquete estadístico para ciencias sociales) (*.sav), Excel (*.xls / *.xlsx), archivo de texto delimitado por tabulaciones (*.txt) y formatos separados por comas (*.csv). Dado que los archivos de datos de Excel pueden contener varias tablas separadas, se debe especificar la tabla que se desea importar.

La primera pantalla en esta vía de análisis le solicitará que importe un archivo de datos de respuesta a IATA. La interfaz de carga de datos se muestra en la Figura 9.2. Las instrucciones comienzan con las palabras **EXAMINEE RESPONSE DATA** para indicar que va a cargar datos que contienen las respuestas individuales de los estudiantes a ítems individuales. Debajo de las instrucciones hay dos cuadros: un resumen de ruta de archivos y un menú desplegable para seleccionar tablas de datos en el archivo seleccionado. A la derecha de estos cuadros está el botón de **Open File.** La tabla de la parte inferior de la interfaz muestra los datos de la fuente de datos seleccionada. Si existen más de 500 filas de datos, solo se mostrarán las primeras 500. Si el formato de datos seleccionado, como Excel o Access, implica la selección de varias tablas, entonces el nombre de la primera tabla del archivo de datos aparecerá en el cuadro de la lista desplegable. De lo contrario, aparecerá el nombre del archivo. En el caso de los archivos de datos de varias tablas es posible que los datos deseados no estén en la primera tabla. Revise el contenido de la tabla de datos que aparece

FIGURA 9.2

Interfaz de carga de datos de respuesta



PILOT1STDID	SCHOOLID	sex	SchoolSize	rural	region	language	MATHC1019	MATHC1027
1	75	2	25	0	4	2	D	C
2	75	2	25	0	4	2	C	B
3	75	2	25	0	4	2	D	B
4	75	2	25	0	4	2	D	A
5	75	2	25	0	4	2	C	C
6	75	2	25	0	4	2	C	B
7	75	2	25	0	4	2	C	D
8	75	1	25	0	4	2	D	B
9	75	2	25	0	4	1	B	B
10	75	1	25	0	4	2	D	C
11	75	1	25	0	4	2	B	D
12	75	1	25	0	4	2	D	A
13	75	1	25	0	4	2	B	A

en el área grande en la parte inferior de la interfaz para verificar que se han seleccionado los datos adecuados. Si la tabla activa no contiene los datos deseados, seleccione una tabla diferente haciendo clic en el menú desplegable.

Para este ejemplo, abra el archivo ***PILOT1.xls.***

1. Haga clic en **Open File** para seleccionar un archivo de datos. En el explorador de archivos, vaya a la carpeta del escritorio que contiene los datos de muestra de IATA.
2. Seleccione (o escriba) ***PILOT1.xls.***
3. Haga clic en **Open** o pulse la tecla **Enter**.

Cuando se abra el archivo, aparecerá un cuadro de diálogo para recordarle que debe confirmar que los datos seleccionados contienen los datos correctos de respuestas de los ítems. Haga clic en **OK** para continuar. Compruebe que los datos piloto de muestra se hayan cargado correctamente; la interfaz debe tener un aspecto similar al de la Figura 9.2. Los datos de la figura muestran los registros de cada estudiante que participó en la prueba piloto. Las primeras siete variables

de la izquierda representan información demográfica y de muestreo de los estudiantes.

- **PILOT1STDID**: código único de identificación del estudiante
- **SCHOOLID**: código único de identificación de la escuela
- **Sex**: sexo del estudiante (1 = femenino, 2 = masculino)
- **SchoolSize**: número total de estudiantes en la escuela
- **Rural**: ubicación de la escuela (0 = urbana, 1 = rural)
- **Region**: identificador numérico de la región geográfica
- **Language**: identificador numérico que indica si el idioma en que se imparte la enseñanza se habla en el hogar del estudiante

El primer ítem de la prueba de matemáticas aparece en la columna 8 y tiene la etiqueta **MATHC1019**. Desplácese por todo el conjunto de datos para comprobar si el archivo contiene los datos de 80 ítems; el ítem de la última columna tiene la etiqueta **MATHC1041**. Los nombres de los ítems son arbitrarios y no representan su posición en la prueba. La mayoría de las celdas tienen los valores A, B, C o D, que indican las opciones elegidas por los estudiantes. Las celdas que contienen el valor 9 indican que un estudiante no respondió al ítem.

Como en la mayoría de muestras piloto, los estudiantes representan una muestra de conveniencia en lugar de una muestra de probabilidad de la población total. En consecuencia, el archivo de datos de respuesta no tiene peso de muestreo.

Haga clic en **Next>>** después de verificar que se ha cargado el archivo correcto de datos de respuestas.

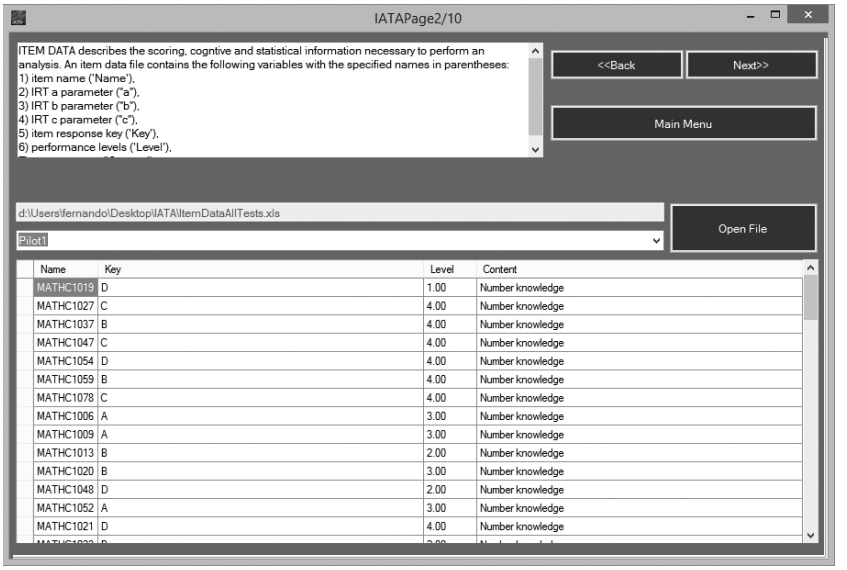
PASO 2: CARGAR LAS CLAVES DE RESPUESTA

Se deben cargar ahora las claves de respuesta de los ítems. Al igual que con los datos de respuesta, los datos de los ítems están en formato Excel en la carpeta de datos de IATA del escritorio.

1. Haga clic en **Open File** para seleccionar un archivo de datos. En el explorador de archivos, vaya a la carpeta del escritorio que contiene los datos de muestra de IATA.
2. Seleccione (o escriba) ***ItemDataAllTests.xls.***
3. Haga clic en **Open** o pulse **Enter**.

FIGURA 9.3

Datos del ítem para los datos de respuesta PILOT1



Cuando se abra el archivo, un cuadro de diálogo emergente le recordará que IATA estimará los parámetros de los ítems que faltan. Haga clic en **OK** para continuar. El archivo de datos seleccionado contiene tablas para todos los ejemplos de este libro. Asegúrese de que se ha seleccionado correctamente la tabla denominada ***PILOT1*** en el menú desplegable. Compruebe que se hayan cargado adecuadamente los datos del ítem correctos; la interfaz debe tener un aspecto similar al de la Figura 9.3. Para encontrar información sobre un ítem específico, ordene los elementos haciendo clic en el encabezado de la columna **Name**.

Después de confirmar que se han cargado los datos del ítem correctos, **haga** clic en **Next>>** para continuar.

PASO 3: ESPECIFICACIONES DE ANÁLISIS

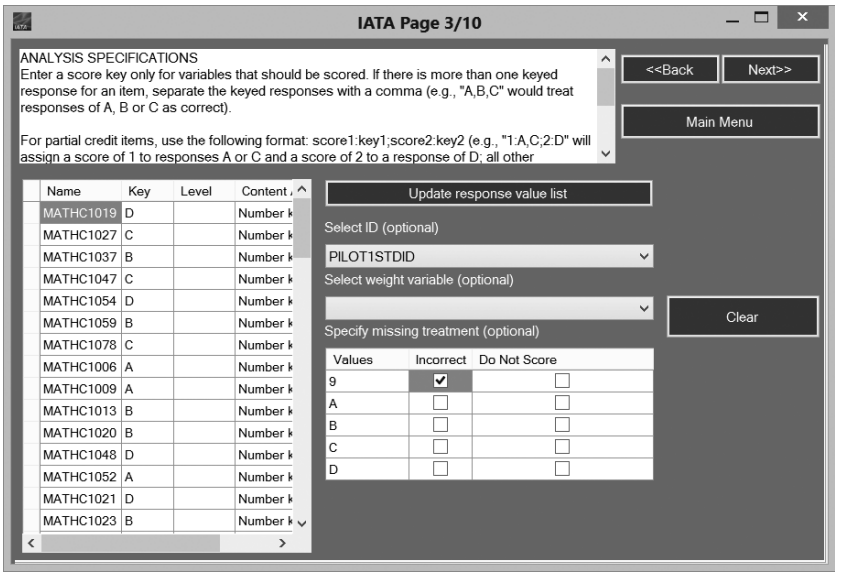
Todos los flujos de trabajo que utilizan datos de respuesta precisan determinadas especificaciones que afectarán a los resultados de todos

los análisis posteriores. Estas especificaciones incluyen información relacionada con las claves de respuesta, la identificación del encuestado, la ponderación del diseño de la muestra y el tratamiento de los códigos de datos faltantes. La interfaz para proporcionar estas especificaciones se muestra en la Figura 9.4. El panel grande de la izquierda contiene una tabla de los ítems de prueba en el archivo de datos de respuesta con los encabezados de las columnas **Name, Key, Level** y **Content**. Si se ha cargado un archivo de datos de ítems, la tabla contendrá sólo las variables que hayan sido identificadas como ítems de prueba; si no, la tabla contendrá todas las variables. Si se ha saltado la carga de un archivo de datos de ítems, deberá introducir las especificaciones de las claves de respuesta de forma manual para cada elemento de esta tabla (véase "Formatos de claves de respuesta" en el capítulo 8 de este volumen).

En la sección central de la interfaz está el botón **Update response value list.** Haga clic en este botón para cambiar las especificaciones de las claves de respuesta, ya sea introduciendo manualmente las claves

FIGURA 9.4

Especificaciones de análisis para los datos PILOT1



de respuesta o eliminando las que ya existen. Al hacer clic en este botón, IATA rellenará los dos menús desplegables con las listas de variables en los datos de respuesta a los que no se haya asignado una clave de respuesta y enumerará todos los valores de respuesta presentes para las variables identificadas como ítems de prueba. Si ha cargado un archivo de datos de ítems, estos menús ya se habrán rellenado con los valores.

Debajo del botón **Update response value list** hay varios controles para introducir especificaciones opcionales: un menú desplegable para especificar la variable de identificación (ID), un menú desplegable para seleccionar la variable de ponderación y una tabla para especificar el tratamiento de los códigos de valores faltantes. Es necesario especificar una variable ID para combinar los resultados de las pruebas producidas por IATA con otras fuentes de datos. La variable ID debe identificar de forma única a cada estudiante; si no se especifica una variable ID, IATA producirá una variable llamada **UniqueIdentifier** para este propósito. La variable de ponderación se utiliza para garantizar que las estadísticas producidas durante el análisis son apropiadas para el diseño de la muestra de la evaluación nacional, y como se ha señalado, no se aplicará en el análisis de los datos de prueba. Si no se proporciona ninguna variable de ponderación, IATA asumirá que todos los estudiantes reciben la misma ponderación (igual a 1).

Para indicar a IATA que un valor de respuesta es un código de respuesta faltante, haga clic en una de las casillas que aparecen junto a los valores en la tabla **Specify missing treatment.** Por defecto, IATA supone que todos los valores de respuesta representan respuestas reales de los estudiantes. Si se marca la casilla de la columna **Incorrect**, entonces IATA tratará ese valor como una respuesta no válida y se puntuará como incorrecta. Si la casilla de la columna **Do Not Score** está marcada, IATA tratará ese valor como omitido y el valor no afectará al resultado de las pruebas de los estudiantes. De forma predeterminada, si los datos de respuesta contienen celdas totalmente vacías o en blanco, IATA los tratará como incorrectos a menos que haya especificado manualmente **Do Not Score.**

Para esta guía, se han introducido tanto las claves como los datos de respuesta, por lo que la lista de ítems que se muestra en la Figura 9.4 contiene sólo las variables con las claves de respuesta en

los datos del ítem. Es recomendable revisar la tabla de claves de respuesta para confirmar que las claves y los demás datos sobre cada ítem sean correctos y estén completos; cualquier error en esta etapa generará más errores en las tareas posteriores del flujo de trabajo. Especifique los detalles de análisis adicionales en el centro de la pantalla. Utilice las siguientes especificaciones:

1. Utilice el primer menú desplegable para seleccionar **PILOT1STDID** como variable del identificador (el identificador que se asignó inicialmente a los estudiantes; véase la Figura 9.2).
2. Ya que estos datos no cuentan con una ponderación de muestreo, puede dejar la segunda casilla con menú desplegable en blanco.
3. Debido a que el valor de 9 se tratará como incorrecto, marque la casilla correspondiente en la tabla de valores de la sección **Specify missing treatment**. Aunque los datos ***PILOT1*** no tienen entradas en blanco, puede dejar la especificación predeterminada para tratar las entradas en blanco como incorrectas.

Una vez que se introduzcan las especificaciones, la interfaz debe tener un aspecto similar al de la Figura 9.4.

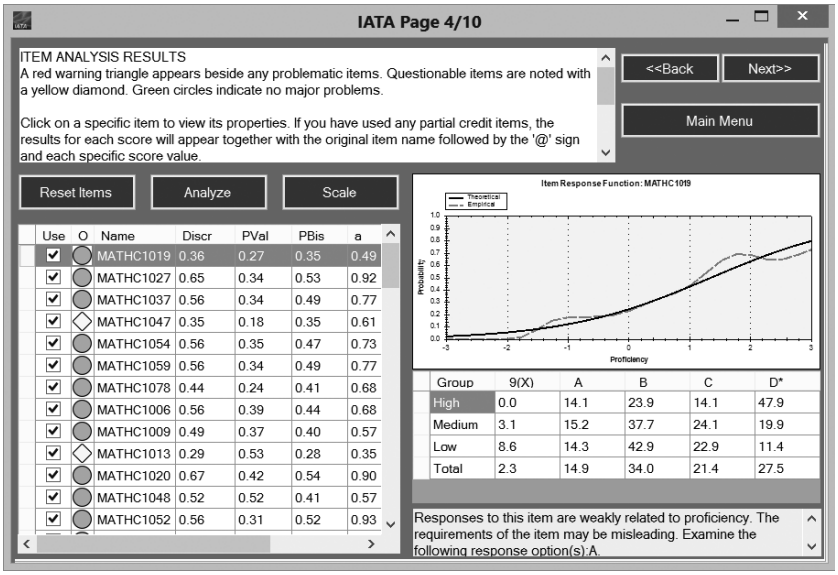
Confirme que las especificaciones son correctas y haga clic en **Next>>** para continuar. Los datos comenzarán a procesarse de forma automática. Las etapas de procesamiento son datos de configuración, puntuación, parámetros de estimación, escalas de la Teoría de Respuesta al Ítem (TRI), cálculo de la puntuación real y análisis factorial. A medida que el procesamiento continúa, la interfaz irá mostrando la etapa en la que se encuentra. En función de la velocidad del equipo utilizado y del tamaño del conjunto de datos, el análisis puede tardar en completarse entre unos segundos y varios minutos. Cuando IATA acaba el procesamiento, muestra los resultados en la interfaz de análisis de los ítems.

PASO 4: ANÁLISIS DE LOS ÍTEMS

Cuando el procesamiento de datos termine, la interfaz de análisis de los ítems se actualizará con los resultados que se muestran en la Figura 9.5. Mediante la interfaz, se puede acceder a estos resultados

FIGURA 9.5

Resultados del análisis de los ítems para los datos PILOT1, MATHC1019



así como ver y guardar información de diagnóstico sobre cada ítem de la prueba. En esta interfaz se muestran cuatro tipos de resultados:

- Estadísticas y parámetros estadísticos que describen cada ítem (a la izquierda)
- Una ilustración gráfica de la relación entre la competencia de los estudiantes y la probabilidad de responder correctamente a un ítem, también conocida como "función de respuesta al ítem" o FRI (en la parte superior derecha)
- Una tabla de contingencia en la que se describen las proporciones de estudiantes con puntuaciones altas, medias y bajas en las pruebas, que seleccionaron cada respuesta para los ítems, también conocida como "análisis de distractores" (en la parte central derecha)
- Un resumen en términos sencillos de los resultados del análisis de los ítems (en la parte inferior derecha)

La tabla en la parte izquierda de la interfaz de análisis de los ítems presenta información estadística, así como un símbolo que describe la adecuación global de cada ítem. El nombre de cada ítem aparece en

la columna **Name**, que se encuentra a la derecha de los símbolos de resumen. Examine los resultados detallados de un ítem individual mediante el uso de las flechas del teclado o el ratón para marcar la fila en la que aparece el ítem. Utilice las casillas de verificación en la columna **Use** para incluir cada fila o excluir ítems del análisis. Desmarque una de estas casillas de ítems para eliminar el ítem del análisis. Después haga clic en el botón **Analyze** para volver a ejecutar el análisis con el grupo de ítems reducido. (Los ítems eliminados seguirán en la lista pero habrá un triángulo rojo al lado de cada uno de ellos). Para devolver todos los ítems a su estado original, haga clic en **Reset Items** y **Analyze.** Tenga en cuenta que al hacer clic en **Reset Items** restablecerá *todos* los ítems. Para eliminar un ítem del análisis de forma permanente, elimine las claves de respuesta en la interfaz de las especificaciones del análisis. El botón **Scale** no recalcula los parámetros de los ítems, sino que solo calcula puntuaciones escaladas de la TRI para los datos de respuesta utilizando los parámetros de los ítems que ya han sido estimados o cargados en IATA desde un archivo de datos externo.

Estadísticas de los ítems

Las tres columnas a la derecha del nombre del ítem en la Figura 9.5 contienen estadísticas clásicas de los ítems: el índice de discriminación del ítem (**Discr**)**;** la facilidad del ítem (**PVal**), a la que se refiere a veces como dificultad del ítem, aunque los valores más grandes indican un ítem de prueba más fácil; y la correlación biserial puntual (**PBis**) (véase, por ejemplo, Crocker y Algina, 2006; Haladyna, 2004). Las últimas tres columnas son estimaciones de los parámetros de la TRI: el parámetro de pendiente (a), el parámetro de localización o umbral (b) y el parámetro de pseudoazar (c). Si estas tres columnas estuvieran ocultas a la vista, tendremos que usar la barra de desplazamiento situada en la parte inferior de la tabla.

En general, las estadísticas clásicas se pueden interpretar directamente. La facilidad del ítem (**PVal**) oscila entre 0 y 1 y describe el grado de facilidad de un ítem para la muestra proporcionada: un valor de 0 indica que ningún estudiante respondió correctamente, mientras

que un valor de 1 indica que todos los estudiantes respondieron correctamente. El índice de discriminación y la correlación biserial puntual proporcionan mediciones alternativas de la misma relación, es decir, la relación entre las respuestas a cada ítem y la puntuación global de la prueba. Para ambas estadísticas, el valor debe ser mayor que 0,2. Estas pautas no deben considerarse como absolutas, ya que los índices se ven afectados por factores ajenos a la discriminación de los ítems, como la precisión de la prueba general. Por ejemplo, la facilidad de los ítems tiende a limitar el valor absoluto del índice de discriminación y la correlación biserial puntual. Si la facilidad del ítem difiere sustancialmente de 0,5 (es menor que 0,2 o mayor que 0,8), el índice de discriminación y la correlación biserial puntual subestimarán la relación entre la competencia y el rendimiento de los estudiantes en un ítem de prueba. Aunque algunos ítems demasiado fáciles o difíciles tienden a reducir las relaciones observadas con la competencia, estos pueden cubrir importantes contenidos curriculares que deben ser incluidos en una prueba o que pueden —por ejemplo, en el caso de ítems sencillos— ser necesarios para mantener la motivación del estudiante durante la prueba. Por estas u otras razones, es conveniente incluir una cantidad relativamente pequeña de ítems tanto demasiado fáciles como difíciles.

Los parámetros TRI no se deben interpretar de forma aislada. Aunque cada uno describe un comportamiento específico de los ítems de las pruebas, la relación entre las respuestas a los ítems y la competencia general es el resultado de las interacciones entre los tres parámetros y los niveles de competencia de estudiantes individuales.

La mayoría de los ítems en el análisis actual tienen un círculo verde que indica que no presentan problemas importantes y son relativamente satisfactorios. Al desplazarse por la lista de ítems de la izquierda encontrará 13 ítems con símbolos de advertencia con forma de rombo (**MATHC1047**, **MATHC1013**, **MATHC1002**, **MATHC1070**, **MATHC1034**, **MATHC1035**, **MATHC1032**, **MATHC1010, MATHC1068, MATHC1046, MATHC1024, MATHC1058 y MATHC1030)**. Uno de los ítems (MATHC1075) tiene un símbolo de advertencia triangular y se considera potencialmente problemático. La práctica óptima consiste en examinar los resultados de todos

los ítems, independientemente del símbolo de resumen que asigne IATA. Esta guía se centra en unos pocos ejemplos.

Los resultados para el primer ítem se muestran por defecto en el gráfico y la tabla de la derecha. IATA ha asignado un círculo verde[1] a este ítem **MATHC1019**. En las siguientes secciones se describe cada uno de los resultados que produce IATA para este ítem.

Item Response Function

IATA muestra la **Item Response Function** de un ítem de prueba seleccionado en la ventana de gráficos que se encuentra a la derecha de la interfaz de análisis de ítems (véase la Figura 9.5). La revisión de la TRI suele ser más intuitiva que el examen de los parámetros TRI o las estadísticas de ítems para determinar la utilidad relativa de un ítem de prueba. Un ítem útil estará fuertemente relacionado con la competencia, indicada por una FRI (función de respuesta al ítem, del inglés *Item Response Function*) que tiene una pronunciada forma de S con una zona estrecha en la que la curva es casi vertical. La pendiente de la **FRI** para **MATHC1019** es sistemáticamente positiva, pero la relación es débil; ningún área tiene una pendiente especialmente pronunciada. Esta pendiente poco pronunciada corrobora el bajo índice de discriminación (**Discr** = 0,36) y la baja correlación biserial puntual (**PBis** = 0,35).

Como con cualquier método de modelado estadístico, la TRI es útil solo si los datos empíricos se ajustan al modelo teórico. IATA compone un gráfico de la FRI teórica producida para cada ítem o puntuación. Para ello utiliza los parámetros estimados así como la FRI empírica estimada directamente de las proporciones de respuestas correctas en cada nivel de competencia. El gráfico se puede utilizar para evaluar la idoneidad del uso de la TRI en la descripción de cada ítem. Si el modelo de TRI es adecuado, la línea roja discontinua será muy similar a la línea continua negra y las desviaciones entre las dos líneas serán inferiores a 0,05, en particular en la zona entre -1 y 1 en la que se encuentran muchos estudiantes. En el caso de **MATHC1019**, las FRI teóricas y empíricas son casi idénticas, lo que indica que, aunque la relación entre el ítem y la competencia puede ser débil, la FRI describe de forma precisa sus propiedades estadísticas.

Análisis de distractores

En la parte inferior de la derecha de la interfaz de análisis de los ítems de la Figura 9.5, IATA produce estadísticas para cada valor de respuesta (que incluye los códigos de valores faltantes y los valores de respuesta incorrectos) y un resumen textual del análisis. Las estadísticas se calculan por separado para grupos de estudiantes con rendimiento bajo, medio y alto, de acuerdo con el porcentaje de puntuación de la prueba. Los datos de la Tabla 9.1 representan un análisis de distractores de un ítem particular.

La relación de discriminación entre un ítem y la competencia puede ser baja o incluso negativa por diferentes razones. Estas pueden ser una mala redacción, instrucciones confusas, errores de muestreo y errores de tipeo o de codificación. El análisis de distractores se puede utilizar para detectar y remediar algunos de estos errores más comunes mediante la observación de los patrones en las respuestas del ítem. El ítem que funcione de forma correcta debe tener las siguientes características:

- La opción correcta de la columna (D), señalada con un asterisco (*), debería presentar un porcentaje elevado para el grupo alto y porcentajes sucesivamente más bajos para los grupos medio y bajo. **MATHC1019** cumple esta condición con unos valores de 47,9, 19,9 y 11,4 para los grupos alto, medio y bajo, respectivamente.
- En el caso del grupo con capacidades más bajas, el porcentaje que eligió la opción correcta (D) debería ser menor que el porcentaje que seleccionó cualquiera de las otras opciones. Todas las demás opciones incorrectas (A, B, C) para **MATHC1019** presentan este patrón.

TABLA 9.1
Análisis de distractores para MATHC1019, datos PILOT1

Grupo	9(X)	A	B	C	D*
Alto	0,0	14,1	23,9	14,1	47,9
Medio	3,1	15,2	37,7	24,1	19,9
Bajo	8,6	14,3	42,9	22,9	11,4
Total	2,3	14,9	34,0	21,4	27,5

Nota: El asterisco indica la columna de respuestas correctas.

- Los porcentajes de las columnas correspondientes a valores de respuesta incorrectos deben ser aproximadamente iguales para cada nivel de destreza y en general, comparados con el resto de valores de respuesta incorrectos. **MATHC1019** no sigue este patrón, ya que el porcentaje de encuestados que no respondió correctamente y que seleccionó la opción es considerablemente mayor que el porcentaje que eligió las opciones A o C.
- En el caso del grupo con capacidades más altas, el porcentaje que escogió la opción correcta (D) debería ser mayor que el porcentaje que seleccionó cualquiera de las otras opciones. **MATHC1019** cumple este patrón: 47,9 es mayor que los valores para las opciones A (14,1), B (23,9) y C (14,1).
- El porcentaje de los códigos de valores faltantes debe estar próximo a cero para todos los grupos. La proporción de estudiantes con respuestas faltantes (código 9) fue mayor para los de rendimiento bajo (8,6) que para los de rendimiento alto (0,0), lo cual indica que fue razonable la decisión de tratar el código como incorrecto (en lugar de omitido).
- Los códigos de respuesta faltante que son tratados como omitidos (indicado con **OMIT**) deberán tener los mismos porcentajes de estudiantes para cada nivel de capacidad. Este código no se utilizó para estos datos.

IATA ofrece un resumen textual sobre el rendimiento del ítem que incluye advertencias si la discriminación es inaceptablemente baja, en cuyo caso se indicará lo que se podrá hacer para mejorarlo. Por ejemplo, IATA identificará los distractores que no sean eficaces para conseguir el apoyo de los encuestados (o que tengan perfiles estadísticos similares a los de las respuestas correctas). Si IATA detecta cualquier problema habitual en los datos, se mostrará un resumen textual de los resultados en el cuadro de texto debajo de la tabla de análisis de los distractores.

En los resultados de **MATHC1019**, el resumen textual de la parte inferior derecha recomienda examinar la opción de respuesta codificada como A. En la tabla de análisis de distractores, se puede observar que la respuesta A está refrendada aproximadamente por la misma proporción de estudiantes de alto rendimiento y de bajo rendimiento, lo que indica que no funciona bien como distractor.

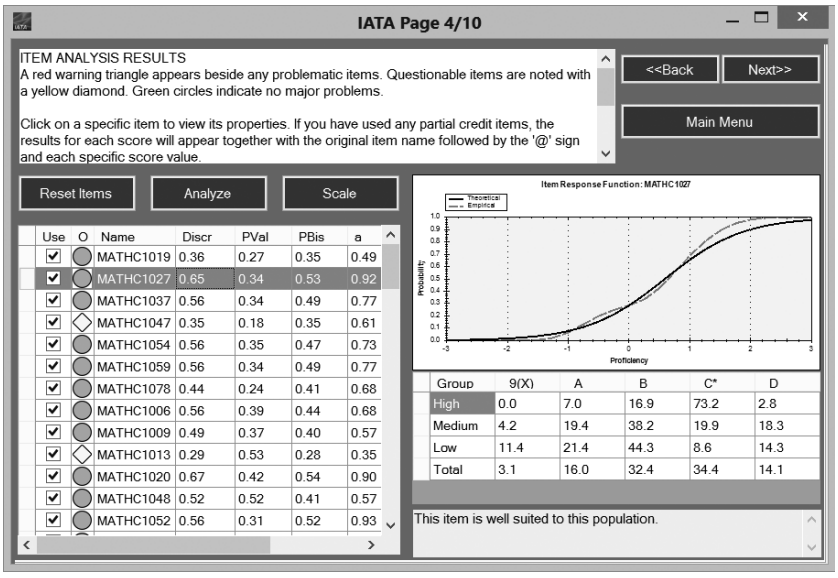
El análisis de distractores de los datos de la evaluación nacional también puede ser útil para los proveedores de cursos de educación para docentes y para el personal encargado del currículo ya que puede ayudar a identificar los errores y confusiones más habituales de los estudiantes. Las autoridades educativas encargadas de elaborar el currículo también pueden utilizar los datos para juzgar la idoneidad de material específico para un nivel particular.

Comparación de ítems

En comparación con el ítem anterior (véase la Figura 9.6), el segundo ítem de la prueba, **MATHC1027**, tiene una relación directa con la competencia. Este hecho se puede apreciar porque la FRI es mucho más pronunciada, y la discriminación (0,65) y los valores de correlación biserial puntual (0,53) son mayores. Las FRI teóricas y empíricas son casi idénticas, lo cual indica que el modelo de respuesta al ítem estadístico es adecuado para los datos de respuesta. La tabla de

FIGURA 9.6

Resultados del análisis de los ítems de los datos PILOT1, MATHC1027

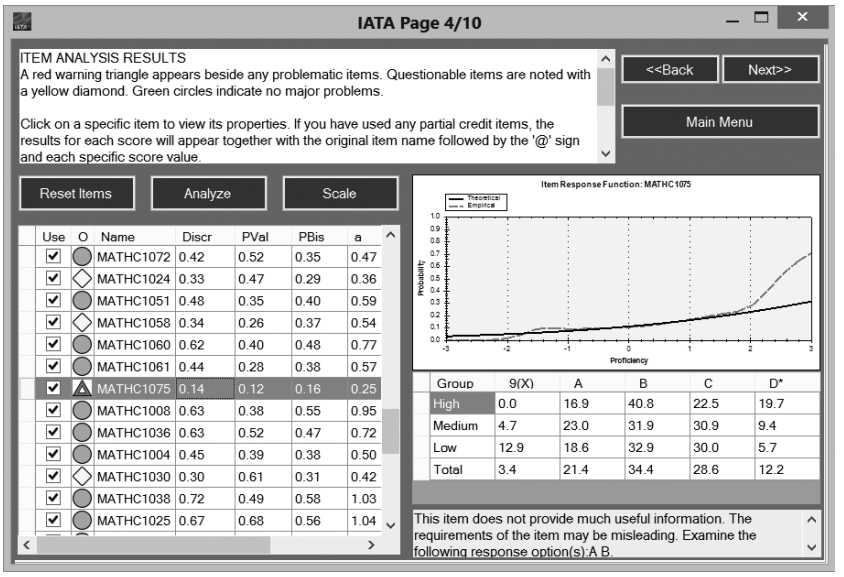


análisis de los distractores muestra que el 73,2 % de los estudiantes en el grupo con capacidad alta seleccionó la opción correcta (C), mientras que en el grupo con capacidad media fue el 19,9 % y en el grupo con capacidad baja el 8,6 %. Los estudiantes con rendimiento bajo tenían más probabilidades de seleccionar cualquiera de los valores de respuesta incorrecta (A, B y D), así como el código de respuesta faltante (9), que los estudiantes con rendimiento alto.

A diferencia de los dos ítems ya analizados, los ítems que tienen un símbolo de advertencia triangular son normalmente ítems mediocres cuya inclusión en las pruebas puede producir resultados erróneos o menos útiles. El número de ítems mediocres que aparece en una prueba piloto como esta se puede minimizar si se siguen las pautas de creación de ítems descritas en el volumen 2 de esta serie (Anderson y Morgan 2008). El único ítem con un símbolo de advertencia en estos datos es **MATHC1075** (véase la Figura 9.7). Al hacer clic en el ítem, se puede ver que los resultados indican una relación casi inexistente con la competencia, ya sea con respuestas correctas o incorrectas.

FIGURA 9.7

Resultados del análisis de los ítems de los datos PILOT1, MATHC1027

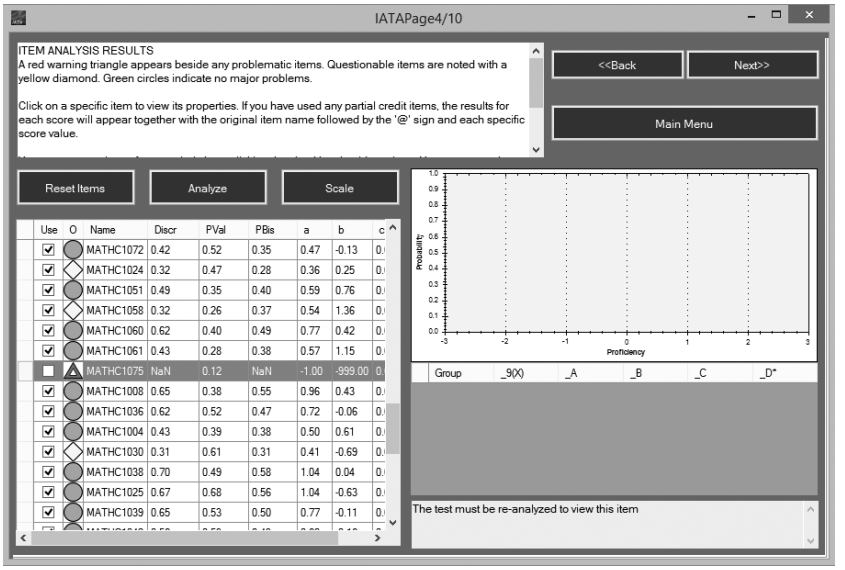


Aunque un código de respuesta faltante sigue estando relacionado con la competencia, el patrón esperado no era evidente. Los estudiantes en el grupo de capacidad baja no fueron los que más probabilidades tuvieron de seleccionar alguna de las tres opciones incorrectas, ni tampoco los estudiantes en el grupo de capacidad alta tenían menos probabilidades de hacerlo. El ítem fue especialmente débil en la discriminación entre los estudiantes con niveles medio y bajo. El índice de discriminación es bajo (0,14), al igual que el de la correlación biserial puntual (0,16). El ítem puede estar relacionado con la competencia pero la relación no se puede calcular porque muy pocos estudiantes contestaron de forma correcta (**PVal** = 0,12). La inclusión del ítem en la prueba tendería a aumentar la influencia de factores aleatorios en las puntuaciones obtenidas debido a que las respuestas al ítem no son claramente dependientes de la competencia. El hecho de incluir este ítem (y otros ítems problemáticos) en el análisis también podría reducir la precisión de los cálculos estadísticos de otros ítems de las pruebas, ya que las estadísticas y parámetros de los ítems se analizan utilizando las puntuaciones de las pruebas.

Los ítems se pueden eliminar del análisis haciendo clic en la casilla que se encuentra a la izquierda del nombre de cada ítem para desactivarlo. Después de eliminar un ítem, los resultados se deberán recalcular. Para ello haga clic en **Analyze** antes de eliminar cualquier otro ítem. La eliminación de un único ítem afectará a los resultados de todos demás ítems. Si hubiera muchos ítems problemáticos, habría que eliminarlos uno a uno ya que algunos ítems podrían aparecer marcados como problemáticos solo por la influencia de ítems mediocres en los resultados de los análisis. Si se eliminaran de forma accidental demasiados ítems, vuelva a comprobar cada ítem o haga clic en **Reset Items** sobre la lista de ítems para restablecer la lista de ítems completa. Para este ejemplo, elimine **MATHC1075** y vuelva a ejecutar el análisis para producir los resultados de la Figura 9.8, en la que los resultados de **MATHC1075** se resaltan después de la eliminación. Tenga en cuenta que los datos de **Discr** y **PBis** de este ítem se han sustituido por **NaN** (del inglés *Not a Number*, valor no numérico) o valores fuera de rango, los cuales no influirán en cálculos posteriores. La tabla de análisis de distractores de la derecha no aparece para los ítems eliminados, y en el resumen textual dice que los datos de la

FIGURA 9.8

Resultados del análisis de los ítems de los datos PILOT1, después de eliminar MATHC1075



prueba se deben volver a analizar. Las estadísticas de los ítems restantes permanecen casi sin cambios ya que solo se ha eliminado un ítem.

