Universidad de Murcia
Facultad de Psicología

Tesis Doctoral

Propiedades distribucionales de un estadístico de medición apropiada

Rosa María Núñez Núñez

Director: Dr. D. José Antonio López Pina
Universidad de Murcia

Murcia, 2002

No hay certeza alguna allí donde no se pueda aplicar
alguna de las ciencias matemáticas o algunas de las
que se relacionan con las ciencias matemáticas.

Leonardo da Vinci, *Cuadernos de notas*

Agradecimientos

Cuando terminé Psicología y me dieron la beca predoctoral recuerdo que los compañeros me preguntaban sobre qué trataría de hacer la tesis y yo les contestaba que no lo sabía, pero seguro que tendría que ver con la Psicometría. Ellos se sorprendían: ¿¡De Psicometría!? ¿Cómo te puede gustar eso? Pues sí, *eso* me gusta. Después de dejar mis estudios en Ciencias Exactas y de darme cuenta, al finalizar la diplomatura de Magisterio en Ciencias Físico-Matemáticas, de que mi vocación no era la enseñanza en la EGB –todavía llamada así por aquel entonces–, desemboqué en la Facultad de Psicología porque el plan de estudios incluía Estadística en el curso puente o pasarela, como se prefiera, y Psicología Experimental y Psicometría en cuarto curso, lo que me llevó a pensar que habría números por algún sitio. El primer parcial de Estadística lo impartió el Dr. José Antonio López Pina; en sus clases disfrutaba y me gustó tanto como profesor que, mira por donde, es el director de esta tesis. Muchas veces no se encuentran palabras para agradecer a alguien lo mucho que ha confiado en tí, te ha enseñado y animado para conseguir una de tus metas, y eso es lo que me ocurre a mí. Por lo tanto, siguiendo el *principio de la parsimonia* de Guillermo de Ockham por el que se rige el método científico, simplemente diré: Gracias, José Antonio.

Durante casi cinco años formándome en el Area de Metodología de las Ciencias del Comportamiento, el apoyo, la ayuda y el afecto de la Dra. Mª Dolores Hidalgo Montesinos han sido inconmensurables. Ella ha sabido siempre darme buenos consejos y orientaciones. Junto con ella, en este área he tenido la suerte de tener por profesores y compañeros al Dr. Manuel Ato García, Dr. Julio Sánchez Meca, Dr. Antonio P. Velandrino Nicolás, Dr. Juan José López García, Dr. Fulgencio Marín Martínez y Dr. Rafael Rabadán Anta, personas con mucha paciencia que han soportado todas mis interrupciones –que no han sido pocas– y con las que he compartido innumerables momentos.

Ahora que he logrado algo con lo que siempre soñé y siendo consciente de lo que ha acontecido, toda la gratitud que dirijo a mis padres, José María y Rosa María, y a mi hermana Patricia, se queda escasa. Ellos mejor que nadie han sabido llevar lo mejor posible mis cambios de humor, mis malos momentos, también han sido partícipes de mis ratos de gloria, a cambio de cariño, comprensión y, como no, de resignación. Pero ya se sabe que la convivencia es compleja y los trueques no son siempre justos.

En mi opinión, la elaboración de una tesis o de cualquier proceso de investigación es un camino arduo y difícil de entender por quien no lo ha llevado a cabo alguna vez. Por eso, aprovecho esta ocasión para aplaudir a mis amigos incondicionales, quienes han demostrado una empatía tremenda; ellos siempre han estado ahí, dándome aliento y tragándose mis *neuras* (algunas, por cierto, de libro de texto).

Gracias al Dr. Francisco Toledo Romero, un amigo muy querido al que admiro y aprecio, que siempre me ha dado muy buenos remedios, a quien le debo mucho y nada de ello canjeable.

No me puedo olvidar de los que han sido o siguen siendo becarios *precarios* a la par que yo, con los que he tenido el placer de coincidir estos años: Dr. Cándido J. Inglés Saura, Juan Pedro Sánchez Navarro, y especialmente, Francisca Galindo Garre y Ana Díaz Beyá con quienes, al ser compañeras de despacho, he pasado muchas horas, de las que tengo recuerdos muy gratos y su amistad, y a las que me acuso de convertir en fumadoras pasivas. Gracias a todos los profesores y alumnos, colegas de otras áreas y departamentos de la Facultad de Psicología que se han interesado por mi trayectoria.

A todos ellos y a los que me dejo en el tintero (o en el tóner): Gracias.

Índice general

Capítulo 1

Introducción

Hoy en día nos vemos implicados, directa o indirectamente, en procedimientos de evaluación. Los sistemas de evaluación más conocidos son los empleados en el contexto educativo, clínico y laboral; las encuestas y los sondeos de opinión son otro medio de valoración sobre una amplia diversidad de temas políticos, económicos, medioambientales... Aunque no existen cánones para la obtención de medidas exactas y precisas en el campo de las Ciencias Sociales y del Comportamiento, lo que sí está bastante generalizado es el uso de instrumentos con los que se consiga una evaluación objetiva, dejando de lado, en cierto modo, todos aquellos que impliquen juicios subjetivos. Se ha asignado el término *test* a todas las pruebas de evaluación y valoración que poseen la característica, o al menos el propósito, de ser medidas objetivas de conocimientos y aptitudes generales o particulares sobre una materia; de niveles de desarrollo sensitivo, perceptivo y motor; de grados de destrezas implicados en una tarea; de actitudes hacia una situación concreta, de rasgos de personalidad y de patologías físicas o psicológicas. En el mercado hay tests de muy diversa índole desde los clásicos de papel y lápiz o con material manipulativo, hasta los que emplean los recursos más avanzados de la informática e incluso algunos de ellos se aplican por medio de Internet. Con los tests se pretende hacer inferencias, clasificar u ordenar a las personas, diligencias éstas que serán más o menos adecuadas en función de cómo se cumplan dos requisitos básicos e imprescindibles que deben poseer estos instrumentos de medida: fiabilidad y validez.

La Psicometría es la rama de la Psicología que se encarga de medir variables psicológicas (habilidades, rasgos de personalidad, aptitudes, actitudes, rendimiento...) para describirlas, explicarlas y predecirlas a través de un elenco de dominios de conductas, las cuales representan las variables factibles y susceptibles de medición. Además, esta disciplina es la encargada de supervisar el

proceso de construcción de tests y de escalas de medida, que para darlo por concluido recoge datos de un estudio piloto a partir del cual se analizará la fiabilidad y la validez del instrumento. Un estudio piloto consiste, entre otras cosas, en reunir las respuestas de una muestra representativa de sujetos para los que el test ha sido diseñado. Sin embargo, no son pocas las ocasiones en las que la estimación del constructo psicológico que se persigue con el test es inadecuada, con los consecuentes perjuicios que supone la veracidad del instrumento de medida, bien sea porque el sujeto se expone a una situación de test con la que no está familiarizado, bien porque deliberadamente decide no contestar al test de manera honesta o no le presta atención suficiente, o bien porque ha sido preparado o entrenado a contestarlo para lograr una determinada puntuación. Si parte de los sujetos de la muestra piloto contestan con estos u otros mecanismos indeseables de respuesta, el test se sometería a un replanteamiento innecesario con consiguientes costos temporales y económicos. Si estos comportamientos ocurren en situaciones de test ya estandarizados, el evaluador obtendrá estimaciones erróneas del constructo y las decisiones que se deriven del empleo del test carecerán de fundamento. Si al conjunto de respuestas de un sujeto a los ítems de un test se le denomina *patrón de respuestas*, al patrón que no representa a la variable psicológica del sujeto se le ha calificado de diferentes formas: *patrón atípico*, *patrón no normal*, *patrón inapropiado*, *patrón no ajustado* o *patrón aberrante*, siendo este último la traducción literal del término inglés *aberrant pattern*. Detectar patrones de respuesta inapropiados o atípicos es muy importante en muchos escenarios; por ejemplo, en selección de personal se suelen diseñar perfiles del trabajador óptimo para el desempeño de un puesto de trabajo, si se elige a sujetos no cualificados esto puede conllevar gastos en vano para una formación posterior; en el ámbito educativo se califica a los alumnos con los resultados de pruebas de conocimientos y aptitudes, en muchos casos útiles para admitirlos o excluirlos de determinadas titulaciones, si los patrones son incoherentes con las capacidades reales de los alumnos se podrían cometer injustas decisiones.

Por lo tanto, además de que los conocimientos psicométricos de los tests son requisitos indispensables para todos aquellos que los utilicen, los usuarios de los mismos deben ser conscientes de que los datos aportados por estas pruebas repercuten, en mayor o menor grado, en las personas evaluadas. La Comisión Internacional de Tests (*International Test Commission* [ITC], 2000; Bartram, 2001) publicó *Directrices Internacionales para el Uso de los Tests* en donde, uno de sus puntos acerca de la política del empleo de los mismos, se puede leer que el evaluador tiene "responsabilidades hacia los evaluados, antes, durante y después de la sesión de tests". Dicho esto, suponiendo que el usuario de tests está documentado en las propiedades y restricciones psicométricas de la prueba

con la que va a evaluar y la aplica en el contexto apropiado, en el momento de analizar e interpretar los resultados, según la ITC, el evaluador debe:

- Inspeccionar los resultados para detectar posibles errores o anomalías en las puntuaciones (apartado 2.6.8).
- Interpretar los resultados a la luz de la información disponible sobre la persona (edad, escolaridad, cultura, etc.), teniendo en cuenta las limitaciones técnicas del test, el contexto de la evaluación, y las necesidades de las personas o instituciones con intereses legítimos en el resultado del proceso evaluativo (apdo. 2.7.5).
- Tomar en consideración cualquier experiencia previa que la persona evaluada haya tenido con el test, en el caso de que se disponga de datos sobre los efectos de dicha experiencia sobre el rendimiento de la prueba (apdo. 2.7.12).

Frary (1993), y Hulin, Drasgow y Parsons (1983, cap. 4) comentan los estudios pioneros sobre el problema de los patrones de respuesta atípicos y el interés por detectarlos, para lo que se valían de procedimientos heurísticos. Los primeros patrones atípicos que se definieron fueron los de azar y copia (véase Frary, 1993) catalogados como *patrones de respuestas espurias*. Un patrón de respuestas espurias altas correspondería al sujeto de baja habilidad que copia las respuestas de los ítems más difíciles del test de otro sujeto más apto que él. Un patrón de respuestas espurias bajas sería característico de sujetos que se enfrentan a un test bidimensional y son muy hábiles en la dimensión medida por los ítems más difíciles del test, pero son menos diestros en la dimensión medida por los ítems más fáciles y optan por contestarlos azarosamente.

El incremento de investigaciones y el planteamiento de métodos y estadísticos para identificar dichos patrones comenzaron hace más de medio siglo con el trabajo Guttman (1950). Este autor elaboró un modelo determinístico según el cual, ante un test cuyos ítems están ordenados en dificultad creciente o, lo que es lo mismo, en orden decreciente a la proporción de aciertos (π_j), los sujetos con baja habilidad contestarían correctamente a los ítems de baja dificultad y erróneamente a los ítems de dificultad media y alta. A partir de este modelo, Meijer (1996) recoge los cuatro tipos de patrones atípicos definidos por Levine y Rubin (1979) y describe tres categorías más:

1. *Torpeza*. El sujeto tiene dificultad para comenzar a contestar el test y necesita tiempo para adaptarse a él. Cuando lo consigue, no presta atención y no comprueba las respuestas que ha dado a algunos de los ítems

más fáciles por lo que, al final, el porcentaje de ítems fáciles acertados es inferior al de los ítems de dificultad media y alta.

2. *Azar*. Un sujeto de baja habilidad no debería acertar los ítems con niveles medios y altos de dificultad. A veces, el sujeto decide contestar al azar estos ítems y, en consecuencia, la proporción de ítems fáciles acertados sería alta mientras que la proporción de aciertos a ítems de dificultad media y alta sería, aproximadamente, equivalente al total del número de ítems medios y difíciles dividido por el número de opciones de respuestas de los ítems.

3. *Copia*. Un sujeto con baja habilidad acierta los ítems más fáciles del test y se arriesga con los ítems de dificultad media; sin embargo, desconoce las respuestas a los ítems de dificultad alta por lo que decide copiar las respuestas de su compañero, el cual tiene un nivel de habilidad más alto que el suyo. El patrón resultante tendrá altos porcentajes de respuestas correctas en los ítems fáciles y difíciles del test.

4. *Tenacidad*. Un sujeto tenaz es aquel que cavila la respuesta al ítem, es metódico, elabora la respuesta cautelosamente y, además, no consiente pasar al siguiente ítem si no ha contestado al anterior. Los patrones de estos sujetos siguen el orden del patrón de Guttman. El problema reside en que estos sujetos no suelen concluir el test cuando hay un tiempo límite para contestarlo, de ahí que su patrón tenga un alto porcentaje de respuestas correctas en ítems fáciles y de media dificultad y bajo en ítems difíciles.

5. *Errores de transcripción*. Algunos tests contienen dos partes: una con los enunciados de los ítems y las opciones de respuesta, la otra formada por una hoja de codificación de respuestas en la que los sujetos deben marcar la opción escogida del ítem. Sea un sujeto de alta habilidad que decide no contestar por el momento o *saltar* uno de los ítems y pasar al siguiente; el error se comete en la hoja de respuestas al marcar la opción elegida en el lugar destinado al ítem precedente, es decir, al ítem que ha saltado. En consecuencia, los ítems de alta dificultad son incorrectos cuando deberían haber sido contestados correctamente.

6. *Ingenio*. Esta situación puede ocurrir en sujetos competentes que encuentran demasiado fáciles los ítems de baja dificultad y los reinterpretan, lo cual conlleva el darles respuestas erróneas. Los ítems de dificultad media y alta los aciertan. La tasa de respuestas a ítems fáciles es baja y la correspondiente a ítems de dificultad media y alta es elevada.

7. *Impericia*. Sea un test bidimensional: una de las habilidades es evaluada por los ítems más fáciles y la segunda habilidad lo es por los ítems de dificultad media y alta. Un sujeto que no sea competente en la habilidad mediada por los ítems fáciles del test los fallará, mientras que tendrá éxito en los ítems de media y alta dificultad.

Un ejemplo de estos patrones atípicos de respuesta se describen en la Tabla 1.1. Sea un test formado por 10 ítems de respuesta dicotómica o dicotomizados, los cuales están ordenados en orden de dificultad creciente. Los valores π_j han sido seleccionados aleatoriamente para que hubiera ítems fáciles (ítems 1-4), de dificultad media (ítems 5-7) e ítems difíciles (ítems 8-10).

Tabla 1.1. Perfiles de patrones de respuesta atípicos

	Item										**Patrón**
Sujeto	1	2	3	4	5	6	7	8	9	10	**atípico**
1	0	0	0	1	1	1	1	1	0	1	Torpeza
2	1	1	1	1	0	0	1	0	0	1	Azar
3	1	1	0	1	0	1	0	1	1	1	Copia
4	1	1	1	1	1	1	1	0	0	0	Tenacidad
5	1	1	1	1	0	1	0	0	0	0	Err.trans.
6	0	0	0	0	1	1	1	0	1	1	Ingenio
7	0	0	1	0	1	1	1	0	1	0	Impericia
π_j	0.90	0.85	0.83	0.82	0.57	0.50	0.48	0.32	0.25	0.15	

La parte de la Psicometría encargada de identificar y tratar los patrones de respuesta atípicos fue definida por Levine y Rubin (1979) como *medición apropiada*, campo de investigación y estudio que, a su vez, es de gran utilidad dentro del contexto de la construcción de tests y el uso de bancos de ítems con propiedades psicométricas, así como en el análisis de la validez de los mismos (Drasgow y Guertler, 1987; Meijer, 1997, 1998; Schmitt, Chan, Sacco, McFarland y Jennings, 1999; Schmitt, Cortina y Whitney, 1993). Las técnicas y estadísticos que evalúan el ajuste de las respuestas del sujeto se les considera *índices de medición apropiada* (Drasgow, Levine y McLaughlin, 1987; Drasgow, Levine y Williams, 1985; Levine y Rubin, 1979). En la literatura aparecen otras designaciones empleadas indistintamente: *índices de escalabilidad* (Reise y Waller, 1993) e *índices de precaución* (Sato, 1975; Tatsuoka, 1984). Meijer y Sijtsma (1999, 2001) aconsejan la denominación *métodos de ajuste de persona* de Trabin y Weiss (1979), ya que con ella se refieren a todos los métodos estadísticos que evalúan la ausencia de ajuste de las respuestas de un sujeto en un test con relación a un modelo de la Teoría de Respuesta al Item (TRI) o a la falta de concordancia entre dicho patrón y los patrones de la

muestra a la que pertenece el sujeto. Esta definición sería más general que la de índices de medición apropiada, con la cual sólo se contemplaban los métodos que identificaban patrones atípicos en el entorno de la TRI. En este trabajo de revisión de Meijer y Sijtsma se clasifican los estadísticos de ajuste de persona por la teoría o el modelo psicométrico bajo el cual se han elaborado:

1. Métodos basados en la comparación de un patrón de respuestas con los de un grupo normativo y procedimientos no paramétricos:
   - El coeficiente de correlación biserial-personal de Donlon y Fischer (1968).
   - El índice de precaución de Sato (1975).
   - El índice U_1 de Meijer (1994) y van der Flier (1980).
   - Los índices de acuerdo, desacuerdo y seguridad de Kane y Brennan (1980).
   - El índice de precaución modificado de Harnisch y Linn (1981).
   - Los índices U_3 y Z_{U_3} de van der Flier (1982).
   - Los índices de conformidad y de consistencia con la norma de Tatsuoka y Tatsuoka (1983).
   - El estadístico H_i^T de Sijtsma (1988), y Sijtsma y Meijer (1992).
2. Métodos para el modelo de Rasch:
   - El análisis de residuales de Wright y Masters (1982), y Wright y Stone (1979).
   - Los índices UB y UW de Smith (1985).
   - El estadístico M de Molenaar y Hoijtink (1990).
   - El índice χ^2_{SC} de Klauer y Rettig (1990).
   - El índice $T(X)$ de Klauer (1991, 1995).
3. Métodos para los modelos de dos y tres parámetros:
   - Los estadísticos basados en la función de verosimilitud: l_0 de Levine y Rubin (1979), l_z de Drasgow, Levine y Williams (1985) y l_{zm} de Drasgow, Levine y McLaughlin (1991).
   - La curva de respuesta de persona de Trabin y Weiss (1983), y Weiss (1973).
   - Los estadísticos ECI basados en el índice de precaución de Tatsuoka (1984).

- Los estadísticos de ajuste óptimo de Drasgow, Levine y McLaughlin (1987).
- El índice $\lambda(\mathbf{u})$ de Levine y Drasgow (1988).

4. Métodos para los Tests Adaptativos Informatizados (TAI):
   - El índice K de Bradlow, Weiss y Cho (1998).
   - El estadístico T de van Krimpen-Stoop y Meijer (2000).
   - El índice Z_c de McLeod y Lewis (1999).

Todos ellos han sido empleados en tests simulados (Li y Olejnik, 1997; Nering, 1995, 1997; Nering y Meijer, 1998; Noonan, Boss y Gessaroli, 1992; Reise, 1990; Rogers y Hattie, 1987; van Krimpen-Stoop y Meijer, 1999) y con patrones de respuesta reales o simulados de tests de actitudes (Parsons, 1983), de personalidad (Meijer y Nering, 1997; Reise, 1995; Reise y Flannery, 1996; Reise y Waller, 1993; Reise y Widaman, 1999; Schmitt *et al.*, 1999; Zickar y Drasgow, 1996), y de tests cognitivos y de aptitudes (Birenbaum, 1986; Drasgow, 1982; Drasgow y Levine, 1986; Drasgow, Levine y McLaughlin, 1987, 1991; Drasgow, Levine y Williams, 1985; Harnisch y Linn, 1981; Harnisch y Tatsuoka, 1983; Klauer y Rettig, 1990; Levine y Drasgow, 1982, 1983a, 1983b; Levine y Rubin, 1979; McLeod y Lewis, 1999; Meijer, 1994; Meijer, Muijtjens y van der Vleuten, 1996; Meijer y Nering, 1997; Meijer, Sijtsma y Smid, 1990; Miller, 1986; Molenaar y Hoijtink, 1996; Rudner, Bracey y Skaggs, 1996; Schmitt *et al.*, 1999; Tatsuoka, 1996; Tatsuoka y Tatsuoka, 1982, 1983; Trabin y Weiss, 1983; van der Flier, 1982; Wollack, 1997; Wollack, Cohen y Serlin, 2001; Wright y Masters, 1982; Wright y Stone, 1979).

Este trabajo se ha centrado en evaluar uno de los estadísticos más exitosos en la identificación de patrones de respuesta atípicos: el índice l_z. La decisión de si un patrón es atípico o no se toma sobre la base de su prueba estadística que compara los valores observados con los teóricos, ya que l_z sigue una distribución normal de media 0 y desviación típica 1 cuando los datos se ajustan al modelo. Si los valores de l_z se aproximan a 0, el patrón de respuesta es apropiado; si l_z tiene valores negativos, el patrón es atípico; si el estadístico l_z es positivo, el patrón de respuesta observado es más apropiado que el pronosticado por el modelo.

Distintas investigaciones han comprobado que la normalidad del estadístico está afectada por factores tales como el método de estimación de la habilidad (Nering, 1995, 1997; Reise, 1995), la longitud del test (Noonan *et al.*, 1992), el tipo de test (van Krimpen-Stoop y Meijer, 1999), el modelo de respuesta ajustado (Noonan *et al.*, 1992) y la dimensionalidad del test (Li y Olejnik,

1997). Lógicamente, si el supuesto de normalidad es violado, la prueba estadística provocará la inexactitud de la clasificación de un patrón normal o atípico. La parte empírica en la que consiste esta tesis intenta esclarecer la adecuación a ley de normalidad del estadístico l_z con un estudio de simulación que manipulará diferentes condiciones experimentales: la longitud del test, la magnitud del parámetro de discriminación, el modelo psicométrico de medida, la distribución de la habilidad de los sujetos y el método de estimación de los parámetros del modelo.