Puede seguir con la revisión de todos los ítems haciendo clic en cada fila en la lista de ítems o mediante las flechas del teclado de arriba y abajo. Tenga en cuenta que los resúmenes textuales proporcionados por IATA se basan únicamente en la evidencia estadística y no se sustentan en el contenido de los ítems. Un ítem al que IATA le otorga una puntuación baja puede no ser un ítem mediocre de forma general. Sin embargo, una calificación baja indica que el ítem no puede proporcionar información útil al utilizar la prueba actual con la población actual.

En general, las recomendaciones que proporciona IATA para editar o eliminar ítems se deben considerar en el contexto de la finalidad de la prueba y las razones iniciales para incluir un ítem específico. Aunque algunos ítems se pueden mantener a pesar de sus propiedades estadísticas, por ejemplo, por la necesidad de representar de forma adecuada los aspectos clave del currículo, se deben eliminar todos los

ítems con índices de discriminación negativos o bien asignarles claves nuevas (si la clave no se ha ingresado correctamente) antes de proseguir con otros análisis. Dichos ítems producen interferencias o variaciones no deseadas en los datos de respuesta al ítem y reducen la precisión de los cálculos de otros ítems. Si se eliminan algunos de los ítems supuestamente débiles durante el análisis de los datos piloto se ayudará a aumentar la exactitud de los resultados estadísticos. Sin embargo, la selección del último conjunto de ítems que sigue a la prueba piloto debe llevarse a cabo de forma conjunta por especialistas en la materia que trabajen en estrecha colaboración con la persona o el equipo responsable de la calidad general de la prueba de evaluación nacional.

Cuando termine de revisar todos los ítems, haga clic en **Next>>** para continuar.

PASO 5: DIMENSIONALIDAD DE LA PRUEBA

Uno de los supuestos estadísticos de la TRI, así como uno de los requisitos para la interpretación válida de los resultados de la prueba, es que el rendimiento en los ítems de las pruebas debe representar un único constructo o dimensión interpretable. Lo ideal es que una prueba nacional de rendimiento académico de un constructo tal como matemáticas o ciencias mida la dimensión o constructo único para el que ha sido diseñada, y no debe medir otros constructos o dimensiones tales como la habilidad lectora. El objetivo de la interfaz de la dimensionalidad de la prueba es detectar si no se cumple el supuesto de que (a) una única dimensión dominante afecta al rendimiento de la prueba y (b) las relaciones entre los rendimientos entre parejas o grupos de ítems se pueden explicar con esta dimensión dominante. En la mayoría de los casos, el segundo supuesto proviene del primero, aunque para pruebas largas (con más de 50 ítems) los grupos pequeños de ítems pueden ser dependientes de forma localizada sin que tengan un efecto evidente sobre la dimensionalidad global de la prueba.

El análisis de la dimensionalidad de la prueba determina el grado en que la prueba mide distintas dimensiones de competencia y en qué

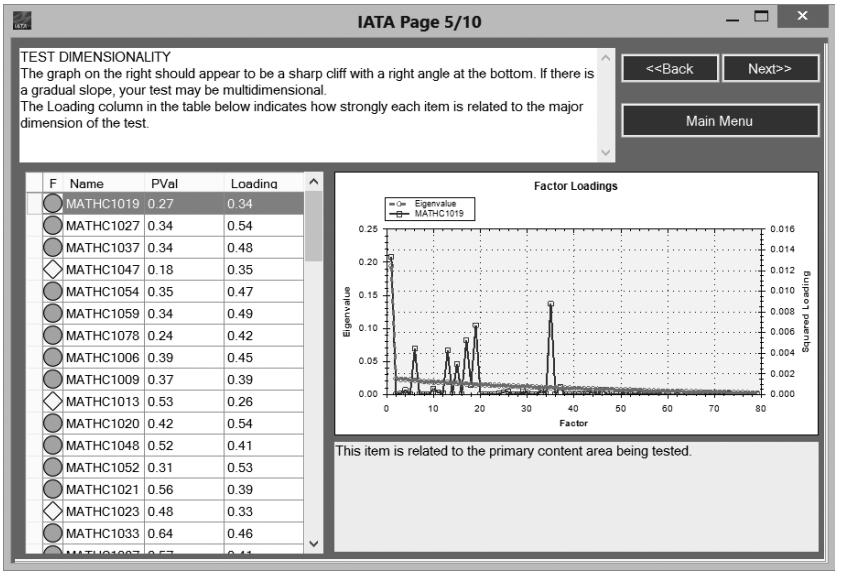
medida cada ítem está relacionado con cada dimensión. Cuanto menor sea el número de dimensiones que tienen una influencia importante en los ítems de la prueba, más válidas serán las interpretaciones de las puntuaciones de la prueba. Aunque esta evidencia es insuficiente para confirmar la validez de la prueba, puede proporcionar información importante sobre el contenido de determinados ítems. Otros aspectos sobre la validez, tales como la validez del contenido (la cual es muy importante en el contexto de una evaluación nacional), se consideran normalmente más importantes que los datos estadísticos para determinar la validez de una prueba o de un ítem. Anderson y Morgan (2008) proporcionan una descripción de procedimientos diseñada para asegurar la adecuación de la validez del contenido de una prueba.

Desde una perspectiva estadística, la estimación de las puntuaciones y los parámetros de TRI depende del concepto de probabilidad, que asume que la probabilidad de que se produzca un evento (por ejemplo, una respuesta correcta) depende de una única dimensión que representa la competencia. Si los ítems dependen de distintas dimensiones, entonces las puntuaciones y parámetros estimados serán incorrectos.

En la Figura 9.9, el gráfico de la derecha muestra el gráfico de sedimentación de la prueba general y las cargas factoriales al cuadrado del primer ítem, **MATHC1019**. En la parte de la izquierda de la interfaz hay una tabla similar a la que aparece en la interfaz de análisis de los ítems. El resumen de símbolos (como se describen en el capítulo 8 de este volumen) en la columna con la etiqueta **F** junto a la columna del ítem **Name** describe la idoneidad general de un ítem en términos de su relación con la dimensión primaria común a la mayoría de los ítems de la prueba. A la derecha de la columna **Name**, se muestra la facilidad clásica del ítem (**PVal**), junto con la carga del ítem en la dimensión primaria (**Loading**). La *carga*, que va de -1 a 1, es la correlación entre el rendimiento de cada ítem y la dimensión primaria de la prueba. Por ejemplo, el valor de 0,34 de **MATHC1019** indica que las respuestas puntuadas de este ítem tienen una correlación de 0,34 con la puntuación de la prueba general (porcentaje de respuestas correctas). No hay un valor ideal,[2] pero las cargas próximas a 1 indican que los ítems son mejores.

FIGURA 9.9

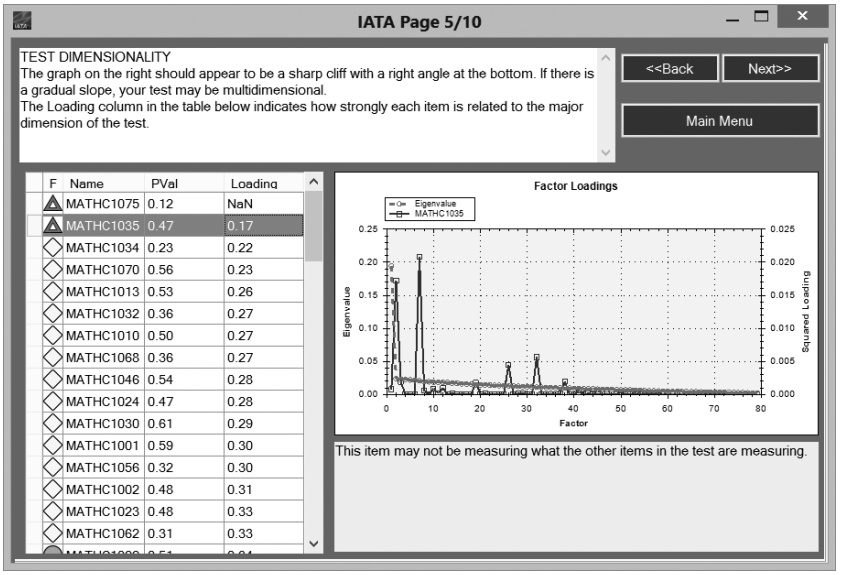
Prueba y dimensionalidad del ítem de los datos PILOT1, MATHC1019



Los resultados de la tabla se deben interpretar junto con los resultados gráficos mostrados en el lado derecho de la interfaz. El resultado principal mostrado en la ventana de gráficos es el gráfico de sedimentación *(scree plot)*, el cual describe la proporción de varianza (valor propio) explicado para cada dimensión potencial (por ejemplo, la habilidad lectora). Los marcadores en forma de círculo representan la influencia relativa de cada dimensión potencial (valor propio) [3] sobre los resultados generales de la prueba, mientras que la línea continua que conecta los marcadores en forma de rombo describe la influencia relativa de cada dimensión potencial sobre los elementos de prueba individuales (carga al cuadrado). La magnitud de los valores propios es menos importante que el patrón del gráfico de sedimentación. El gráfico de sedimentación de la prueba general debería tener un único valor propio grande a la izquierda seguido del resto de valores propios, los cuales deben ser relativamente pequeños y similares en magnitud (Figura 9.10). El patrón en forma de L del gráfico con solo dos segmentos de línea diferenciados indica que una única dimensión

FIGURA 9.10

Resultados de la dimensionalidad del ítem de los datos PILOT1, MATHC1035



común es responsable de los resultados de la prueba ***PILOT1.*** Cuanto mayor sea el número de segmentos de línea distintos necesarios para conectar el punto superior izquierdo con la línea casi horizontal inferior, en más dimensiones se basará el rendimiento de la prueba.

Al seleccionar cada ítem en la lista de la izquierda se mostrará el gráfico de sedimentación específico de cada ítem a la derecha. En teoría, el gráfico para ítems individuales debería ser similar al de la prueba general: el valor más alto en la línea específica de cada ítem en el gráfico debería estar en el extremo izquierdo (que corresponde a la dimensión principal de la prueba). Sin embargo, las características específicas de cada ítem pueden introducir diferentes patrones que no son necesariamente problemáticos. Por ejemplo, el ítem **MATHC1019** en la Figura 9.9 no es problemático; aunque algunas cargas diferentes de cero se producen en otras dimensiones, la carga más fuerte sucede en la dimensión primaria. En general, los resultados específicos del ítem solo se deben consultar si el rendimiento de la prueba se basa claramente en más de una dimensión (es decir, son visibles más de dos

segmentos de línea distintos). En ese caso, se deben identificar y examinar los ítems para los que los gráficos específicos tienen valores de carga al cuadrado correspondientes a las mismas dimensiones que los valores propios problemáticos.

Existe una salvedad al interpretar el gráfico de sedimentación: el efecto de la facilidad del ítem. En las pruebas en las que la mayoría de los ítems tienen una facilidad similar, los ítems con un grado de facilidad mucho mayor o menor que el resto de los ítems tienden a generar factores de dificultad artificial, en particular, con distribuciones anormales del porcentaje de puntuación de la prueba. Puede parecer que los ítems con una facilidad extrema definen un factor independiente simplemente debido a que algunos estudiantes (por ejemplo, con rendimiento alto o bajo) generarán unos patrones de respuesta que, aparentemente, tendrían una relación inusualmente estrecha en comparación con las relaciones entre los demás ítems. Sin embargo, estos factores de dificultad no son problemáticos en sí. La revisión de las cargas de los ítems puede ayudar a determinar si los factores secundarios son artefactos (errores de observación) o problemas reales. Para determinar si un factor secundario es un factor de dificultad, se deben examinar las cargas de los ítems con facilidad baja (< 0,2) o alta (> 0,8) (**PVal**). Si las cargas de estos ítems presentan unos picos que se corresponden con la posición del factor secundario, lo más probable es que el factor secundario sea un factor de dificultad y pueda ignorarse.

Cargas del ítem

El modelo TRI asume independencia local entre los ítems, lo cual significa que las respuestas a un ítem no deben depender de las respuestas a otro ítem. En teoría, conforme a la TRI, una prueba debe tener preguntas que sean independientes en todas las dimensiones excepto en la dimensión de la prueba principal. Una dependencia local significativa del ítem puede dar lugar a una estimación inexacta de los parámetros de ítems, las estadísticas de la prueba y la competencia de los estudiantes. Por ejemplo, una prueba de matemáticas que incluye una pregunta compleja de resolución de un problema puede asignar puntuaciones diferentes a cada uno de los pasos lógicos

requeridos para calcular la respuesta final. Si quien realiza la prueba contestó el paso 1 de forma incorrecta, esto influirá en la probabilidad de una respuesta correcta en cada uno de los pasos posteriores. Este conjunto de ítems de prueba dependientes no sería apropiado para el modelo de TRI. En este caso, el ítem debe tratarse de forma apropiada como ítem de crédito parcial único.

Debido a que la dependencia local es normalmente problemática solo en los ítems que tienen una escasa relación con la dimensión primaria, para utilizar esta interfaz de la forma más eficaz se deben ordenar los ítems en la columna **Loading** haciendo clic una vez en el encabezado de la columna[4] (véase la Figura 9.10) y comparar los ítems con cargas mediocres para identificar los picos comunes en los gráficos de carga del ítem. Si muchos ítems con cargas mediocres tienen picos en sus gráficos de carga que se corresponden con la misma dimensión, pueden tener cierta dependencia local. Debido a que estas estadísticas tienden a ser sensibles a los errores de muestreo, los resultados de la revisión estadística se deberían utilizar para motivar a revisiones más detalladas del contenido de los ítems en vez de para tomar decisiones definitivas.

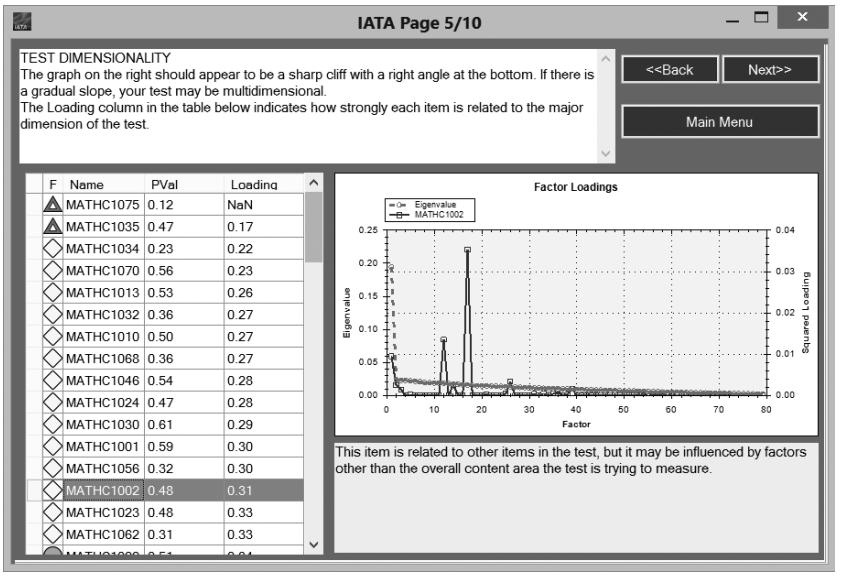
El ítem seleccionado después de la clasificación de los ítems es **MATHC1075**. Debido a que este ítem se eliminó del análisis en el paso previo del análisis del ítem, su carga es **NaN**, y no se muestran resultados para este ítem (el gráfico muestra solo el gráfico de sedimentación para toda la prueba). IATA asigna un símbolo de advertencia triangular para cualquier ítem cuya dimensionalidad pueda ser problemática por lo que respecta al cálculo de otras estadísticas. Tenga en cuenta que IATA ha marcado solo otro ítem (**MATHC1035**) con el símbolo de advertencia triangular (Figura 9.10). El ítem está relacionado débilmente con la dimensión primaria y tiene una relación notablemente más fuerte con la dimensión secundaria, lo cual indica que se puede medir una dimensión que es distinta de la dimensión de la mayoría de los ítems. Sin embargo, los resultados por sí mismos no son pruebas concluyentes para justificar la eliminación de este ítem de la prueba. Los expertos en currículos y profesores experimentados deberán revisar cualquier ítem estadísticamente problemático para determinar si una cuestión relacionada con el contenido justifica su eliminación o revisión.

IATA asigna un símbolo de advertencia con forma de rombo para cada ítem que tenga una carga más fuerte en una dimensión secundaria que en una dimensión de prueba primaria; estos ítems no serán problemáticos en cálculos posteriores. Se muestra un ejemplo típico en la Figura 9.11 para **MATHC1002**. Este ítem está relacionado con varias dimensiones, pero debido a que estas dimensiones tienen tan poca influencia en los resultados generales de la prueba, como indican los relativamente pequeños valores propios (línea discontinua) que se corresponden con los picos de las cargas fuertes (línea continua), la determinación de si la dimensionalidad del ítem es aceptable o no debería ser una cuestión del contenido de la prueba en lugar de las estadísticas.

Todas las pruebas son multidimensionales hasta cierto punto porque no todos los ítems pueden probar exactamente lo mismo. Por lo tanto, si el gráfico de sedimentación no indica ningún problema, los efectos de cualquier multidimensionalidad a nivel del ítem o codependencia serán probablemente insignificantes. Para

FIGURA 9.11

Resultados de la dimensionalidad del ítem de los datos PILOT1, MATHC1002



este ejemplo, se conservan todos los ítems para los análisis posteriores ya que el gráfico de sedimentación general no indica ningún problema.

Una vez que haya terminado la revisión de los ítems, haga clic en **Next>>** para continuar con la interfaz del análisis del funcionamiento diferencial de los ítems (FDI).

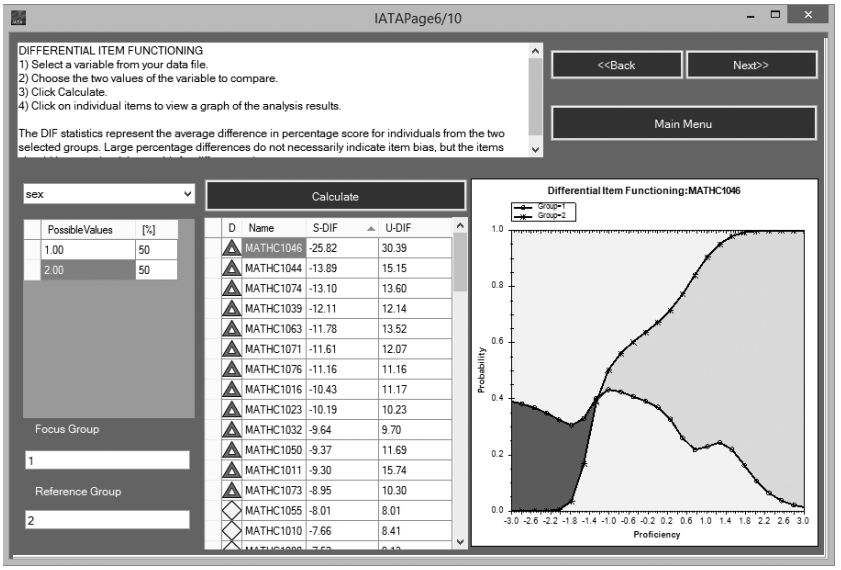
PASO 6: FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

El análisis del FDI examina el grado en el que la FRI de un ítem es estable a través de diferentes grupos de estudiantes. Si la FRI es diferente para dos grupos (por ejemplo, hombres y mujeres), las puntuaciones calculadas mediante su uso pueden estar sesgadas para un grupo o para estudiantes dentro de intervalos específicos de competencias en el grupo. El análisis del FDI comprueba las diferencias en la competencia de la media del grupo, lo que significa que las ventajas y desventajas relativas expresadas por los resultados del FDI son independientes de las diferencias en la competencia promedio en los grupos. Por ejemplo, si se está interesado en el alcance del sesgo del género en un ítem de prueba en particular, los resultados del análisis del FDI indicarán si el ítem está influenciado en favor de los niños o las niñas *después* de tener en cuenta la diferencia entre los géneros de la puntuación de la prueba general.

La interfaz de análisis del FDI se muestra en la Figura 9.12. En la parte izquierda, se encuentra el conjunto de cuatro controles utilizados para especificar el análisis. El menú desplegable de la parte superior le permite seleccionar una variable de la lista en los datos de respuesta que no sean ítems de la prueba. Una vez que seleccione la variable, IATA presentará sus valores exclusivos en la tabla **Possible Values**, así como el porcentaje (no ponderado) de estudiantes con cada valor. Para seleccionar los grupos que desee comparar, haga clic en primer lugar en el valor del grupo muestra elegido que desee y, a continuación, en el valor que represente al grupo de referencia. La especificación del grupo muestra elegido y el grupo de referencia determina la forma en que se calcularán los resúmenes estadísticos; las estimaciones utilizan la distribución de competencia de muestra

FIGURA 9.12

Resultados del análisis del FDI de los datos PILOT1 por sexos, MATHC1046



del grupo muestra elegido para calcular las estadísticas de estabilidad y sesgo medias. Para cambiar los grupos muestra elegidos y de referencia, haga clic en los valores de la tabla **Possible Values**; los valores asignados a los grupos muestra elegidos y de referencia se actualizarán en los cuadros de texto de la parte inferior izquierda. Las estadísticas del FDI son las más sensibles a los grupos muestra elegidos, debido a que la práctica habitual consiste en asegurarse de que el grupo muestra elegido sea un grupo minoritario o tradicionalmente desfavorecido.

Para este ejemplo, se llevó a cabo un análisis de FDI con la variable **Sex** para comprobar si las estudiantes (código 1) se encuentran en desventaja en comparación con sus homólogos varones (código 2). Para especificar este análisis y revisar los resultados, realice los siguientes pasos:

1. En el menú desplegable de la izquierda, seleccione la variable **Sex**. La tabla que se encuentra debajo se rellenará con los valores **1.00** y **2.00**, con valores del 50 por ciento para cada valor, lo que indica

que la muestra cuenta con la misma representación de ambos sexos.

2. En la tabla de valores, haga clic en el valor **1.00**. Esto hará que el valor de **1.00** (que representa a mujeres) se presente como grupo de enfoque en el cuadro de texto que se encuentra debajo.
3. En la tabla de valores, haga clic en el valor **2.00**. Esto hará que el valor de **2.00** (que representa a hombres) se presente como grupo de referencia en el cuadro de texto que se encuentra debajo.
4. Haga clic en **Calculate** y espere a que se complete el cálculo.
5. Una vez que se haya realizado el cálculo, en la lista de ítems, haga clic en el encabezado de la columna **S-DIF** para ordenar todos los ítems por el valor de la estadística S-DIF.

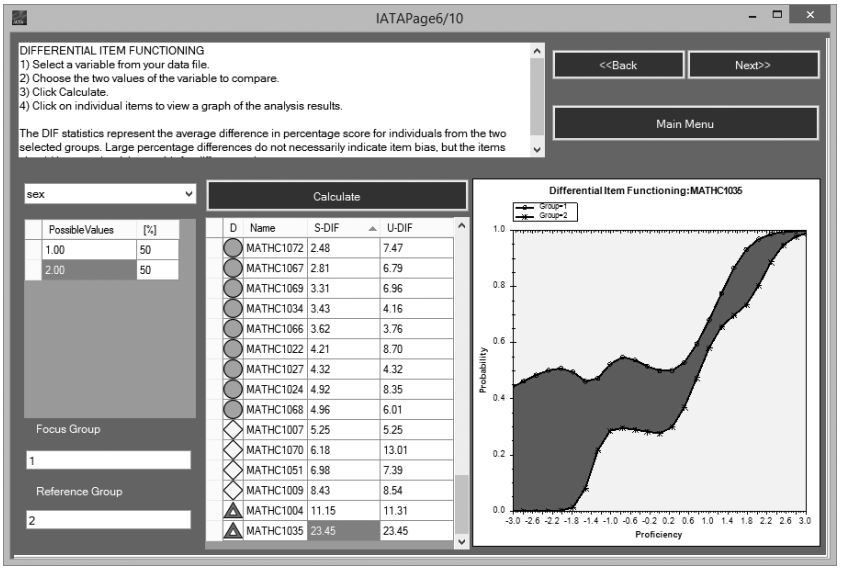
Una vez que realice estos pasos, podrá ver una interfaz similar a la de la Figura 9.12. En este ejemplo, IATA marca 15 ítems con un símbolo de advertencia o precaución. Para cada ítem, se calculan dos estadísticas: S-DIF y U-DIF. S-DIF describe la diferencia vertical media entre los grupos (muestra menos referencia), mientras que U-DIF describe las diferencias absolutas medias entre los grupos. El valor de la estadística U-DIF siempre es positivo y el valor absoluto es mayor que el de S-DIF. Incluso aunque un grupo no refleje una ventaja sistemática (el valor de S-DIF es próximo a 0), un ítem puede estar más estrechamente relacionado con la competencia en un grupo, lo que haría que la estadística U-DIF fuese mayor.

MATHC1035 es un ejemplo de ítem con FDI constante, donde los valores absolutos de S-DIF y U-DIF son idénticos (véase la Figura 9.13). En este ítem, la ventaja femenina es evidente en todo el intervalo de competencia. La diferencia constante sugiere que las mujeres son más propensas a presentar un rendimiento más adecuado en este ítem que los hombres, a pesar de que tengan el mismo nivel de competencia. La estadística S-DIF indica que, en promedio, la probabilidad de que las mujeres respondieran correctamente era más de un 23 por ciento mayor que en el caso de los hombres con una competencia comparable.

Con el análisis de FDI, las estadísticas y las figuras tienden a ser muy sensibles a los errores de muestreo, que pueden hacer que los ítems mostrados presenten diferencias que no aparecerían en muestras de

FIGURA 9.13

Resultados del análisis de FDI de datos PILOT1 por sexos, MATHC1035

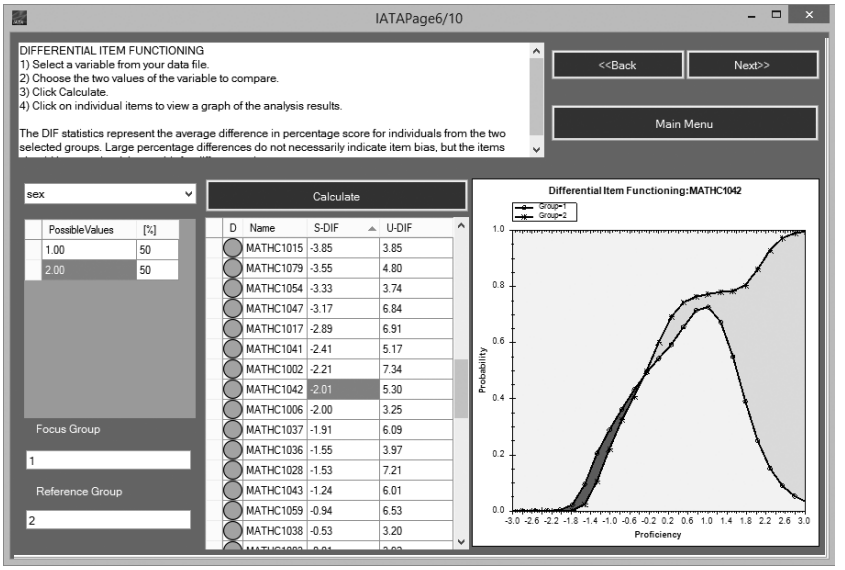


mayor tamaño. IATA asigna un símbolo de advertencia cuando el coeficiente de variación de muestreo[5] es inferior a 0,2 (lo que indica que es muy probable que la diferencia observada no se deba a un error de muestreo) o cuando exista una diferencia muy amplia en S-DIF o U-DIF que debería examinarse en muestras pequeñas.

Debido a la sensibilidad a errores de muestreo, en ocasiones, los resultados gráficos pueden ser engañosos. Cuando el número de encuestados en los extremos superior e inferior de la escala de competencia es reducido, las respuestas de uno o dos estudiantes pueden decidir el aspecto de los gráficos en estos extremos. Debido a que los resúmenes estadísticos ponderan el cálculo según el número de estudiantes de los grupos muestra elegidos en cada nivel de competencia, no se ven afectados por los errores aleatorios en la misma medida que los gráficos. El gráfico de los resultados de **MATHC1042** en la Figura 9.14 ofrece un ejemplo de la forma en que los resultados gráficos pueden ser engañosos en algunos casos. Aunque el gráfico sugiere una desventaja muy acentuada

FIGURA 9.14

Resultados del análisis de FDI de los datos PILOT1 por sexos, MATHC1042



para las mujeres (la zona ligeramente sombreada), la estadística S-DIF real (-2,01) indica que se trata de una desventaja relativamente débil.

También pueden encontrarse pruebas del FDI cuando el contenido específico del ítem no coincide de la misma manera con la dimensión de prueba primaria que los demás ítems. Por ejemplo, en matemáticas, un objetivo de aprendizaje común para los estudiantes más jóvenes consiste en reconocer las herramientas de medición para las diferentes unidades (por ejemplo, centímetros, kilogramos o grados centígrados). Los estudiantes de zonas remotas o desfavorecidas, incluso aunque sean buenos en matemáticas, pueden no estar expuestos a estas herramientas en la misma medida que los estudiantes de zonas urbanas. Por lo tanto, es posible que se encuentren sistemáticamente en desventaja con los ítems de la prueba que requieran estos conocimientos específicos. No obstante, esta desventaja no es una propiedad de los ítems de la prueba; es consecuencia de una desventaja específica en la competencia. Antes de llegar

a ninguna conclusión sobre sesgos para estudiantes concretos, es necesario que expertos en contenidos curriculares que tengan en cuenta las posibles diferencias étnicas, geográficas o de sexo examinen los ítems de la prueba para confirmar que los indicios sobre el sesgo de las pruebas estadísticas coinciden con las evidencias del análisis del contenido.

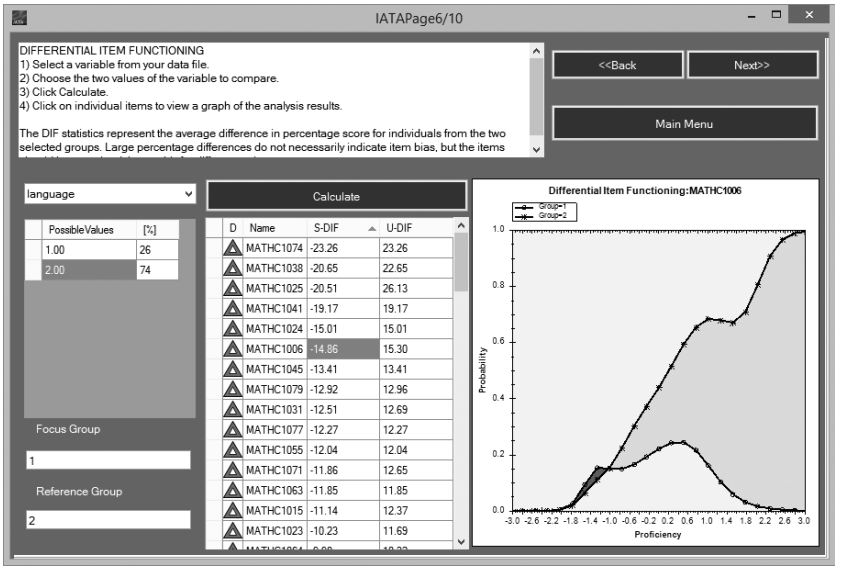
El análisis de FDI debe llevarse a cabo para todos los grupos y características demográficos que se compararán en los principales análisis de los resultados. La presencia del FDI en relación con una característica no suele estar relacionada con la presencia o ausencia del FDI con respecto a otra característica. Por lo general, las variables más importantes que deben tenerse en cuenta para el FDI son las variables de estratificación de la muestra (como **Region**) o las variables del cuestionario de contexto. Los datos ***PILOT1*** cuentan con tres variables demográficas: **Sex**, **Language** y **Region**. A modo de ejercicio independiente, puede llevar a cabo análisis de FDI similares para **Language** y **Region** realizando los mismos pasos que para el análisis de FDI de **Sex**, y asegurándose de seleccionar el grupo minoritario como grupo muestra elegido y de hacer clic en **Calculate** para actualizar los resultados.

La Figura 9.15 muestra un resultado de FDI común cuando una prueba se lleva a cabo en el idioma materno de algunos estudiantes, pero no de otros. Los resultados proceden de un análisis de FDI del ítem **MATHC1006**. Este ítem es un ejemplo extremo de FDI en el que la respuesta correcta está estrechamente relacionada con la competencia lingüística de una población (en este caso, **Language** = 2) y tiene una relación débil o inexistente en el caso de la otra (**Language** = 1).

El análisis de FDI en IATA puede servir de herramienta de investigación para determinar si grupos de estudiantes específicos tienen problemas con subdominios concretos. El análisis de FDI también puede facilitar la comprensión de las diferencias que pueden introducirse en las versiones en diferentes idiomas de una prueba que se haya traducido. Las pruebas estadísticas de FDI pueden utilizarse para ayudar a los traductores a corregir los errores de traducción revelados durante la prueba piloto.

FIGURA 9.15

Resultados del análisis de FDI de los datos PILOT1 por idioma de los estudiantes, MATHC1006



El propósito primario del análisis de FDI consiste en fomentar el debate y la revisión de los ítems de las pruebas piloto y actuar como guía para la interpretación de los resultados. Para cada análisis de FDI que se lleve a cabo, IATA guarda los resultados en una tabla de datos.[6] Estos resultados, así como los gráficos que resulten especialmente interesantes, deben copiarse,[7] guardarse y compartirse con especialistas en contenido curricular para determinar posibles explicaciones para el patrón de diferencias entre los grupos muestra elegidos y los grupos de referencia. Si se establece de forma clara que un ítem está sesgado, debe eliminarse de las especificaciones del análisis de la página 2 de IATA. Asimismo, sería necesario repetir los análisis anteriores de IATA. Finalmente, debido a que los resultados de los análisis de FDI son muy susceptibles a los errores de muestreo, cualquier decisión relacionada con incluir o no un ítem de prueba concreto en la versión final de una prueba basándose en la sospecha de un sesgo debe contar con una justificación sólida en relación con el plan de estudios o el

contenido. En esta guía, continuaremos sin eliminar ninguno de los ítems de la prueba.

Tras llevar a cabo los análisis de FDI y revisar los resultados, haga clic en **Next>>**.

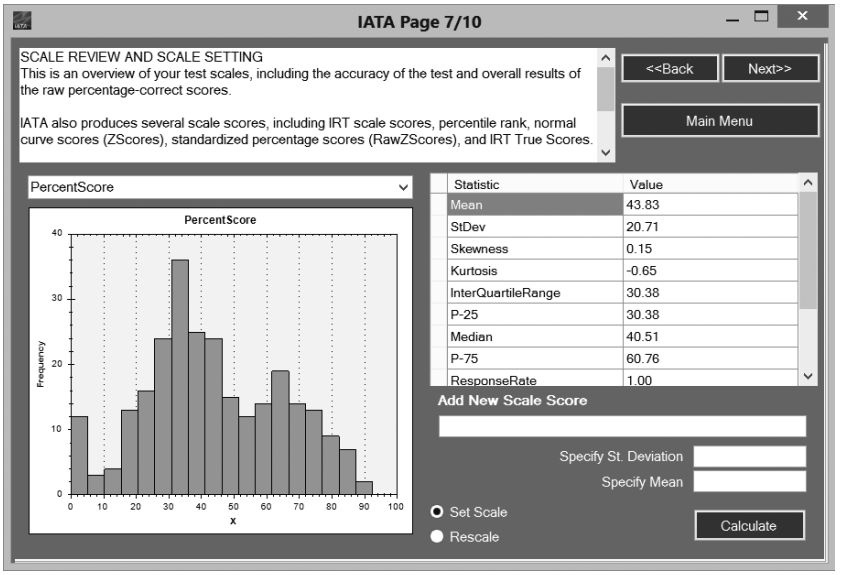
PASO 7: REVISIÓN DE ESCALA

La técnica que consiste en el desarrollo de métricas numéricas para la interpretación del rendimiento de las pruebas se conoce como *"creación de escalas"*. IATA genera informes de los resultados de las pruebas con las siguientes puntuaciones escaladas: **PercentScore, Percentile, RawZScore, ZScore, IRTscore** y **TrueScore**. Estas escalas se encuentran descritas de forma más detallada en la Tabla 8.1. El rendimiento según estas escalas predeterminadas se resume en una escala del 0 al 100 o en la escala estándar, que tiene una media de 0 y una desviación estándar de 1. Utilice la escala que le resulte más útil para comunicar los resultados. Es posible que las diferentes entidades interesadas prefieran distintos tipos de escalas. Por lo general, **IRTscore** es la más útil para la mayoría de los propósitos, pero presenta el inconveniente para la comunicación de que la puntuación de aproximadamente la mitad de los estudiantes es inferior a cero. Ya que es posible que muchas de las entidades interesadas no sepan interpretar las puntuaciones en escalas negativas, es preferible crear una nueva escala para que el valor de las puntuaciones de los estudiantes no sea inferior a cero.