Pero dos capítulos preceden al que contiene la praxis. En el capítulo 2 se agrupan los procedimientos de ajuste de personas que, siguiendo la clasificación de Meijer y Sijtsma (2001), o bien toman como criterio a un grupo normativo, bien se han desarrollado para modelos no paramétricos, bien pertenecen a la Teoría de la Generalizabilidad (TG). Dado que el estadístico l_z se elaboró bajo los supuestos de la TRI, ésta se expone en el capítulo 3 junto con los índices de medición apropiada que en ella se han propuesto, así como los más novedosos en la aplicación de los TAI. Al final de este capítulo, se hace una reseña a un índice de ajuste de persona planteado dentro del Análisis de Estructura de Covarianza (AEC).

Capítulo 2

Estadísticos para detectar patrones atípicos tomando como criterio un grupo normativo

Dentro de este capítulo se exponen aquellas técnicas desarrolladas para la detección de patrones atípicos no basadas en los supuestos de la TRI. En general, el procedimiento que siguen es escoger un patrón de respuestas de un sujeto en un test de ítems dicotómicos o dicotomizados y compararlo con el patrón esperado según el grupo normativo al que pertenece. Un patrón de respuestas individual se define por una secuencia de 1s y 0s en función de si la respuesta es afirmativa o negativa, acierto o fallo, respectivamente. Estos estadísticos son:

1. Los estadísticos obtenidos a partir del escalograma de Guttman (1950):
   - El coeficiente de correlación biserial-personal de Donlon y Fischer (1968).
   - El índice de precaución de Sato (1975).
   - Los índices basados en el número de errores Guttman de Meijer (1994) y van der Flier (1977).
   - El índice de precaución modificado de Harnisch y Linn (1981).
   - El índice de conformidad con la norma de Tatsuoka y Tatsuoka (1982).
   - El coeficiente de escalabilidad de Sijtsma y Meijer (1992).

2. El índice g_2 de Frary, Tideman y Watts (1977).
3. Los estadísticos basados en modelos de respuesta al ítem no paramétricos: los índices U_3 y Z_{U_3} de van der Flier (1980, 1982).
4. Los estadísticos desarrollados dentro de la TG: los índices de acuerdo y desacuerdo de Kane y Brennan (1980).

Antes de comentar los aspectos más importantes de los estadísticos que se han enumerado, se describen dos índices no derivados de ninguna teoría de medida psicológica, cronológicamente anteriores a los citados y que Hulin, Drasgow y Parsons (1983, cap. 4) incluyeron en una revisión de los procedimientos de medición apropiada basados en la TRI: el *índice de predicción de la previsibilidad* de Ghiselli (1960) y el *promedio ponderado* de Jacobs (1963). Son dos métodos heurísticos que inicialmente no tenían como finalidad la identificación de patrones atípicos, pero que podían ser utilizados para tal propósito.

El índice de predicción de la previsibilidad de Ghiselli

El empleo de tests en los ámbitos académico y profesional para pronosticar las puntuaciones en el criterio –ya sea éste la nota media de un estudiante a final de curso o el rendimiento laboral– está sujeto a errores, por lo que los pronósticos son imperfectos en algunas ocasiones. Esta fue la razón que llevó a Ghiselli (1960) a idear un procedimiento que redujera el error de predicción test-criterio, mediante la identificación de aquellos individuos para quienes las puntuaciones en el test no pronosticaban correctamente las puntuaciones en el criterio. La relación test-criterio estaría representada por una recta de regresión que se debería verificar:

$$\hat{y}_i = a + bX_i$$

donde $\hat{y}_i$ es la puntuación pronosticada en el criterio y X_i es la puntuación del sujeto i en el test. Sin embargo, cuando se obtienen las puntuaciones verdaderas en el criterio (y_i) y éstas se representan junto con la recta de regresión del pronóstico, se aprecia que algunas de las puntuaciones verdaderas en el criterio se alejan de la recta. La distancia entre las puntuaciones en el criterio observadas y pronosticadas para un sujeto i (D_i) sería un indicador del error cometido en el pronóstico y se obtiene con la expresión:

$$D_i = |y_i - \hat{y}_i|$$

Ghiselli intentó encontrar otra variable Z relacionada con D_i, a la que denominaría *predictor de previsibilidad*, encargada de identificar sujetos con mayor probabilidad de contener errores de predicción elevados. Los valores de Z eran obtenidos con un proceso de validación cruzada mediante la partición de una muestra en dos subgrupos. En uno de los subgrupos se realiza el análisis de ítems y se calculan las D_i; tras el análisis de ítems se seleccionan aquellos que discriminan bien entre los sujetos que son previsibles –las distancias son pequeñas– y los imprevisibles –distancias grandes–. Los ítems válidos para prever las puntuaciones en el criterio son los que definen Z y los que después son empleados en el segundo subgrupo; las puntuaciones Z pronosticarían los sujetos de este subgrupo cuyas puntuaciones serían previsibles. Se comprobó que la validez de los ítems del test aumentaba al eliminar de la muestra a los sujetos cuyas puntuaciones son imprevisibles.

El predictor de previsibilidad Z es una variable que depende del test y de la muestra, por lo que hay que calcularlo en cada nueva situación de test, convirtiéndolo en un procedimiento poco rentable e incluso, a veces, difícil de calcular. Además, Z se define por métodos empíricos que dejan sin explicación la imprevisibilidad de algunos de los sujetos.

El promedio ponderado de Jacobs

El promedio ponderado (J_i) de Jacobs (1963) es una estimación cuantitativa de un patrón de respuesta atípico aunque, al igual que el anterior, no fue ideado para tal objetivo. Para calcularlo se ordenan los ítems del test desde el menos difícil al más difícil y se agrupan en quintiles; J_i tiene por expresión:

$$J_i = \frac{Q_2 + 2Q_3 + 3Q_4 + 4Q_5}{X_i}$$

donde Q_m es el número de respuestas correctas del sujeto i en cada quintil ($m = 1, 2, 3, 4, 5$) y $X_i = \sum_{m=1}^{5} Q_m$, es decir, el número de respuestas correctas total del sujeto en el test. El índice J_i toma valores entre 0 y 4; si $J_i = 0$ indica que todas las respuestas correctas están en el primer quintil (en los ítems más fáciles del test); si $J_i = 4$ el sujeto habría contestado correctamente a los ítems más difíciles del test, por lo que su patrón de respuestas sería calificado de atípico, al igual que aquellos otros sujetos que obtuvieran altos valores de J_i.

Los inconvenientes de este índice son: a) la pérdida de información que produce la agrupación de los ítems en quintiles; b) la ponderación de los quintiles en la ecuación es arbitraria, ya que supone que la dificultad de los ítems de Q_1 es 0, la de Q_2 es 1 y así sucesivamente; c) J_i es un promedio de la dificultad

de los ítems acertados, por lo que un valor J_i alto debería interpretarse como que el sujeto ha acertado muchos ítems.

2.1. El escalograma de Guttman y sus derivaciones estadísticas

Las primeras técnicas utilizadas para identificar patrones de respuesta atípicos comparaban el patrón de respuesta observado de un sujeto con el patrón esperado según el modelo determinístico de Guttman (1944, 1950). Este modelo fue elaborado con el propósito de construir escalas unidimensionales de actitudes, en las que los ítems estarían ordenados de modo que la afirmación de un ítem implicaría la asertividad de los ítems jerárquicamente inferiores. Si se considera una muestra de N sujetos y un test de actitudes con n ítems dicotómicos, cada sujeto tendría un vector de respuestas de la forma:

$$\mathbf{U}_i = (u_{i1}, u_{i2}, \ldots, u_{ij}, \ldots, u_{in})$$

donde u_{ij} es la respuesta del sujeto i al ítem j, la cual sería $u_{ij} = 1$ si la respuesta es afirmativa y $u_{ij} = 0$ si es negativa. Hay en total 2^n vectores o patrones de respuesta diferentes posibles con n ítems dicotómicos. Si cada sujeto posee un nivel de actitud subyacente desconocido θ_i, el análisis de su patrón de respuestas observado dará información acerca de ese nivel. Por el modelo determinístico general, la probabilidad de afirmar un ítem depende del nivel de actitud del sujeto $[P(u_{ij} = 1|\theta_i)]$.

Guttman (1950) definió el concepto de *ítem perfecto*, *ítem-Guttman* o *ítem-G* como aquel que a lo largo del continuo de actitud tiene asociado un valor crítico desconocido δ_j, tal que para valores $\theta_i \geq \delta_j \Rightarrow P(u_{ij} = 1|\theta_i) = 1$ y para $\theta_i < \delta_j \Rightarrow P(u_{ij} = 1|\theta_i) = 0$. A cada ítem j le corresponde un δ_j al que se denomina *dificultad del ítem*, un factor condicionante de la respuesta del sujeto, el cual provoca que la probabilidad de afirmar el ítem también se subordine a la dificultad del mismo y por ello:

$$\begin{aligned} \text{si } \theta_i \geq \delta_j &\Rightarrow P(u_{ij} = 1|\theta_i, \delta_j) = 1 \\ \text{y si } \theta_i < \delta_j &\Rightarrow P(u_{ij} = 1|\theta_i, \delta_j) = 0 \end{aligned}$$

Cuando n ítems perfectos o ítems-G se ordenan en dificultad creciente ($\delta_j < \delta_k \, \forall j, k \in n$) dan origen a una *escala perfecta* y, en consecuencia, a $n+1$ *patrones*

de respuesta perfectos de los 2^n patrones posibles; además, en ella, los sujetos están ordenados por la cantidad de 1s que presentan sus patrones desde aquel que más 1s tiene al que menos. Un ejemplo de una escala perfecta para una muestra de 15 sujetos que responde a 10 ítems se ilustra en la Tabla 2.1.

Tabla 2.1. Ordenación de ítems y sujetos en una escala perfecta

	Item										
Sujeto	1	2	3	4	5	6	7	8	9	10	$X_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10
2	1	1	1	1	1	1	1	1	1	0	9
3	1	1	1	1	1	1	1	1	0	0	8
4	1	1	1	1	1	1	0	0	0	0	6
5	1	1	1	1	1	1	0	0	0	0	6
6	1	1	1	1	1	1	0	0	0	0	6
7	1	1	1	1	1	0	0	0	0	0	5
8	1	1	1	1	1	0	0	0	0	0	5
9	1	1	1	1	1	0	0	0	0	0	5
10	1	1	1	1	1	0	0	0	0	0	5
11	1	1	1	1	0	0	0	0	0	0	4
12	1	1	1	1	0	0	0	0	0	0	4
13	1	1	1	0	0	0	0	0	0	0	3
14	1	1	0	0	0	0	0	0	0	0	2
15	1	0	0	0	0	0	0	0	0	0	1
$X_{.j}$	15	14	13	12	10	6	3	3	2	1	79
$p_{.j}$	1	0.93	0.87	0.80	0.67	0.40	0.20	0.20	0.13	0.07	

En una muestra de N sujetos cuyos niveles de actitud están ordenados según una escala de intervalo, el valor observado $p_{.j}$ es la *facilidad* o *popularidad* del ítem y su expresión es:

$$p_{.j} = \frac{X_{.j}}{N}$$

donde $X_{.j}$ es el número de sujetos que han afirmado el ítem. Por la terminología empleada, δ_j y $p_{.j}$ podrían generar cierta confusión, pero hay una diferencia entre ellos: mientras que la dificultad del ítem δ_j es independiente de la distribución de la población sobre el eje de actitud, la facilidad del ítem $p_{.j}$ depende de la distribución de la población para dicha actitud y, en general, para dos ítems j y k ordenados según $\delta_j < \delta_k$ en la escala perfecta de Guttman, se verifica que $p_{.j} > p_{.k}$.

El modelo de Guttman es un modelo determinístico que, por definición, no refleja la presencia de variables aleatorias. Así, los sujetos que tengan el

mismo patrón de respuestas es porque tienen el mismo θ y pertenecen a la misma categoría dentro de la escala perfecta siempre que su patrón sea uno de los $n+1$ que ésta contempla. Si un sujeto de una determinada categoría de θ presenta un patrón de respuestas inclasificable en la escala Guttman [e.g., (1110100)], sería un patrón atípico.

El escalograma de actitudes se ha extrapolado al ámbito de las habilidades, del rendimiento, de la evaluación de conocimientos... En estos contextos, el término de facilidad o popularidad del ítem se entendería como el valor observado que permite estimar la proporción de respuestas correctas de cada uno de los sujetos a un ítem del test, $\hat{\pi}_j = X_{.j}/N$. La fórmula general de los estadísticos para la detección de patrones atípicos por contraste con los patrones de un grupo normativo es:

$$G_i \equiv \frac{\sum_{j=1}^{X_{.j}} w_j - \sum_{j=1}^{n} X_{.j} w_j}{\sum_{j=1}^{X_{.j}} w_j - \sum_{j=n-X_{.j}+1}^{k} w_j} \tag{2.1}$$

donde $X_{.j} = \sum_{i=1}^{N} u_{ij}$ es el número de respuestas correctas al ítem y w_j es el valor que pondera al ítem para restringir el rango de valores del estadístico con el fin de facilitar la toma de decisiones así como la interpretación del mismo. Para un modelo de respuesta al ítem en el que el parámetro de habilidad del sujeto (θ_i) es conocido y δ_j es el parámetro de dificultad del ítem medido en la misma escala que la habilidad, entonces:

$$\begin{aligned} \theta_i \geq \delta_j &\quad \Leftrightarrow \quad P_j(\theta_i) = 1 \\ \theta_i < \delta_j &\quad \Leftrightarrow \quad P_j(\theta_i) = 0 \end{aligned}$$

El planteamiento de Guttman (1944, 1950) ha sido el punto de partida de la elaboración de los siguientes estadísticos para identificar patrones atípicos de respuesta:

- La correlación biserial-personal de Donlon y Fischer (1968).
- El índice de precaución de Sato (1975).
- Los índices basados en el número de errores Guttman de Meijer (1994) y van der Flier (1977).
- El índice de precaución modificado de Harnisch y Linn (1981).
- El índice de conformidad con la norma de Tatsuoka y Tatsuoka (1982).
- El coeficiente de escalabilidad de Sijtsma y Meijer (1992).

2.1.1. El coeficiente de correlación biserial-personal de Donlon y Fischer

Este coeficiente valora la relación existente entre el grado de dificultad que tiene un sujeto para acertar un ítem y la que teóricamente debería tener por pertenecer a una determinada muestra.

Asumiendo que el rasgo subyacente al sujeto (θ_i) se distribuye según la ley normal, entonces las respuestas del sujeto a los ítems del test también se ajustarán a una distribución normal. Ordenados los ítems en dificultad creciente ($\delta_j < \delta_k$) –especificada a partir de las respuestas a los ítems de la muestra normativa–, se fija un punto de corte que divide el continuo de los ítems en aquellos que son más fáciles para el sujeto y los acierta, y los que son más difíciles y los falla. Bajo estos supuestos, el coeficiente de correlación biserial-personal de Donlon y Fischer (1968) se obtiene comparando el patrón de respuestas de un sujeto (formado por 0s y 1s) y los índices de dificultad de los ítems en el grupo normativo. El índice de dificultad de los ítems se normaliza con la escala Δ del Educational Testing Service (ETS; Angoff, 1982):

$$\hat{\Delta}_j = 4\hat{z}_j + 13$$

siendo $\hat{z}_j = \hat{z}_{1-p_{.j}}$ la transformación del índice de dificultad del ítem j en la muestra normativa. Entonces, la correlación biserial-personal es:

$$r_{bisper} = \frac{\bar{\Delta} - \bar{\Delta}_c}{s_{\hat{\Delta}}} \frac{p_{i.}}{h} \tag{2.2}$$

donde

$\bar{\Delta}$ es la media de los $\hat{\Delta}_j$;

$\bar{\Delta}_c$ la media de los $\hat{\Delta}_j$ acertados;

$s_{\hat{\Delta}}$ la desviación típica los $\hat{\Delta}_j$ del test;

$p_{i.}$ la proporción de ítems acertados por el sujeto;

h la altura de la curva normal estandarizada que le corresponde a $p_{.j}$.

Los valores altos de r_{bisper} indican que el patrón del sujeto no es atípico, es decir, está de acuerdo o es similar a cualquier otro patrón de respuestas de un sujeto de esa muestra con el mismo o aproximado valor de rasgo subyacente. Si r_{bisper} es negativo, informaría de que el patrón de respuestas en cuestión

Tabla 2.2. Tabla S-P

Sujeto	Item 1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$p_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10	1.0
2	1	1	1	1	1	1	1	1	1	0	9	0.9
3	1	1	1	1	1	0	1	1	0	1	8	0.8
4	1	0	1	1	1	1	0	1	0	0	6	0.6
5	1	1	1	1	0	1	0	0	1	0	6	0.6
6	1	1	1	0	1	0	1	0	1	0	6	0.6
7	1	1	1	1	0	0	1	0	0	0	5	0.5
8	1	1	1	0	1	1	0	0	0	0	5	0.5
9	1	0	0	1	0	1	0	1	1	0	5	0.5
10	1	1	0	1	0	0	1	0	0	1	5	0.5
11	0	1	1	1	1	0	0	0	0	0	4	0.4
12	1	0	0	0	1	1	0	0	0	1	4	0.4
13	1	1	0	0	0	1	0	0	0	0	3	0.3
14	1	0	1	0	0	0	0	0	0	0	2	0.2
15	0	1	0	0	0	0	0	0	0	0	1	0.1
$X_{.j}$	13	11	10	9	8	8	6	5	5	4	$x_{..} = 79$	
$p_{.j}$	0.87	0.73	0.67	0.60	0.53	0.53	0.40	0.33	0.33	0.27	$p_{..} = 0.527$	

— Curva-S
- - Curva-P

está inversamente relacionado con la dificultad de los ítems en la muestra normativa y se calificaría de patrón atípico.

2.1.2. El índice de precaución de Sato

El índice de precaución de Sato (1975; Tatsuoka y Linn, 1983) es aplicable tanto a los ítems del test como a los sujetos de la muestra. Para su cálculo es necesario construir una matriz $N \times n$ de respuestas (u_{ij}) de N sujetos de la muestra a n ítems dicotómicos del test. Esta matriz se denomina *tabla S-P*. En sus filas se colocan los sujetos ordenados de modo descendente en función de su puntuación total en el test, i.e., se ordenan de arriba hacia abajo comenzando por el sujeto con la puntuación más alta. En las columnas aparecen los ítems dispuestos en dificultad creciente de acuerdo con el escalograma de Guttman, a la izquierda se situaría el ítem más fácil de todos los que componen el test. Sea una muestra de 15 sujetos que contestan un test de 10 ítems dicotómicos (Tatsuoka y Linn, 1983), cuyas respuestas han sido recogidas en una tabla conforme a la ordenación de filas y columnas indicada por Sato (Tabla 2.2).

Tabla 2.2. Tabla S-P

	Item											
Sujeto	1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$p_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10	1.0
2	1	1	1	1	1	1	1	1	1	0	9	0.9
3	1	1	1	1	1	0	1	1	0	1	8	0.8
4	1	0	1	1	1	1	0	1	0	0	6	0.6
5	1	1	1	1	0	1	0	0	1	0	6	0.6
6	1	1	1	0	1	0	1	0	1	0	6	0.6
7	1	1	1	1	0	0	1	0	0	0	5	0.5
8	1	1	1	0	1	1	0	0	0	0	5	0.5
9	1	0	0	1	0	1	0	1	1	0	5	0.5
10	1	1	0	1	0	0	1	0	0	1	5	0.5
11	0	1	1	1	1	0	0	0	0	0	4	0.4
12	1	0	0	0	1	1	0	0	0	1	4	0.4
13	1	1	0	0	0	1	0	0	0	0	3	0.3
14	1	0	1	0	0	0	0	0	0	0	2	0.2
15	0	1	0	0	0	0	0	0	0	0	1	0.1
$X_{.j}$	13	11	10	9	8	8	6	5	5	4	$x_{..} = 79$	
$p_{.j}$	0.87	0.73	0.67	0.60	0.53	0.53	0.40	0.33	0.33	0.27	$p_{..} = 0{,}527$	

— Curva-S
- - Curva-P

Esta tabla contempla los siguientes datos:

$X_{i.}$ es el número de ítems que acierta el sujeto i;

$p_{i.}$ la proporción de respuestas correctas del sujeto i;

$X_{.j}$ el número de sujetos que aciertan el ítem j;

$p_{.j}$ la proporción de sujetos que aciertan el ítem j;

$x_{..}$ la puntuación total en el test;

$p_{..}$ la proporción de respuestas correctas total.