En la Figura 9.16 se muestra la interfaz para la revisión de las puntuaciones escaladas y la creación de puntuaciones escaladas adicionales. En la parte izquierda, aparece un menú desplegable y una ventana de gráficos. Puede seleccionar cualquiera de los tipos de puntuación escalada del menú desplegable para representar la distribución de la puntuación escalada seleccionada. En la figura se muestra el gráfico de la puntuación escalada seleccionada: **PercentScore**. En la parte derecha, aparece un panel en el que se presentan los resúmenes estadísticos de la puntuación seleccionada. En la parte inferior derecha aparece un conjunto de controles para ajustar la escala de **IRTscore** mediante la aplicación de una desviación estándar y una desviación

FIGURA 9.16

La interfaz de revisión y establecimiento de escalas



media nueva. El procedimiento de ajuste de la escala solo se aplica a **IRTscore**, que es la salida de puntuación primaria de IATA.

Distribuciones de puntaciones de pruebas e información de pruebas

IATA muestra las distribuciones de puntuaciones en forma de histogramas, donde cada barra representa un intervalo de puntuaciones y la altura de cada barra representa la proporción de estudiantes con puntuaciones dentro de dicho intervalo. En el caso de los tipos de puntuaciones expresados en escalas con medias aproximadas de 0 y desviaciones estándar aproximadas de 1 (**StandardizedZscore**, **RawZScore** e **IRTscore**), IATA también representa la función de información de la prueba en forma de línea continua. La función de información de la prueba describe la precisión de la prueba en diferentes niveles de competencia en la escala estándar que se ha utilizado para los ítems. Está inversamente relacionada con el error estándar de medición; si la información de la prueba es alta, el error

estándar de medición será bajo. La función de información de la prueba debe interpretarse en relación con las necesidades o el propósito específicos de la prueba. Por ejemplo, si el propósito de la prueba es la identificación de estudiantes con competencias bajas, una prueba que resulte más precisa para estudiantes con niveles de competencia altos no sería adecuada y no serviría como medición apropiada para la identificación de estudiantes con competencias bajas. Por lo general, el error medio de medición para todos los estudiantes se minimiza si la función de información de una prueba es ligeramente mayor, pero tiene aproximadamente la misma forma y ubicación que la distribución de la competencia para los estudiantes sometidos a la prueba. La comparación de la función de información de la prueba con la distribución de las puntuaciones de la prueba puede indicar si el diseño de la prueba se beneficiaría al modificar el equilibrio de los ítems con una mayor precisión para los que tengan un rendimiento alto o bajo.

Resúmenes estadísticos

IATA produce los siguientes resúmenes estadísticos para cada puntuación de la prueba:

1. Media
2. Desviación estándar
3. Asimetría
4. Curtosis
5. Rango intercuartil
6. Percentil 25
7. Mediana
8. Percentil 75
9. Tasa de respuesta
10. Confiabilidad
11. Número total de encuestados
12. Número de ítems en la prueba
13. Número de ítems incluidos en el análisis

Las ocho primeras estadísticas describen la distribución de las puntuaciones estimadas. Para ver las últimas cinco filas, utilice la

barra de desplazamiento que se encuentra a la derecha del cuadro de estadística y valor en la Figura 9.16. Estas estadísticas le ayudan a determinar la idoneidad de las puntuaciones escaladas para diferentes propósitos (por ejemplo, análisis estadístico secundario o informes por cuantiles). Las últimas cinco estadísticas describen las condiciones bajo las cuales se llevó a cabo el análisis y ofrecen una estimación holística de la prueba, que debe comprobarse para confirmar que el análisis se realizó siguiendo las especificaciones correctas. La *tasa de respuesta* describe el promedio de respuestas válidas (con valores) en cada uno de los ítems. La *confiabilidad* es una medición de resumen general de la precisión media de una prueba para una muestra determinada de estudiantes. Tanto la tasa de respuesta como la confiabilidad van de 0 a 1 y deben ser tan altas como sea posible. El número total de ítems incluidos en el análisis refleja el hecho de que algunos ítems pueden eliminarse del análisis si se consideran inadecuados a causa de una mala redacción, debido a que resulten confusos para los estudiantes o por otras insuficiencias técnicas. Para esta guía, el número de encuestados es 262, el número de ítems es 80 y el número de ítems aceptables es 79. (**MATHC1075** se eliminó del análisis).

La interfaz de escalas resulta más útil para las administraciones de evaluaciones finales que para las pruebas piloto. Debido a que la muestra de la prueba piloto no ponderada no es representativa, las distribuciones de resultados no deben generalizarse al rendimiento de la población. Además, debido a que no se crearán informes de las puntuaciones de las pruebas, no es necesario generar puntuaciones escaladas derivadas. Los resultados adicionales de la interfaz de la escala no son pertinentes para el análisis de los datos ***PILOT1***. Haga clic en **Next>>** para continuar a la siguiente tarea.

PASO 8: SELECCIÓN DE ÍTEMS DE PRUEBA

La selección óptima de ítems mediante IATA se encuentra disponible cada vez que se carga o se crea un archivo de datos de ítems durante un análisis de datos de respuesta. IATA puede seleccionar automáticamente los ítems en función de las características del ítem estadístico para generar la prueba más eficiente para la duración y el propósito de

una prueba determinada. El principio básico que subyace a la creación de pruebas basadas en TRI es que el diseñador de la prueba tiene ciertas expectativas en relación con el grado de errores de medición que una prueba debería tener en diferentes niveles de competencia, así como los requisitos relacionados con el equilibrio del contenido que debe incluirse en la prueba.

Por lo general, cuantos más ítems haya en una prueba, más información se puede generar sobre los niveles de competencia de los examinados. No obstante, por desgracia, las pruebas con demasiados ítems no suelen resultar prácticas ni deseables; pueden generar perturbaciones innecesarias en las escuelas, así como fatigar a la persona que lleva a cabo la prueba y deteriorar la motivación del estudiante, lo que hace que los resultados sean menos precisos. El desarrollo, la administración, la puntuación y el procesamiento de unas pruebas excesivamente largas también supone unos costes elevados. Con el fin de maximizar la eficiencia, las pruebas solo deben incluir los ítems de prueba más informativos del grupo de ítems disponibles. IATA puede ayudar a desarrollar pruebas con el menor número posible de ítems de prueba necesarios para cumplir los propósitos de los responsables políticos y otras entidades interesadas.

La determinación de un nivel de error estándar aceptable depende del propósito de las evaluaciones. Lo ideal sería crear una prueba que ofrezca un alto nivel de precisión en todos los niveles de competencia. No obstante, esto requeriría un gran número de ítems, lo que aumentaría el tiempo que cada estudiante debe dedicar para completar la prueba. Esto podría provocar la reducción de la validez de los resultados de la prueba al permitir que la fatiga y el aburrimiento influyan en las puntuaciones de la prueba. Si se trata de una prueba referida a una norma, se precisa información detallada (y un menor número de errores de medición) para todos los niveles de competencia. Por el contrario, si se trata de una prueba referida a un criterio, solo se requiere información en relación con los umbrales de competencia en los que se toman las decisiones.

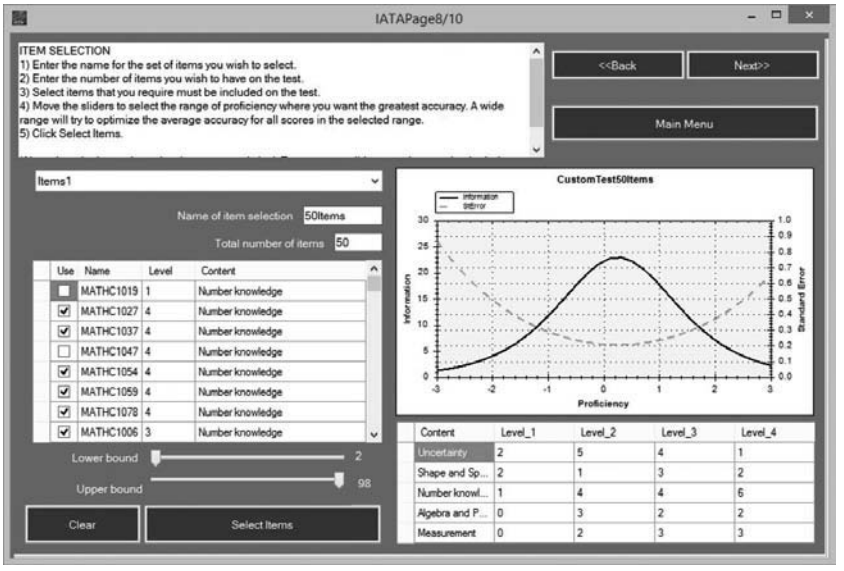
No obstante, la selección de ítems en la etapa piloto no debe determinarse únicamente en función de los resultados de los análisis estadísticos. La validez de la interpretación de los resultados es el aspecto más importante que es necesario tener en cuenta al crear pruebas de

evaluación nacional. Las puntuaciones de las pruebas deben representar de forma adecuada y precisa el dominio que se esté midiendo. Las herramientas más importantes para el mantenimiento de la validez de la prueba son los marcos teóricos y la tabla de especificaciones o documento de especificaciones técnicas de la prueba. Un plan rector ayuda a determinar el equilibrio en los niveles de cualificaciones cognitivas y contenido que es necesario incluir en una prueba (Anderson y Morgan, 2008).

En la Figura 9.17 se muestra la interfaz para la selección de los ítems de prueba óptimos. En la parte izquierda se encuentra un menú desplegable que permite elegir un origen para la selección de ítems de una lista de orígenes de datos disponibles que IATA genera automáticamente en función de los datos cargados y los análisis realizados (véase la Tabla 8.5). En este ejemplo, se encuentra disponible la tabla ***Items1***, que contiene los resultados del análisis actual.[8] Debajo de la selección de orígenes de datos aparecen los campos que permiten especificar el nombre que se aplicará a la selección de ítems y el

FIGURA 9.17

Resultados de selección de ítems de datos PILOT1, 50 ítems



número total de ítems que se seleccionará de los datos del ítem. La tabla que aparece debajo de estos campos contiene una lista con todos los ítems calibrados en el origen de datos seleccionado, así como el nivel de competencia (**Level**) y el área de contenido (**Content**) asociados a cada ítem. Aunque los dos últimos campos de datos suelen leerse en IATA en un archivo de datos del ítem, los datos también pueden editarse directamente en la tabla de forma manual. El proceso de selección estadística no requiere las especificaciones **Level** y **Content**, pero contar con información detallada sobre cada ítem le ayudará a optimizar la selección de ítems y a mantener la representación de contenido deseada. Al hacer clic en la casilla que se encuentra a la izquierda del nombre de un ítem para marcarla, se obliga a IATA a seleccionar el ítem, con independencia de sus propiedades estadísticas.

Debajo de la tabla de ítems hay dos controles deslizantes que le permiten especificar el intervalo de competencia en el que desee maximizar la precisión de la prueba. Los controles se establecen de forma que el valor mínimo se corresponda con el percentil 2 de la competencia y el valor máximo se corresponda con el percentil 98 (el valor seleccionado actualmente aparece a la derecha de cada control deslizante). Puede especificar un intervalo más reducido en el que maximizar la información mediante la modificación de los límites superior e inferior para reflejar los objetivos de la evaluación. IATA seleccionará los ítems para minimizar el error estándar de medición medio en el rango de competencia entre los límites inferior y superior, suponiendo una distribución de competencia normal en la muestra de estudiantes que se evaluará.

El principal propósito de los ítems de la prueba piloto consiste en determinar los ítems que resultarán más útiles en la administración final de la evaluación nacional. Si se observa que la muestra piloto de estudiantes se encuentra por encima de la media en términos de competencia, es necesario tener en cuenta esta expectativa al seleccionar los ítems. Teniendo en cuenta que desea crear una prueba final de 50 ítems, introduzca las siguientes especificaciones en IATA:

1. En la casilla **Name of item selection,** escriba **50Items** (el nombre es arbitrario; el nombre se utiliza aquí para poder comparar los

resultados obtenidos con los resultados de la carpeta de datos de muestra de IATA).

2. Introduzca el número 50 en la casilla **Total number of items.**
3. Mueva el control deslizante de **Upper bound** para que su valor sea 80. Esta especificación indica que la selección de ítems no intentará maximizar la precisión por encima del percentil 80 en la distribución de competencia de la muestra. Este ajuste se elige para compensar la posibilidad de que la competencia de la muestra piloto sea mayor que la de la población en general.
4. Haga clic en **Select Items**.

Una vez que IATA haya llevado a cabo la tarea, la interfaz debe tener un aspecto similar al de la Figura 9.17. En la parte izquierda de la lista de ítems, puede ver los 50 ítems reales seleccionados (el último es **MATHC1041**). En la parte derecha, el gráfico muestra la información colectiva y el error de medición esperado de los elementos seleccionados si se administraron como una prueba. Los resultados indican que la selección de ítems es más precisa en torno a la puntuación de competencia de cero (competencia media de la muestra actual). En la tabla que aparece debajo del gráfico se resume la distribución de ítems seleccionados en las áreas de contenido y los niveles cognitivos (para estos datos, se ha utilizado 1 como valor predeterminado en todos los ítems; los valores pueden editarse directamente en la tabla de ítems o cargarse en el archivo de datos del ítem inicial). Si los datos de esta tabla indican que la selección estadísticamente óptima no se ajusta adecuadamente al modelo de la prueba, podrá modificar el equilibrio del contenido mediante la selección y eliminación manual de ítems específicos utilizando las casillas que se encuentran junto a cada uno de los nombres de los ítems en la tabla de la izquierda. Al seleccionar ítems de forma manual, se cargará automáticamente el resumen de las propiedades de la prueba que aparece en la parte derecha.

La selección de ítems también se registra como una tabla de datos del ítem en IATA con el nombre **CustomTest50ItemsA**. Al igual que ocurre con todos los datos generados por IATA, puede

avanzar a la interfaz final del flujo de trabajo para ver y exportar esta tabla de datos (véase "Paso 10: visualización y guardado de resultados"). Los ítems de la tabla se ordenan según la idoneidad para los criterios de selección, con los ítems más adecuados en la parte superior.

A causa del reducido número de elementos en el análisis actual, los usuarios pueden utilizar IATA simplemente para ordenar todos los ítems según la idoneidad para el rango de competencia deseado (es decir, por debajo del percentil 80 en la muestra actual). Posteriormente, el equipo de desarrollo de pruebas puede revisar el archivo de datos del ítem y, al seleccionar ítems para la prueba final, clasificar los ítems en términos de idoneidad, así como garantizar el mantenimiento del equilibrio adecuado del contenido. Para crear una nueva selección de ítems, ejecute los siguientes pasos:

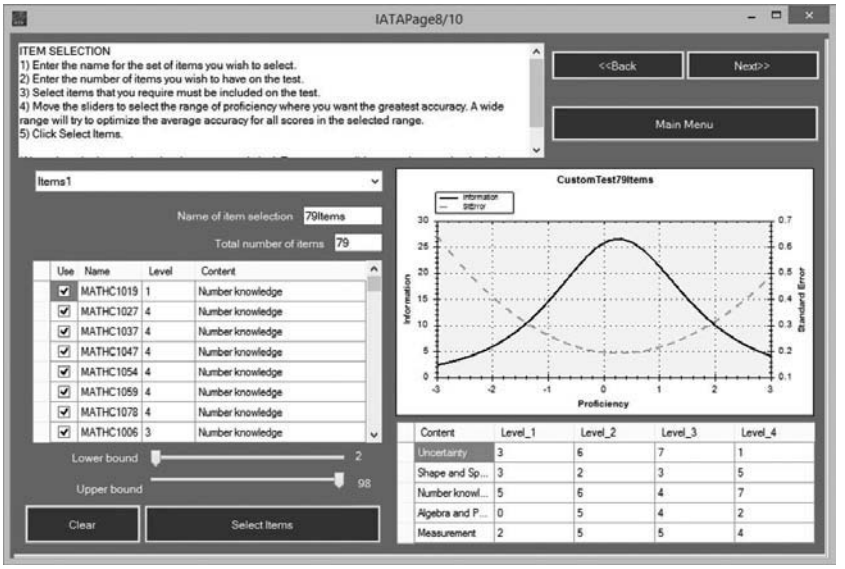
1. Haga clic en **Clear** para eliminar todas las selecciones anteriores de la lista de ítems.
2. Ingrese un nombre nuevo para la selección de ítems: **79Items** (si el nombre ya está en uso, puede sobrescribir los resultados anteriores).
3. Ingrese el número máximo de ítems disponibles (79) como número total de ítems. Si ingresa un número mayor que el número de ítems disponibles, IATA solo llevará a cabo la selección entre los ítems disponibles.
4. Deje el límite superior en 80, ya que el intervalo objetivo de competencia no ha cambiado.
5. Haga clic en **Select Items**.

En la Figura 9.18 se muestran algunos de los resultados del análisis de la prueba piloto de 79 ítems. Se ha agregado una tabla de resultados (llamada ***CustomTest79Items***) al conjunto de resultados de IATA, que puede consultar en la interfaz final del flujo de trabajo. Los desarrolladores de pruebas pueden utilizar esta información para ayudar a mejorar la calidad de los ítems mediante la identificación y la corrección de los ítems menos efectivos.

El proceso de selección de ítems depende de la calidad de los ítems disponibles. IATA no puede mejorar la precisión en áreas de

FIGURA 9.18

Resultados de selección de ítems de datos PILOT1, 79 ítems



competencia específicas si no hay ítems disponibles con información para dichas áreas. El proceso automatizado puede ayudar a seleccionar los ítems más adecuados que estén disponibles, pero no puede mejorar la precisión de los ítems en cuestión.

Una vez que revise los resultados, haga clic en **Next>>** para continuar.

PASO 9: ESTÁNDARES DE RENDIMIENTO

En la etapa de prueba piloto, no hay pruebas suficientes para respaldar la configuración de los estándares de rendimiento. Aunque hay información disponible sobre las propiedades de los ítems estadísticos y las especificaciones utilizadas para crear los ítems, todavía no se dispone de información detallada sobre la distribución de la competencia en la población de estudiantes. Por lo tanto, cualquier intento de establecer estándares de rendimiento en la etapa piloto sería innecesario y potencialmente engañoso.

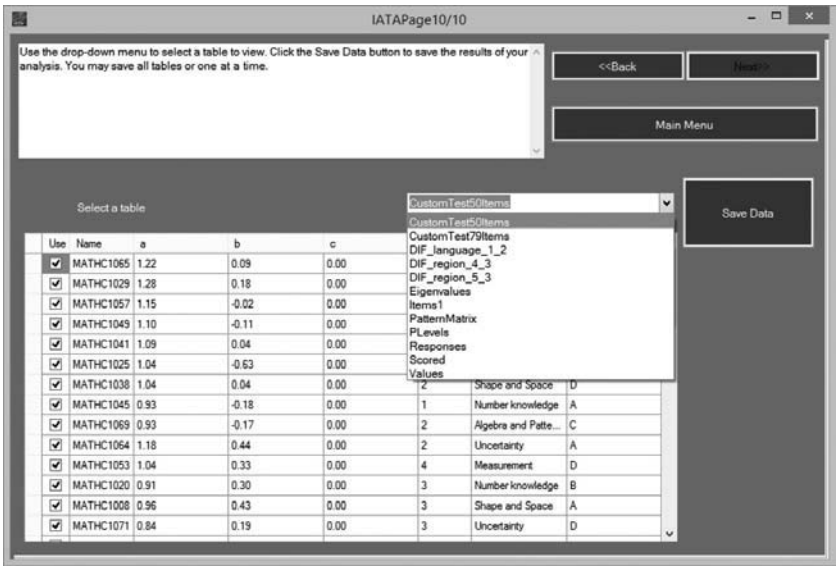
Puesto que esta guía de ejemplo del análisis de datos de pruebas piloto no requiere ninguna configuración estándar, haga clic en **Next>>** para continuar a la interfaz de visualización y guardado de resultados.

PASO 10: VISUALIZACIÓN Y GUARDADO DE RESULTADOS

Para todos los flujos de trabajo de análisis, IATA produce una serie de resultados diferentes en forma de tablas de datos. Estos resultados pueden visualizarse y guardarse en la interfaz final de cada flujo de trabajo. Esto le permitirá revisar cada una de las tablas de datos de los resultados generados durante el flujo de trabajo de análisis. En la interfaz se muestra la tabla de datos seleccionada en el menú desplegable. Para cambiar el origen de datos, seleccione una tabla diferente del menú desplegable, como se muestra en la Figura 9.19. (En la Tabla 8.5 se muestra una lista completa y una descripción de las tablas de datos disponibles generadas por IATA).

FIGURA 9.19

Visualización de resultados del análisis de datos PILOT1



Tenga en cuenta que, aunque no se haya especificado la creación de estándares de rendimiento, la tabla **PLevels** se crea automáticamente con los valores de especificaciones predeterminados.

Puede guardar estas tablas de resultados en un solo archivo de salida o en varios archivos al hacer clic en **Save Data**. Puede guardar una sola tabla o todas las tablas a la vez en diferentes formatos. Se recomienda utilizar dos formatos de archivo para guardar las salidas de IATA: Excel (*.xls/*.xlsx) y SPSS (*.sav).

Por lo general, es preferible utilizar Excel, ya que todas las tablas de datos pueden guardarse en un solo archivo de datos. El formato de Excel también puede abrirse con software gratuito como OpenOffice (que se puede descargar de http://www.openoffice.org). No obstante, las versiones antiguas de Excel están limitadas a 255 variables como máximo. Si los archivos de datos tienen más variables, IATA solo guardará las primeras 255 en el archivo *.xls. Para guardar archivos de datos de mayor tamaño, utilice los formatos *.sav o *.xlsx. La ventaja de los archivos de SPSS es que pueden almacenar tablas de datos de mayor tamaño y permiten almacenar metadatos (si se editan en el paquete de software de SPSS). No obstante, tenga en cuenta que SPSS tiene una limitación principal: cada tabla de datos se guardará en un archivo independiente. Un cuadro de diálogo de archivo le solicitará que especifique el nombre de archivo y la ubicación de los resultados, así como el formato de salida. Elija el formato de datos deseado y haga clic en **Save** para terminar de guardar la tabla o las tablas.[9] Los archivos resultantes contienen todos los resultados tabulares generados durante el flujo de trabajo de análisis completo, lo que proporciona la documentación del análisis.

A modo de referencia, los resultados de esta guía de análisis de la tabla de resultados ***Items1*** están incluidos en el archivo ***ItemDataAllTests.xls***. Se ha cambiado el nombre de la hoja de cálculo que contiene los datos de la tabla ***Item1*** del análisis actual a ***ReferenceP1***. En los resultados guardados, los valores "True" y "False" de la columna E (OK) indican los ítems incluidos en el análisis final. En estos resultados, solo el valor de **MATHC1075** es "False".

En el caso de los análisis de pruebas piloto reales (es decir, en los que no se utilicen datos simulados), es necesario proporcionar las tablas de resultados y los gráficos copiados y pegados durante el flujo

de trabajo de análisis a los desarrolladores de pruebas, que pueden utilizar la información para modificar la prueba mediante la selección, ordenación y adición de ítems, en caso necesario, con el fin de maximizar la precisión y la utilidad del formulario de prueba final.

NOTAS

1. Consulte la Tabla 8.6 para obtener una descripción de los símbolos y sus significados.

2. Una carga igual a 1 es inadmisible, ya que precisaría que todos los encuestados consiguieran la misma puntuación en todos los ítems. Este requisito implica que la prueba solo podría producir dos valores de puntuación diferentes, lo que no ofrece demasiada información.

3. Los valores mostrados en IATA están estandarizados para expresar la proporción de la variación total representada por cada valor propio.

4. Al hacer clic en el encabezado dos veces, se organizará la columna en orden descendente.

5. El coeficiente de varianza del muestreo se calcula como el error estándar de la estadística S-DIF dividido entre el valor absoluto de la estadística S-DIF.

6. Todos los resultados de esta guía se encuentran disponibles para su consulta y comparación en la carpeta de datos de muestra de IATA en la tabla de Excel llamada ***ReferencePILOT1.xls***. Las tablas de resultados de FDI se encuentran en las hojas de cálculo con nombres que empiezan por DIF_.

7. Para copiar cualquiera de los gráficos de análisis de FDI, coloque el cursor sobre el gráfico y utilice las funciones **Copy** y **Paste** del menú contextual.

8. En el caso de los análisis que incluyan vinculaciones, seleccione los datos del ítem previamente calibrados (***Items2***) o el conjunto de ítems comunes para ambos orígenes de datos de ítems (***MergedItems***).

9. Si guarda todas las tablas y selecciona el formato de salida de SPSS (*.sav), se exportará cada tabla de resultados como un archivo de datos *.sav independiente, con el nombre proporcionado como prefijo para todos los nombres de tablas.

CAPÍTULO 10

REALIZAR EL ANÁLISIS INTEGRAL DE LOS DATOS DE LA ADMINISTRACIÓN DE UNA PRUEBA FINAL

Para realizar este ejercicio, utilice el conjunto de datos de muestra ***CYCLE1***. La clave de respuestas para esta prueba se incluye en el libro de Excel ***ItemDataAllTests.xls***, en la hoja ***CYCLE1***.

Los análisis de este capítulo se basan en el rendimiento de los alumnos en una evaluación nacional de matemáticas administrada a una muestra nacional de alumnos. La prueba final tuvo 50 ítems, que representaban cinco áreas de contenido (conocimiento de números, forma y espacio, relaciones, resolución de problemas e incertidumbre) en proporciones determinadas por las especificaciones de la prueba. La muestra final adoptó un diseño estratificado y por conglomerados, en el que la unidad primaria de muestreo estuvo constituida por las escuelas y la muestra objetivo estuvo constituida por 30 alumnos de cada escuela. La muestra comprendió 79 escuelas seleccionadas para que fueran representativas de cinco regiones nacionales y estratificadas según el idioma de enseñanza y la condición de ruralidad. El número total de alumnos de la muestra es 2242, representando a una población de alrededor de 86 000 alumnos.

Esta guía sigue los mismos pasos que el análisis de los datos de la prueba piloto del capítulo 9. No obstante, debido a que el objetivo primordial de la prueba final es producir e interpretar puntajes, el

análisis de los ítems comúnmente se realiza sin el enfoque exploratorio característico de un análisis de los datos de la prueba piloto. En consonancia, estas guías se centran en los aspectos propios del análisis de los datos de la prueba final que no están presentes en un análisis de los datos de una prueba piloto. Además de los análisis realizados con los datos de la prueba piloto, los análisis de los datos de la prueba completa de este capítulo comprenden el cálculo de puntajes escalares y estándares de rendimiento. Cuando los pasos del análisis sean idénticos a los descritos en el capítulo 9, consulte la información detallada en ese capítulo.

Para comenzar el análisis, haga clic en **Response data analysis** en el menú principal de IATA (Item and Test Analysis).

PASO 1: CONFIGURACIÓN DEL ANÁLISIS

Los procedimientos para configurar el análisis son similares a los descritos en el capítulo 9. En primer lugar, cargue un archivo de respuestas, luego cargue un archivo de datos de ítems y después especifique el análisis. Si se requiere más información, consulte las instrucciones detalladas en los pasos 1 a 3 del capítulo 9. La carpeta de datos de muestra de IATA contiene:

- El archivo de datos de respuesta para este capítulo que se llama ***CYCLE1.xls***. (Este archivo tiene 2242 registros y 58 variables.)
- El archivo de datos de ítems que está en la tabla ***CYCLE1*** del archivo de Excel ***ItemDataAllTests.xls.*** Asegúrese de seleccionar el nombre de tabla correcta en la interfaz de carga de datos de ítems. (El archivo de datos de ítems ***CYCLE1*** tiene 50 registros.)

Los ítems de la prueba de la evaluación nacional son un subconjunto de los ítems de la prueba piloto descritos en el capítulo 9.

Las especificaciones para este análisis son ligeramente diferentes de las del análisis de los datos de la prueba piloto, principalmente debido al uso de muestreo probabilístico en la administración plena de la evaluación nacional. La primera diferencia es el nombre de la variable de identificación: **CYCLE1STDID**. La segunda, que afectará los resultados del análisis, es la presencia de una ponderación de diseño

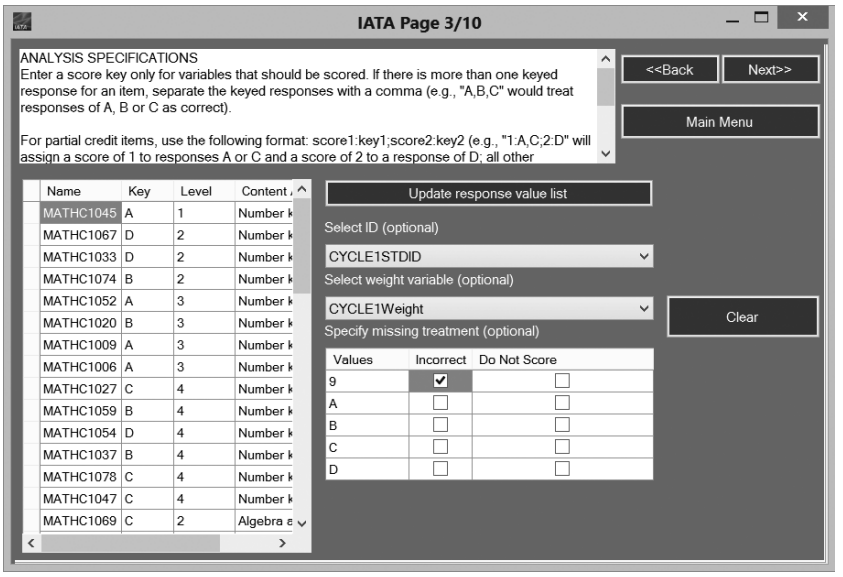
de la muestra llamada **CYCLE1weight**. Estas especificaciones de variables se deben seleccionar en los menús desplegables. En estos datos, el valor 9 representa respuestas faltantes que se tratan como incorrectas. La Figura 10.1 muestra cómo deben configurarse todas las especificaciones.

Observe que los datos de ítems para la evaluación final también incluyen los datos del campo **Level**, la tercera columna de la tabla sobre la izquierda. Estos datos son números naturales (1 o mayor) que representan el nivel esperado de rendimiento o competencia que los especialistas en el contenido del currículo asignaron a cada ítem de la prueba: el nivel 1 representa el nivel de rendimiento más bajo (es decir, competencia mínima), y el nivel 4 representa el nivel más alto. Si bien a cada ítem se asigna un nivel, puede que algunos alumnos no alcancen ni siquiera el nivel más bajo.

Luego de comprobar que las especificaciones y los datos sean correctos, haga clic en **Next>>** para continuar. El análisis comenzará automáticamente y la interfaz se actualizará periódicamente conforme avance. Si trabaja con conjuntos de datos más grandes u ordenadores

FIGURA 10.1

Especificaciones para el análisis de los datos de CYCLE1



más lentos, puede que el análisis parezca avanzar lentamente en la etapa de estimación de parámetros, que es la que más tiempo consume. No cierre el programa; IATA continuará ejecutándose y mostrará una actualización cuando el análisis haya finalizado.

PASO 2: RESULTADOS BÁSICOS DEL ANÁLISIS

Debido a que los ítems problemáticos ya se identificaron y eliminaron en el análisis de los datos de la prueba piloto, no queda ningún ítem problemático en el conjunto de datos totales. Para confirmar que el comportamiento de los ítems es el adecuado, revise los resultados de (a) el análisis de ítems (IATA Page 4/10) y (b) la dimensionalidad de la prueba (IATA Page 5/10). Si desea instrucciones sobre cómo realizar estas tareas, consulte los pasos 4 y 5 del capítulo 9. Observe que todos los ítems enumerados en IATA Page 4/10 tienen círculos (verdes) con excepción de **MATHC1046**, que se identificó como parcialmente problemático en el capítulo 9, pero que se mantuvo en la prueba. Cuando finalice, pase a la interfaz de funcionamiento diferencial del ítem (DIF) (IATA Page 6/10).

PASO 3: ANÁLISIS DE FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM

Si bien ya se realizó un análisis de FDI sobre los datos de la prueba piloto, conviene replicar los análisis con la muestra completa, debido a que los resultados del análisis de FDI suelen ser sensibles a errores de muestreo. Otras razones para realizar un análisis de FDI son que puede que haya nuevas variables en la muestra completa respecto de las presentes en la muestra piloto y que la muestra ofrece un número más satisfactorio de casos para el análisis de FDI.

Aquí el análisis de FDI se realiza para examinar la posibilidad de sesgo urbano, es decir, si los alumnos rurales están desfavorecidos respecto de los alumnos urbanos. Para los datos de ***CYCLE1***, el valor 1 para este indicador significa que el alumno asiste a una escuela rural, en tanto que el valor 0 señala que el alumno concurre a una

escuela urbana. Para especificar este análisis y revisar los resultados, siga los pasos a continuación:

1. Seleccione la variable **Rural** del menú desplegable de la izquierda. Esto hará que la tabla debajo del menú desplegable se cargue con los valores 0.00 y 1.00, con los valores 56 por ciento para 0.00 y 44 por ciento para 1.00, lo que indica que el 44 por ciento de los alumnos (sin ponderar) de la muestra asisten a escuelas rurales.
2. En la tabla de valores, haga clic en **1.00**. De esta manera, el valor 1.00 (que representa a los alumnos rurales) se ingresará como el grupo de interés en el cuadro de texto que aparece debajo.
3. En la tabla de valores, haga clic en **0.00**. Así se ingresará el valor 0.00 (que representa a los alumnos urbanos) como el grupo de referencia en el cuadro de texto que está debajo.
4. Haga clic en **Calculate** y espere a que termine el cálculo.
5. Cuando haya finalizado, haga clic en el encabezado de la columna S-DIF en la lista de ítems para organizarlos según el valor de la estadística S-DIF.

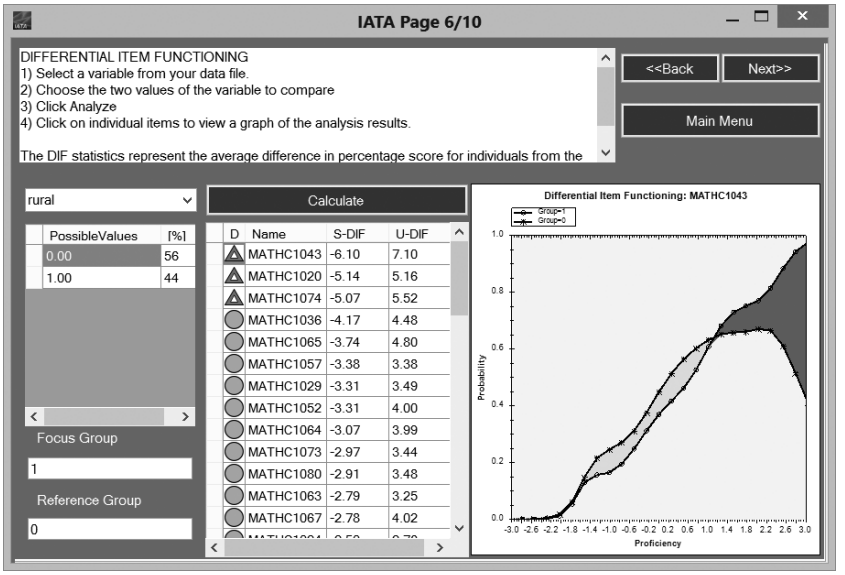
Una vez completados estos pasos, la interfaz aparecerá tal como se muestra en la Figura 10.2. La mayoría de las estadísticas S-DIF y U-DIF tienen un valor inferior a 5, lo que indica que, luego de controlar las diferencias de competencia de los alumnos rurales y urbanos, las diferencias de rendimiento de ítems entre los alumnos de estas zonas suelen ser insignificantes.

El propósito de realizar un análisis de FDI en la etapa de prueba final de una evaluación nacional es determinar si se debe dejar de lado algún ítem para calcular los puntajes de los alumnos. A esta altura del análisis, sería apropiado compartir los resultados del análisis estadístico con el comité director de la evaluación nacional; el comité determinará si se deben quitar o mantener los ítems potencialmente problemáticos. Si se quita un ítem, se debe volver a ejecutar el análisis luego de borrar la clave de respuesta de ese ítem, en la interfaz de especificaciones del análisis, o después de desmarcar el ítem en la interfaz de análisis de ítems. El presente ejemplo se basa en el supuesto de que se mantuvieron todos los ítems.

Luego de revisar todos los ítems, presione **Next>>** para continuar.

FIGURA 10.2

Resultados del análisis de FDI para los datos de CYCLE1 por zona, MATHC1043



PASO 4: CONFIGURACIÓN DE ESCALA

La escala predeterminada que se emplea para calcular los resultados para los puntajes escalares en la Teoría de Respuesta al Ítem (TRI) es la estándar, o escala Z, que tiene una media de 0 y una desviación estándar de 1. Muchos actores parecen experimentar problemas con los puntajes expresados en esta escala, debido a que la mitad de los alumnos tendrán puntajes negativos. Del mismo modo, los puntajes entre 0 y 100 presentan desafíos comunicacionales; distintos públicos tienden a suponer que un puntaje de 50 representa un puntaje aprobatorio, lo que puede no ser realidad, según las especificaciones de la prueba.

A los efectos de su comunicación, puede que no sea aconsejable informar resultados de pruebas con un puntaje promedio menor al 50 por ciento o inferior a 0. Algunas evaluaciones a gran escala convierten sus puntajes calculados a escalas que tienen una media de 500, 100 o 50 y desviaciones estándar de 100, 20 y 10, respectivamente.

Cada equipo de evaluación nacional debe elegir el tipo de puntaje que mejor facilite la comunicación efectiva de los resultados.

En IATA se pueden realizar dos tipos de ajustes con escalas: definición de escala con Set Scale y cambio de escala con Rescale. Set Scale permite especificar la media y la desviación estándar deseada para los puntajes escalares. Rescale permite aplicar una transformación lineal simple a las variables IRT score, lo que es útil si se necesita comparar los puntajes escalares con una escala que ya ha sido establecida en un análisis anterior. En ese caso, se pueden emplear los parámetros de ítems del ciclo previo para estimar los puntajes de la prueba o equiparar resultados de los datos de los alumnos en el nuevo ciclo a fin de que las variables IRT score que IATA calcula se puedan comparar con los calculados en el ciclo anterior. Luego, se puede cambiar la escala de los resultados calculados con Rescale para que sean comparables con la escala informada en el ciclo previo.

En cualquier caso, para crear el nuevo puntaje escalar, se debe ingresar el nombre del nuevo puntaje y especificar la desviación estándar y la media en los recuadros correspondientes. Cuando haga clic en el botón **Calculate**, IATA producirá los nuevos puntajes escalares y mostrará la distribución y los extractos estadísticos.

A diferencia del análisis de la prueba piloto, cuya función principal es ayudar a establecer el diseño de la prueba, el objetivo primordial del análisis de los datos de la prueba de una evaluación nacional es producir puntajes. En consecuencia, estas guías requieren examinar más de cerca y especificar las propiedades de los puntajes de la prueba; ambas tareas se realizan desde la interfaz de configuración de escala. En primer lugar, comparar la distribución de los puntajes de competencia con la exactitud de la prueba en cada puntaje de competencia (que también se conoce como *información de la prueba*) indica la calidad de la deducción que se puede inferir sobre las distintas escalas de competencia. En segundo lugar, crear una escala comunicacional para los resultados de la prueba establece una métrica de comunicación de esos resultados a los actores.

El gráfico de la Figura 10.3 indica que la información de la prueba, representada por la línea negra sólida, está bien distribuida respecto de la distribución de la competencia en la muestra. El pico de frecuencia a la izquierda del gráfico en aproximadamente -3 de la escala

de competencia corresponde a los alumnos que no respondieron correctamente ningún ítem de la prueba. La prueba no tiene suficiente información para determinar de forma precisa cuán bajos son los niveles de competencia de estos alumnos; en consecuencia, se les asigna a todos un mismo puntaje arbitrariamente bajo.

Para revisar la distribución de la variable IRTscore, seleccione **IRTscore** del menú desplegable en la parte superior izquierda de la interfaz. Como muestra la Figura 10.3, la interfaz se actualizará con detalles descriptivos sobre las variables IRTscore y la información de la prueba. La media de la distribución **IRTscore** es –0.02, y la desviación estándar es 1.08. Estos valores no son significativos en sí mismos debido a que representan la escala arbitraria en la que se calibraron los ítems.

Compare estos resultados con la forma estadísticamente ideal de la función de la información de la prueba, en términos de maximización de la confiabilidad general para una población con distribución normal, como ilustra la Figura 10.4. A los efectos de la comparación,

FIGURA 10.3

Distribución de la competencia (variable IRT score) e información de la prueba; datos de CYCLE1

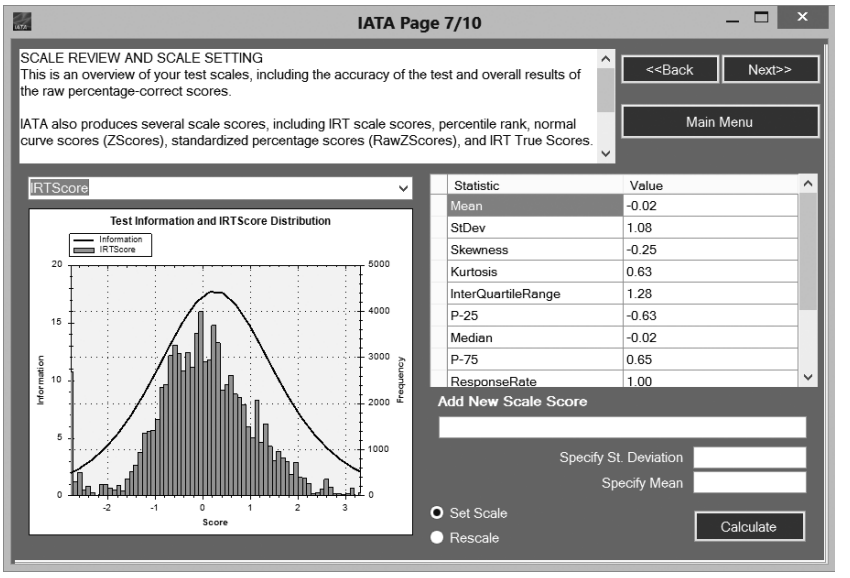
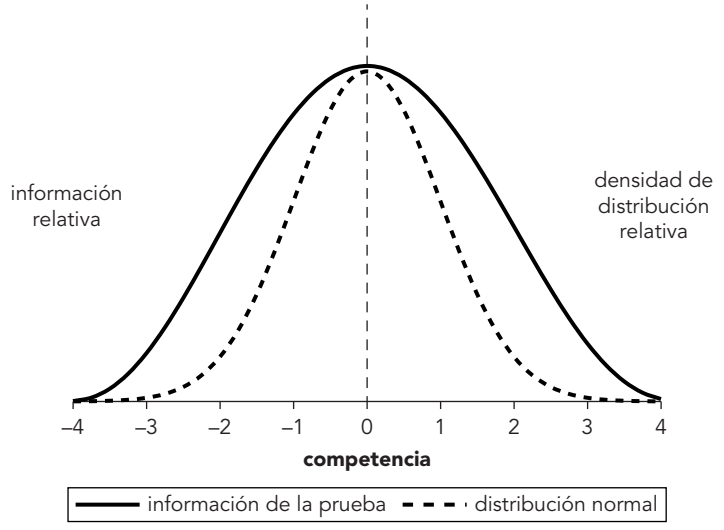


FIGURA 10.4

Comparación entre la información ideal de la prueba y la distribución normal



la línea de la distribución normal estándar es punteada. La información ideal para una muestra debe brindar la mayor cantidad de información en los niveles de competencia que representen a un gran número de alumnos, pero también necesita tener suficiente información para distinguir entre los alumnos en los extremos más alto y más bajo de la competencia.

Estos resultados también indican que la prueba resultó bastante difícil para los alumnos. El pico de la función de información suele estar ubicado en la región de competencia en la que es más probable que los alumnos obtengan un puntaje del 50 por ciento. En la Figura 10.3, este pico está apenas por encima del puntaje promedio de –0,02, lo que indica que los alumnos por encima del promedio en general solo obtuvieron un puntaje de respuestas correctas del 50 por ciento.

Para producir una escala de reporte más útil que se base en la variable IRT score, utilice las funciones de **Add New Scale Score** en la parte inferior derecha de la interfaz (véase la Figura 10.3). Para este ejemplo, se debe suponer que el comité director nacional solicitó una

nueva escala que requiere configurar la media en 500 y la desviación estándar en 100.

Esta escala se configurará en el primer ciclo de la evaluación nacional y se utilizará en los ciclos posteriores para comunicar cambios de rendimiento a través del tiempo. El nombre de este puntaje será **NAMscore** (acrónimo para puntaje de evaluación nacional en matemáticas). Para indicar estas especificaciones, siga los pasos a continuación:

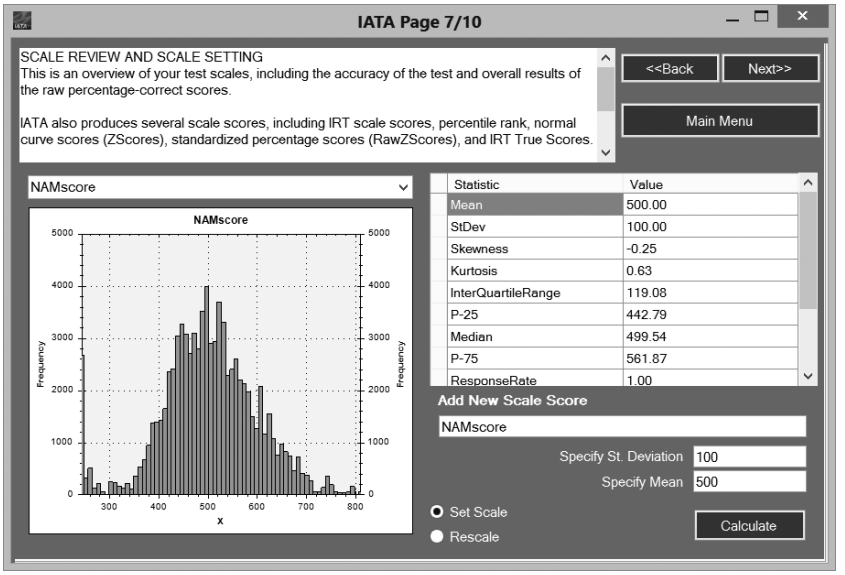
1. Escriba **NAMscore** en el campo que está debajo de la etiqueta **Add New Scale Score**.
2. Ingrese el valor 100 para la desviación estándar en el campo **Specify St. Deviation**.
3. Ingrese el valor 500 para la media en el campo **Specify Mean**.
4. Asegúrese de que esté seleccionada la opción **Set Scale**. Esto garantiza que el puntaje escalar producido tendrá una media exactamente igual a 500 y una desviación estándar equivalente a 100 para la muestra. La opción **Rescale** simplemente ajustará la variable IRT score existente según la media y la desviación estándar especificadas.
5. Haga clic en **Calculate**.

Cuando IATA termine de procesar lo solicitado, se actualizará la interfaz con el gráfico de extracto y las estadísticas para el puntaje escalar recién creado, tal como muestra la Figura 10.5.

Existen relativamente pocas limitaciones para seleccionar un puntaje escalar derivado. Se puede utilizar cualquier nombre válido para este puntaje siempre y cuando no haya sido usado en los datos de respuestas (consulte en el capítulo 8 las convenciones sobre nomenclatura y los nombres restringidos para las variables). La media puede ser cualquier número real, y la desviación estándar puede ser cualquier número real mayor a cero. No obstante, es importante asegurar que los puntajes más bajos comunicados para los alumnos no sean inferiores a cero. Debido a que el puntaje más bajo suele estar alrededor de tres a cuatro desviaciones estándar más abajo que la media, conviene configurar la media al menos cuatro desviaciones estándar por encima de cero. La elección de una escala de reporte se debe

FIGURA 10.5

Distribución y extractos estadísticos para el nuevo puntaje escalar (NAMscore); datos de CYCLE1



tratar con el comité director de la evaluación nacional en las etapas iniciales de planificación a fin de que todos los actores puedan comprender cómo interpretar los resultados informados.

Una vez creado el nuevo puntaje escalar, presione **Next>>** para continuar.

PASO 5: SELECCIÓN DE ÍTEMS DE LA PRUEBA

Los datos de ***CYCLE1*** representan el ciclo inicial de un programa de evaluación nacional. Si la prueba ha de usarse en ciclos posteriores para fines de comparación, se deberá establecer un vínculo con los resultados del ciclo inicial. Para hacerlo, seleccione un subconjunto de ítems que sean exactos y que representen el continuo de competencia.

Una práctica adecuada para mantener un vínculo sólido entre pruebas es que las evaluaciones adyacentes compartan alrededor del 50 por ciento de los ítems. Estos se conocen como *ítems de anclaje*. Para facilitar el proceso de selección de estos ítems de anclaje, use la

opción de selección de ítems en IATA, que permite clasificar cada ítem de la prueba actual según su idoneidad para maximizar la exactitud en toda la escala de competencia. Para hacer esta selección, siga los pasos a continuación:

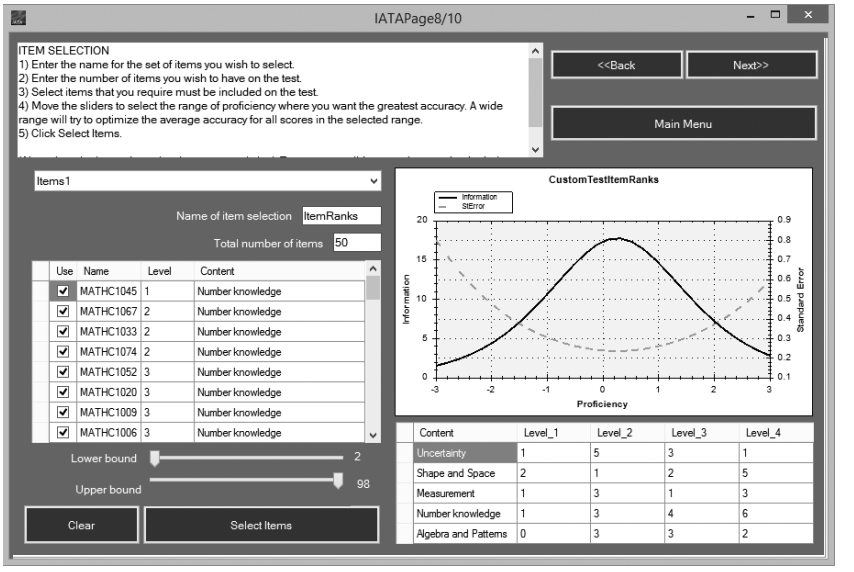
1. Escriba el nombre **ItemRanks** en el campo **Name of item selection**.
2. Ingrese el número 50 en el campo **Total number of items** para seleccionar todos los ítems.
3. Para los límites inferior y superior, mantenga los valores predeterminados de 2 y 98.
4. Haga clic en **Select Items**.

La Figura 10.6 muestra los resultados completos. Todos los ítems disponibles se seleccionaron y categorizaron según contenido y nivel cognitivo identificados en sus especificaciones originales.

Los resultados producidos por estas especificaciones se agregan al conjunto de resultados del análisis actual en forma de tabla de datos

FIGURA 10.6

Selección de ítems; datos de CYCLE1



de ítems de IATA. Esta tabla se debe entregar a los desarrolladores de la prueba responsables de la modificación de la evaluación nacional ciclo 2 (o siguiente) para que puedan seleccionar un conjunto de ítems compartidos, teniendo en consideración la información sobre el contenido y el valor psicométrico de cada ítem de la prueba de la evaluación nacional ciclo 1 (o primero). El ideal sería que un conjunto de ítems de anclaje tenga la mitad de ítems que la prueba completa y represente el contenido y las especificaciones cognitivas en las mismas proporciones que la prueba completa. Como regla general, es poco probable que las vinculaciones estadísticas donde los ítems de vinculación constituyen menos del 20 por ciento del total de la prueba ofrezcan un vínculo significativo, independientemente de la exactitud o el grado de representación del contenido de los ítems de vinculación. Un método pragmático para seleccionar ítems sería comenzar con los ítems más deseables y asignarlos a las celdas de las especificaciones de la nueva prueba según su contenido y nivel cognitivo hasta alcanzar el número deseado en cada celda o hasta agotar la lista de ítems.

Una vez que IATA haya completado este análisis, presione **Next>>** para continuar.

PASO 6: DETERMINACIÓN DE ESTÁNDARES DE RENDIMIENTO

La mayoría de las evaluaciones actuales informan sus resultados por niveles. Evaluaciones internacionales como el Estudio sobre el Progreso Internacional de la Competencia en Lectura (PIRLS), el Programa para la Evaluación Internacional de Alumnos (PISA) y el Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS), así como también muchas evaluaciones nacionales como la Evaluación Nacional del Progreso Educativo (NAEP) de los Estados Unidos, informan los puntajes de rendimiento de los alumnos en términos de niveles de referencia o desempeño (véase Greaney y Kellaghan, 2008; Kellaghan, Greaney y Murray, 2009). El TIMSS, por ejemplo, utiliza cuatro niveles de referencia: bajo, intermedio, alto y avanzado (Martin, Mullis y Foy, 2008). Los estándares de rendimiento

deben tener umbrales estadísticos significativos y no arbitrarios, como por ejemplo percentiles, debido a que son la principal herramienta para resumir e informar el desempeño de los alumnos. El proceso de definición de estándares de rendimiento significativos se conoce como *determinación de estándares*.

IATA facilita los procedimientos de determinación de estándares, ya que permite especificar probabilidades de respuestas (RP) correctas para cada ítem para luego calcular los niveles de competencia (valores RP) asociados con las RP especificadas. Por ejemplo, si se determina una RP del 50 por ciento, el valor RP de un ítem será el nivel de competencia asociado con una probabilidad del 50 por ciento de que se responda correctamente. En evaluaciones a gran escala, se emplea una amplia variedad de RP que van, en general, desde el 50 hasta el 80 por ciento. Una práctica usual es utilizar un 67 por ciento, que suele ser óptimo desde el punto de vista estadístico para clasificar ítems. No obstante, la elección de una RP también debe basarse en información sobre las definiciones normativas sobre qué probabilidad de éxito constituye el dominio en determinado nivel de grado y en el conocimiento de las consecuencias de la forma en que se emplearán los estándares. Por ejemplo, en un ámbito educativo, donde las consecuencias de informar fracasos tienden a ser mayores que aquellas de informar éxitos, puede que se prefieran RP más bajas.

Antes de analizar los datos, un panel de partes interesadas, que incluya a expertos en currículo y enseñanza, debe decidir, con el asesoramiento del comité director de la evaluación nacional, la cantidad de niveles de dominio que se han de utilizar. Algunas evaluaciones nacionales eligen dos niveles, como por ejemplo aceptable y no aceptable; otras optan por tres niveles, como bajo, adecuado y avanzado, en tanto que otras usan cuatro o más. Si el panel de partes interesadas decide usar más de dos niveles, se deberá definir cada nivel, con excepción del más bajo, mediante un conjunto de ítems que se considere que pueden ser respondidos por los alumnos que reflejan ese nivel de desempeño. En general, a menos que una evaluación incluya cientos de ítems (lo que requeriría un diseño con rotación de cuadernillo), solo habrá suficientes ítems como para definir adecuadamente tres o cuatro niveles.

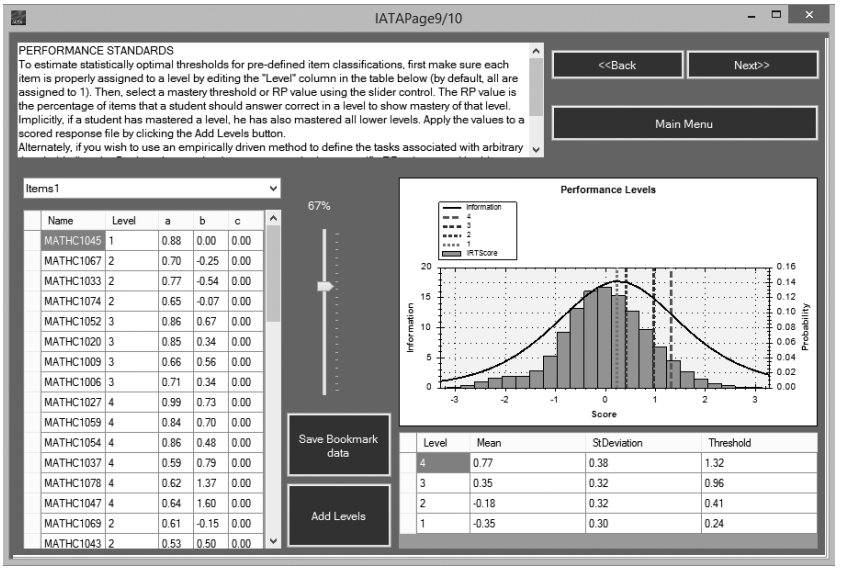
La Figura 10.7 muestra la interfaz desde donde se realiza este análisis. A la izquierda, un menú desplegable permite seleccionar el origen de

los ítems que se desea seleccionar. Al igual que en la interfaz de selección de ítems, tiene la opción de elegir cualquier fuente de datos de ítems disponible en el flujo de trabajo actual. Para los análisis actuales, solo está disponible la tabla ***Items1***.[1] Los ítems de la fuente seleccionada aparecen en la tabla debajo del menú desplegable. Los valores de la columna **Level** se pueden editar directamente en cada fila. Para estimar umbrales estadísticamente óptimos de acuerdo con la clasificación actual de ítems, mueva la barra de desplazamiento vertical desde el centro de la interfaz hasta la RP deseada. Cuando se abre la interfaz, la RP predeterminada es 67 por ciento, lo que indica que el criterio empleado para clasificar ítems o estimar umbrales óptimos es una probabilidad de respuesta correcta del 67 por ciento para cada ítem.

Al hacer clic en la barra de desplazamiento vertical y arrastrarla para ajustar su valor, IATA actualizará los umbrales óptimos y producirá los resultados en la ventana de gráfico sobre la derecha y en la tabla de resultados situada debajo. El gráfico ilustra la posición de cada umbral con líneas verticales en relación con la distribución de la

FIGURA 10.7

Interfaz de estándares de rendimiento predeterminados; datos de CYCLE1



competencia y la función de información de la prueba. Esta información demuestra la utilidad de los niveles. Por ejemplo, si un nivel tiene muy pocas unidades con respuesta, cualquier extracto estadístico que describa a los alumnos de ese nivel será demasiado reducido o inestable para permitir su interpretación. Del mismo modo, si la prueba no es exacta en el umbral de un nivel, la clasificación de los alumnos de ese nivel será inexacta.

La tabla debajo de la ventana del gráfico en la Figura 10.7 describe los ítems que representan cada nivel con la media y desviación estándar de los parámetros **b** del ítem. La columna más a la derecha contiene el umbral estimado para cada nivel. Por ejemplo, la media y la desviación estándar de los parámetros **b** para el nivel 4 son 0.77 y 0.38, respectivamente. El valor 0.77 indica que el promedio de los parámetros **b** para los ítems del nivel 4 corresponde a un puntaje de competencia de 0.77 en la escala de la TRI. El umbral RP67 para el nivel 4 es 1.32. Estas estadísticas son útiles para determinar si la asignación de los ítems es lógica. Por ejemplo, si la desviación estándar de los ítems de un nivel es mayor que la distancia entre las medias o los umbrales de niveles adyacentes, puede que la base estadística para definir los niveles sea débil. En estos resultados, la desviación estándar dentro de los niveles tiende a ser de alrededor de 0.35, en tanto que la distancia entre los niveles adyacentes varía entre 0.17 y 0.53, lo que indica que los niveles están bastante bien definidos.

Se emplean varios métodos para determinar los puntos de corte o umbrales más adecuados entre los niveles de competencia. Uno de los métodos se denomina *bookmark*, procedimiento basado en la TRI que se beneficia de tener en la misma dimensión latente la dificultad de los ítems y la capacidad de la persona. Implica la participación de un comité de expertos para la determinación de estándares (como especialistas en currículo y docentes experimentados) para que revise cuidadosamente todos los ítems de la prueba a la luz de la información disponible en las especificaciones de la prueba, los currículos, el desempeño de los alumnos en la prueba y las definiciones normativas acerca de lo que los alumnos saben y pueden hacer en cada nivel de competencia (véase Karantonis y Sireci, 2006; Mitzel et al, 2001).

Los procedimientos que regulan el funcionamiento de los comités en relación a la selección, la capacitación y las interacciones de sus

miembros y el uso que hacen de los datos de distintas fuentes varían, pero no se tratan en el presente documento. Por el contrario, el foco se pone en la forma en que los datos generados por IATA pueden contribuir a la definición de niveles de competencia.

En primer lugar, el comité prepara una versión especialmente diseñada del cuadernillo de la prueba de la evaluación nacional con un ítem de opción múltiple o respuesta cerrada por página, según sus valores RP. La tarea del comité es identificar ítems que se encuentren en los límites entre niveles o grupos de ítems distintos desde el punto de vista cognitivo. Luego, el comité coloca señaladores o marcadores en esos límites en ese cuadernillo. Los ítems seleccionados para el grupo de nivel más alto (nivel 4, por ejemplo) son los que los alumnos en ese nivel tienen más probabilidades de responder correctamente comparados con los alumnos de niveles inferiores. De manera similar, es más probable que los ítems seleccionados para el nivel 3 sean respondidos correctamente por los alumnos de ese nivel (y del nivel 4) que por los alumnos de niveles inferiores.

A modo de ejemplo, supongamos que el comité de expertos decidió utilizar una RP del 50 por ciento para validar la clasificación inicial de los ítems realizada por los desarrolladores. Para obtener pruebas de la validación, complete los siguientes pasos:

1. Haga clic en la barra de desplazamiento vertical para configurar la RP en 50 por ciento, como se muestra en la Figura 10.8.
2. Haga clic en **Save Bookmark data**. IATA presentará un cuadro de diálogo de confirmación para notificarle que los datos han sido guardados.
3. Presione **Next>>** para avanzar hacia la pantalla de visualización de resultados.
4. Seleccione la tabla **BookmarkData** del menú desplegable.

La Figura 10.9 muestra los resultados del análisis con el método *bookmark*. Los datos incluyen el nombre del ítem (**Name**), los parámetros TRI (**a, b, c**), la clasificación original de nivel (**Level**), el archivo fuente de las estadísticas del ítem (**Source**) y los valores RP 50 para cada ítem. Por ejemplo, el tercer ítem de la Figura 10.9 (**MATHC1025**) tiene valores de 0.90 y –0.78 para los parámetros **a** y **b** respectivamente

FIGURA 10.8

Interfaz de estándares de rendimiento, RP = 50 por ciento; datos de CYCLE1

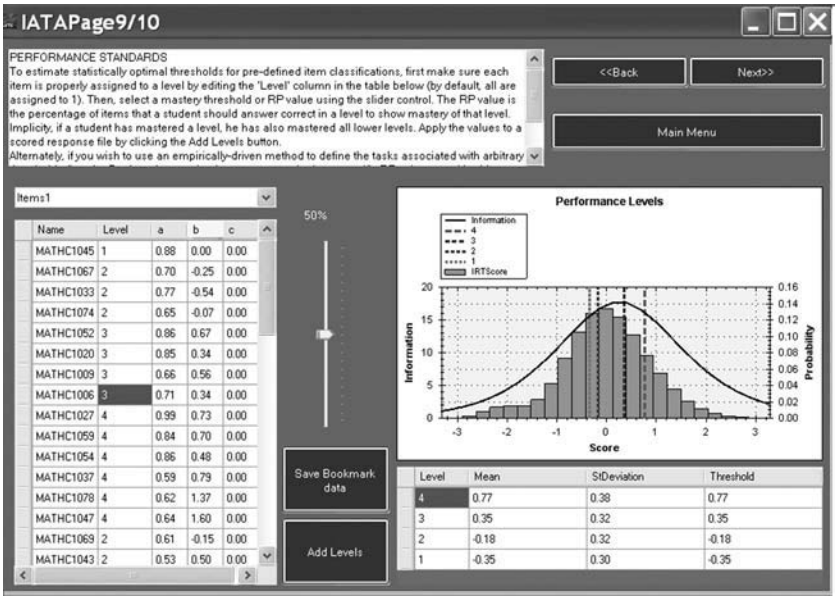
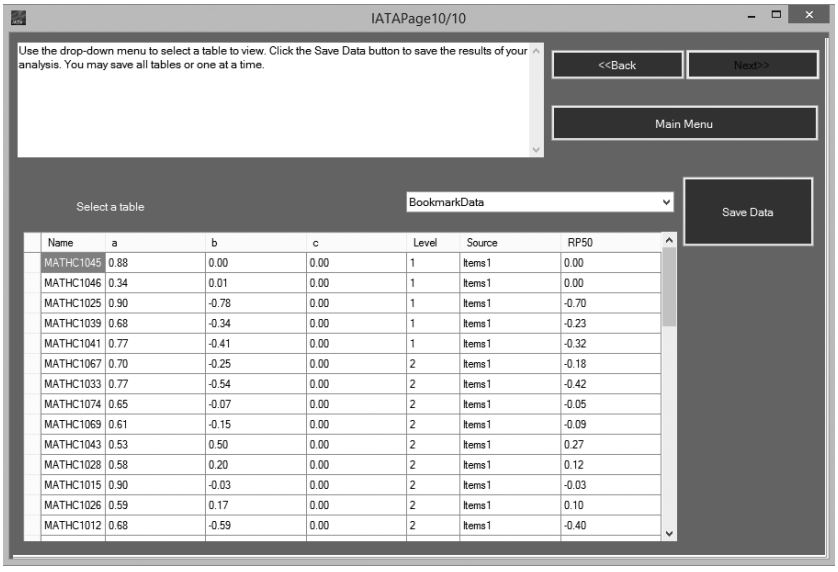


FIGURA 10.9

Datos de marcadores, RP = 50 por ciento; datos de CYCLE1



y se clasificó inicialmente como nivel 1, con un valor de RP 50 de –0.70 (lo que indica que tiene una probabilidad del 50 por ciento de tener un puntaje de competencia de –0.70). (En este caso, el ítem tiene solo una columna de valores RP, pero una tabla de datos de marcadores puede incluir varias columnas de valores RP.) La tabla de resultados seleccionada se debe exportar y entregar al comité de expertos. Los valores en la columna **RP50** informan el orden de presentación de los ítems en el método *bookmark* para clasificar los ítems en niveles de competencia y definir puntos de corte.

Con el procedimiento *bookmark*, los miembros del comité revisan cada ítem siguiendo el orden de sus valores RP. Cuando encuentran un ítem que creen que representa un estándar de rendimiento más alto, agregan un "marcador" en esa hoja del cuadernillo especialmente diseñado. Los valores RP inmediatamente anteriores a los marcadores representan los umbrales propuestos para determinar los niveles de competencia.

En general, se emplea una combinación de debate en grupo y cálculo estadístico de promedio para fusionar los umbrales producidos por diferentes revisores, a fin de generar umbrales finales, incluso si no son estadísticamente óptimos. Para el desarrollo de descripciones cualitativas de cada nivel de competencia, los ítems se clasifican según los umbrales finales.

El comité responsable de la determinación de estándares debe recibir distintos tipos de información simultáneamente, tales como especificaciones de ítems, referencias del currículo y definiciones normativas de lo que los alumnos saben y pueden hacer en cada nivel de competencia. El comité debe conciliar las distintas fuentes de información para determinar los puntos de corte más útiles y la asignación de los ítems de la prueba a los niveles. Los miembros del comité podrán, según su criterio, optar por usar clasificaciones de ítems definidas por adelantado por los desarrolladores, en vez de reclasificar los ítems sobre la base de los resultados del procedimiento *bookmark*. En ambos casos, los umbrales calculados por IATA representan los umbrales estadísticamente óptimos para las clasificaciones de ítems especificadas. Para generar umbrales óptimos desde el punto de vista estadístico, solo debe ajustar la RP a un porcentaje deseado, luego IATA realizará automáticamente el cálculo con la RP y la clasificación de nivel de los ítems y guardará los resultados en la tabla ***PLevels*** en

el conjunto de resultados del análisis. Por defecto, a menos que ingrese manualmente los valores en la tabla, IATA guarda los umbrales correspondientes a un RP de 67 por ciento. Observe que IATA no actualiza automáticamente el nivel asignado a un ítem; si la clasificación de un ítem se modifica con el método *bookmark* o algún otro procedimiento de clasificación de ítems, en los datos de entrada del ítem o directamente en IATA se deberá ingresar el nuevo nivel de clasificación.

Puede modificar de forma manual el nivel de umbral editando los umbrales directamente en la tabla de resultados. Una vez modificados los valores, se actualizará automáticamente el gráfico. Los ajustes más frecuentes incluyen la equidistancia entre los umbrales y la asignación de umbrales que, tras aplicar constantes de configuración de escala, se establecerán a incrementos enteros (por ejemplo, 5 o 10). Deberá adoptarse un criterio profesional al conciliar las pruebas del análisis estadístico y de contenido con la necesidad de comunicar los resultados a públicos no expertos. Se deberá lograr un equilibrio entre la simplicidad y la comunicación exacta de diferencias significativas que hubiera en el rendimiento de los alumnos.

Para el ejemplo actual, supongamos que el comité, tras considerar los datos iniciales de marcadores y otras fuentes de información, propone definir los niveles con el siguiente conjunto de puntos de corte: –0.85, –0.25, 0.35 y 0.95. Los alumnos con puntajes inferiores a –0.85 se clasificarían como de nivel 1; los alumnos con puntajes en la categoría de –0.85 a –0.24 se clasificarían como de nivel 2, y así sucesivamente.

Haga clic en **<<Back** para volver a la interfaz de estándares de rendimiento, para poder registrar estos puntos de corte en el archivo de datos de resultados y asignar los alumnos a los niveles adecuados. Siga los pasos a continuación:

1. Ingrese los valores recomendados producidos por el panel de partes interesadas en las filas correspondientes debajo de la columna con la etiqueta **Threshold**. Presione **Enter** cuando termine de cargar los valores para asegurarse de que IATA actualice la interfaz correctamente.
2. Presione el botón **Add Levels**. IATA asignará los alumnos a los niveles apropiados sobre la base de sus variables IRT score.

La Figura 10.10 muestra la asignación de los umbrales para los niveles de rendimiento. Estos niveles son equidistantes, lo que permite asignar una proporción lógica de alumnos a cada nivel. A pesar de que no existe ninguna razón matemática que respalde esta decisión, una práctica usual en la mayoría de las evaluaciones nacionales e internacionales es emplear umbrales equidistantes, ya que resultan más intuitivos a los públicos no expertos, que son los destinatarios principales de los resúmenes de niveles de competencia.