A partir de esta matriz se trazan dos curvas:

1. La curva del sujeto o curva-S (*student curve* o *S-curve*) es una línea que comienza a la derecha de la n-ésima celdilla de la matriz del primer sujeto, siendo ésta la que coincide con el número de aciertos del sujeto ($X_{i.}$). Al ir conectando las sucesivas celdillas de cada uno de los sujetos, se obtiene una curva con la forma de la función de distribución de ogiva. Sato (1975) definió la *curva-S perfecta* como aquella que, manteniendo la curva-S observada invariante, se obtiene al permutar los 0s de la izquierda de la curva por 1s y los 1s de la derecha de la misma por 0s (Tabla 2.3). En la matriz de la nueva tabla están las respuestas modificadas de los sujetos a los ítems del test, denotadas M^S_{ij}. Las puntuaciones totales iniciales de

Tabla 2.3. Curva-S perfecta

	Item											
Sujeto	1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$M_{i.}^{S}$
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	0	9	9
3	1	1	1	1	1	1	1	1	0	0	8	8
4	1	1	1	1	1	1	0	0	0	0	6	6
5	1	1	1	1	1	1	0	0	0	0	6	6
6	1	1	1	1	1	1	0	0	0	0	6	6
7	1	1	1	1	1	0	0	0	0	0	5	5
8	1	1	1	1	1	0	0	0	0	0	5	5
9	1	1	1	1	1	0	0	0	0	0	5	5
10	1	1	1	1	1	0	0	0	0	0	5	5
11	1	1	1	1	0	0	0	0	0	0	4	4
12	1	1	1	1	0	0	0	0	0	0	4	4
13	1	1	1	0	0	0	0	0	0	0	3	3
14	1	1	0	0	0	0	0	0	0	0	2	2
15	1	0	0	0	0	0	0	0	0	0	1	1
$X_{.j}$	13	11	10	9	8	8	6	5	5	4	79	79
$M_{.j}^{P}$	15	14	13	12	10	6	3	3	2	1		

los sujetos en el test ($X_{i.}$) y la puntuación total tras la transformación en la curva-S perfecta ($M_{i.}^{S}$) no varían; sin embargo, sí cambia el número de respuestas acertadas por ítem debido a dicha reestructuración de la matriz ($X_{.j} \neq M_{.j}^{P}$).

Tabla 2.3. Curva-S perfecta

	Item											
Sujeto	1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$M_{i.}^{S}$
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	0	9	9
3	1	1	1	1	1	1	1	1	0	0	8	8
4	1	1	1	1	1	1	0	0	0	0	6	6
5	1	1	1	1	1	1	0	0	0	0	6	6
6	1	1	1	1	1	1	0	0	0	0	6	6
7	1	1	1	1	1	0	0	0	0	0	5	5
8	1	1	1	1	1	0	0	0	0	0	5	5
9	1	1	1	1	1	0	0	0	0	0	5	5
10	1	1	1	1	1	0	0	0	0	0	5	5
11	1	1	1	1	0	0	0	0	0	0	4	4
12	1	1	1	1	0	0	0	0	0	0	4	4
13	1	1	1	0	0	0	0	0	0	0	3	3
14	1	1	0	0	0	0	0	0	0	0	2	2
15	1	0	0	0	0	0	0	0	0	0	1	1
$X_{.j}$	13	11	10	9	8	8	6	5	5	4	79	79
$M_{.j}^{P}$	15	14	13	12	10	6	3	3	2	1		

Tabla 2.4. Curva-P perfecta

Sujeto	Item 1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$M_{i.}^{S}$
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	1	9	10
3	1	1	1	1	1	1	1	1	1	1	8	10
4	1	1	1	1	1	1	1	1	1	1	6	10
5	1	1	1	1	1	1	1	1	1	0	6	9
6	1	1	1	1	1	1	1	0	0	0	6	7
7	1	1	1	1	1	1	0	0	0	0	5	6
8	1	1	1	1	1	1	0	0	0	0	5	6
9	1	1	1	1	0	0	0	0	0	0	5	4
10	1	1	1	0	0	0	0	0	0	0	5	3
11	1	1	0	0	0	0	0	0	0	0	4	2
12	1	0	0	0	0	0	0	0	0	0	4	1
13	1	0	0	0	0	0	0	0	0	0	3	1
14	0	0	0	0	0	0	0	0	0	0	2	0
15	0	0	0	0	0	0	0	0	0	0	1	0
$X_{.j}$	13	11	10	9	8	8	6	5	5	4	79	
$M_{.j}^{P}$	13	11	10	9	8	8	6	5	5	4	79	

2. La curva problema o curva-P (*problem curve* o *P-curve*) se construye por columnas comenzando por el primer ítem de la tabla. El origen se encuentra en la n-ésima celdilla en la que se igualan el número de sujetos y el número de aciertos en el ítem ($X_{.j}$). De modo similar, el autor definió la *curva-P perfecta* como la resultante del cambio de los 1s que están por debajo de la curva-P observada por 0s y los 0s que están por encima de la misma por 1s (Tabla 2.4). La notación de las respuestas modificadas a los ítems es M_{ij}^{P}. La transformación altera las puntuaciones totales iniciales de los sujetos ($X_{i.} \neq M_{i.}^{S}$), mientras que deja invariante el número de respuestas por ítem ($X_{.j} = M_{.j}^{P}$).

Tabla 2.4. Curva-P perfecta												
	Item											
Sujeto	1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$M_{i.}^{S}$
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	1	9	10
3	1	1	1	1	1	1	1	1	1	1	8	10
4	1	1	1	1	1	1	1	1	1	1	6	10
5	1	1	1	1	1	1	1	1	1	0	6	9
6	1	1	1	1	1	1	1	0	0	0	6	7
7	1	1	1	1	1	1	0	0	0	0	5	6
8	1	1	1	1	1	1	0	0	0	0	5	6
9	1	1	1	1	0	0	0	0	0	0	5	4
10	1	1	1	0	0	0	0	0	0	0	5	3
11	1	1	0	0	0	0	0	0	0	0	4	2
12	1	0	0	0	0	0	0	0	0	0	4	1
13	1	0	0	0	0	0	0	0	0	0	3	1
14	0	0	0	0	0	0	0	0	0	0	2	0
15	0	0	0	0	0	0	0	0	0	0	1	0
$X_{.j}$	13	11	10	9	8	8	6	5	5	4	79	
$M_{.j}^{P}$	13	11	10	9	8	8	6	5	5	4	79	

El *índice de precaución para el sujeto i* (C_i) se calcula a partir del análisis de las tablas que contemplan la curva-S observada y la curva-S perfecta. Su expresión:

$$C_i = \frac{\sum_{j=1}^{n}(u_{ij} - p_{i.})(X_{.j} - p_{..})}{\sum_{j=1}^{n}(M_{ij}^{S} - p_{i.})(X_{.j} - p_{..})} \tag{2.3}$$

donde M_{ij}^{S} es la respuesta del sujeto i al ítem j por la curva-S perfecta. Si $C_i = 0$ el patrón está de acuerdo con el patrón Guttman; valores elevados de C_i indican que el patrón es atípico. Pero C_i no tiene límite superior en su rango de valores, lo cual dificulta su interpretación al no poder decidir apropiadamente acerca de si el patrón es atípico o no. Por ello, Harnisch (1983) sugirió que se debería considerar a un patrón como atípico si $C_i > 0{,}60$.

En el caso de que se quisiera estudiar el ítem j, el índice de precaución del ítem, C_j, se obtiene a partir de las tablas que contienen la curva-P observada y la curva-P perfecta:

$$C_j = \frac{\sum_{i=1}^{N}(u_{ij} - p_{.j})(X_{i.} - p_{..})}{\sum_{i=1}^{N}(M_{ij}^{P} - p_{.j})(X_{i.} - p_{..})} \tag{2.4}$$

donde M_{ij}^{P} es la respuesta del sujeto i al ítem j por la curva-P perfecta. El índice C_j coincide con la razón del coeficiente de correlación ítem-test o índice

de discriminación observado (r_j) y el coeficiente de correlación ítem-test estandarizado o índice de discriminación por la curva perfecta (r'_j), ya que el numerador de C_j es la covarianza entre el vector columna que contiene las respuestas de los sujetos al ítem (u_{ij}) y el total de aciertos del sujeto ($X_{i.}$), y el denominador es la covarianza entre el vector suma de las respuestas de cada sujeto a los n ítems ($X_{i.}$) y cada una de las respuestas a los ítems por la curva-P perfecta (M_{ij}^P). La equivalencia entre ambos índices es:

$$C_j = \frac{Cov(u_{ij}, X_{i.})}{Cov(M_{ij}^P, X_{i.})} = \frac{\frac{Cov(u_{ij}, X_{i.})}{\sigma_j(u_{ij})\sigma_j(X_{i.})}}{\frac{Cov(M_{ij}^P, X_{i.})}{\sigma_j(M_{ij}^P)\sigma_j(X_{i.})}} = \frac{r_j}{r'_j}$$

2.1.3. Los índices basados en el número de errores Guttman

Guttman (1950, p. 70) definió el número de errores de un patrón según el modelo determinístico como el número de respuesta erróneamente pronosticadas para un sujeto a partir de su puntuación. Van der Flier (1977) elaboró un índice muy sencillo de calcular para detectar patrones atípicos a partir de una tabla S-P. Con este índice, denotado U_{1i}^*, se obtendría la desviación del patrón de respuestas del sujeto i respecto del patrón de la curva-S perfecta:

$$U_{1i}^* = \frac{U_i}{X_{i.}(n - X_{i.})} \tag{2.5}$$

donde U_i es el resultado de añadir un 1 a la derecha de cada 0 del patrón observado hasta el último acierto y sumar esa cantidad de 1s añadidos a los 1s que aparecen al derecha de la curva-S observada, valor que varía entre 0 y $X_{i.}(n - X_{i.})$, número máximo de errores Guttman correspondientes al patrón del sujeto i; $X_{i.}$ es la puntuación total del sujeto i y n el número de ítems. Al dividir U_i por $X_{i.}(n - X_{i.})$, lo cual estrecha el rango del índice U_{1i}^* a $(0, 1)$, se pueden comparar patrones con diferente puntuación total y determinar si es un patrón atípico o no; cuanto más se aleje U_{1i}^* de 0, más atípico será el patrón.

Meijer (1994) modificó el índice U_{1i}^* y sugirió que el uso del número de errores Guttman podría ser un buen índice para detectar patrones atípicos, siempre y cuando éste no se dejara influir ni por el número de ítems del test ni por el número de aciertos del sujeto en el test. Para ello, dividió el número de errores Guttman de un patrón (G) por el número máximo de errores que le corresponde al número de respuestas correctas de ese patrón; la ecuación de este estadístico denotado G_i^* es:

$$G_i^* = \frac{G}{X_{i.}(n - X_{i.})} = \frac{\sum_{j=1}^{n-1} \sum_{k=j+1}^{n} f_{jk}}{X_{i.}(n - X_{i.})} \tag{2.6}$$

donde

j y k son los índices de los ítems, tal que $p_{.j} \geq p_{.k}$ ($j = 1, 2, \ldots, n-1$ y $k = j+1, \ldots, n$);

n es el número de ítems del test;

$f_{jk} = 1$ si el sujeto tiene un error Guttman en los ítems j y k ($u_{ij} = 0$ y $u_{ik} = 1$);

$f_{jk} = 0$ en cualquier otro caso.

Si $G_i^* = 0$, el patrón es un vector Guttman y cuanto más se aleje de este valor más atípico será. Meijer (1995) aclaró que, aunque se emplee el número de errores Guttman, se adoptó la definición de error dada por Loevinger (1947, 1948), la cual contempla una versión probabilística del modelo determinístico de Guttman (1944, 1950). Por esta definición, el número de errores se contabiliza por pares de ítems; por ejemplo, sea el patrón de respuestas (01011), por la definición de Guttman se detectarían cuatro errores, mientras que por la de Loevinger se identificarían cinco errores.