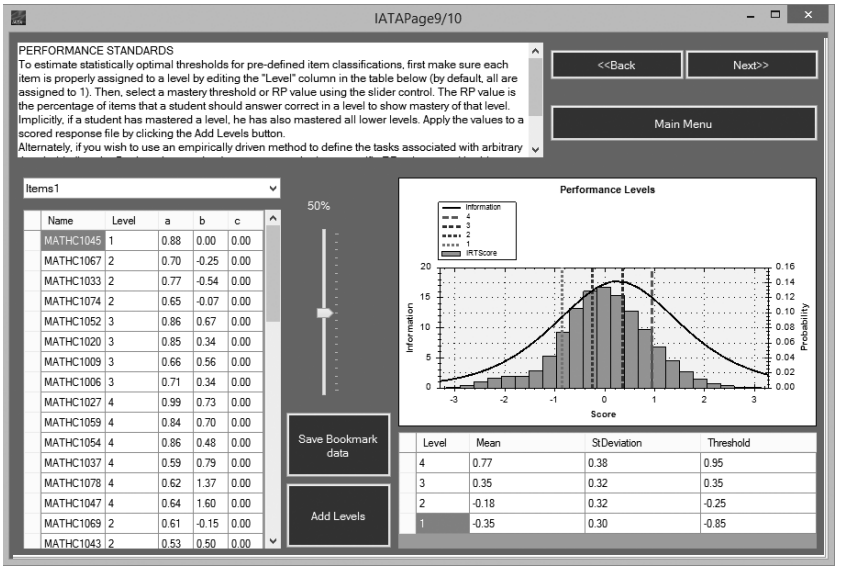
En la tabla de datos ***Scored***, que se puede visualizar en la pantalla final del flujo de trabajo del análisis, el registro de cada alumno también incluye una variable llamada **Level**. Esta variable contiene el nivel de estándar de rendimiento al que se asigna a cada alumno sobre la base de los umbrales que se muestran en la Figura 10.10. Por ejemplo, se clasificó al primer alumno de la lista debajo de **CYCLE1STDID** como de nivel 4, con un puntaje en percentil de 85.13 y un valor variable TRI de 1.06. Observe que, en este ejemplo, no se modificaron las medias y desviaciones estándar de los parámetros **b** para cada nivel en la tabla de resumen. Debido a que estos valores son resúmenes de estadísticas de ítems en vez de puntajes de alumnos, solo se modificarán si se actualiza la clasificación de los ítems; esto se puede realizar tanto en el archivo clave de ítems como directamente en la columna con la etiqueta **Level** de la tabla que aparece a la izquierda en la Figura 10.10.

Ya que el comité de expertos empleó umbrales para clasificar ítems en niveles de competencia o rendimiento, ahora debe efectuar descripciones cualitativas de los niveles, que especifiquen los conocimientos y las habilidades que indican competencia en cada nivel. Puede examinar ejemplos de descripciones y competencias de nivel en otros volúmenes de esta serie, en particular en el volumen 1, *Evaluación de los niveles nacionales de rendimiento académico*, Figura B3.3 referente a PISA (Greaney y Kellaghan, 2008) y el volumen 5, *Utilización de los resultados de una evaluación nacional del rendimiento académico*, Tabla 2.6 referente a NAEP; Tabla 2.7 referente a Vietnam; y Tabla 6.2 referente a Mozambique (Kellaghan, Greaney y Murray, 2009).

Luego de definir los umbrales de los estándares de rendimiento y aplicarlos a los puntajes de los alumnos, presione **Next>>** para pasar a la interfaz de visualización y almacenamiento de resultados.

FIGURA 10.10

Interfaz de estándares de rendimiento con umbrales definidos manualmente; datos de CYCLE1



Name	Level	a	b	c
MATHC1045	1	0.88	0.00	0.00
MATHC1067	2	0.70	-0.25	0.00
MATHC1033	2	0.77	-0.54	0.00
MATHC1074	2	0.65	-0.07	0.00
MATHC1052	3	0.86	0.67	0.00
MATHC1020	3	0.85	0.34	0.00
MATHC1009	3	0.66	0.56	0.00
MATHC1006	3	0.71	0.34	0.00
MATHC1027	4	0.99	0.73	0.00
MATHC1059	4	0.84	0.70	0.00
MATHC1054	4	0.86	0.48	0.00
MATHC1037	4	0.59	0.79	0.00
MATHC1078	4	0.62	1.37	0.00
MATHC1047	4	0.64	1.60	0.00
MATHC1069	2	0.61	-0.15	0.00
MATHC1043	2	0.53	0.50	0.00

Level	Mean	StDeviation	Threshold
4	0.77	0.38	0.95
3	0.35	0.32	0.35
2	-0.18	0.32	-0.25
1	-0.35	0.30	-0.85

PASO 7: ALMACENAMIENTO DE RESULTADOS

En la interfaz para visualizar y guardar los resultados, puede ver lo que se ha generado con esta guía de ejemplo. Deben guardarse todas las tablas por razones de documentación del proyecto y también para facilitar la vinculación de la prueba con ciclos posteriores de datos. Como referencia, los resultados de los datos de ítems de esta guía se incluyen en el archivo ***ItemDataAllTests.xls***, en la planilla llamada ***ReferenceC1***.

NOTA

1. La tabla ***Items2*** también está disponible para los flujos de trabajo de análisis que emplean vinculaciones.

CAPÍTULO 11

ANÁLISIS DE LA ROTACIÓN DE CUADERNILLOS

Use el conjunto de datos de muestra ***PILOT2*** para completar este ejercicio. La lista de respuestas de esta prueba está en el cuaderno de ejercicios de Excel ***ItemDataAllTests*** de la hoja ***PILOT2***.

Con los diseños que usan rotación de cuadernillos, se puede evaluar una gran cantidad de ítems agrupándolos en diferentes cuadernillos de prueba. Se emplean diferentes cuadernillos para administrar las pruebas, de modo que ningún alumno complete todos los ítems. Con excepción de las especificaciones iniciales de análisis, el flujo de trabajo restante sigue los procedimientos que se describieron en las guías anteriores.

PASO 1: CARGA DE DATOS

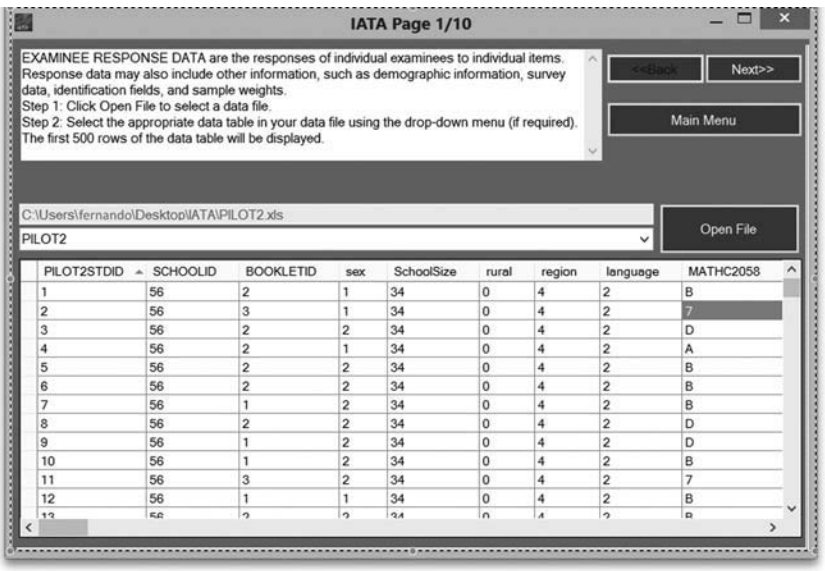
El análisis comienza con el flujo de trabajo **Response data analysis.** En la interfaz de datos de las respuestas del examinado, seleccione el archivo de datos de muestra ***PILOT2***. Este archivo cuenta con un diseño de tres cuadernillos que se administrará a 712 participantes. El archivo de datos contiene un total de 107 variables, con 99 ítems. No se incluyen todos los ítems en cada uno de los cuadernillos, pero

podría ocurrir si el comité director nacional de evaluación solicitó que la prueba sea lo suficientemente larga para cubrir un plan de estudios extenso. A fin de reducir la fatiga de los alumnos, cada uno de ellos completará un cuadernillo con un subconjunto de ítems. En la Figura 11.1, la tercera columna contiene una variable llamada **BOOKLETID**. Allí se incluyen valores como 1, 2 o 3, que hacen referencia a los cuadernillos que completaron los alumnos. Además de las respuestas del ítem alfabético y el código 9 de respuesta faltante (que no se muestran en la figura), hay un valor 7 que aparece con frecuencia y que indica que no se incluyó un ítem en el cuadernillo asignado al alumno. Por ejemplo, los datos de la figura indican que en el cuadernillo del alumno con **PILOT2STDID = 2** no se incluyó **MATHC2058**. El código 7 se trata como "omitido" y no afecta los resultados de la prueba de un alumno. Haga clic en **Next>>** para continuar.

En la interfaz de carga de datos del ítem, cargue los datos del ítem ***PILOT2*** del archivo ***ItemDataAllTests.xls***. La tabla ***PILOT2*** contiene 99 registros y cuatro variables. Tenga en cuenta que **MATHC2047**

FIGURA 11.1

Respuestas del alumno, datos de PILOT2



PILOT2STDID	SCHOOLID	BOOKLETID	sex	SchoolSize	rural	region	language	MATHC2058
1	56	2	1	34	0	4	2	B
2	56	3	1	34	0	4	2	7
3	56	2	2	34	0	4	2	D
4	56	2	1	34	0	4	2	A
5	56	2	2	34	0	4	2	B
6	56	2	2	34	0	4	2	B
7	56	1	2	34	0	4	2	B
8	56	2	2	34	0	4	2	D
9	56	1	2	34	0	4	2	D
10	56	1	2	34	0	4	2	B
11	56	3	2	34	0	4	2	7
12	56	1	1	34	0	4	2	B

tiene los siguientes valores: **Key** = C, **Level** = 1,00, y **Content** = number knowledge item.

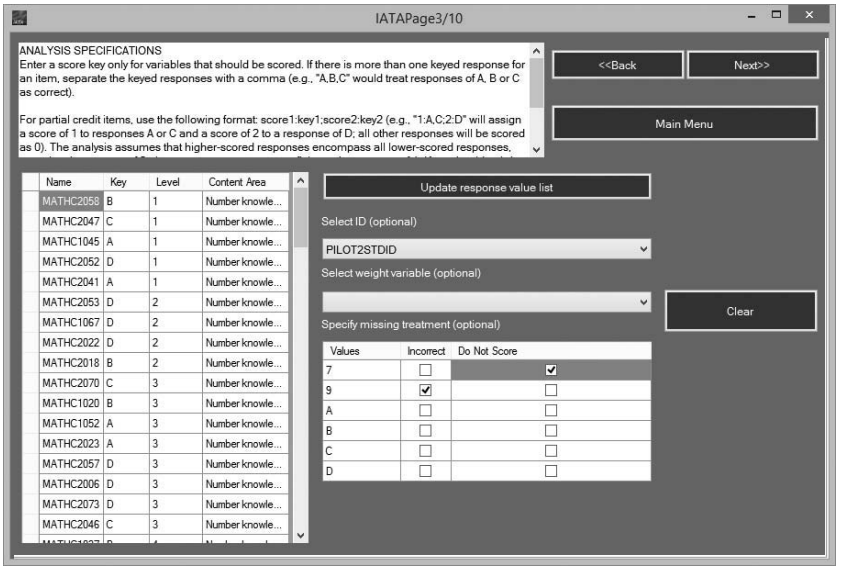
Confirme que se hayan cargado los datos correctos de la respuesta y el ítem, y luego haga clic en **Next>>** para continuar con las especificaciones de análisis.

PASO 2: ESPECIFICACIONES DE ANÁLISIS

Debajo de **Select ID (optional)**, seleccione **PILOT2STDID**. En las especificaciones **Specify missing treatment (optional)** de los ejemplos anteriores, la casilla de la columna titulada **Incorrect** se marcó únicamente en los casos de datos faltantes. Por el diseño de rotación de cuadernillos, debe incluir un código de omisión en la tabla **Specify missing treatment (optional)** para indicar que no se deben puntuar algunas de las respuestas. Marque el valor 7 en la columna **Do Not Score**, como se indica en la Figura 11.2. Cuando se observe un valor

FIGURA 11.2

Especificaciones de análisis, rotación de cuadernillos, datos de PILOT2



7 en los datos de respuesta de algún alumno, IATA (*Item and Test Analysis*) ignorará el ítem, de modo que no afectará los resultados del alumno. Asimismo, el cálculo de las estadísticas o parámetros no se verá afectado en los participantes que cuenten con códigos de respuesta 7 en un ítem.

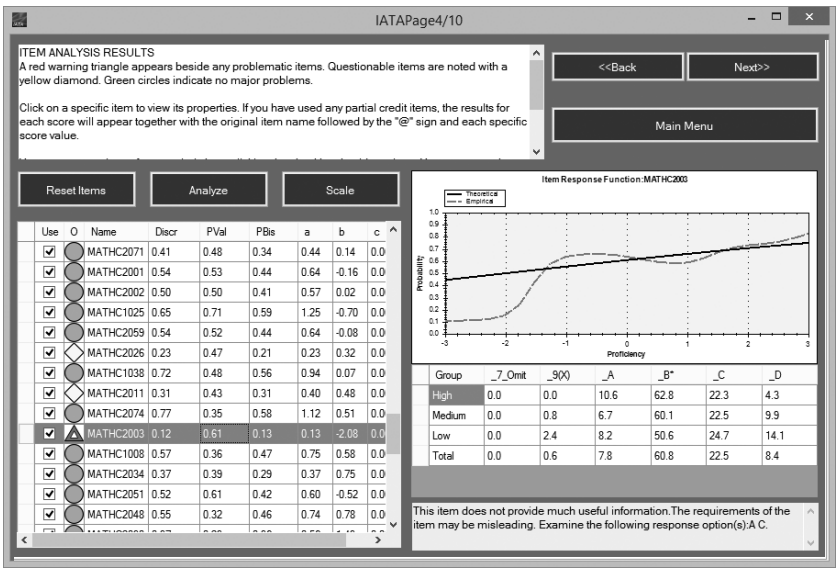
Luego de ingresar las especificaciones de análisis, haga clic en **Next>>** para continuar. El análisis se iniciará automáticamente. Dado que el tiempo computacional se ve afectado más por la cantidad de ítems de prueba que por la cantidad de alumnos, el análisis llevará más tiempo de ejecución que en el resto de las guías.

PASO 3: RESULTADOS DEL ANÁLISIS DEL ÍTEM

Cuando IATA termine de ejecutar el análisis, aparecerán los resultados en la Figura 11.3, donde la tabla se desplegó hacia abajo para mostrar **MATHC2003**, al que se le asignó un símbolo de advertencia. Los resultados indican que este ítem está poco relacionado con la

FIGURA 11.3

Resultados del análisis del ítem, datos de PILOT2, MATHC2003



competencia. Los alumnos suelen tener una probabilidad de 0,61 de seleccionar la respuesta correcta, independientemente de su nivel de dominio. IATA sugiere que esta debilidad puede derivar de requisitos confusos, que habitualmente están asociados con distractores ineficientes. Cuando los alumnos no comprenden los requisitos de un ítem o cuando no hay una respuesta inequívocamente correcta, tienden a adivinar. Sobre el lado derecho de la interfaz, en la tabla de análisis de los distractores, se presenta un resumen de la información, a partir del cual se observa que la opción D es el único distractor que produce la conducta deseada. La columna vacía debajo del título **Omit** es para recordar que el código 7 no debe influir sobre el cálculo. A fin de evitar que **MATHC2003** reduzca potencialmente la precisión de los resultados del análisis, desmarque la casilla que se encuentra al lado del nombre del ítem de la columna **Use** y haga clic en **Analyze** para actualizar los resultados.

Luego de revisar los resultados del análisis, proceda a replicar los análisis que se explicaron en los capítulos anteriores. La especificación e interpretación de las tareas restantes del flujo de trabajo son básicamente las mismas que las presentadas en las guías previas.

Si para administrar la prueba se emplean múltiples cuadernillos con diferentes combinaciones de ítems, IATA puede efectuar el análisis dimensional solo si los diferentes cuadernillos comparten una cierta cantidad de ítems. Por ejemplo, con tres bloques de ítems de prueba (A, B, C) y tres formularios de prueba (1, 2, 3), el formulario de prueba 1 debe incluir los bloques A/B; el formulario de prueba 2, los bloques B/C; y el formulario de prueba 3, los bloques C/A. Dado que las secciones se rotan por completo, no quedan ítems huérfanos. Por el contrario, si el formulario de prueba 1 contiene el bloque A/B, el formulario de prueba 2 contiene el bloque A/B, y el formulario de prueba 3 contiene el bloque B/C, entonces el bloque C queda huérfano del bloque A, y así no pueden calcularse las correlaciones entre los ítems del bloque A y el bloque C. Si tiene bloques de ítems huérfanos y desea realizar un análisis dimensional, debe eliminar los ítems huérfanos del análisis o efectuar un análisis dimensional en cada formulario de prueba por separado. No obstante, cuando haya suficiente evidencia para confirmar que el conjunto colectivo de ítems está evaluando una única dimensión, puede incluir la totalidad de los ítems en

el análisis para calcular los parámetros y puntuaciones del ítem según la Teoría de Respuesta al Ítem.

Si la rotación es compleja, el principio general para un diseño eficiente es garantizar que ningún ítem de prueba aparezca solamente en un único cuadernillo. Si un ítem no aparece en un único cuadernillo solamente, los cálculos del parámetro del ítem pueden estar sujetos a un incremento de los errores de muestreo ya que los cálculos estarán asociados específicamente con los ítems de ese cuadernillo y los alumnos a quienes se les asignó. A medida que aumenta la cantidad de cuadernillos en los que aparece un ítem, también lo hace la precisión de los cálculos del parámetro de ese ítem.

Pueden repetirse los análisis tratados en los capítulos anteriores con los datos de ***PILOT2*** como ejercicio independiente. Se pueden consultar los resultados de los datos del ítem de esta guía de análisis en el archivo ***ItemDataAllTests.xls***, en la hoja de cálculo ***ReferenceP2***.

CAPÍTULO 12

ANÁLISIS DE LOS ÍTEMS DE CRÉDITO PARCIAL

Use el conjunto de datos de muestra ***PILOT2PartialCredit*** para completar este ejercicio. La lista de respuestas de esta prueba está en el cuaderno de ejercicios de Excel ***ItemDataAllTests*** del documento llamado ***PILOT2PartialCredit***.

Al igual que en el escenario de evaluación nacional que se describió en capítulos anteriores de este volumen, esta guía sigue el desarrollo continuo del instrumento de prueba con la introducción de análisis para los ítems de respuesta corta que se puntuaron utilizando créditos parciales. Excepto las especificaciones iniciales de análisis, en los últimos cuatro ítems de ***PILOT2*** (detallados a continuación), el flujo de trabajo restante sigue los procedimientos que se describieron en las guías anteriores. Esta guía se centra en los requisitos únicos del análisis del ítem de crédito parcial (véase Anderson y Morgan, 2008).

PASO 1: CARGA DE DATOS

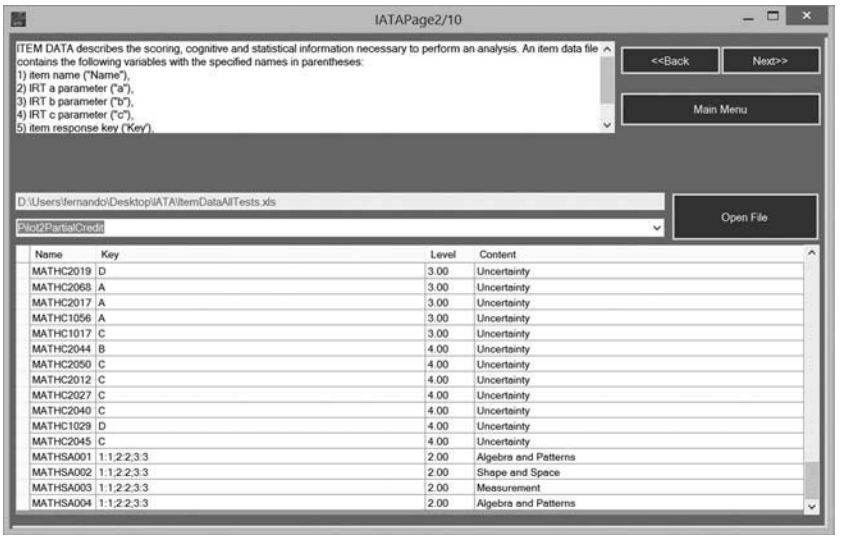
El análisis comienza con el flujo de trabajo **Response data analysis** (Análisis de los datos de las respuestas). En la interfaz de datos de las respuestas del examinado, seleccione el archivo de datos de muestra ***PILOT2PartialCredit***. Estos datos contienen 111 variables y

712 participantes. Los datos presentan un diseño compuesto por tres cuadernillos que incluyen 103 ítems de prueba, los mismos que se utilizaron en el análisis del diseño con rotación equilibrada de cuadernillos del capítulo 11, a los que se le sumaron cuatro ítems de crédito parcial. En cada uno de estos ítems, los alumnos pueden haber obtenido puntuaciones de 0, 1, 2 o 3, según la calidad de sus respuestas. IATA (*Item and Test Analysis*) producirá estadísticas de cualquier valor superior a 0. Los nombres de los ítems adicionales son **MATHSA001, MATHSA002, MATHSA003** y **MATHSA004**. Haga clic en **Next>>** para continuar.

En la interfaz de carga de datos del ítem, cargue los datos del ítem ***PILOT2PartialCredit*** del archivo ***ItemDataAllTests.xls***, como se indica en la Figura 12.1, donde se deslizó el panel de datos hasta el fondo. Las últimas cuatro filas del archivo de datos contienen información de los ítems de crédito parcial. La clave de puntuación de estos ítems contiene la información necesaria para asignar diferentes valores numéricos a los diversos códigos en el archivo de respuesta, según la calidad de la respuesta del alumno. En este ejemplo, la clave

FIGURA 12.1

Lista de respuestas de ítem y metadatos, datos de PILOT2



Name	Key	Level	Content
MATHC2019	D	3.00	Uncertainty
MATHC2068	A	3.00	Uncertainty
MATHC2017	A	3.00	Uncertainty
MATHC1056	A	3.00	Uncertainty
MATHC1017	C	3.00	Uncertainty
MATHC2044	B	4.00	Uncertainty
MATHC2050	C	4.00	Uncertainty
MATHC2012	C	4.00	Uncertainty
MATHC2027	C	4.00	Uncertainty
MATHC2040	C	4.00	Uncertainty
MATHC1029	D	4.00	Uncertainty
MATHC2045	C	4.00	Uncertainty
MATHSA001	1:1;2:2;3:3	2.00	Algebra and Patterns
MATHSA002	1:1;2:2;3:3	2.00	Shape and Space
MATHSA003	1:1;2:2;3:3	2.00	Measurement
MATHSA004	1:1;2:2;3:3	2.00	Algebra and Patterns

de puntuación refleja el puntaje manual de cada ítem: el código 1 se puntúa como 1, el código 2, como 2 y el código 3, como 3. No es necesario proporcionar una especificación de puntuación para un valor de 0. Si un código de respuesta no está en la lista de respuestas y no se trata como faltante, se le asignará un puntaje de 0.

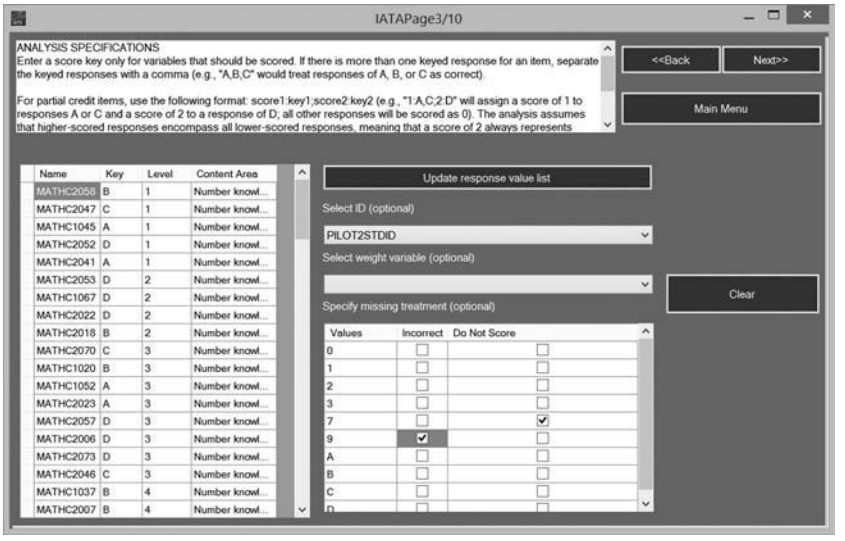
Confirme que se hayan cargado los datos correctos de la respuesta y el ítem, y luego haga clic en **Next>>** para continuar con las especificaciones de análisis.

PASO 2: ESPECIFICACIONES DE ANÁLISIS

Puesto que estos datos son prácticamente idénticos a los utilizados en el capítulo 11, use las mismas especificaciones de análisis. Debajo de **Select ID (optional)**, ingrese **PILOT2STDID**, y en la sección **Specify missing treatment (optional)**, marque el valor 7 en la columna **Do Not Score**, como se indica en la Figura 12.2. Tenga en cuenta que

FIGURA 12.2

Especificaciones de análisis, rotación de cuadernillos con ítems de crédito parcial, datos de PILOT2



todas las puntuaciones de los ítems correspondientes a los ítems de crédito parcial también aparecen en la tabla de valores de los ítems.

Luego de ingresar las especificaciones de análisis, haga clic en **Next>>** para continuar. El análisis se iniciará automáticamente. Dado que el tiempo computacional se ve afectado más por la cantidad de ítems de prueba que por la cantidad de alumnos en los datos, el análisis llevará más tiempo de ejecución que en el resto de las guías.

PASO 3: RESULTADOS DEL ANÁLISIS DEL ÍTEM

Cuando IATA termine de ejecutar el análisis, los resultados indicarán que **MATHC2003** es problemático (Figura 12.3). Elimine el ítem del análisis desmarcando la casilla que se encuentra al lado del nombre del ítem de la columna **Use** y haciendo clic en **Analyze** para actualizar los resultados. IATA le preguntará si debe recalibrar los ítems de crédito parcial. Haga clic en **Yes** para proceder.

En la tabla izquierda de la Figura 12.4, puede observar las filas que IATA creó automáticamente para cada puntuación de un ítem en

FIGURA 12.3

Resultados del análisis del ítem, datos de PILOT2, MATHC2003

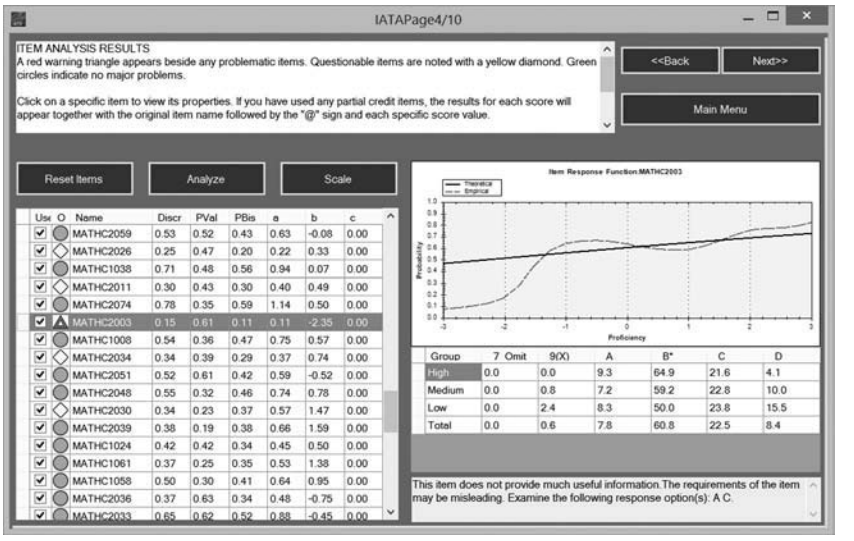
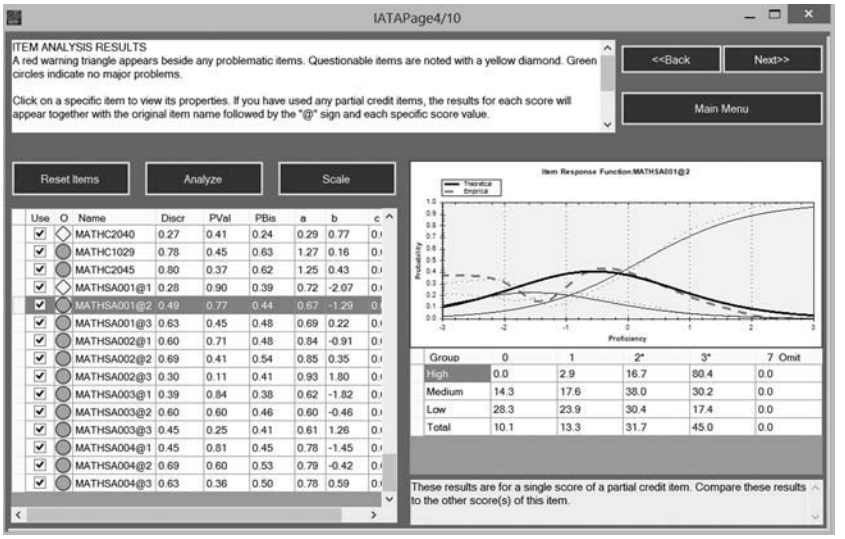


FIGURA 12.4

Función de respuesta del ítem de crédito parcial, datos de CYCLE2, MATHSA001, puntuación = 2



cada uno de los ítems de crédito parcial. En las filas que representan puntuaciones de ítems de crédito parcial (donde la columna **Name** contiene el símbolo @ seguido de un número entero), las estadísticas se calculan como si cada puntuación fuera un único ítem correcto/incorrecto en el que la respuesta correcta es cualquier valor superior o igual a la puntuación seleccionada. IATA creará un conjunto adicional de resultados estadísticos para cada puntuación de créditos parciales incluida en la clave de puntuaciones de un ítem. Por ejemplo, con un ítem de crédito parcial que tenga puntuaciones (que no sean cero) de 1 y 2, la facilidad del ítem para la puntuación de 1 (**ItemName@1**) describiría la proporción de alumnos con puntajes mayores o iguales a 1, mientras que la facilidad del ítem para la puntuación de 2 (**ItemName@2**) describiría la proporción de alumnos con un puntaje de 2.

En la tabla de análisis de distractores de la Figura 12.4, tenga en cuenta que **MATHSA001@2** usa los códigos 2 y 3 como respuestas correctas. Por otro lado, **MATHSA001@1**, usaría los códigos 1, 2 y 3 como respuestas correctas. La facilidad del ítem siempre es mayor con

los puntajes más bajos de un ítem ya que incluye a todos los alumnos a quienes se les asignaron puntuaciones más altas. Por ejemplo, en la Figura 12.4, se resaltan los resultados de **MATHSA001** con la puntuación de 2 seleccionada. En este ítem, la puntuación de 1 (**MATHSA001@1**) tiene un valor de **PVal** (facilidad del ítem) de 0,90, la puntuación de 2 (**MATHSA001@2**) tiene un valor de **PVal** de 0,77, y la puntuación de 3 (**MATHSA001@3**) tiene un valor de **PVal** de 0,45.

Si bien las estadísticas describen cada puntuación por separado, las funciones de respuesta del ítem (FRI) correspondientes a los ítems de crédito parcial señalan que a un alumno solo se le puede asignar un único valor de puntuación. La función de respuesta del ítem de un ítem de crédito parcial se representa como un conjunto de curvas características de la categoría del ítem (CCCI), una por cada puntuación del ítem. Cuando se selecciona una fila correspondiente a una puntuación específica, en el gráfico aparecen las CCCI de la puntuación seleccionada en negrita. En cada nivel de desempeño, la CCCI expresa la probabilidad de que a un participante con un determinado nivel de desempeño se le asigne una puntuación específica, y no el resto. Como puede observarse en la Figura 12.4, a medida que aumenta el desempeño, la probabilidad de cada puntaje primero aumenta, luego disminuye, a medida que se incrementa la probabilidad de que los alumnos obtengan puntuaciones más altas.

Si bien no hay reglas simples aplicables al análisis de una función de respuesta del ítem para un ítem de crédito parcial, un esquema útil para la puntuación de crédito parcial suele tener la propiedad de que cada puntuación tendrá la máxima probabilidad de ser seleccionada en un cierto intervalo de desempeño. Por ejemplo, al primer valor de puntuación de **MATHSA001** se le asignó un símbolo de precaución. La CCCI indica que la probabilidad de que se asigne un puntaje de 1 no es superior a ningún otro puntaje en ningún nivel de desempeño. El puntaje de 2 de **MATHSA001**, que se representa como una curva con forma de campana y un pico de aproximadamente –0,5, es el valor más probable para todos los alumnos con un desempeño por debajo del promedio. Estos resultados indican que la puntuación de 1 no aporta información útil ya que es estadísticamente indistinguible de la puntuación de 2. Con la mayoría de los ítems de crédito parcial,

la categoría más alta es, en general, la más útil, ya que los marcadores humanos tienden a poder identificar con claridad las respuestas completamente incorrectas o completamente correctas, pero tienen menos capacidad para distinguir con precisión entre diferentes grados de corrección parcial.

Una de las tareas principales del análisis de los ítems de crédito parcial en las pruebas piloto es determinar si todos los valores de puntuación son útiles y cómo mejoran el proceso de asignación de puntajes. Por ejemplo, si una categoría de puntuación (como 1) tiene una probabilidad baja de ser asignada, existen dos posibilidades principales: ningún participante (o unos pocos) produjo respuestas que correspondan con ese puntaje, o los correctores no lograron identificar a qué respuestas se les debería asignar el puntaje. En el primer caso, las respuestas asociadas con la puntuación deben consolidarse con una categoría de puntuación adyacente (como fusionar los puntajes 1 y 2 en una categoría de puntuación de 1 o 2). En el segundo caso, el problema se puede resolver con una capacitación más intensiva o estandarizada de los correctores. Por consiguiente, aunque los resultados de los análisis de los ítems dicotómicos son principalmente relevantes para los redactores de ítems y los creadores de la prueba, los resultados del análisis de crédito parcial también deben ser compartidos con los equipos responsables de la puntuación de los cuadernillos de los alumnos.

Luego de revisar los resultados, proceda a replicar los análisis que se explicaron en los capítulos anteriores. La especificación e interpretación de las tareas restantes del flujo de trabajo son básicamente las mismas que se presentaron en las guías previas. La única diferencia se encuentra en la especificación requerida para los ítems de crédito parcial en la interfaz de selección de ítems de IATA.

Para completar la selección adecuadamente con los ítems de crédito parcial, primero debe ingresar manualmente todos los puntajes de los ítems de crédito parcial marcando las puntuaciones del ítem correspondiente y luego contar cada puntaje como un ítem separado al ingresar la cantidad total de ítems. Por lo tanto, si desea ingresar 10 ítems y uno de esos ítems es uno de crédito parcial con dos categorías de puntuación, debe especificar una selección de 11 ítems y preseleccionar de forma manual los valores de puntuación de ítem

para el ítem de crédito parcial deseado. Pueden repetirse los análisis tratados en los capítulos anteriores con los datos de ***PILOT2*** como ejercicio independiente.

Se pueden consultar los resultados de los datos del ítem de esta guía de análisis en el archivo ***ItemDataAllTests.xls***, en la hoja de cálculo ***ReferenceP2PC***.

CAPÍTULO 13

COMPARACIÓN DE EVALUACIONES

Utilice la recopilación de datos ***CYCLE2*** para llevar a cabo este ejercicio. La clave de respuestas para esta prueba está en el libro de trabajo de Excel ***ItemDataAllTests.xls*** solapa ***CYCLE2***.

Se utilizan ítems comunes para vincular la Teoría de Respuesta al Ítem (TRI) y las escalas de informes públicos de los datos de ***CYCLE2*** con las respectivas escalas utilizadas en los datos de ***CYCLE1***. Los resultados vinculados se utilizan para escalar calificaciones y para aplicar los estándares de rendimiento.

A menudo, los gobiernos y otras partes están interesados en determinar si los niveles de rendimiento de los estudiantes se elevaron, decayeron o permanecieron constantes en el tiempo. El interés en modificar los estándares es particularmente importante en los períodos en los cuales se modifican los currículos o se producen reformas substanciales en el sistema (por ejemplo, un cambio en los niveles de financiamiento). Los gobiernos también pueden estar interesados en los efectos sobre el rendimiento de los estudiantes de un rápido crecimiento de la matrícula causado por la implementación de programas tales como Educación para Todos o Iniciativa de Vía Rápida, ahora conocida como Alianza Mundial por la Educación. Con un fuerte vínculo, las puntuaciones de una evaluación nacional pueden compararse

con una evaluación llevada a cabo en un momento precedente, permitiendo concluir si se produjo un cambio concomitante en el rendimiento estudiantil (véase Mislevy, 1992).