2.1.4. El índice de precaución modificado de Harnisch y Linn

Harnisch y Linn (1981) hicieron una modificación del índice de precaución C_i de Sato (1975), para solventar las dificultades de interpretación que éste presenta cuando su magnitud es elevada. El *índice de precaución modificado* (C_i^*) para el sujeto i se calcula con la siguiente expresión (la nomenclatura se corresponde con la descrita en las secciones precedentes):

$$C_i^* = \frac{\sum_{j=1}^{X_{i.}} (1 - u_{ij}) X_{.j} - \sum_{j=X_{i.}+1}^{n} u_{ij} X_{.j}}{\sum_{j=1}^{X_{i.}} X_{.j} - \sum_{j=n-X_{i.}+1}^{n} X_{.j}} \tag{2.7}$$

El índice C_i^* toma valores entre 0 y 1, facilitando así su interpretación; si $C_i^* = 0$ el sujeto i tiene un patrón de respuestas perfecto y cuanto más se aleje de este valor cuanto más atípico será, llegando al máximo de atipicidad si $C_i^* = 1$, lo cual indicaría que el sujeto presenta un patrón Guttman inverso. El punto crítico para clasificar a un patrón como atípico es $C_i^* > 0{,}30$.

2.1.5. El índice de conformidad con la norma de Tatsuoka y Tatsuoka

Un aspecto a tener en cuenta en los tests que evalúan los procesos cognitivos empleados en la resolución de problemas es la *consistencia* de las respuestas a lo largo del tiempo. La importancia de la identificación de patrones inconsistentes no sólo radica en detectar sujetos que fallan los ítems por emplear procesos cognitivos erróneos, sino también en localizar a aquellos sujetos que, aunque responden correctamente a los ítems, no han utilizado las operaciones mentales adecuadas. La consecuencia directa que se deriva de estos últimos es el cambio de la didáctica para la resolución de problemas. Tatsuoka y Tatsuoka (1982) desarrollaron un índice para dicho propósito al que denominaron *índice de conformidad con la norma* (*Norm Conformity Index*, NCI), el cual valora el grado de aproximación entre el patrón de respuestas observado de un sujeto y el que le correspondería en la escala Guttman.

Para calcular NCI se requiere de una matriz $N \times n$ en la que N es el tamaño muestral y n el número de ítems del test ordenados en dificultad creciente ($\delta_j < \delta_k \, \forall j, k \in n$). El índice NCI confronta el patrón de respuestas de un sujeto y un patrón del grupo normativo ajustado a la escala Guttman con el mismo número de respuestas correctas que aquel. Este índice es un coeficiente de correlación entre la dificultad de los ítems ordenados según el grupo normativo y el patrón de respuestas del sujeto en estudio:

$$NCI_i = \frac{2 \sum_{j<k}^{n} u_{ij}}{\sum_{j=1}^{n} u_{ij}} - 1 \tag{2.8}$$

El orden de los ítems condiciona a NCI. Su intervalo de valores es $(-1, +1)$; si $NCI = +1$, el patrón de respuestas es un patrón ajustado a la escala Guttman; si $NCI = -1$ indica que el patrón es un vector de respuestas inverso o invertido al que le correspondería en dicha escala. Cuanto más se aleje NCI de $+1$, cuanto más atípico será el patrón. El primer sumando de la Ecuación 2.8 es equivalente al valor del índice U_1^* de van der Flier (1977), por lo que también se podría expresar NCI del siguiente modo:

$$NCI_i = 2U_{1i}^* - 1 \tag{2.9}$$

Además, NCI se puede obtener a partir del coeficiente gamma de Goodman-Kruskal (1954) mediante el cómputo del número de concordancias e inversiones que existen entre el orden de los índices de dificultad de los ítems en el grupo normativo, y las puntuaciones de 0s y 1s del patrón individual:

$$\gamma = \frac{f_c - f_i}{f_c + f_i}$$

donde f_c es el número de concordancias y f_i el número de inversiones.

2.1.6. El coeficiente de escalabilidad de Sijtsma y Meijer

Sijtsma y Meijer (1992) propusieron una extensión del *coeficiente de escalabilidad H* de Loevinger (1948) para identificar patrones atípicos a partir de una matriz $N \times n$ ordenada según el escalograma de Guttman (Tabla 2.5). A este estadístico lo denotaron H^T.

En la última columna de la tabla aparecen los valores $p_{i.} = X_{i.}/n$, estimadores de la proporción esperada de ítems que acierta el sujeto i a través de medidas repetidas localmente independientes. Para calcular H^T es necesario conocer la proporción esperada de ítems acertados por un sujeto i y por un sujeto g, esto es, $p_{ig.} = x_{ig}/n$ donde x_{ig} es el número de ítems acertados coincidentes de los sujetos i y g. Siguiendo el escalograma Guttman, para dos sujetos i y g se verifica que $p_{i.} \geq p_{g.}$. Si la covarianza de las respuestas de ambos sujetos a los n ítems del test se obtiene con la expresión:

$$\sigma_{ig} = p_{ig.} - p_{i.}p_{g.}$$

entonces el valor máximo de la covarianza se obtiene cuando $p_{ig.} = p_{g.}$ y, por lo tanto, $\sigma_{\text{máx},ig} = p_{g.}(1 - p_{i.})$.

El coeficiente H^T para comparar el patrón de un sujeto con el de otro de la misma muestra se define como:

$$H^T = \frac{\sum\sum_{i>g}\sigma_{ig}}{\sum\sum_{i>g}\sigma_{\text{máx},ig}} \tag{2.10}$$

Pero si lo que se pretende es comparar el patrón de un sujeto con el resto de la muestra, entonces el coeficiente H^T es:

$$H_i^T = \frac{\sum\sum_{i\neq g}\sigma_{ig}}{\sum\sum_{i\neq g}\sigma_{\text{máx},ig}} \tag{2.11}$$

Tanto H^T como H_i^T varían entre 0 y 1; si son iguales a 0 el patrón del sujeto es atípico; en el caso opuesto, cuando H^T o H_i^T valen 1, indicaría que

es un patrón Guttman perfecto. Sin embargo, Sijtsma (1986) observó que en presencia de un patrón de respuestas perfecto estos coeficientes no siempre son iguales a 1.

Tabla 2.5. Ordenación de ítems y sujetos en una escala perfecta

	Item											
Sujeto	1	2	3	4	5	6	7	8	9	10	$X_{i.}$	$p_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10	1.0
2	1	1	1	1	1	1	1	1	1	0	9	0.9
3	1	1	1	1	1	1	1	1	0	0	8	0.8
4	1	1	1	1	1	1	0	0	0	0	6	0.6
5	1	1	1	1	1	1	0	0	0	0	6	0.6
6	1	1	1	1	1	1	0	0	0	0	6	0.6
7	1	1	1	1	1	0	0	0	0	0	5	0.5
8	1	1	1	1	1	0	0	0	0	0	5	0.5
9	1	1	1	1	1	0	0	0	0	0	5	0.5
10	1	1	1	1	1	0	0	0	0	0	5	0.5
11	1	1	1	1	0	0	0	0	0	0	4	0.4
12	1	1	1	1	0	0	0	0	0	0	4	0.4
13	1	1	1	0	0	0	0	0	0	0	3	0.3
14	1	1	0	0	0	0	0	0	0	0	2	0.2
15	1	0	0	0	0	0	0	0	0	0	1	0.1
$X_{.j}$	15	14	13	12	10	6	3	3	2	1	79	
$p_{.j}$	1	0.93	0.87	0.80	0.67	0.40	0.20	0.20	0.13	0.07		

2.2. El estadístico g_2 de Frary, Tideman y Watts

A partir de los supuestos de la Teoría Clásica de Test (TCT), Frary, Tideman y Watts (1977) elaboraron el índice g_2 para identificar patrones con respuestas copiadas. Para ello, contrastaron el número observado y esperado de ítems con la misma respuesta de dos sujetos. Para cada par de sujetos, uno de ellos es el sujeto que copia (sujeto C) y el otro es el sujeto del que se copia la respuesta o sujeto *fuente* (F). Si las respuestas de éste se fijan, el número esperado de ítems con respuestas idénticas de los sujetos C y F es la suma de probabilidades:

$$E(h_{CF}|\mathbf{U}_F) = \sum_{j=1}^{n} P_C(u_{jF})$$

donde

u_{jF} es la respuesta del sujeto F al ítem j;

$P_C(u_{jF})$ la probabilidad de que C escoja la misma respuesta que F;

$\mathbf{U}_F$ el patrón de respuestas del sujeto F;

h_{CF} el número de respuestas iguales de C y F.

ya que las respuestas de los dos sujetos pueden ser iguales o no, el procedimiento de comparación de respuestas es un ensayo de Bernoulli. Por lo tanto, suponiendo que las covarianzas de los ítems son insignificantes, la varianza del número de ítems con idéntica respuesta es:

$$\sigma^2_{h_{CF}|\mathbf{U}_F} = \sum_{j=1}^{n} P_C(u_{jF})[1 - P_C(u_{jF})]$$

Entonces, el índice g_2 es la diferencia entre el número de respuestas idénticas observado y esperado, dividido por la desviación típica de la diferencia:

$$g_2 = \frac{h_{CF} - E(h_{CF}|\mathbf{U}_F)}{\sigma_{h_{CF}|\mathbf{U}_F}} \tag{2.12}$$

un índice que sigue una distribución normal. Para obtener $P_C(u_{jF})$, los autores implementaron un algoritmo que se desarrolla conociendo a priori: la dificultad de los ítems calculada según la TCT, las dificultades de los distractores –alternativas de respuesta incorrectas–, y la razón entre la puntuación de C –número de respuestas correctas– y la media de las puntuaciones en el test de todos los sujetos. Las ventajas de g_2 son las siguientes:

- Describe un modelo para calcular la probabilidad de que un sujeto elija cualquiera de las opciones de respuesta del ítem, aunque carezca de justificación teórica y se sirva de la puntuación del sujeto en el test.
- Utiliza la información de todos los ítems del test.
- Tiene una prueba estadística con distribución conocida.
- Puede ser calculado con cualquier tamaño muestral.

No obstante, Frary *et al.* (1977) afirmaron que la distribución de g_2 podría alejarse de la normalidad si los algoritmos utilizados para calcular $P_C(u_{jF})$ no son los adecuados. Las ecuaciones inmersas en él dependen de, por un lado, si la puntuación en el test de C es mayor o menor a la media de todas las puntuaciones y, por otro lado, si el sujeto F ha acertado o fallado el ítem. Pero los principales inconvenientes de g_2 son los supuestos que requiere para poder ser aplicado, éstos son:

- Asumir que la discriminación de los ítems es constante para todos los sujetos, basándose en las formas de las rectas de regresión ítem-test.
- La probabilidad de escoger un distractor es constante e independiente del nivel de habilidad del sujeto.
- Al no tener en cuenta el nivel de habilidad, g_2 infraestima la correlación biserial-puntual de cada una de las alternativas del ítem.
- Las formas de las rectas de regresión dependen tanto del ítem que se está analizando como de las características de todos los ítems del test, por lo tanto, dichas formas varían en función de los parámetros del modelo de respuesta.

2.3. Estadísticos basados en modelos de respuesta al ítem no paramétricos: los índices U_3 y Z_{U_3} de van der Flier

Los modelos de respuesta al ítem no paramétricos relajan algunas de las restricciones supuestas para el empleo de los modelos de respuesta paramétricos. Sea una muestra de N sujetos y un test de n ítems, donde θ_i es el valor de la variable aleatoria de habilidad del sujeto i ($i = 1, 2, \ldots, N$), δ_j es el valor de dificultad del ítem j ($j = 1, 2, \ldots, n$) y $P(\theta_i, \delta_j) \equiv P_j$ es la probabilidad de acertar el ítem j por un sujeto de habilidad θ_i; los principios generales de los modelos no paramétricos son:

- Las funciones de respuesta al ítem no están definidas paramétricamente.
- La variable aleatoria θ está medida en una escala ordinal.
- No se asume que la variable θ siga una distribución a priori.

Van der Flier (1980, 1982) desarrolló el índice U_3 para probar el ajuste de personas dentro del campo de los modelos de medida no paramétricos. El punto de referencia del autor fue el Modelo de Homogeneidad Monótona (MHM) de Mokken (1971), el cual incorpora algunos supuestos más a los tres anteriores:

- El test es unidimensional.
- Se cumple la independencia local de los ítems.

- El rasgo o habilidad θ es una variable monótona, de manera que a mayor valor de θ mayor es la probabilidad de acertar el ítem y, en consecuencia, si $\theta_i < \theta_{i'} \Rightarrow P(\theta_i, \delta_j) \leq P(\theta_{i'}, \delta_j) \, \forall i \neq i' \in N$.
- $0 \leq P(\theta, \delta) \leq 1$.
- La función de respuesta al ítem $P(\theta, \delta)$ es monótona creciente a lo largo del continuo de θ.

El MHM permite ordenar a los sujetos con respecto a los valores de θ y, por consiguiente, el ordenamiento por dificultad de los ítems depende de θ y varía con ella, lo cual pone de manifiesto que la jerarquía de los ítems varía con las muestras de una población. Sin embargo, el modelo no permite estimaciones de los valores de habilidad, por lo que para ordenar a los sujetos se recurre a las puntuaciones verdaderas de los mismos, i.e., a las puntuaciones observadas en el test ($X_{i.} = \sum_{j=1}^{n} u_{ij}$), las cuales son estimaciones de θ y mantienen el orden de las puntuaciones observadas.

El índice U_3 toma como criterio a un grupo normativo para identificar patrones atípicos, el cual es el punto de referencia para comparar un patrón de respuestas observado con puntuación $X_{i.}$, con el patrón esperado en dicho grupo con la misma puntuación. Su expresión es la siguiente:

$$U_3 = \frac{\sum_{j=1}^{X_{i.}} \lg\left(\frac{P_j}{1-P_j}\right) - \sum_{j=1}^{n} u_{ij} \lg\left(\frac{P_j}{1-P_j}\right)}{\sum_{j=1}^{X_{i.}} \lg\left(\frac{P_j}{1-P_j}\right) - \sum_{j=n-X_{i.}+1}^{n} \lg\left(\frac{P_j}{1-P_j}\right)} \tag{2.13}$$

El rango de valores de U_3 oscila entre 0 y 1, de modo que si $U_3 = 0$ indica que el patrón de respuestas es un patrón Guttman perfecto y si $U_3 = 1$ el patrón de respuestas es el patrón inverso al que le correspondería en el escalograma. Cuanto más se aleja U_3 de 0, más se aleja el patrón de ser un patrón Guttman perfecto y más atípico resulta. Pero U_3 depende de los valores de habilidad, por lo que el resultado podría llevar a errores en su interpretación. Para eliminar esta supeditación, van der Flier (1982) lo estandarizó y logró otro estadístico, denotado Z_{U_3}, distribuido según la ley normal:

$$Z_{U_3} = \frac{U_3 - E(U_3)}{Var(U_3)^{1/2}} \tag{2.14}$$

en donde

$$E(U_3|X_{i.}) = \frac{\sum_{j=1}^{X_{i.}} \lg\left(\frac{P_j}{1-P_j}\right) - \eta}{\sum_{j=1}^{X_{i.}} \lg\left(\frac{P_j}{1-P_j}\right) - \sum_{j=n-X_{i.}+1}^{n} \lg\left(\frac{P_j}{1-P_j}\right)}$$

$$\eta = \sum_{j=1}^{n} P_j \lg\left(\frac{P_j}{1-P_j}\right) + \frac{\sum_{j=1}^{n} P_j(1-P_j)\lg\left(\frac{P_j}{1-P_j}\right)}{\sum_{j=1}^{n} P_j(1-P_j)}\left(X_{i.} - \sum_{j=1}^{n} P_j\right)$$

$$Var(U_3|X_{i.}) = \frac{\beta}{\sum_{j=1}^{X_{i.}} \lg\left(\frac{P_j}{1-P_j}\right) - \sum_{j=n-X_{i.}+1}^{n} \lg\left(\frac{P_j}{1-P_j}\right)}$$

$$\beta = \sum_{j=1}^{n} P_j(1-P_j)\left[\lg\left(\frac{P_j}{1-P_j}\right)\right]^2 - \frac{\left[\sum_{j=1}^{n} P_j(1-P_j)\lg\left(\frac{P_j}{1-P_j}\right)\right]^2}{\sum_{j=1}^{n} P_j(1-P_j)}$$

2.4. Estadísticos desarrollados dentro de la Teoría de la Generalizabilidad (TG): los índices de acuerdo y desacuerdo de Kane y Brennan

La Teoría de la Generalizabilidad (TG) fue elaborada por Cronbach, Gleser, Nanda y Rajaratnam (1972) para resolver las limitaciones de medida de la TCT. En la TG se retoman los conceptos de tests paralelos, errores de medida y coeficiente de fiabilidad para definirlos de nuevo y eliminar las restricciones que, en la práctica, suponían un problema para emplear la TCT cuando se utilizaban tests referidos al criterio y tests de dominio o maestría, ya que éstos estudian al sujeto y las conclusiones se elaboran respecto a él y no respecto a un grupo de referencia estandarizado.

El modelo general de la TG descansa en la formulación del Análisis de Varianza (ANOVA):

$$X_{ij} = \mu + \pi_i + \beta_j + (\pi\beta, e)_{ij,e}$$

donde

X_{ij} es la puntuación observada del sujeto i en el ítem j;

μ la media total;

$\pi_i = (\mu_i - \mu)$ la variable aleatoria del efecto de sujeto y se distribuye normalmente $[N(0, \sigma_i)]$;

$\beta_j = (\mu_j - \mu)$ la variable aleatoria del efecto del ítem y sigue una distribución normal $N(0, \sigma_j)$;

$(\pi\beta, e) = (X_{ij} - \mu_i - \mu_j + \mu)$ la variable aleatoria del efecto de la interacción S$ujeto \times$ Item, que se confunde con el error en los diseños ANOVA de medida única y se distribuye normalmente $[N(0, \sigma_e)]$.

Kane y Brennan (1980) desarrollaron tres medidas de fiabilidad para tests referidos al dominio o tests de maestría. Para evitar la confusión con el concepto de fiabilidad empleado en la TCT, los autores hacen referencia al término *seguridad (dependability)*[1]. Aplicando los conceptos de la TG, sea un test de maestría (J) compuesto por n ítems escogidos de un universo de ítems y contestado por una muestra de N sujetos seleccionados de una población. La puntuación del sujeto i a los n ítems del test J es, según el modelo lineal general:

$$X_{iJ} = \mu + S_i + \beta_J + (S\beta, e)_{iJ}$$

donde

X_{iJ} es la puntuación media observada del sujeto i en el test J;

μ la media de las puntuaciones de la población en el universo de ítems;

S_i el efecto del sujeto i;

β_J el efecto del test J;

$(S\beta, e)_{iJ}$ el efecto de la interacción del sujeto con el test, confundido con el error residual e.

Todos los efectos del modelo son aleatoriamente independientes.

Sea otro test de maestría (K) formado por n' ítems extraídos del mismo universo de ítems que el anterior para medir el mismo dominio o nivel de maestría. Los tests J y K son aleatoriamente paralelos porque han sido construidos desde el mismo universo, son independientes, tienen igual número de ítems $(n = n')$ e igual puntuación de corte (λ) para evaluar la maestría. Entonces, la *función de acuerdo referida al dominio* del sujeto i es:

$$A(S_{iJ}, S_{iK}) = A_i = (S_{iJ} - \lambda)(S_{iK} - \lambda)$$

[1]Este índice ha sido traducido por otros autores como *índice de dependencia-interdependencia* (Blanco, 1989) o *coeficiente de generalizabilidad* (Martínez, 1995)

La función A_i es positiva si en ambos tests se clasifica al sujeto maestro o no maestro, es decir, si la desviación de la puntuación media del sujeto en ambos tests es grande y en el mismo sentido; A_i es negativa si en ambos tests la clasificación es diferente, i.e., existe desacuerdo en las puntuaciones medias del sujeto y, en consecuencia, las desviaciones son grandes y de sentido contrario. Si los recesos de la puntuación de corte son próximos a 0, A_i es próxima a 0 y se clasifica al sujeto de caso *borderline*.

A partir de la función de acuerdo esperado (A_i) se define el *índice de seguridad* ϑ, el cual evalúa el grado de concordancia esperado entre dos puntuaciones medias cualesquiera de la población en dos tests aleatoriamente paralelos e independientes. La expresión de la *función de acuerdo esperado* es:

$$\begin{aligned} A_i &= E_{i,J,K}[(S_{iJ} - \lambda)(S_{iK} - \lambda)] \\ &= E[(\mu - \lambda)^2] + E[S_i^2] + E[\beta_J \beta_K] + E[(S\beta, e)_{iJ}(S\beta, e)_{iK}] \end{aligned}$$

Como los dos tests son independientes:

$$\begin{aligned} E[\beta_J \beta_K] &= 0 \\ E[(S\beta, e)_{iJ}(S\beta, e)_{iK}] &= 0 \end{aligned}$$

y por definición, $E[S_i^2] = \sigma^2(S)$, resultando la función de acuerdo esperado:

$$A_i = (\mu - \lambda)^2 + \sigma^2(S) \tag{2.15}$$

La *función de acuerdo máximo esperado* mide el grado de acuerdo de la puntuación media de un sujeto en un test consigo misma:

$$\begin{aligned} A_{\text{máx},i} &= E_{i,J}[(S_{iJ} - \lambda)^2] \\ &= (\mu - \lambda)^2 + \sigma^2(S) + \frac{\sigma^2(\beta_J)}{n} + \frac{\sigma^2(S\beta, e)_{iJ}}{n} \end{aligned} \tag{2.16}$$

El objetivo de los tests de maestría es hacer clasificaciones exactas de los sujetos en un dominio. Para valorar la exactitud de la clasificación o de la decisión tomada acerca del sujeto se calcula el *índice de seguridad sobre decisiones de maestría*, cuya formulación es:

$$\vartheta_i = \frac{A_i}{A_{\text{máx},i}}$$

$$= \frac{(\mu-\lambda)^2+\sigma^2(S)}{(\mu-\lambda)^2+\sigma^2(S)+\frac{\sigma^2(\beta_J)}{n}+\frac{\sigma^2(S\beta,e)_{iJ}}{n}} \tag{2.17}$$

Cuanto mayor sea la desviación de la población frente a la puntuación de corte $[(\mu-\lambda)^2]$ mayor será ϑ, incluso cuando todos los sujetos tienen la misma puntuación en el universo de ítems, con lo cual $\sigma^2(S)=0$ y $(\mu-\lambda)^2$ cuantificaría la validez de la clasificación. Entonces, cuanto mayor sea $(\mu-\lambda)^2$ más fácil será tomar una decisión para clasificar a los sujetos con total seguridad, aunque el test no aporte información fiable sobre las diferencias individuales presentes entre las puntuaciones de la población (Brennan y Kane, 1977).