La vinculación de los resultados de las pruebas es útil en otras situaciones. Además, al comparar los resultados a través de los ciclos, son comunes los siguientes escenarios. Un país con una serie de jurisdicciones educativas puede crear una prueba para cada jurisdicción con contenidos curriculares específicos de cada jurisdicción. Si las pruebas para las diferentes jurisdicciones comparten ítems comunes, las puntuaciones de las diferentes pruebas pueden ser utilizadas para comparar el rendimiento de las diferentes jurisdicciones. La vinculación de pruebas también puede utilizarse para comparar los resultados de una evaluación nacional y una internacional si la evaluación nacional incluye ítems utilizados y calibrados previamente en una evaluación internacional. Por ejemplo, si un país participó previamente en el Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS), incluir un número de ítems TIMSS en una evaluación nacional posterior puede contribuir a detectar si el desempeño cambió desde la administración del TIMSS precedente. En este escenario, el procedimiento de vinculación utilizará los parámetros de ítems de TIMSS en el archivo de los datos de ítems de referencia que se estimaron, utilizando los datos de respuesta específicos de los estudiantes de cada país. Alternativamente, un procedimiento de vinculación puede ayudar a identificar cómo se compara el desempeño del país en una evaluación nacional con el desempeño de otros países en TIMSS. En este caso, los archivos de datos de ítems deberían incluir los parámetros de ítems TIMSS que fueron estimados a partir de los datos TIMSS internacionales.

En el transcurso de este capítulo se describe los métodos necesarios para implementar un ciclo de seguimiento en un programa de evaluación nacional. Tras ejecutar el programa, necesitará seguir este proceso de trabajo para asegurar que las interpretaciones que hacen los sectores interesados utilizando los resultados de un nuevo ciclo de evaluación sean consistentes y comparables con las hechas durante el primer ciclo. Los datos ***CYCLE2*** de este ejemplo representan el segundo ciclo de una evaluación nacional, posterior al primer ciclo, ***CYCLE1***, que se analizó en el capítulo 10. Es posible una

FIGURA 13.1

Análisis de datos de respuesta con proceso de vinculación

vinculación entre estas dos evaluaciones porque la prueba ***CYCLE2*** contiene varios ítems de anclaje que también se utilizaron en la prueba ***CYCLE1***. La vinculación le permitirá monitorear cambios en el desempeño de los estudiantes a lo largo del tiempo. Si los diferentes ciclos de evaluación están vinculados a ítems comunes, los nombres de ítems únicos y permanentes ayudarán a rastrear qué ítems se utilizan en cada ciclo de evaluación a efectos de establecer vinculaciones.

Los contenidos del capítulo se centran en las interfaces y especificaciones que son distintivos de este proceso.

Repase los capítulos precedentes para explicaciones más detalladas de las interfaces de procesos comunes.

En el menú principal, haga clic en la primera opción de menú, **Response data analysis with linking**, para entrar al proceso de análisis, tal como se muestra en la Figura 13.1. Este proceso necesita datos de respuesta, datos de ítems (claves de respuesta) para los datos de respuesta que serán analizados, y un archivo de datos de respuesta de referencia que se usa para anclar los resultados vinculados.

PASO 1: CONFIGURACIÓN DEL ANÁLISIS

El proceso comienza con los mismos pasos de captura de datos iniciales, como el análisis de datos de respuesta.

1. En la primera interfaz del proceso, capture los datos de respuesta desde el archivo llamado ***CYCLE2.xls*** en la solapa de muestra de *Item and Test Analysis* (IATA). Estos datos incluyen 2484 registros y 61 variables. El primer caso tiene los siguientes valores: **SchoolID** = 2; **Sex** = 2; **SchoolSize** = 21; y **Rural** = 0. Haga clic en **Next >>**.

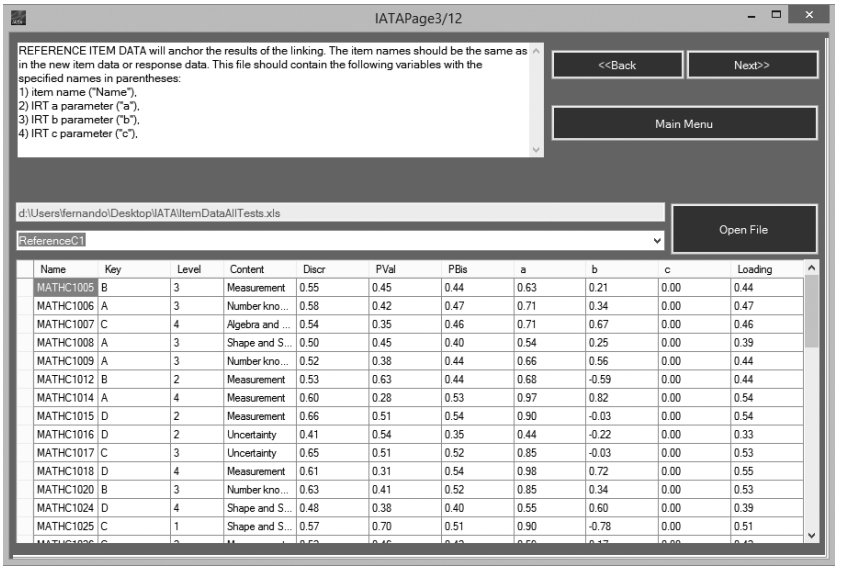
2. En la segunda interfaz, capture los correspondientes datos de ítems. Al abrir el archivo ***ÍtemDataAllTests.xls***, asegúrese de que esté seleccionada la tabla ***CYCLE2***. La tabla tiene 53 registros y cuatro variables, incluyendo tres ítems de crédito parcial. **MATHC2047** tiene los siguientes valores: **Key** = C, **Level** = 1.00, y **Content** = conocimiento de los números. Haga clic en **Next >>**.
3. Utilice una nueva interfaz de captura de datos que no haya sido utilizada en revisiones previas. La interfaz requiere un archivo que contenga datos de ítems de referencia. Los ítems de referencia contienen los parámetros TRI **a**, **b**, y **c** (si existe la opción), que se estimaron a partir de una muestra de referencia, tal como un ciclo previo de evaluación nacional o internacional.

El archivo de datos de ítems de referencia debe contener por lo menos algunos ítems que estén incluidos en la evaluación nacional en curso. Para este ejemplo, utilice los resultados producidos a partir del análisis del archivo de datos **CYCLE1**. Estos datos se proveen en la solapa de datos de muestra del archivo llamado ***ItemDataAllTests.xls***, en la hoja llamada ***ReferenceC1*** (también pueden haberse guardado los resultados de los ejercicios del capítulo 10), Al abrir este archivo, asegúrese de que la tabla seleccionada sea la tabla llamada ***ReferenceC1***. Esta tabla, que tiene 50 registros y once variables, contiene los resultados estadísticos que describen todos los ítems del primer ciclo de la evaluación nacional. Estos datos de referencia se ilustran en la Figura 13.2. En el presente ejemplo de evaluación nacional, la prueba ***CYCLE2*** incluye 25 ítems que tienen sus parámetros de ítems en el archivo de datos de ítems de referencia ***CYCLE1***. Es importante mantener la coherencia de los nombres de los ítems en todos los archivos de datos, porque en el procedimiento de vinculación IATA hace coincidir ítems utilizando sus nombres.

Nótese que este archivo también contiene varios campos de datos que se calcularon durante el análisis de los datos de ***CYCLE1*** además de las variables **a**, **b** y **c** (por ejemplo, **Level**, **Content**, **Discr**, **PVal**, **PBis** y **Loading**). Estas variables pueden conservarse en el archivo de datos pero no se utilizan en el análisis de vinculación. Asimismo, aunque los datos de ítems de referencia contienen información para los

FIGURA 13.2

Datos de ítems de referencia de CYCLE1 a ser vinculados con datos de CYCLE2



Name	Key	Level	Content	Discr	PVal	PBis	a	b	c	Loading
MATHC1005	B	3	Measurement	0.55	0.45	0.44	0.63	0.21	0.00	0.44
MATHC1006	A	3	Number kno...	0.58	0.42	0.47	0.71	0.34	0.00	0.47
MATHC1007	C	4	Algebra and ...	0.54	0.35	0.46	0.71	0.67	0.00	0.46
MATHC1008	A	3	Shape and S...	0.50	0.45	0.40	0.54	0.25	0.00	0.39
MATHC1009	A	3	Number kno...	0.52	0.38	0.44	0.66	0.56	0.00	0.44
MATHC1012	B	2	Measurement	0.53	0.63	0.44	0.68	-0.59	0.00	0.44
MATHC1014	A	4	Measurement	0.60	0.28	0.53	0.97	0.82	0.00	0.54
MATHC1015	D	2	Measurement	0.66	0.51	0.54	0.90	-0.03	0.00	0.54
MATHC1016	D	2	Uncertainty	0.41	0.54	0.35	0.44	-0.22	0.00	0.33
MATHC1017	C	3	Uncertainty	0.65	0.51	0.52	0.85	-0.03	0.00	0.53
MATHC1018	D	4	Measurement	0.61	0.31	0.54	0.98	0.72	0.00	0.55
MATHC1020	B	3	Number kno...	0.63	0.41	0.52	0.85	0.34	0.00	0.53
MATHC1024	D	4	Shape and S...	0.48	0.38	0.40	0.55	0.60	0.00	0.39
MATHC1025	C	1	Shape and S...	0.57	0.70	0.51	0.90	-0.78	0.00	0.51

50 ítems en la evaluación del primer ciclo, solo se utilizan para estimar la vinculación de la información de los 25 ítems comunes con la evaluación del segundo ciclo.

Luego de capturar los tres archivos de datos, haga clic en **Next>>** para continuar con la interfaz de especificaciones de análisis (IATA Page 4/12). Las especificaciones del análisis son similares a las de los datos de ***CYCLE1***. Ingrese o seleccione los siguientes detalles en la interfaz de especificaciones de análisis, y haga clic en **Next>>** para completar el análisis y ver los resultados:

- La variable de identificación del estudiante es **CYCLE2STDID**.
- La variable de ponderación es **CYCLE2Weight**.
- El código 9 es tratado como incorrecto.

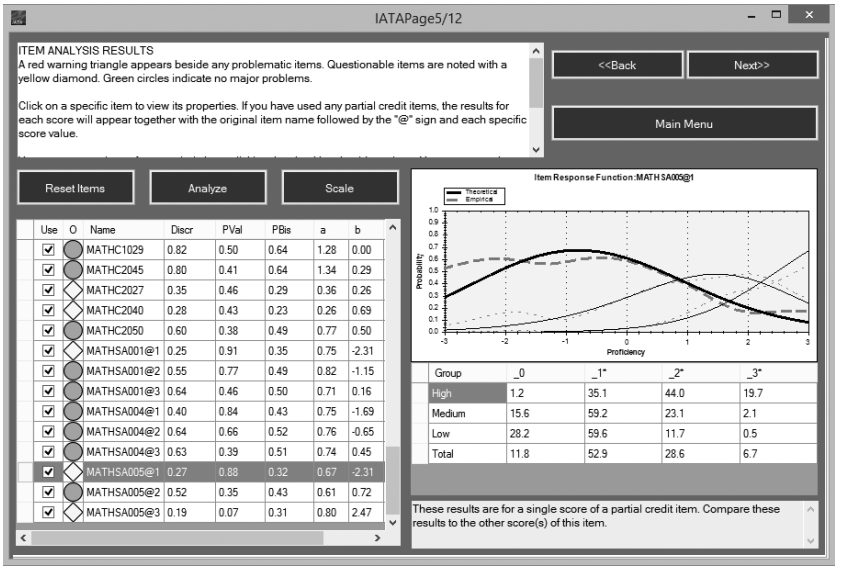
Al proceder con la interfaz de análisis de ítems (IATA Page 5/12) automáticamente comenzará el análisis. No hay ítems problemáticos en los datos. Al revisar cada ítem, nótese que, aunque los ítems de

crédito parcial tienen múltiples puntuaciones, algunos de ellos pueden seguir siendo "fáciles", en los que muchos alumnos alcanzan las categorías de puntuaciones más altas (tales como **MATHSA004**), y otros pueden ser "difíciles", en los que relativamente pocos alumnos alcanzan las categorías más altas de puntuación, tales como **MATHSA005** (Figura 13.3).

Avance a través del proceso de trabajo para revisar los resultados de dimensionalidad de la prueba y para llevar a cabo los análisis de funcionamiento diferencial de ítems (FDI) que puedan ser de interés sobre las variables demográficas (localización, sexo, idioma) siguiendo los mismos procedimientos descritos en capítulos previos. Aunque muchos de los ítems tienen símbolos de advertencia para uno o más análisis FDI, para los fines de este ejemplo se asume que todos los ítems son correctos. Luego de revisar el análisis FDI, haga clic en **Next>>** para continuar con la interfaz de vinculación.

FIGURA 13.3

Resultados del análisis de ítems para los datos de CYCLE2, MATHSA005, Score = 1

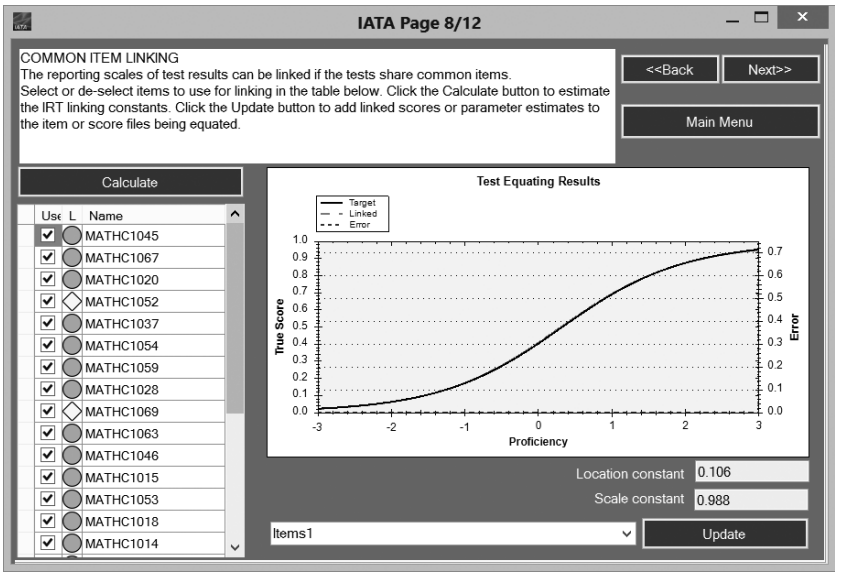


PASO 2: VINCULACIÓN DE ÍTEMS COMUNES

Haga clic en **Calculate** en el sector izquierdo de la interfaz (Figura 13.4). Esto automáticamente seleccionará y listará una tabla que contiene los 25 ítems que son comunes a los datos de ítems de referencia y a los nuevos datos que provienen de la evaluación nacional en curso. En la tabla de ítems, la primera columna, **Use**, especifica si debe incluirse o no el ítem en la estimación de las constantes de vinculación (por defecto, se incluyen todos los ítems que aparecen en los datos de referencia así como los nuevos datos). La columna **L** contiene un símbolo de diagnóstico sumario para cada ítem; el símbolo de precaución por defecto (el diamante amarillo) se actualiza después de que IATA haya calculado los resultados de vinculación. La manera más efectiva de utilizar esta interfaz es, primero, calcular los resultados con todos los ítems, y luego examinar la información diagnóstica para identificar y eliminar todos los ítems con resultados anómalos. Repita estos dos pasos hasta que el vínculo sea estable.

FIGURA 13.4

Resultados de la vinculación de ítems comunes. CYCLE2 con CYCLE1



Haga clic en el botón **Calculate** para estimar las constantes de vinculación y para evaluar la calidad estadística de la vinculación. Cuando finaliza el cálculo, IATA muestra un resumen de la calidad de la vinculación en el gráfico que se encuentra a la derecha y actualiza los símbolos de diagnóstico sumario en la tabla de ítems que se encuentra a la izquierda (Figura 13.4). El gráfico muestra tres líneas: una línea sólida, una línea discontinua y una línea punteada. Las líneas sólida y discontinua muestran las curvas características de la prueba (CCP). La CCP de una prueba resume el comportamiento estadístico del conjunto de los ítems, proveyendo información similar a la que provee una función de respuesta a ítems (FRI), pero para varios ítems simultáneamente. Idealmente, las vinculaciones y las CCP de referencia deberían ser idénticas (si solo hay una línea visible, las dos están presumiblemente superpuestas de manera perfecta), indicando que las diferencias en la magnitud y variabilidad entre la escala de vinculación y la escala de referencia se registran a través de la escala de competencia mostrada. La línea punteada muestra la diferencia absoluta entre las dos CCP, expresada como una proporción de la puntuación total de la prueba. El valor de la diferencia varía habitualmente a lo largo de la escala de competencia, indicando que la vinculación puede no ser estable para todas las escalas de puntuación. Para escalas de puntuación con amplias diferencias, los resultados vinculados no estarán en la misma escala que los datos de referencia; por lo tanto, no serán comparables. Sin embargo, si la diferencia promedio es pequeña (por ejemplo, < 0,01), el error puede ser considerado insignificante.

En la Figura 13.4, la curva de referencia (sólida) representa los ítems de prueba de ***CYCLE1***, y la curva vinculada (discontinua) representa los ítems de prueba de ***CYCLE2*** luego de la aplicación de la vinculación. Ver ambas curvas en la figura es difícil porque la curva de referencia y las CCP de referencia son prácticamente idénticas, algo que también presenta la curva de error, que tiene un valor constante de aproximadamente cero a lo largo de la escala mostrada de competencia.[1]

Debajo del gráfico se muestran las constantes de vinculación estimadas en los dos cuadros de texto. La *constante de localización* ajusta diferencias en la magnitud de las escalas originales de los nuevos datos (la evaluación nacional en curso) y los datos de referencia

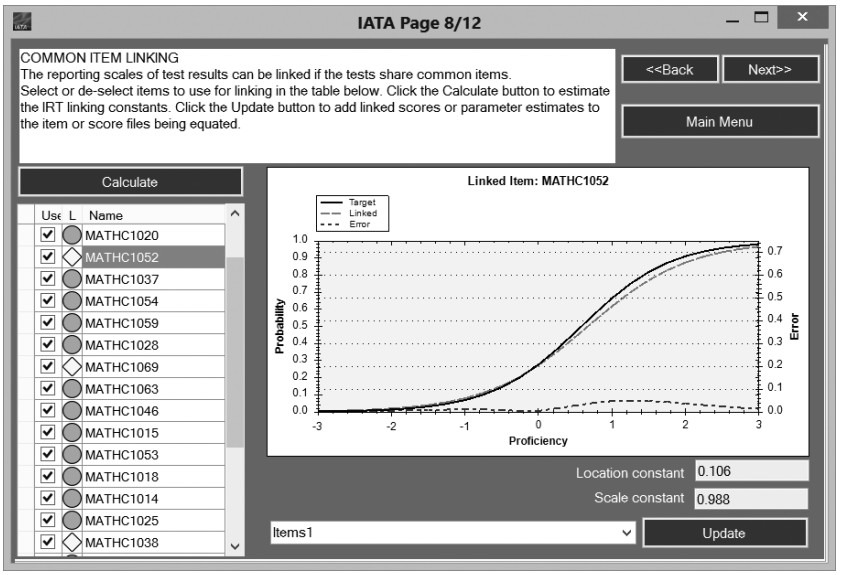
(la evaluación precedente), y la *constante de escala* ajusta diferencias en la variabilidad entre escalas. Hablando de manera general, las dos constantes pueden ser interpretadas en conjunto de modo tal que cualquier valor en la escala TRI bruta (por ejemplo, la puntuación TRI de un estudiante que surge de los resultados del análisis en curso) por la constante de escala, adicionando luego la constante de localización hará que el resultado vinculado sea directamente comparable con la escala TRI en ***CYCLE2***. Esta comparabilidad significa que, luego de que se haya aplicado la vinculación de escala, cualquier diferencia remanente entre los resultados de ***CYCLE1*** y los resultados transformados de ***CYCLE2*** representa diferencias en el desempeño en la prueba más que diferencias entre las pruebas en sí mismas.

En la tabla de ítems a la izquierda de la Figura 13.4, los símbolos de diagnóstico se actualizan luego de realizados los cálculos para indicar cualquier vinculación potencialmente problemática a nivel de ítem. Un ítem vinculado es problemático si su FRI vinculado es muy diferente del FRI de referencia. Hacer clic en cualquier ítem de la lista de ítems le permitirá ver los resultados de la función de vinculación aplicada a cada ítem de la prueba. Como sucede en las comparaciones CCP totales, el FRI vinculado debería ser similar al FRI de referencia.[2] Aun cuando los resultados para la prueba total parezcan muy buenos, la función de vinculación puede no funcionar muy bien para algunos ítems. Sin embargo, puesto que muchos errores de muestreo se verifican a nivel de ítem, las diferencias entre FRI son habitualmente problemáticas solo si el error entre FRI vinculado y FRI de referencia es mayor que 0,05[3] (véase el ejemplo para **MATHC1052** en la Figura 13.5).

Un ejemplo común de una situación que puede causar que un ítem particular de una prueba demuestre un comportamiento idiosincrásico en un análisis de vinculación se da cuando un área específica de contenido medida por un ítem de vinculación es utilizada como base para intervenciones instruccionales entre los dos períodos de prueba (tales como un énfasis creciente en la utilización de un tipo particular de gráficos en matemáticas o en un aspecto de la gramática en lengua). Ya que el desempeño en ese ítem específico de prueba probablemente mejore de manera idiosincrásica, las constantes de vinculación estimadas a partir de la totalidad de los ítems no representarán los

FIGURA 13.5

Resultados de la vinculación de ítems comunes, CYCLE2 con CYCLE1, MATHC1052



cambios específicos en los ítems entre la primera y la segunda administración.

MATHC1052, marcado con un símbolo de precaución, es un ejemplo moderado de este fenómeno. Los resultados para el ítem se muestran en la Figura 13.5. Aunque las constantes de vinculación parecen ajustarse exitosamente a la diferencia del ítem en localización (esto es, la dificultad del ítem respecto del ejemplo dado), se verifican algunas diferencias entre las dos líneas, particularmente en el nivel más alto de competencia. Las FRI de referencia y los vinculados son diferentes entre sí, y la línea punteada en la parte inferior, que expresa la diferencia entre las dos, se ubica hasta valores de 0,08 pero generalmente se ubica por debajo de 0,05. Estas diferencias son irrelevantes y, en la mayoría de las situaciones prácticas, este valor de error no es problemático.

En el caso de que las diferencias entre las FRI de referencia y los vinculados fueran lo suficientemente amplias como para ser

problemáticas (por ejemplo, si son sistemáticamente superiores a 0,05 a lo largo de una amplia escala de competencia), habría que eliminar el ítem ofensivo desmarcando la casilla contigua al nombre del ítem en la columna **Use** y haciendo clic en **Calculate**. Si bien se puede eliminar uno o dos ítems sin introducir con ello problemas de validez, si se eliminan muchos ítems de las estimaciones de las funciones de vinculación, la validez de la vinculación se debilita porque los ítems de anclaje pueden no reflejar adecuadamente el equilibrio pretendido en el contenido. Tenga presente que la validez de la vinculación depende de la estabilidad estadística de los ítems y de la coherencia de la representación del contenido entre las dos evaluaciones. Si el análisis estadístico de resultados sugiere que algunos ítems deberían eliminarse de la vinculación, debería elevarse la recomendación al comité directivo de la evaluación nacional antes de tomar una decisión. Cuantos menos ítems haya en común entre las dos evaluaciones que se va a vincular, tanto más débil será la vinculación. En el caso de la presente muestra, los resultados indican una vinculación muy estable. Como consecuencia, el equipo de la evaluación nacional puede confiar en que la prueba utilizada en la evaluación en curso (***CYCLE2***) es apropiada para monitorear los cambios en los niveles de rendimiento de los estudiantes desde la evaluación nacional precedente.

Dos controles bajo las constantes de vinculación —un menú desplegable y un botón etiquetado **Update**— le permitirán aplicar las constantes de vinculación directamente a los resultados del actual análisis. Vea los resultados del análisis en la interfaz final (IATA Page 12/12) del proceso de trabajo de análisis. Para aplicar las constantes de los parámetros de ítems a los resultados actuales (IATA Page 8/12), siga los siguientes pasos:

1. Seleccione **Items1** en el menú desplegable inferior.
2. Haga clic en **Update**. En los resultados del análisis IATA agregará parámetros de ítems vinculados —cuya escala representa ahora la escala establecida por los parámetros de ítem de ***CYCLE1***—, a la tabla ***Items1***. Los parámetros de ítems vinculados están identificados como **a_link** y **b_link**. IATA indicará cuándo se han actualizado los resultados.

Para actualizar las puntuaciones estimadas TRI para los resultados actuales, siga los siguientes pasos:

1. Seleccione **Scored** del menú desplegable.
2. Haga clic en **Update**. IATA agregará una variable de puntuación a la tabla puntuada en los resultados de los análisis, identificada como **LinkedScore**. Esta puntuación se expresa en la escala establecida por los parámetros de ítem establecidos por ***CYCLE1***.

Una vez actualizados los resultados, tanto de **Items1** como de **Scored**, haga clic en **Next>>** para continuar.

PASO 3: REAJUSTE DE RESULTADOS VINCULADOS

Si aplicó las constantes de vinculación a la tabla de datos puntuada tal como se describe en el paso 2, la interfaz de IATA **Scale Review and Scale Setting** (IATA Page 9/12) incluirá el nombre **LinkedScore** en el menú desplegable ubicado en la parte superior derecha. Seleccione **LinkedScore** para mostrar los resúmenes gráficos y estadísticos para los resultados vinculados ***CYCLE2***, que están expresados en la escala establecida por los datos de ítem de referencia de ***CYCLE1***. El valor de **LinkedScore** para este ejemplo tiene una media de 0,10 y una desviación estándar de 1,07. Nótese que los resultados predeterminados pueden mostrar las estadísticas resumen de los resultados de **PercentScore**. Para mostrar los resultados corregidos de **LinkedScore,** selecciónelos utilizando el menú desplegable.

Para convertir la puntuación vinculada TRI en una puntuación escalada que pueda ser comparada con la variable **NAMscore** que fue producida durante el análisis de los datos de ***CYCLE1***, siga los siguientes pasos:

1. Ingrese **NAMscore** en el cuadro de texto que se encuentra debajo de la etiqueta **Add New Scale Score**.
2. Ingrese 100 en el cuadro **Specify St. Deviation**, el valor originalmente establecido para los datos de ***CYCLE1***.
3. Ingrese 500 para **Specify Mean**, el valor original establecido para los datos de ***CYCLE1***.

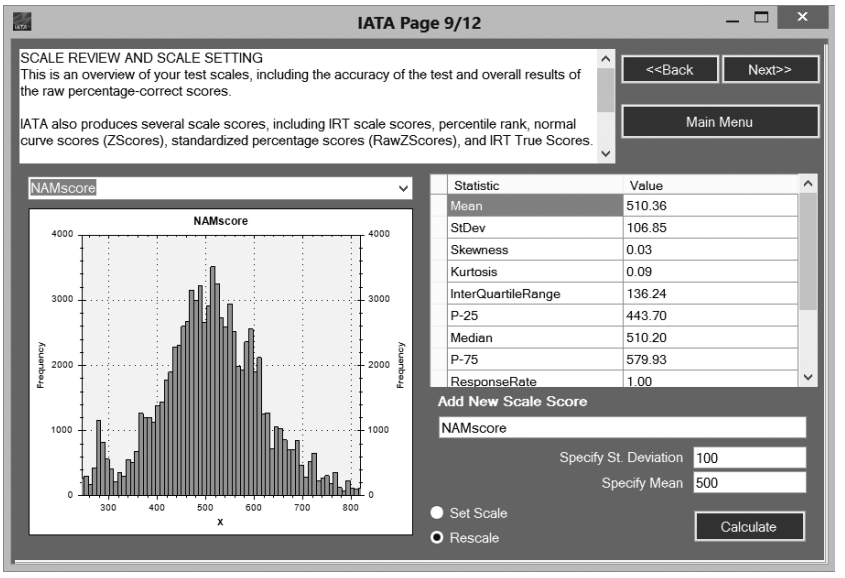
4. Seleccione la opción **Rescale**. Esta opción asegura que la nueva puntuación escalada mantiene la vinculación estimada en la interfaz previa.
5. Haga clic en **Calculate**. IATA creará la nueva puntuación escalada y mostrará la distribución y las estadísticas descriptivas, tal como se muestra en la Figura 13.6. La media de 510,36 indica que los resultados del año en curso muestran una mejora de 10,36 puntos respecto de la evaluación precedente.

Puesto que el flujo de trabajo es específico para la vinculación, IATA automáticamente produce la nueva puntuación escalada utilizando la puntuación TRI vinculada. Debido a que existen ítems de vinculación apropiados, el procedimiento de vinculación pudo producir **NAMscores** para dos evaluaciones separadas, que pueden ser comparadas ya que están en una escala común.

Después de agregar la nueva puntuación escalada a los resultados, haga clic en **Next>>** para continuar.

FIGURA 13.6

Puntuaciones de prueba de CYCLE2 expresadas en la escala de CYCLE1 (NAMscore)



PASO 4: ASIGNACIÓN DE ESTÁNDARES DE DESEMPEÑO

La mayoría de las tareas en el flujo de trabajo **Response data analysis with linking** están especificadas de la misma manera que en procesos de trabajo previos. El análisis de selección de ítems en IATA Page 10/12 puede llevarse a cabo como un ejercicio independiente. Sin embargo, los desempeños estándar son tratados de manera diferente. Luego de la primera evaluación nacional, la determinación del estándar debería llevarse a cabo solo como un ejercicio de validación. La revisión periódica de los umbrales es útil para determinar si es necesario establecer nuevos estándares de desempeño (por ejemplo, si la calidad de la educación está mejorando), pero el establecimiento de nuevos umbrales para los niveles de competencia habitualmente debería coincidir con importantes cambios de política, tales como una reforma curricular.

Cuando se ejecuta la interfaz **Developing and assigning performance standards** en IATA (IATA Page 11/12), pueden utilizarse dos fuentes de parámetros de ítems para guiar el proceso de establecimiento de estándares: los ítems utilizados en la evaluación actual (***Items1***) o los ítems de referencia utilizados en evaluaciones previas (***Items2***). Estas fuentes de parámetros de ítems están disponibles en el menú desplegable que está sobre la tabla, a la izquierda. En la tabla que se ubica en el ángulo inferior derecho de acuerdo con la fuente de parámetro de ítem seleccionada se muestran las medidas estadísticas de síntesis que describen parámetros de ítems y umbrales. Las medidas estadísticas de síntesis mostradas al entrar en esta interfaz describen las estimaciones desde los datos actuales (***CYCLE2***) en la tabla ***Items1***. La mejor práctica consiste en utilizar la fuente ***Items1***, que se selecciona por defecto, porque la tabla ***Items2*** contiene todos los ítems utilizados para producir las puntuaciones de prueba actuales.

El ejercicio de determinación de estándares debería llevarse a cabo solo (a) si *no* se llevó a cabo un procedimiento de determinación de estándares en un ciclo de evaluación previo o (b) si existe una razón para sospechar que los estándares de desempeño cambiaron, en el sentido de que las expectativas normativas acerca de los tipos de habilidades o acerca de la calidad de la educación pueden

haber cambiado a lo largo del tiempo. Nótese que el desempeño estudiantil difiere de los estándares de desempeño. Por ejemplo, el desempeño estudiantil puede estar incrementándose si los estudiantes tienen mejor desempeño en las pruebas, pero los estándares de desempeño pueden estar decreciendo si las partes interesadas tienen menores expectativas acerca de lo que constituye un desempeño "aceptable". Replicar el proceso de determinación de estándares utilizando las diferentes series de parámetros de ítems puede contribuir a identificar si los estándares de desempeño permanecen estables a través de las evaluaciones. Si los estándares son estables, replicar el procedimiento de determinación de estándares utilizando diferentes series de ítems debería producir series de umbrales similares. Al replicar cualquier procedimiento de determinación de estándares, antes de aplicar cualquier vinculación tenga en cuenta que los parámetros de ítems en la tabla ***Items1*** se refieren a la distribución de competencia en la evaluación actual, y que los parámetros de ítem en la tabla ***Items2*** se refieren a la distribución de competencia en la evaluación previa. Para comparar los umbrales producidos por el análisis de las fuentes de parámetros de ítem, aplique las constantes de vinculación (por ejemplo, según se muestra en la Figura 13.5).

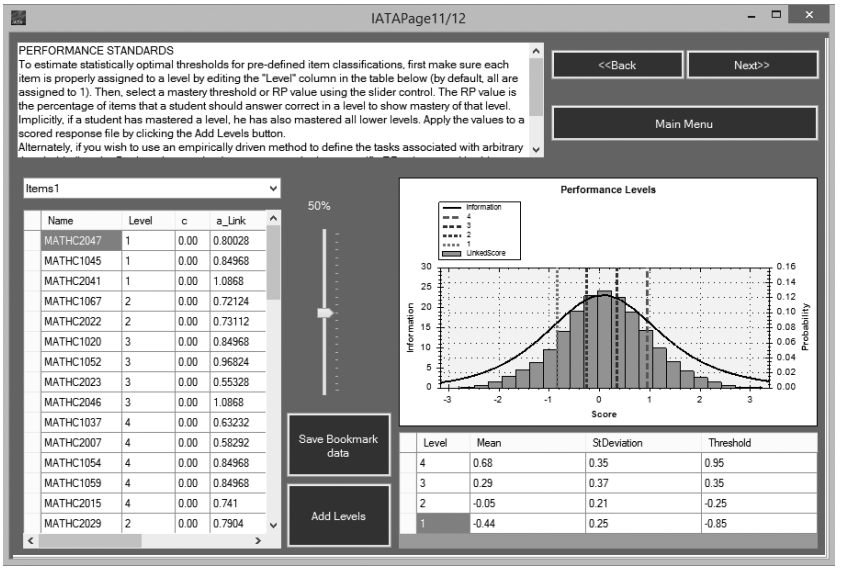
Ya que el desempeño se incrementó respecto de los datos en ***CYCLE1***, las medias de los parámetros del ítem **b** aparecerán más bajas que las que se refieren a los datos de ***CYCLE1*** para los ítems clasificados al mismo nivel. Sin embargo, no debería estimar nuevos umbrales en el flujo de trabajo actual, ya que los estándares de desempeño se determinaron y asignaron inicialmente en el análisis de los datos de ***CYCLE1***. Los siguientes umbrales (que se establecieron en la primera evaluación nacional) se utilizan para esta evaluación nacional:

- Nivel 4: 0,95
- Nivel 3: 0,35
- Nivel 2: –0,25
- Nivel 1: –0,85

Aplique los estándares de desempeño a los datos de ***CYCLE2*** ingresando manualmente en la tabla los umbrales en la columna

FIGURA 13.7

Asignación de estándares de desempeño, datos de CYCLE2



Threshold, tal como se muestra en la Figura 13.7, y haciendo clic en el botón **Add Levels**. Para actualizar estos umbrales, primero ajuste la probabilidad de respuesta utilizando el control deslizante ubicado en el medio de la interfaz. Esto hará que IATA genere una tabla con valores por defecto, que pueden ser reemplazados por los valores asignados desde el recorrido ***CYCLE1***. La variable **Level** se añadirá, entonces, a la tabla de datos de los estudiantes, ***Scored***. Aunque las especificaciones de los estándares de desempeño no cambien en cuanto al proceso de trabajo, esta interfaz —al igual que la interfaz escalada— reconocerá que se está utilizado un flujo de trabajo de vinculación y distribuirá los estudiantes en niveles basados en la puntuación TRI vinculada más que en la puntuación TRI bruta.

Los resultados de datos de ítems de este recorrido de análisis se incluyen en el archivo ***ItemDataAllTests.xls*** en la hoja de cálculo llamada ***ReferenceC2***.

NOTAS

1. Para obtener más resultados interesantes con mayor error, replique este análisis con los resultados de PILOT1 y PILOT2 como un ejercicio independiente. Tenga en cuenta que para alcanzar la meta de minimizar los errores en las vinculaciones, las pruebas vinculadas deben tener un número de ítems de anclaje y un tamaño de muestra suficientes para producir estadísticas precisas.

2. Si tiene datos de respuesta de ambas evaluaciones, puede llevar a cabo un análisis más sensible analizando FDI en el proceso de trabajo **Response data analysis** en los archivos de datos de respuesta combinados, utilizando los datos fuente para identificar la variable FDI. Las puntuaciones TRI que se produzcan se vincularán automáticamente, pero no serán interpretables en la escala de ninguna de las pruebas a menos que se anclen los parámetros de ítems.

3. Se verifica una excepción solo con ítems altamente discriminatorios. Amplias diferencias entre FRI (por ejemplo, cuando la línea de error es mayor que 0,05) generalmente indican problemas. Si estas diferencias se verifican solo en una escala de competencia reducida (por ejemplo, en una escala de menos de 0,4 puntos de competencia, es decir dos marcas de verificación en el eje x del gráfico por defecto), no afectarán negativamente a la calidad de la vinculación.

CAPÍTULO 14

MÉTODOS ESPECIALIZADOS EN IATA

Este capítulo describe el empleo de cuatro aplicaciones especiales en los flujos de trabajo de IATA (Item and Test Analysis). Cada una de las interfaces ya ha sido descrita en detalle en capítulos anteriores de este volumen. Las primeras tres partes del capítulo: "Vinculación de datos del ítem", "Selección de ítems óptimos de la prueba" y "Desarrollo y asignación de estándares de rendimiento" presentan un repaso de los tres flujos de trabajo que utilizan archivos de datos de ítems IATA como sus principales datos de entrada. Estos flujos de trabajo resultan más apropiados cuando ya se han analizado los datos de respuestas de los alumnos y se han estimado los parámetros de los ítems de la prueba. Estos flujos permiten desarrollar algunos análisis utilizando los parámetros del ítem sin que se requiera volver a analizar los datos de respuestas originales del alumno. La capacidad de realizar análisis utilizando datos del ítem resulta útil en las situaciones en las cuales no se puede acceder a los datos de respuestas del alumno (por ejemplo por razones de seguridad o por limitaciones en la capacidad de transferencia de datos), pero sí se cuenta con parámetros de ítems de fuentes tales como informes técnicos o pequeñas tablas de datos. En la sección final "Análisis de datos de respuesta con parámetros de

ítem anclados" se utilizan los parámetros de ítems anclados de evaluaciones nacionales anteriores para analizar los datos de respuestas del alumno provenientes de evaluaciones posteriores que comparten muchos ítems comunes.

Dado que en los capítulos anteriores ya se han descrito las principales interfaces de IATA que se emplean para acceder a estas funciones, este capítulo se concentrará en los pasos específicos para cada uno de los siguientes temas, en lugar de brindar una guía completa.

VINCULACIÓN DE DATOS DEL ÍTEM

El capítulo 13 de este volumen mostró cómo se calculaban los parámetros vinculados en el mismo flujo de trabajo que el análisis de los datos de respuesta al ítem. Sin embargo, en la práctica, estas dos actividades pueden ocurrir en distintos momentos. El ministerio de educación, por ejemplo, podría llevar a cabo una evaluación nacional inicial (***CYCLE1***) en el año uno y una evaluación nacional de seguimiento (***CYCLE2***) en el año cinco, y luego ejecutar o encargar un ejercicio de vinculación utilizando las estimaciones de parámetros de ítem de ambas evaluaciones en el año seis. Los analistas pueden utilizar las constantes de vinculación calculadas en IATA y actualizar los datos de ***CYCLE2*** produciendo nuevos puntajes y parámetros vinculados. Estos nuevos puntajes pueden demostrar la dimensión del cambio en los niveles de rendimiento del alumno entre el año uno y el año cinco.

Para vincular los datos del ítem aplicando este enfoque, debe contar con ítems anclados comunes a dos análisis diferentes de datos de evaluaciones nacionales. Para este ejemplo se vincularán los resultados de las evaluaciones nacionales ***CYCLE1*** y ***CYCLE2***. Tienen 25 ítems en común. El proceso es similar al flujo de trabajo para **Análisis de datos de respuesta con proceso de vinculación** descrito en el capítulo 13.

En el capítulo 10 se analizaron los datos de la primera evaluación nacional (***CYCLE1***). Antes de iniciar la vinculación de las dos evaluaciones, debe analizar los datos de ***CYCLE2*** siguiendo los

procedimientos descritos en el capítulo 10. A continuación un resumen de esos procedimientos:

1. Haga clic en **Response data analysis**.
2. Cargue los datos del alumno de ***CYCLE2*** (2484 registros).
3. Cargue los datos del ítem de ***CYCLE2*** (53 registros).
4. Seleccione 9 para representar los datos faltantes que se califican como incorrectos y emplee las flechas desplazables para resaltar los valores de la muestra y de la ponderación de ***CYCLE2*** (IATA Page 3/10). Recuerde del capítulo 13 que los valores numéricos adicionales (0, 1, 2, 3) de la columna de valores representan los puntajes de los ítems de crédito parcial.
5. No elimine ningún ítem de la interfaz de análisis del ítem (IATA Page 4/10).
6. Omita el análisis de la dimensionalidad y del funcionamiento del ítem diferencial, así como la revisión y la determinación de la escala, la selección del ítem y las funciones de la determinación de estándares (IATA Page 5/10 y 9/10) con el fin de lograr los objetivos de este ejercicio.
7. Para confirmar la correcta realización del análisis, en la interfaz final (IATA Page 10/10), seleccione **Scored** del menú desplegable y desplácese hasta el final. El primer alumno de la lista (**CYCLE2STDID**) recibe un puntaje de teoría de respuesta al ítem **(IRT) score** de 2.58, un puntaje **percentile score** de 99.37 y un puntaje **TrueScore** de 93.26.
8. Guarde el conjunto total de resultados como ***CYCLE2_UNLINKED.xls***.

Después de completar el análisis de los datos de ***CYCLE2***, seleccione **Linking item data** del menú principal. Realice los pasos que se describen a continuación:

1. Cargue el archivo de datos del alumno con los resultados de las evaluaciones actuales (en este ejemplo, ***CYCLE2***) que se vincularán con los resultados de la evaluación anterior

(en este ejemplo, ***CYCLE1***). El archivo de datos debe incluir la variable denominada **IRTscore**. No obstante, en caso de vincular resultados empleando resultados que han sido modificados, asegúrese de que la variable que contenga el puntaje IRT sin escala reciba la denominación **IRTscore**.[1] Para este ejemplo, emplee el archivo de datos denominado ***CYCLE2_UNLINKED.xls*** y cargue la tabla ***Scored***. (El primer alumno de la lista, **CYCLE2STDID** = 1, tiene puntajes **IRTscore** de 2.58 y **Percentile score** de 99.37 al finalizar el registro.) Haga clic en **Next>>**.

2. En IATA Page 2/5, cargue los parámetros del ítem estimados a partir de los resultados de la evaluación más reciente. La tabla de datos debe incluir los nombres de los ítems y los parámetros de IRT. En este ejemplo, para vincular los resultados de ***CYCLE2*** con la escala de ***CYCLE1***, cargue la tabla ***Items1*** del archivo ***CYCLE2_UNLINKED.xls***. Esta tabla incluye los parámetros de los ítems de la evaluación nacional del ***CYCLE2***. Asimismo, estos datos están disponibles en la tabla ***ReferenceC2*** que se encuentra en el archivo de datos de muestra de IATA ***ItemDataAllTests.xls***. El primer ítem de la lista, **MATHC1008**, tiene los siguientes valores: **Key** = A, **Level** = 3, **a** = 0.58 y **b** = 0.20. Haga clic en **Next>>**.

3. En IATA Page 3/5, cargue los parámetros del ítem a partir de los resultados de la evaluación anterior del ***CYCLE1***. El archivo debe incluir los nombres de los ítems y los parámetros de IRT. En este ejemplo, para vincular los resultados de ***CYCLE2*** con la escala del ***CYCLE1***, cargue la tabla ***ReferenceC1*** del archivo ***ItemDataAllTests.xls***. El primer ítem de la lista, **MATHC1005**, tiene los siguientes valores: **Key** = B, **Level** = 3, **a** = 0.63 y **b** = 0.21. Haga clic en **Next>>**.

4. Haga clic en **Calculate** para estimar las constantes de vinculación de IRT (IATA Page 4/5). Las siguientes constantes de vinculación, **Location** y **Scale**, describen la transformación lineal que convierte la escala IRT de ***CYCLE2*** en la escala IRT de ***CYCLE1***. Revise la calidad de la vinculación, específicamente, los ítems con funciones de respuesta al ítem (IRF) con vinculación idiosincrásica. En este ejemplo, la mayoría de los ítems muestran buenas

vinculaciones, las que se indican en círculos verdes. Revise los ítems problemáticos; estos tienen indicadores de precaución o advertencia (diamantes amarillos o triángulos rojos). Examine los gráficos de los ítems individuales, en particular, las líneas punteadas azules de error. Elimine todo ítem de la lista con una estimación de vinculación (sobre la izquierda) sistemáticamente mayor que 0.05 (en el eje derecho) dentro de la escala de –2.0 +2.0 (en el eje de competencia). Si se identifica más de un ítem problemático, elimine un solo ítem a la vez. Vuelva a calcular las constantes de vinculación después de cada eliminación haciendo clic en **Update**. Eliminar una pequeña cantidad de ítems problemáticos puede ser suficiente para mejorar la estimación de las vinculaciones para el resto de los ítems. Por el contrario, si se eliminan muchos ítems se puede mejorar la estimación estadística, aunque también puede reducir la validez general de la vinculación. Aplique las constantes de vinculación a los parámetros del ítem y a los puntajes de la prueba. Para este ejemplo, las constantes de ubicación (location) y escala (scale), estimadas empleando el conjunto completo de ítems comunes son 0.106 y 0.988, respectivamente. Estas constantes de escalas se pueden utilizar para convertir un puntaje IRT de la evaluación de ***CYCLE2*** a la escala de la evaluación de ***CYCLE1***, multiplicando la variable IRTscore de ***CYCLE2*** por 0.988 y sumándole 0.106. Estas constantes de vinculación indican que los puntajes de las pruebas en ***CYCLE2*** son, en promedio, un poco menos variados y un poco más altos que los puntajes en ***CYCLE1***. Haga clic en **Next>>**.

5. Observe que los parámetros vinculados **a** y **b** de **MATHC1008** son 0,58 y 0,20, respectivamente. Los parámetros vinculados describen el comportamiento estadístico de cada ítem de ***CYCLE2*** en la escala del ***CYCLE1***. Esta información puede ser útil en caso de que un equipo de evaluación nacional desee determinar un estándar o desarrollar una nueva prueba empleando el conjunto combinado de ítems. Las siguientes secciones de este capítulo analizan la manera de realizar estas dos funciones empleando únicamente datos del ítem. Guarde los resultados del análisis (IATA Page 5/5).

SELECCIÓN DE ÍTEMS ÓPTIMOS DE LA PRUEBA

Hasta el momento, los ítems de la prueba se han ido seleccionando a medida que se analizaban los datos de respuestas del alumno. No obstante, un ministerio de educación apropiadamente podría solicitar que se identifique un conjunto de ítems de evaluaciones completadas para evaluaciones futuras. Esta solicitud puede tener lugar después de haber analizado los datos de respuestas del alumno y de haber finalizado las actividades del ciclo de la evaluación nacional. Los desarrolladores y los analistas de la prueba pueden identificar y seleccionar los ítems de anclaje de evaluaciones anteriores sin necesidad de encontrar y volver a analizar el conjunto completo de datos del alumno de esta evaluación. Estos ítems de anclaje formarán el componente central o, al menos, el más importante de la nueva evaluación, a la que se agregarán los nuevos ítems. Los ítems de anclaje se emplearán para vincular los resultados de la nueva evaluación con los de la evaluación existente. Para sentar las bases de una prueba exacta y una vinculación estable entre las dos pruebas, los ítems de anclaje deben ofrecer la mayor exactitud posible en toda la escala de competencias esperada del alumno. En este caso, la tarea de selección del ítem solo requiere parámetros de ítems existentes en la evaluación anterior.

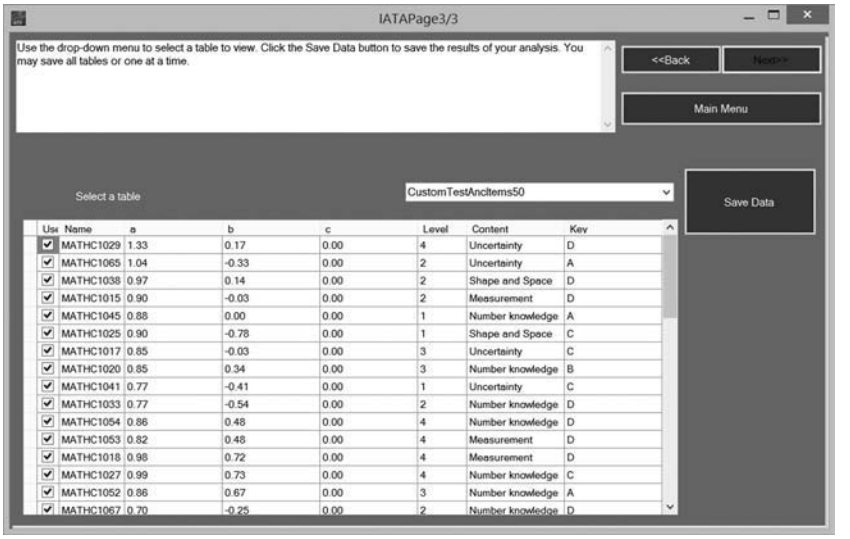
El siguiente ejercicio muestra cómo seleccionar ítems óptimos para vincular dos evaluaciones empleando los resultados de los parámetros del ítem guardados y obtenidos del análisis de los datos de una evaluación anterior.

1. Seleccione el flujo de trabajo **Selecting optimal test items** del menú principal.
2. Cargue el archivo de datos del ítem, la tabla ***ReferenceC1*** del archivo ***ItemDataAllTests.xls*** de la carpeta de IATA que se encuentra en el escritorio (IATA Page 1/3). Los datos deben incluir los nombres de los ítems y los parámetros IRT. Además, deben contener la información del ítem **Level** y **Content**. (El primer ítem de la lista, el ítem con el número más bajo, debe ser **MATHC1005**, el cual tiene los siguientes valores: **Key** = B, **Level** = 3, **a** = 0.63 y **b** = 0.21.) Haga clic en **Next>>**.

3. Para proporcionar los resultados más útiles para la selección de ítems, la cantidad especificada debe ser igual a la cantidad total de ítems en el archivo de datos de ítems. En este caso, dado que la tabla con datos del ítem ***ReferenceC1*** tiene 50 ítems, para la selección, se deben especificar 50 ítems. IATA producirá una tabla en la que todos los ítems disponibles se clasifican en función de su pertinencia para medir a los alumnos dentro de los límites inferior y superior especificados. Estos límites describen los rangos de percentiles aproximados de la muestra original empleada para calibrar los ítems (0 representa el alumno con el rendimiento más bajo y 100, el alumno con el mayor rendimiento). Como regla general, mantenga los valores predeterminados de 2 y 98. No obstante, si en la población de alumnos de la nueva evaluación nacional se prevé una distribución de la competencia significativamente diferente a la de la evaluación anterior, los límites inferior y superior pueden ajustarse para minimizar la inclusión de ítems innecesariamente fáciles o difíciles. Por ejemplo, si espera que la competencia de la nueva población sea muy superior a la de la población original, ajuste el umbral inferior hacia arriba a un valor, x, superior a 2 para reflejar que el alumno con el rendimiento más bajo de la nueva población pueda tener un puntaje equivalente al rendimiento de un alumno en el rango de percentil xth de la población original. De esta manera, reducirá las posibilidades de seleccionar de forma inapropiada ítems fáciles para la nueva población. Ingrese un título, como **AncItems50** (para los ítems de anclaje) en la casilla **Name of item selection** y 50 para la cantidad total de ítems. Haga clic en **Select Items**. IATA generará una tabla de contenido de nivel *X* de 50 ítems (IATA Page 2/3). Haga clic en **Next>>**. Observe (IATA Page 3/3, mostrada en la Figura 14.1) que **MATHC1029** tiene los siguientes valores: **a** = 1.33, **b** = 0.17, **Level** = 4 y **Key** = D.
4. Haga clic en **Save Data**. IATA asigna el prefijo de **CustomTest** a las tablas de selección del ítem, seguido del nombre único especificado para el ejercicio de selección del ítem en particular. En este caso, el archivo de datos de 50 ítems se guardará como ***CustomTestAncItems50*** junto con los demás datos de IATA.

FIGURA 14.1

Selección de ítems óptimos para la prueba, datos de CYCLE1



Puede intentar con una cantidad de ítems seleccionados empleando esta interfaz (como pueden ser 30 o 40). Asigne un nombre diferente a cada selección de ítem (en la casilla **Name of item selection**) si desea guardar los resultados.

DESARROLLO Y ASIGNACIÓN DE ESTÁNDARES DE RENDIMIENTO

La determinación de los estándares de rendimiento (por ejemplo, inferior al básico, básico, competente; avanzado o los niveles 1, 2 y 3) es una etapa importante para permitir el acceso a los resultados de la evaluación nacional a una variedad de público interesado. El capítulo 10 de este volumen describe cómo determinar los estándares de rendimiento como un ejercicio relativamente sencillo. Sin embargo, en la práctica, generalmente se requiere de un trabajo iterativo que implica la revisión del contenido de los ítems y de los resultados estadísticos. Esto debe incluir la aportación de múltiples fuentes (como el personal a cargo del currículo), muchas de las cuales

pueden haber tenido poca experiencia con análisis de datos o con estadísticas. Antes de comenzar el proceso de determinación de los estándares, el análisis de los datos de respuesta debe estar terminado. Para realizar este ejercicio, el comité de revisión debe emplear los parámetros del ítem finales.

El flujo de trabajo **Developing and assigning performance standards** de IATA permite emplear los resultados de análisis anteriores para facilitar el proceso de determinación de estándares. En esta sección se determinan los estándares de rendimiento para los datos de ***CYCLE1***, empleando los parámetros de los ítems de la evaluación de ***CYCLE1***. Para utilizar este flujo de trabajo, debe haber completado un análisis de datos de respuesta al ítem y haber guardado las tablas de resultados ***Items1*** y ***Scored***. Si no ha guardado los datos, repita el análisis descrito en el capítulo 10. Preferentemente, desde la perspectiva del desarrollo de estándares de rendimiento, tanto los parámetros del ítem como los puntajes IRTscore deben estar cargados en IATA. No obstante, los puntajes se utilizan únicamente como referencia para la estimación de los umbrales, y el análisis se puede realizar empleando solo parámetros del ítem sin cargar el puntaje IRT. Examine la distribución de los puntajes por nivel de umbral para determinar las proporciones de alumnos clasificados dentro de cada nivel de competencia. La determinación de los estándares, como se señaló anteriormente (véase la descripción sobre *bookmark* en el capítulo 10), es un procedimiento iterativo y consiste en la revisión de los ítems de la prueba por expertos en currículo y docentes con experiencia en función de la evidencia estadística disponible. Como se analizó en el capítulo 10, en este capítulo el propósito de determinar estándares no es justificar puntos de corte existentes o forzar los datos en los niveles de competencia establecidos previamente, sino determinar los niveles de competencia más útiles y los puntos de corte asociados en función de los ítems disponibles. Es posible que se necesiten varias rondas de revisión y análisis para establecer los niveles umbral. Después de establecer los umbrales, la carga del puntaje IRT en IATA facilitará la incorporación de los niveles de competencia directamente en los resultados de la evaluación de los alumnos.

Para completar este flujo de trabajo, realice los pasos que se indican a continuación. Tenga en cuenta que los pasos 1-4 se pueden repetir

varias veces antes de completar el conjunto de umbrales de nivel de rendimiento.

1. Seleccione el flujo de trabajo **Developing and assigning performance standards** del menú principal. La primera página (IATA Page 1/4) le solicita que cargue los datos de respuesta de los alumnos evaluados. Este paso es opcional, dado que el ejercicio de determinación de estándares puede realizarse solo con los datos del parámetro del ítem. Si cuenta con datos de los alumnos evaluados, puede emplearlos para conocer la utilidad de un conjunto de puntos de corte propuesto, ya que permite estimar la proporción de alumnos que pertenecen a cada nivel de competencia propuesto. No obstante, los datos del alumno evaluado son útiles solo después de completar las principales iteraciones del procedimiento de determinación de estándares.
2. Dado que no se completaron las definiciones de puntos de corte, haga clic en **Next>>** para pasar a IATA Page 2/4.
3. Cargue el archivo de datos del ítem que contiene los parámetros IRT y el nivel preasignado para cada ítem. En este ejemplo, los resultados se utilizan en la tabla de datos de ítems ***Items1*** que IATA genera en forma automática. (Los datos son los mismos que los de ***ReferenceC1*** que se encuentran en ***ItemDataAllTests.xls***.) Tenga en cuenta que cada ítem tiene un nivel de rendimiento preasignado. Haga clic en **Next>>**.
4. Realice los procedimientos de determinación de estándares *bookmark* descritos en el capítulo 10, que incluyen determinar el valor de la probabilidad de respuesta (RP), guardar los datos de marcadores y emplear estos datos para facilitarles a los docentes, especialistas en currículo y otros actores del ámbito educativo la revisión de los ítems de la prueba. Tenga en cuenta que en el mundo real puede encontrarse con una variedad de valores de RP, por lo general, en la escala de 0,50 a 0,80. Durante el proceso de establecimiento del rendimiento, las asignaciones de nivel del ítem pueden ser (y probablemente serán) modificadas durante los análisis realizados por el equipo de evaluación nacional. Es posible que se necesite repetir los pasos 1-4 varias veces, empleando el procedimiento *bookmark*, hasta

que los participantes del ejercicio de determinación de estándares estén convencidos de que los umbrales acordados facilitarán la interpretación significativa de los resultados de la evaluación.

5. Una vez acordados los estándares de rendimiento entre las partes presentes en el comité de revisión responsable de establecerlos, repita los pasos 1-4, asegurándose de cargar los datos del puntaje IRT del alumno del paso 1. Para este ejemplo, cargue los datos del alumno de ***CYCLE1*** con los puntajes IRT que debieron guardarse al final del capítulo 10 (IATA Page 1/4). Cuando IATA guarda los resultados del alumno, el puntaje IRT queda en una variable llamada **IRTscore**, que se encuentra en la tabla de datos denominada ***SCORED***. Esta variable incluye puntajes que IATA ha estimado directamente a partir de los parámetros de los ítems, sin aplicar ninguna modificación en la escala o en la vinculación. Después de cargar los datos, desplácese a través de las respuestas del alumno hasta los ítems individuales (IATA Page 1/4) para ver los valores individuales del puntaje IRT. (El segundo alumno de la lista, **CYCLE1STDID**, tuvo un puntaje IRT de 1.764.) Ingrese los puntos de corte finales en la columna **Threshold** en la parte inferior derecha de la pantalla (IATA Page 3/4) Verifique, a través de la comparación de los puntos de corte verticales con el área de la distribución del puntaje, que cada nivel de competencia incluya una proporción de alumnos que aparecen en el informe. En los casos en que muy pocos alumnos se encuentren en los niveles más altos o más bajos de competencia, a los fines de la elaboración de informes es más práctico combinar los niveles de competencia adyacentes. Opcionalmente, puede aplicar los umbrales de los datos evaluados a través de la interfaz de los estándares de rendimiento haciendo clic en el botón **Add Levels**. Haga clic en **Next>>**.

6. Guarde los resultados con **Save** (Page 4/4 de IATA) empleando un nombre de archivo distintivo, como ***NAMPerfStand***. Como regla general, guarde todas las tablas que contengan datos modificados. Como por ejemplo la tabla ***PLevels***, que se ha actualizado con los nuevos umbrales; la tabla ***Items1***, que puede tener nuevas asignaciones de nivel para los ítems; y la tabla ***Scored***, con los niveles de rendimiento de los alumnos.

ANÁLISIS DE DATOS DE RESPUESTA CON PARÁMETROS DE ÍTEM ANCLADOS

En capítulos anteriores, todos los parámetros IRT de los ítems se consideraron desconocidos y se debieron calcular a partir de los datos de respuestas del alumno. Los parámetros de ítems de la prueba se calcularon para cada evaluación nacional (por ejemplo, evaluación ***CYCLE1*** y evaluación ***CYCLE2***) y se utilizaron para calcular las constantes de vinculación.

IATA también cuenta con una función para importar parámetros fijos de ítems provenientes de evaluaciones anteriores, que se pueden utilizar para vincular sus resultados con los de evaluaciones posteriores. Inicialmente, el programa realiza una estimación de parámetros para la evaluación actual y los vincula con los parámetros sin ajustar de la evaluación anterior. Los ítems utilizados en el proceso de vinculación se denominan parámetros de ítem *anclados*.

Estos parámetros son los parámetros **a**, **b** y (opcionalmente) **c**, a los que se les han asignado valores para algunos ítems de la prueba en un archivo de datos de ítems antes de analizar el archivo de datos de respuesta, al igual que los ítems de anclaje utilizados en el proceso de vinculación formal. Cuando se analizan los datos de respuesta utilizando parámetros de ítem anclados, se calculan los parámetros de los ítems nuevos o no anclados, mientras que los anclados permanecen fijos en sus valores preespecificados. Los resultados estimados, es decir, de los parámetros IRT para ítems no anclados y los puntajes IRT del alumno, se expresan en la escala definida por los parámetros de los ítems anclados. Se prefiere este método en lugar del proceso de vinculación formal cuando las estimaciones de parámetros de ítems generadas con los datos actuales de respuesta pueden ser inferiores a las generadas a partir de estimaciones ya existentes. Esta situación podría darse si la muestra actual fuera mucho más pequeña o menos representativa que la muestra empleada para estimar los parámetros de ítems existentes. Este método también resulta adecuado cuando la mayoría de los ítems de la evaluación actual (más del 70 por ciento) ya tienen estimaciones de parámetros existentes. La única diferencia entre el uso de ítems anclados y las guías descritas en capítulos anteriores es que algunos ítems ya tendrán parámetros en el archivo de entrada de datos del ítem.

Imagine un escenario en el que el comité director de la evaluación nacional decide utilizar una prueba proveniente de un ciclo de evaluación nacional anterior, con solo mínimas modificaciones en el conjunto de ítems de la prueba. En este caso, no es necesario realizar el procedimiento de vinculación completo que se describe en el capítulo 13. Para los relativamente pocos ítems nuevos que se emplean en la evaluación actual, IATA calibrará automáticamente los parámetros IRT de esos ítems y los ubicará en la misma escala que los parámetros de los ítems anclados. Los puntajes IRT finales de los alumnos se basarán tanto en los parámetros de ítems anclados como en los ítems recientemente calibrados y se expresarán en la misma escala que los primeros.

Utilice el conjunto de datos de muestra de ***CYCLE3*** para realizar este ejercicio. Los datos del ítem para esta prueba se encuentran en el libro de ejercicios de Excel, ***ItemDataAllTests***. Estos datos representan el tercer ciclo del programa de evaluación nacional que se analizó en capítulos anteriores. Para este ciclo, el comité director de la evaluación decidió utilizar los ítems de la prueba ***CYCLE2*** luego de modificar mínimamente el contenido de algunos ítems y reemplazando solo ocho de los ítems de opción múltiple, y todos los ítems de respuesta corta. En lugar de volver a hacer una estimación para parámetros nuevos y constantes de vinculación, el comité decidió emplear parámetros de ítems provenientes de ***CYCLE2*** a fin de anclar las estimaciones de parámetro para los ítems nuevos.

Para realizar el análisis con ítems anclados, complete los siguientes pasos.

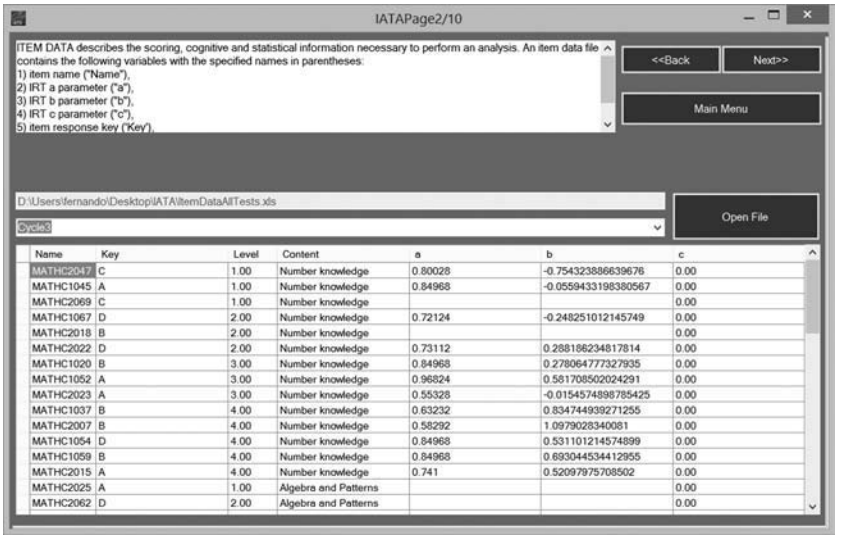
1. Seleccione el flujo de trabajo **Response data analysis** en el menú principal para analizar los datos de respuesta.
2. Desde la carpeta de datos de muestra de IATA (IATA Page 1/10) cargue el archivo de datos de respuesta del alumno ***CYCLE3.xls*** (que contiene 2539 registros y 61 variables). Verifique que el primer alumno de la lista en el archivo de datos tenga los siguientes valores: **SCHOOLID** = 30, **Sex** = 2, **SchoolSize** = 21, **MATHC2047** = C. Haga clic en **Next>>** para continuar.
3. Cargue el archivo ***ItemDataAllTests.xls*** y seleccione la tabla ***CYCLE3*** como datos del ítem (IATA Page 2/10). La tabla contiene

53 registros y siete variables. Tenga en cuenta que el ítem **MATHC2047**, ítem de conocimiento de números, tiene valores de 0.80 y -0.75 para los parámetros **a** y **b**, respectivamente. A diferencia de los archivos de datos de ítems empleados en análisis anteriores, hay valores para los parámetros **a** y **b** para algunos ítems, no para todos, como se muestra en la Figura 14.2. Los parámetros de ítem con valores asignados son los de ítems anclados. Estos valores se generaron durante el análisis de datos de ***CYCLE2*** y se vincularon a la escala original establecida para la evaluación ***CYCLE1.*** Se asignaron nuevos parámetros de ítem estimados a partir de datos de respuesta a varios ítems con claves de respuesta especificada que no los tienen (por ejemplo, **MATHC2069**). Como los parámetros anclados ya se encontraban vinculados a la escala de ***CYCLE1*** del análisis anterior, los parámetros estimados en el análisis actual de los datos de ***CYCLE3*** también se vincularon a la escala ***CYCLE1.*** Haga clic en **Next>>** para pasar a las especificaciones del análisis.

4. Establezca la variable de identificación como **CYCLE3STDID**, la variable de ponderación como **CYCLE3Weight**, y verifique que no haya errores en el valor 9 (IATA Page 3/10). Tenga en cuenta

FIGURA 14.2

Datos del ítem para la evaluación CYCLE3 con parámetros de ítem anclados



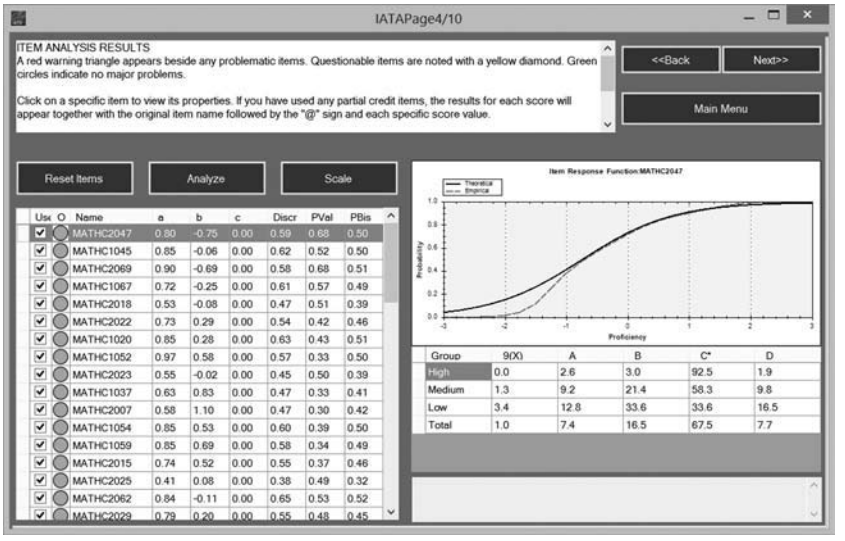
Name	Key	Level	Content	a	b	c
MATHC2047	C	1.00	Number knowledge	0.80028	-0.754323886639676	0.00
MATHC1045	A	1.00	Number knowledge	0.84968	-0.0559433198380567	0.00
MATHC2069	C	1.00	Number knowledge			0.00
MATHC1067	D	2.00	Number knowledge	0.72124	-0.248251012145749	0.00
MATHC2018	B	2.00	Number knowledge			0.00
MATHC2022	D	2.00	Number knowledge	0.73112	0.288186234817814	0.00
MATHC1020	B	3.00	Number knowledge	0.84968	0.278064777327935	0.00
MATHC1052	A	3.00	Number knowledge	0.96824	0.581708502024291	0.00
MATHC2023	A	3.00	Number knowledge	0.55328	-0.0154574898785425	0.00
MATHC1037	B	4.00	Number knowledge	0.63232	0.834744939271255	0.00
MATHC2007	B	4.00	Number knowledge	0.58292	1.0979028340081	0.00
MATHC1054	D	4.00	Number knowledge	0.84968	0.531101214574899	0.00
MATHC1059	B	4.00	Number knowledge	0.84968	0.693044534412955	0.00
MATHC2015	A	4.00	Number knowledge	0.741	0.52097975708502	0.00
MATHC2025	A	1.00	Algebra and Patterns			0.00
MATHC2062	D	2.00	Algebra and Patterns			0.00

que los valores numéricos adicionales (0, 1, 2, 3) de la columna de valores representan los puntajes de los ítems de crédito parcial. Haga clic en **Next>>** para comenzar el análisis.

5. Los resultados generados se muestran en la Figura 14.3. Observe que ahora todos los ítems tienen parámetros pero los ítems de anclaje mantienen sus valores originales (véase los valores **a** y **b** para **MATHC2047** en las figuras 14.2 y 14.3). A diferencia de la vinculación según el nivel de la prueba, ahora se puede ver cómo los parámetros de ítems anclados se ajustan a los datos de respuesta actuales mediante la comparación teórica y empírica de las FRI para cada uno de ellos. Por ejemplo, el ítem **MATHC2047** utilizó parámetros de ítem anclados; la función IRF (FRI) etiquetada como **Theoretical** (teórica) en la Figura 14.3 deriva de los datos de ***CYCLE2***, mientras que la función etiquetada como **Empirical** (empírica) deriva de los datos de ***CYCLE3***. En general, el ajuste de los ítems nuevos, cuyos parámetros se calculan a partir de los datos actuales, tiende a ser más preciso que el de los anclados. Si el ajuste entre las funciones individuales IRF teóricas y

FIGURA 14.3

Resultados del análisis de ítems con parámetros de ítem anclados, datos de CYCLE3, MATHC2047



empíricas es deficiente (es decir, la magnitud del margen vertical entre las FRI teóricas y empíricas es uniformemente mayor a 0.05), y la muestra de datos de respuesta nuevos es grande, no deben emplearse los ítems como parámetros anclados. Sin embargo, si la muestra es pequeña (por ejemplo, menos de 500), entonces la deficiencia en el ajuste entre las FRI teóricas y empíricas se debe simplemente a un error aleatorio y se puede ignorar.

Algunos ítems (como **MATHC1046** y **MATHC2034**) tienen banderas con forma de rombo (amarillas). Estos ítems suelen guardar poca relación con la competencia pero se mantuvieron para los fines de este ejercicio. Tenga en cuenta que el puntaje 1 para el ítem de crédito parcial **MATHA006** (**MATHA006@1**) tiene una bandera triangular (roja). Cerca del 99 por ciento de los alumnos logró un puntaje 1 o más en este ítem. Sin embargo, retener el puntaje 1 no influye demasiado sobre la calidad de los resultados. La bandera indica simplemente que el puntaje no es claramente distinto de 0, mientras que los puntajes de 2 y 3 de este ítem (**MATHA006@2** y **MATHA006@3**, respectivamente) son claramente distintos de los otros valores de los puntajes. Generalmente, lo mejor sería revisar el esquema de puntajes para que este ítem considere como válidos los puntajes 2 y 3 (en IATA, la entrada clave correspondiente sería "2:1; 3:2"). En este ejemplo, sin embargo, mantener el esquema mencionado no implica una desventaja ya que IATA se encargó de ajustar la estimación de puntajes a fin de reflejar el bajo nivel de competencia asociado con el puntaje de crédito parcial 1.

6. Las páginas 5/10 y 6/10 de IATA, en las que se realiza el análisis de la dimensionalidad y del funcionamiento del ítem diferencial de la prueba, pueden saltarse; ambas tareas son idénticas a otras similares realizadas en guías anteriores. Haga clic en **Next>>** para continuar.

7. Debido a que los resultados se vinculan automáticamente a la escala de ***CYCLE1,*** la media y la variación estándar de la variable **IRTscore** (IATA Page 7/10) de los datos de ***CYCLE3*** pueden variar significativamente de 0 a 1 en la muestra actual (en este caso, media = 0.02, variación estándar = 1.04). Una consideración

importante con respecto a la realización de la escala de resultados que emplean parámetros de ítem anclados es que, como los puntajes IRT están anclados a los parámetros vinculados de ***CYCLE2,*** se debe utilizar la opción **Rescale** para generar puntajes escalares, especificando los valores de la media y de la variación estándar empleados al determinar la escala **NAMscore** en ***CYCLE1***. Inserte **NAMscore** y los valores originales de la variación estándar (100) y la media (500). Haga clic en **Calculate** para calcular. Los valores de la media y de variación estándar reajustados a escala para la variable **NAMScore** son 501.71 y 103.96, respectivamente. Haga clic en **Next>>** para continuar.

8. Salte la página de selección del ítem (IATA Page 8/10) Debido a que los puntajes IRT están expresados según la escala establecida con los datos de ***CYCLE1***, aplique los umbrales estándar de rendimiento provenientes de ***CYCLE1*** (Nivel 4 = 0.95, Nivel 3 = 0.35, Nivel 2 = –0.25, Nivel 1 = –0,85) al conjunto de datos de ***CYCLE3*** (IATA Page 9/10). No es necesario establecer el nivel de RP porque los puntos de corte ya se han establecido. Esto ayudará a asegurar que se califique a los alumnos que participaron de la evaluación ***CYCLE3*** en relación con los valores umbrales de ***CYCLE1***. Presione **Enter** después de ingresar los valores umbrales. Haga clic en **Add Levels** para asignar alumnos a los estándares o niveles de rendimiento. Haga clic en **Next>>** para continuar.

9. Haga clic en **Save Data** para guardar todas las tablas de resultados de la evaluación **CYCLE3**. Nótese que en los archivos de datos ***SCORED***, el resultado del primer alumno tiene puntajes **IRT** y **NAM** de 1.41 y 641.10, respectivamente. Como referencia, los resultados de los datos de los ítems correspondientes al presente análisis (***Items1***) están incluidos en la hoja de cálculo con el nombre ***Reference*C3** del archivo ***ItemDataAllTests.xls***.

Por último, observe que los parámetros de ítems anclados resultan particularmente útiles en las situaciones en las que el tamaño de la muestra de la nueva evaluación nacional es pequeño, las pruebas se superponen considerablemente, o se cuenta con datos de respuesta de ambas pruebas. En este último caso, los datos de las respuestas deben

incluir todas las respuestas de ambos ciclos con el fin de facilitar el análisis del funcionamiento del ítem diferencial entre las dos pruebas; los datos del ítem deberán incluir las claves de respuesta para todos los ítems y solo se asignarán valores a los parámetros de los ítems utilizados en el ciclo anterior.

Durante el transcurso de los distintos tipos de análisis que acabamos de tratar, el equipo de evaluación puede realizar modificaciones, como por ejemplo eliminar ítems o ajustar los niveles de competencia o las categorías de contenido del currículo. Cuando la cantidad o la dimensión de esas modificaciones fueran considerables, no deberá utilizarse la vinculación de ítems.

Dada la probabilidad de que los analistas y demás personal relevante cambien entre las evaluaciones nacionales, es importante guardar todas las tablas de datos y también conservar una explicación clara de las decisiones y cambios clave realizados en los archivos de datos del ítem. Como contribución para las futuras evaluaciones nacionales, el analista debería escribir en un archivo de texto ***ReadMe*** una breve descripción de cualquier cambio que se efectuara en un archivo de datos del ítem durante el análisis en ejecución (véase Freeman y O'Malley, 2012).

NOTA

1. La carga de estos datos es opcional, porque la estimación del vínculo estadístico solo requiere parámetros del ítem. Si en esta etapa no se ha de cargar el puntaje IRT, aplique los resultados de la vinculación utilizando un paquete de software diferente (como por ejemplo SPSS [Paquete estadístico para ciencias sociales] o Excel).

CAPÍTULO 15

RESUMEN DE LAS GUÍAS DE IATA

Las guías de ejemplo completadas en la segunda parte de este volumen ofrecen descripciones detalladas de los procedimientos estadísticos más comunes necesarios para crear, implementar y mantener un sistema de evaluación nacional. Una vez que complete estas guías, debería disponer de los conocimientos necesarios para llevar a cabo las siguientes tareas:

- Cargar datos de respuestas de estudiantes
- Cargar datos de parámetros de ítems
- Especificar claves de respuesta a ítems, niveles de rendimiento de ítems y clasificaciones de contenido de ítems
- Especificar el tratamiento de los datos faltantes
- Revisar estadísticas clásicas de ítems
- Interpretar funciones de respuesta a ítems
- Interpretar resultados de dimensionalidad de pruebas
- Interpretar resúmenes estadísticos de pruebas
- Generar puntuaciones escaladas para informes
- Generar e interpretar análisis de funcionamiento diferencial de los ítems
- Estimar y definir umbrales para niveles de competencia

- Seleccionar subconjuntos de ítems para objetivos de medición específicos
- Guardar resultados en un equipo

Estas tareas representan prácticamente todos los requisitos normales para los análisis de pruebas en la implementación de una evaluación nacional. No obstante, limitarse a replicar los ejemplos tal y como aparecen en estos capítulos no es lo mismo que dominar la habilidad para utilizar estas funciones con sus propios datos de la evaluación nacional. Los siguientes pasos del proceso de aprendizaje deberían consistir en revisar cada ejemplo varias veces mientras sigue las instrucciones exactas de los capítulos.

Una vez que haya dominado la interfaz de Item and Test Analysis (IATA), debería estar preparado para experimentar con algunas de las opciones que componen los análisis en IATA. Una vez más, debería revisar cada una de las guías, pero, en lugar de seguir todas las instrucciones de forma precisa, experimente con las opciones disponibles. Por ejemplo, ¿qué ocurre con los resultados de los análisis cuando el número de ítems de la prueba es demasiado reducido? ¿Qué ocurre con los resultados equivalentes cuando el conjunto de ítems con vinculación común solo incluye ítems muy sencillos o muy complicados? Existen muchos otros métodos técnicos para analizar los datos de pruebas que están fuera del alcance de este volumen. No obstante, si experimenta con los datos de muestra de IATA y compara sus resultados con los obtenidos anteriormente durante los diferentes ejercicios de la guía, debería comprender mejor las opciones adecuadas para las diferentes situaciones.

Al analizar los datos de la evaluación nacional, debe identificar en primer lugar el flujo de trabajo más adecuado para su situación en el menú de IATA. Es muy probable que sea muy similar a uno de los ejemplos presentados en uno de los capítulos anteriores. Es posible que algunas situaciones requieran combinaciones de flujos de trabajo, donde los resultados de un flujo de trabajo o análisis se utilizan como datos de entrada para otro.

Finalmente, a medida que aumente su pericia, observará que rara vez existen respuestas o soluciones únicas y perfectas para los problemas de las evaluaciones nacionales. En el mejor de los casos, los

métodos estadísticos empleados en las evaluaciones modernas sirven para minimizar la influencia de los errores que inevitablemente se generan a causa de los desafíos reales de las mediciones educativas. La forma en que los equipos nacionales eligen e implementan estos métodos estadísticos depende de sus objetivos. ¿Cuáles son las necesidades de las entidades interesadas? ¿Cuáles son las consecuencias de las decisiones basadas en los resultados? IATA solo es una herramienta (muy útil, no obstante) para reducir la carga de estos métodos estadísticos y esclarecer los pros y contras de las opciones analíticas disponibles en una evaluación nacional.

ANEXO

II.A

TEORÍA DE RESPUESTA AL ÍTEM

En el capítulo 9 se describen dos aspectos del método de la Teoría Clásica de las Pruebas (TCP) para la medición de competencias, es decir, la facilidad (o dificultad) del ítem y la discriminación del ítem. En este anexo se presta atención a un método alternativo, la Teoría de Respuesta al Ítem (TRI), que unifica los conceptos de facilidad y discriminación del ítem. La TRI también se ha descrito como "teoría de rasgos latentes". Se trata del método más utilizado en las evaluaciones a gran escala.

Un buen punto de partida para comprender la TRI es contrastar qué constituye un ítem de prueba adecuado desde las perspectivas de la TCP y la TRI. Las estadísticas clásicas de los ítems de facilidad y discriminación se centran en estimar y comparar la probabilidad de que diferentes estudiantes elijan la respuesta correcta. En cambio, la TRI caracteriza a los estudiantes según el tipo de respuesta al ítem que generarán con mayor probabilidad e intenta describir las distribuciones de competencia para los estudiantes que respondan de diferentes formas. Aunque un ítem de prueba adecuado desde la perspectiva de la TCP presenta grandes diferencias en la probabilidad de que estudiantes con diferentes niveles de competencia elijan la respuesta correcta, un ítem de prueba adecuado desde la perspectiva de la TRI sería uno en el que la distribución de competencia de los estudiantes

que han respondido correctamente difiera de la distribución de competencia de los estudiantes que no lo han hecho. Mientras que la TCP se centra en la probabilidad de responder correctamente, la TRI presta atención a la estimación de las distribuciones de competencia. Aunque ambas perspectivas suelen estar de acuerdo, la perspectiva de la TRI describe los ítems de una forma mucho más completa y útil.

El programa Item and Test Analysis (IATA) calcula los resultados utilizando diferentes métodos estadísticos. La mayoría de los cálculos utilizan ecuaciones de forma cerrada, lo que significa que el cálculo utiliza los datos de las respuestas de los estudiantes en una progresión ordenada de pasos para producir la estadística deseada, como la media aritmética. En las ecuaciones de forma cerrada, a pesar incluso de que los cálculos cuenten con varios pasos, los valores de cada paso están basados en los datos originales y los resultados de los pasos anteriores. La mayoría de los libros de texto de estadística (por ejemplo, Crocker y Algina, 2006) ofrecen descripciones detalladas de los métodos de forma cerrada para el cálculo de estadísticas clásicas de ítems y otros resúmenes estadísticos básicos.

Algunos cálculos requieren que IATA estime una estadística, x, que está basada en otra estadística, y, donde el valor de y también está basado en el valor de x. En estos casos, ya que x e y no pueden estimarse juntos, IATA debe utilizar un *algoritmo iterativo*. Por lo general, un algoritmo iterativo asume en primer lugar varios valores iniciales para y y los utiliza para estimar los valores de x. A continuación, el algoritmo utiliza los resultados de x para calcular los nuevos valores de y. Después, los nuevos valores de y se utilizan para actualizar los valores de x, y el proceso se repite hasta que las nuevas iteraciones no cambian sustancialmente los valores de las estimaciones. Este método de cálculo se utiliza para analizar la dimensionalidad de las pruebas y los ítems, así como para estimar los parámetros de los ítems de la TRI (Lord y Novick, 1968). Ambos cálculos requieren la estimación de las propiedades de los ítems, como las cargas y los parámetros de TRI.

El análisis de la dimensionalidad utiliza un algoritmo iterativo común, conocido como "descomposición en valores singulares" o "DVS" (véase http://es.wikipedia.org/wiki/Descomposición_en_valores_singulares), pero la estimación de los parámetros de TRI requiere el uso de algoritmos iterativos especializados (Baker y Kim, 2004).

Estos algoritmos deben estimar en primer lugar la probabilidad de que cada estudiante responda correctamente a cada ítem y, posteriormente, deben encontrar los parámetros de ítems que reproduzcan de forma más adecuada dichas probabilidades. A continuación, se utilizan los nuevos parámetros para actualizar las probabilidades estimadas, que, a su vez, se utilizan para actualizar las estimaciones de los parámetros de los ítems, etc., hasta que las estimaciones de cada etapa no mejoran de forma considerable las estimaciones anteriores. La estimación de parámetros de IATA utiliza una variación de este método general que es informáticamente más rápido y estadísticamente más sólido que otros algoritmos. Permite el uso de métodos de TRI con una mayor variedad de datos de muestra que otros tipos de software por lo general.

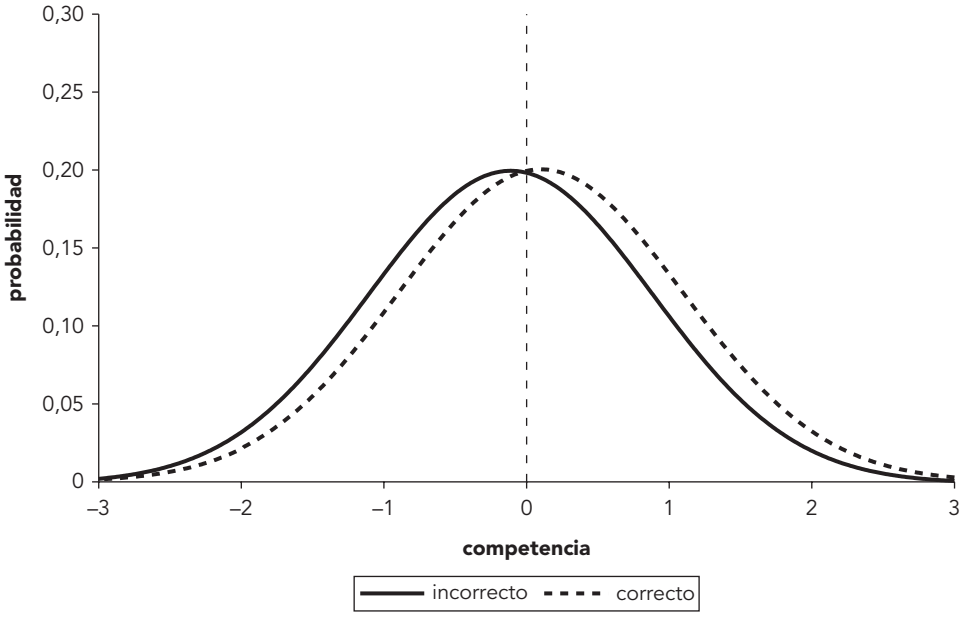
Con el algoritmo de IATA, la primera etapa (estimación de probabilidades) requiere el cálculo de dos distribuciones de competencia para cada ítem: la distribución de los encuestados que respondieron correctamente y la distribución de aquellos que no. Se asume que la forma de estas distribuciones es normal y, para cada ítem, ambas distribuciones comparten la misma variación, pero se diferencian en sus magnitudes relativas y medias. Por ejemplo, si el número de estudiantes que respondieron correctamente es mayor que el de encuestados que no, la magnitud de la distribución de encuestados que respondieron correctamente será mayor que la de encuestados que no. La suma de ambas distribuciones en cada nivel de competencia describe la distribución de competencia para todos los estudiantes, y la proporción de encuestados que respondieron correctamente para la distribución sumada genera las estimaciones de la probabilidad de respuestas correctas en cada nivel de competencia. Este método ofrece ventajas frente a los demás métodos por dos motivos: (a) describe la probabilidad de cada respuesta en todos los niveles de competencia, en lugar de en una muestra arbitraria de niveles de competencia, y (b) pueden describirse las distribuciones de encuestados que respondieron correctamente y los que no utilizando la media de encuestados que respondieron correctamente y la proporción de respuestas correctas para un ítem, ya que la media total de la muestra está limitada a cero y la proporción de respuestas incorrectas es igual a uno menos la proporción de respuestas correctas. En cambio, la mayoría de los métodos solo describen las probabilidades para una muestra de niveles de competencia definidos

arbitrariamente y pueden necesitar cientos de estadísticas calculadas de forma independiente para estimar las diferentes probabilidades. Por lo general, también requieren la especificación de limitaciones arbitrarias o reglas para corregir los errores de estimación.

Las dos distribuciones de la Figura II.A.1 muestran algunas de las características fundamentales de la TRI. Las dos curvas representan las distribuciones de competencia[1] de los encuestados para un solo ítem de prueba. La línea continua de la izquierda describe la competencia de los estudiantes que no respondieron correctamente, y la segunda línea curvada discontinua (---------) describe la competencia de los estudiantes que respondieron correctamente. Este ítem tiene una facilidad de 0,50, lo que refleja la altura idéntica de ambas distribuciones en el eje vertical; el número de encuestados que respondieron correctamente coincide con el número de encuestados que no. La competencia media

FIGURA II.A.1

Distribuciones de competencia para encuestados que respondieron correctamente y encuestados que no respondieron correctamente a un único ítem de prueba (facilidad = 0,50; competencia media de estudiantes que respondieron correctamente = 0)

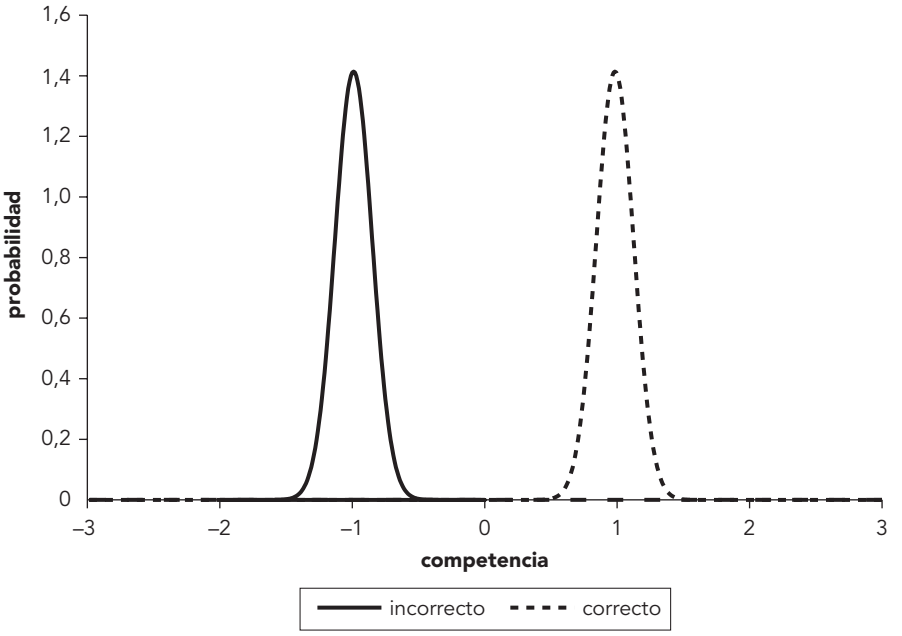


de los encuestados que respondieron correctamente es 0,10, lo que se refleja en el gráfico mediante el pico de la distribución de estudiantes que respondieron correctamente que se encuentra por encima del 0,10 en el eje de competencia. Ya que la media general de ambas poblaciones es 0 y tienen el mismo tamaño, la competencia media de los encuestados que no respondieron correctamente es simétrica en -0,10. Ambas distribuciones son muy similares en términos de tamaño y ubicación, lo que indica que hay muy pocas diferencias en la competencia entre el tipo de estudiantes que respondieron correctamente y el que no. Si no hubiera ninguna diferencia, ambas distribuciones serían idénticas con una media igual a 0, y las respuestas a los ítems no estarían relacionadas con la competencia.

En la Figura II.A.2 se muestra un ítem de prueba mucho más preciso, con una facilidad de 0,50. Este ítem ilustra la relación más estrecha

FIGURA II.A.2

Distribuciones de competencia para encuestados que respondieron correctamente y encuestados que no respondieron correctamente a un único ítem de prueba (facilidad = 0,50; competencia media de los estudiantes que respondieron correctamente = 0,99)

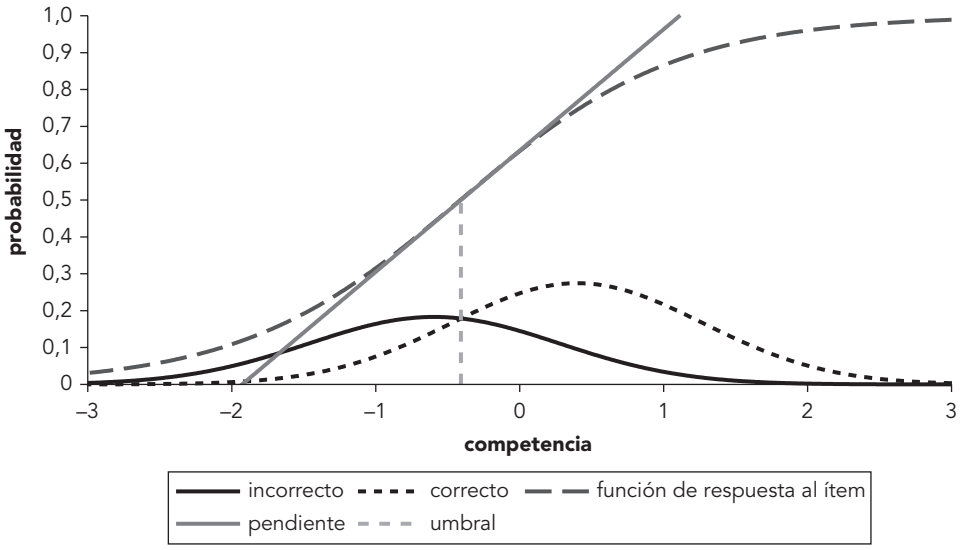


entre la respuesta al ítem y la competencia, donde la competencia media de los encuestados que respondieron correctamente tiene un valor de 1 aproximadamente y la competencia media de los que no respondieron correctamente tiene un valor de -1 aproximadamente. Las distribuciones no se superponen, lo que indica que, en términos de competencia, los encuestados que respondieron correctamente son totalmente diferentes de los que no.

En la práctica, es sumamente improbable que los encuestados que respondieron correctamente sean totalmente diferentes de los que no. Por lo general, existe un área de competencia amplia en la que ambas distribuciones se superponen. Se produce una transición fluida a medida que disminuye la probabilidad de que los estudiantes con mayor competencia sean miembros de la distribución de respuestas incorrectas y aumenta la probabilidad de que lo sean de la distribución de respuestas correctas. Esta transición está ilustrada en la Figura II.A.3 para un ítem con una facilidad de 0,60 (lo que indica

FIGURA II.A.3

Distribuciones de competencia para encuestados que respondieron correctamente y encuestados que no respondieron correctamente a un único ítem de prueba y probabilidad condicional de responder correctamente (facilidad = 0,60; competencia media de estudiantes que respondieron correctamente = 0,40)



que la distribución de encuestados que respondieron correctamente es mayor que la de encuestados que no) y una competencia media de 0,40 para los encuestados que respondieron correctamente. La línea curvada discontinua, también conocida como "*función de respuesta al ítem*" o "FRI", describe el tamaño de la distribución de encuestados que respondieron correctamente en comparación con la distribución de encuestados que no.

Dicho de otro modo, en las áreas de la competencia donde la altura de la distribución de respuestas correctas es menor que la altura de la distribución de respuestas incorrectas, la FRI es inferior a 0,5; cuando ocurre lo contrario, el valor es superior a 0,5. La transición de respuesta incorrecta a respuesta correcta aparece marcada mediante un umbral (la línea vertical discontinua de la Figura II.A.3) que corresponde al punto en el que la distribución de encuestados que no respondieron correctamente se cruza con la distribución de encuestados que sí.

La FRI puede interpretarse como la probabilidad de que un estudiante con un nivel de competencia determinado pertenezca al grupo de encuestados que respondió correctamente. Los valores exactos de la FRI pueden calcularse al dividir la probabilidad de la distribución de encuestados que respondió correctamente entre la suma de probabilidades de ambas distribuciones. Por ejemplo, con el nivel de competencia de -1, el valor de probabilidad de los encuestados que respondieron correctamente es 0,06 aproximadamente y el valor de los encuestados que no respondieron correctamente es aproximadamente 0,15; 0,06/(0,06 + 0,15) = 0,29. Ya que la proporción de encuestados que no respondieron correctamente es la opuesta a la proporción de encuestados que sí, y la competencia media de los encuestados que no respondieron correctamente puede calcularse a partir de la competencia media de los encuestados que sí (debido a que la media general es igual a 0), la FRI es una función de la facilidad del ítem y la competencia media de los encuestados que respondieron correctamente.

Una FRI puede describirse utilizando un modelo estadístico con tres parámetros: a, b y c:

$$P(u = 1) = c + (1 - c)/(1 + \text{Exp}(D * a * (theta - b))),$$

donde P(u = 1) es la probabilidad de que un estudiante elija una respuesta correcta. D representa una constante utilizada para escalar los parámetros del ítem; suele establecerse en -1,7 para que la escala coincida con la escala normal estándar. La variable *theta* representa la competencia de los estudiantes. El mismo modelo describe los ítems de crédito parcial, donde P($u \geq x$) representa cualquier puntuación mayor o igual que una puntuación de crédito parcial específica, x. En el caso del crédito parcial, cada puntuación mayor que 0 contaría con un conjunto de parámetros.

Aunque todos los parámetros interactúan para describir el comportamiento estadístico de un ítem, el parámetro a refleja principalmente la distancia entre las medias de las distribuciones de respuestas correctas e incorrectas; el parámetro b refleja principalmente la facilidad del ítem; y el parámetro c refleja la probabilidad de que se incluya por error a un estudiante de la distribución de respuestas incorrectas en la distribución de respuestas correctas (por ejemplo, si un estudiante adivina una respuesta).

Debido a que el proceso de la TRI es iterativo e informáticamente intensivo, los diferentes paquetes de software pueden producir estimaciones ligeramente diferentes y necesitar mucho tiempo para completar los cálculos. El algoritmo de estimación de IATA tiende a ser más sólido con muestras de diferentes tamaños y es considerablemente más rápido que otros programas de estimación de TRI. Mientras que otros métodos utilizan algoritmos de aproximación iterativa para llevar a cabo el paso de estimación de parámetros de los ítems, IATA calcula los parámetros de los ítems algebraicamente con las siguientes ecuaciones:

$$a = -(\mu^*_{correcto}/(-1 + p^* + p^* \mu^{*2}_{correcto}))/1{,}7\ (1 + q/(q + q_{correcto}))$$

$$b = (\mu_{incorrecto} + \mu^*_{correcto} - (2^*\sigma^{2*}\mathrm{LOG}(q^*/p^*))/(\mu_{incorrecto} - \mu^*_{correcto}))/2$$

$$c = q/(q + q_{correcto}),$$

donde

$$p^* = (1 - (1 - p)/(1 - c))$$

$$q^* = q + q_{correcto}$$

$$\mu^*_{correcto} = (-\mu_{incorrecto} * (1 - p^*))/p^*$$

$$\sigma^2 = 1 - (p^* \mu^{*2}_{correcto} + (q^*)^* \mu^2_{incorrecto} + p^* \mu^*_{correcto} + (q^*) \mu_{incorrecto})$$

$\mu_{correcto}$ = la competencia media de los estudiantes que respondieron correctamente

$\mu_{incorrecto}$ = la competencia media de los estudiantes que no respondieron correctamente

p = la proporción de estudiantes que respondieron correctamente

q = la proporción de estudiantes que no respondieron correctamente

$q_{correcto}$ = la proporción de estudiantes que adivinaron la respuesta correcta aunque no la sabían (esta estadística debe calcularse mediante la aproximación de la asíntota más baja de la función de respuesta empírica del ítem). Tenga en cuenta que no es necesario estimar $q_{correcto}$ si el parámetro c está limitado a 0 (una práctica óptima en una amplia variedad de situaciones de evaluación).

Las nuevas estimaciones de parámetros se utilizan en cada ciclo de estimación para generar funciones de competencia actualizadas para cada estudiante, utilizando los métodos descritos por Baker y Kim (2004). Aunque el algoritmo sigue necesitando un gran número de ciclos iterativos para generar las estimaciones finales, la solidez de las ecuaciones anteriores en el paso de estimación de los parámetros de los ítems reduce en gran medida el tiempo de cálculo y aumenta la estabilidad de las estimaciones.

NOTA

1. En la TRI, la competencia de los estudiantes se describe mediante una escala (por lo general, llamada *"zeta"*) que es similar a la escala de puntuación Z: el nivel de competencia medio teórico es 0 y la desviación estándar es 1. Por lo general, las puntuaciones de la mayoría de los estudiantes se encuentran entre -2 y 2, y menos de uno de cada mil estudiantes conseguirá una puntuación inferior a -3 (o superior a 3).

REFERENCIAS

Anderson, P., y G. Morgan. 2008. *Developing Tests and Questionnaires for a National Assessment of Educational Achievement*. Washington, DC: Banco Mundial.

Baker, F. B., y S.-H. Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. 2.ª ed. Nueva York: Marcel Dekker.

Bullock, J. G., D. P. Green, y S. E. Ha. 2010. "Yes, but What's the Mechanism? (Don't Expect an Easy Answer)." *Journal of Personality and Social Psychology* 98 (4): 550–58.

Crocker, L., y J. Algina. 2006. *Introduction to Classical and Modern Test Theory*. Pacific Grove, CA: Wadsworth.

Cronbach, L. J. 1970. "Test Validation." En *Educational Measurement*, 2.ª ed., editado por R. L. Thorndyke, 443–507. Washington, DC: Consejo Estadounidense sobre la Educación

De Ayala, R. J. 2009. *The Theory and Practice of Item Response Theory*. Nueva York: Guilford Press.

DeMars, C. 2010. *Item Response Theory*. Nueva York: Oxford University Press.
Dumais, J., y J. H. Gough. 2012a. "School Sampling and Methodology", en *Implementing a National Assessment of Educational Achievement*, editado por V. Greaney y T. Kellaghan, 57–106. Washington, DC: Banco Mundial.

———. 2012b. "Weighting, Estimating, and Sampling Error." En *Implementing a National Assessment of Educational Achievement*, editado por V. Greaney y T. Kellaghan, 181–257. Washington, DC: Banco Mundial.

Fan, X. 1998. "Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics." *Educational and Psychological Measurement* 58(3): 357–81.

Freeman, C., y K. O'Malley. 2012. "Data Preparation, Validation and Management." En *Implementing a National Assessment of Educational Achievement*, editado por V. Greaney y T. Kellaghan, 107–79. Washington, DC: Banco Mundial.

Goldstein, H., y R. Wood. 1989. "Five Decades of Item Response Modelling." *British Journal of Mathematical and Statistical Psychology* 42 (2): 139–67.

Greaney, V., y T. Kellaghan. 2008. *Assessing National Achievement Levels in Education*. Washington, DC: Banco Mundial.

———, eds. 2012. *Implementing a National Assessment of Educational Achievement*. Washington, DC: Banco Mundial.

Haladyna, T. M. 2004. *Developing and Validating Multiple-Choice Test Items*. 3.ª ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., H. Swaminathan, y H. J. Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hutcheson, G., y N. Sofroniou. 1999. *The Multivariate Social Scientist*. Londres: Sage.

Karantonis, A., y S. G. Sireci. 2006. "The Bookmark Standard Setting Method: A Literature Review." *Educational Measurement: Issues and Practice* 25 (1): 4–12.

Kellaghan, T., y V. Greaney. 2001. *Using Assessment to Improve the Quality of Education*. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Instituto Internacional de Planificación Educativa.

Kellaghan, T., V. Greaney, y T. S. Murray. 2009. *Using the Results of a National Assessment of Educational Achievement*. Washington, DC: Banco Mundial.

Lord, F. M., y M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Martin, M. O., I. V. S. Mullis, y P. Foy (con J. F. Olson, E. Erberber, C. Prewschoff, y J. Galia). 2008. *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at*

the Fourth and Eighth Grades. Chestnut Hill, MA: Centro de Estudios Internacionales TIMSS y PIRLS, Boston College.

Mislevy, R. J. 1992. *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Servicio de Pruebas Educativas

Mitzel, H. C., D. M. Lewis, R. J. Patz, y D. R. Green. 2001. "The Bookmark Procedure: Psychological Perspectives." En *Setting Performance Standards: Concepts, Methods, and Perspectives*, editado por G. J. Cizek, 249–81. Mahwah, NJ: Lawrence Erlbaum Associates.

OCDE (Organización para la Cooperación y el Desarrollo Económicos). 2007. *PISA 2006: Competencias en ciencias para el mundo del mañana. Volumen 1: Análisis.* París: OCDE.

Raudenbush, S. W., y A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2da. ed. Thousand Oaks, CA: Sage.

Snijders, T. A. B., y R. J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage.

ECOAUDITORÍA

Declaración de beneficios medioambientales

El Grupo Banco Mundial tiene el compromiso de reducir su huella ambiental. En apoyo a dicho compromiso, la División de Publicaciones y Conocimiento impulsa las opciones de edición electrónica y la tecnología de impresión por encargo, desde centros regionales distribuidos por todo el mundo. En conjunto, estas iniciativas permiten reducir las tiradas y las distancias de envío, lo que redunda en un menor consumo de papel, menor uso de productos químicos, menores emisiones de gases de efecto invernadero y menor cantidad de residuos.

La División de Publicaciones y Conocimiento sigue las normas recomendadas sobre el uso de papel establecidas por la Green Press Initiative (Iniciativa de Prensa Ecológica). La mayor parte de nuestros libros se imprime con papel certificado por el Consejo de Administración de Bosques (FSC), y el contenido en papel reciclado de casi todos ellos oscila entre el 50 y el 100 por ciento. La fibra reciclada del papel de nuestros libros es o bien sin blanquear o blanqueada mediante procesos totalmente libres de cloro (TCF), procesos de fabricación sin cloro (PCF) o procesos de blanqueo libre de cloro elemental mejorado (EECF).

Puede encontrarse más información sobre la filosofía ambiental del Banco en http://www.worldbank.org/en/about/what-we-do/crinfo.

Evaluaciones nacionales del rendimiento académico

La evaluación efectiva del desempeño de los sistemas educativos es un componente clave en la formulación de políticas para optimizar el desarrollo del capital humano en todo el mundo. Los cinco libros de la serie *Evaluaciones nacionales del rendimiento académico* presentan conceptos clave de las evaluaciones nacionales de los niveles de rendimiento estudiantil, desde las cuestiones normativas que deben abordarse cuando se diseña y se lleva a cabo las evaluaciones hasta el desarrollo de las pruebas, el muestreo, la depuración de datos, las estadísticas, la redacción de informes y el uso de los resultados para mejorar la calidad de la educación.

Análisis de los datos de una evaluación nacional del rendimiento académico es el cuarto de cinco volúmenes de la serie *Evaluaciones nacionales del rendimiento académico*. Otros volúmenes han descrito los procedimientos de una evaluación hasta llegar a la fase de preparación de los datos para su análisis estadístico, el tema del presente volumen. Los análisis precisos que se lleven a cabo dependerán de las necesidades de información de los responsables políticos y los gestores educativos. En la mayoría de las evaluaciones nacionales, estas necesidades tienen que ver con la calidad del aprendizaje estudiantil, los factores relacionados con el aprendizaje, las cuestiones de equidad, y en algunos casos, la transformación de los resultados educativos con el paso del tiempo.

El volumen 4, que consta de dos partes, explica detalladamente los pasos necesarios para el análisis de los datos recopilados en una evaluación nacional. La 1.ª parte ofrece una introducción general a los análisis estadísticos llevados a cabo normalmente en las evaluaciones a gran escala, midiendo la tendencia central y la dispersión de los puntajes de los alumnos, así como las relaciones entre variables. La 2.ª parte describe el programa IATA (Item and Test Analysis), que utiliza la Teoría Clásica de las Pruebas y la Teoría de Respuesta al Ítem para establecer escalas sobre las que determinar los puntajes de los alumnos. Se describe en detalle los pasos del análisis de las administraciones de las pruebas piloto y las pruebas definitivas. Un CD complementario contiene ejercicios especialmente diseñados y ficheros de datos de respaldo para ambas partes del volumen. Este libro será de interés para los especialistas en evaluaciones en gobiernos nacionales, regionales y locales, instituciones de investigación y universidades.

READ
Russia Education Aid for Development
When children learn, nations prosper

Irish Aid
Rialtas na hÉireann
Government of Ireland

GRUPO BANCO MUNDIAL

ISBN 978-1-4648-0749-7

SKU 210749