Kane y Brennan (1980) consideraron que también era importante evaluar el grado de coherencia de un sujeto para contestar un test de maestría, ya que cuando éstos se emplean no interesa la posición relativa de un sujeto con respecto al resto de la población, sino la puntuación del sujeto en términos absolutos. Por definición, la puntuación universo de un sujeto es:

$$\mu_i = E_J[S_{iJ}] = \mu + S_i$$

y el error asociado a ella es:

$$\varrho_i = (S_{iJ}-\lambda)-(\mu_i-\lambda) = \beta_J + (S\beta,e)_{iJ}$$

entonces, la varianza de ϱ_i es la *función de desacuerdo esperado* (L_i), que se obtiene sustrayendo del acuerdo máximo esperado el acuerdo esperado:

$$L_i = A_{\text{máx},i} - A_i \tag{2.18}$$

$$= \frac{\sigma^2(\beta)}{n} + \frac{\sigma^2(S\beta,e)}{n} = \varrho_i \tag{2.19}$$

Además, los autores desarrollaron un índice para el caso en el que el sujeto responda al azar a los ítems, al que denominaron *índice de seguridad corregido por la presencia de aciertos por azar*; su expresión es:

$$\vartheta_{ic} = \frac{A_i - A_{ic}}{A_{\text{máx},i} - A_{ic}} \tag{2.20}$$

donde A_{ic} es la *función de acuerdo debido a la presencia del azar*, la cual se obtiene cuando la probabilidad de acertar cada ítem del test es $n_{i.}/J$, donde $n_{i.}$

es el total de respuestas correctas del sujeto i. El índice ϑ_{ic} valora en qué medida están de acuerdo el patrón observado y el esperado una vez eliminado el efecto del azar. Su interpretación es la misma que ϑ_i (Ecuación 2.17).

Capítulo 3

Estadísticos de medición apropiada basados en la Teoría de Respuesta al Item

3.1. La Teoría de Respuesta al Item (TRI): introducción, conceptos básicos y supuestos

Cuando Binet y Simon (1916/1973) definieron las escalas de edad mental y el concepto de *curva característica del ítem*, estaban construyendo la base de una teoría psicométrica. Años más tarde, este término fue retomado por Lawley (1943, 1944) para expresar el rendimiento de un grupo de sujetos en un ítem y elaborar nuevos procesos para estimar los parámetros de las curvas. En 1950, Lazarsfeld propuso el concepto de *rasgo latente* para definir al constructo psicológico que podía ser medido a través de un conjunto de ítems, rasgo subyacente en todos los sujetos, medible pero no observable directamente. El deseo de obtener una medida de este rasgo dio origen a un conjunto de modelos enmarcados como Modelos de Rasgo Latente (MRL) y Modelos de Clase Latente (Lazarsfeld y Henry, 1968).

El auge del que gozaba la TCT a mediados del siglo XX propició el escaso interés práctico sobre los MRL, hasta que Lord (1952, 1953a, 1953b) formuló el *modelo de ogiva normal de dos parámetros*, que años después Birnbaum (1968) sustituyó por el *modelo logístico de dos parámetros* (2-p) más fácil de emplear. En Dinamarca, Rasch (1960) define el *modelo de respuesta al ítem de un parámetro* y, en EE.UU., Birnbaum propuso un modelo equivalente en su función matemática y en sus resultados al de Rasch pero partiendo de

supuestos diferentes. Lord (1980) denominó a la teoría que engloba a estos y a otros modelos como *Teoría de Respuesta al Item* (TRI), cuyo objetivo era el estudio de los ítems que componían los tests y en donde también se incluían los MRL.

Pero fue el libro de Lord y Novick (1968) el que estimuló la considerable cantidad de investigaciones sobre la TRI, constituyéndose como una entidad sólida a finales de los años 70, a lo que se le sumaron los avances de la informática, facilitando así la aplicación de los modelos y la estimación de los parámetros de los sujetos y de los ítems en tests que, en lo que se llevaba investigando, estaban formados por ítems de respuesta dicotómica –acierto o fallo, afirmación o negación, acuerdo o desacuerdo– y en los que subyacía un único rasgo o habilidad.

No hay que olvidar las aportaciones de Samejima (1969) a la TRI. Ella amplió el campo de estudio y aplicación con la incorporación de nuevos modelos que contemplaban la posibilidad de que la respuesta al ítem fuera politómica y continua, y generalizó los modelos unidimensionales ya existentes al ámbito multidimensional. Hambleton y Swaminathan (1985), entre otros, popularizaron la TRI haciéndola más comprensible y operable.

El concepto de rasgo latente según la TRI se define como un rasgo o habilidad no observable ni directamente medible, pero que se puede predecir y explicar a partir del desarrollo de un test o ítem (Lord y Novick, 1968). Para ello, se establece una relación matemática entre el desarrollo observable o puntuación en un test y el rasgo o habilidad latente. Esta relación es la base de los diversos modelos matemáticos que se han postulado en función de dos supuestos básicos acerca de la estructura del test, del rasgo latente y de la relación entre éste y la respuesta al ítem:

- La unidimensionalidad del espacio latente.
- La independencia local de los ítems.

A continuación se describen ambos supuestos.

3.1.1. La unidimensionalidad del espacio latente

En una teoría general de rasgos latentes se postula que en la realización de un ítem el sujeto está empleando uno o más rasgos o factores. Por esto, algunos de los modelos de respuesta al ítem asumen que es un solo rasgo latente o habilidad el que se hace constatar en la realización de un ítem aunque, a

pesar de esta suposición, los defensores de la unidimensionalidad del modelo aclaran que este rasgo no se manifiesta de forma única y hay otros factores que están influyendo a la vez en dicha realización. La justificación de los modelos unidimensionales reside en los estudios experimentales o estadísticos que han comprobado dicha unidimensionalidad; cabe señalar la revisión de Hattie (1985) sobre los 88 métodos que se utilizaron para comprobar dicho supuesto, llegando a la conclusión de que el análisis factorial no lineal (Fraser y McDonald, 1988; McDonald, 1985; McDonald y Ahlawat, 1974) y el análisis de residuales eran las técnicas más prometedoras para evaluar la unidimensionalidad del test en esos momentos. Cuesta (1996) y López (1995) destacan:

- El análisis factorial de la información completa (Bock y Aitkin, 1981; Bock, Gibbons y Muraki, 1988).
- La prueba de unidimensionalidad, monotonía e independencia condicional basada en tablas de contingencia (Holland y Rosenbaum, 1986).
- El procedimiento estadístico DIMTEST basado en la independencia esencial y dimensionalidad esencial de Stout (1987) modificado por Nandakumar (1991, 1993; Nandakumar y Stout, 1993).
- Los métodos basados en el principio de independencia local (Roznowsky, Tucker y Humphreys, 1991).

La dificultad para comprobar la dimensionalidad del test o del ítem ha llevado a que algunos autores (Hambleton y Swaminathan, 1985; Mulaik, 1972; Samejima, 1974; Stout, 1987, 1990) hayan preferido relajar el supuesto y asumir que, más que un simple rasgo o habilidad, son k rasgos o habilidades los que se manifiestan en la realización de un test o ítem de los cuales uno de ellos es el rasgo *dominante*. Si se asume este supuesto, entonces se podrá ajustar un modelo de respuesta unidimensional a los datos. Con el objetivo de eliminar restricciones a la aplicabilidad y ajuste de los modelos de respuesta al ítem, otros autores han optado por el uso de modelos multidimensionales (e.g., Ackerman, 1992; Ansley y Forsyth, 1985; Kok, 1988; Reckase, Carlson, Ackerman y Spray, 1986; Shealy y Stout, 1993; yen, 1984).

Pero a pesar de la difícil tarea de encontrar modelos estrictamente unidimensionales, este supuesto es la expectativa que se mantiene a la hora de construir un test bajo la TRI ya que, los tests unidimensionales, son cómodos de interpretar y válidos para ser aplicados en las posibles muestras de una población.

3.1.2. La independencia local de los ítems

El segundo supuesto fundamental de la TRI está muy relacionado con el anterior y afirma que la respuesta de un sujeto a un ítem de un test es estadísticamente independiente de la respuesta de ese sujeto al resto de ítems del test. Para que esto se verifique los ítems del test deben cumplir dos propiedades (Hambleton y Swaminathan, 1985):

1. El orden de presentación de los ítems no debe mediar en la realización del test, aunque algunas investigaciones sostienen que sí puede influir en las respuestas del sujeto (Hambleton y Traub, 1973; yen, 1981).
2. Los datos que se obtienen del test deben ser unidimensionales, esto es, para un nivel de habilidad fijado, los ítems del test no correlacionan entre sí, en caso contrario dos o más habilidades son medidas por los ítems del test.

Si un test está formado por n ítems ($j = 1, 2, ..., n$) con formato de respuesta dicotómico y U_j es la variable aleatoria correspondiente al vector de respuesta de un sujeto a los n ítems del test ($U_j = u_1, u_2, \ldots, u_n$), la función de respuesta del sujeto al ítem se ajusta a una distribución binomial:

$$\begin{aligned} P_j(\theta) &= P(U_j = 1|\theta) \\ Q_j(\theta) &= P(U_j = 0|\theta) \end{aligned}$$

donde $P_j(\theta)$ es la probabilidad de que el sujeto acierte el ítem j y $Q_j(\theta)$ es la probabilidad de que el sujeto responda erróneamente al ítem j [$Q_j(\theta) = 1 - P_j(\theta)$]. El supuesto de independencia local de los ítems evidencia que:

$$\begin{aligned} & P(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n|\theta) = \\ = \; & P(U_1 = u_1|\theta)P(U_2 = u_2|\theta) \cdots P(U_n = u_n|\theta) = \\ = \; & [P_1(\theta)^{u_1} Q_1(\theta)^{1-u_1}][P_2(\theta)^{u_2} Q_2(\theta)^{1-u_2}] \cdots [P_n(\theta)^{u_n} Q_n(\theta)^{1-u_n}] \end{aligned}$$

y, por lo tanto, asegura que la probabilidad de la puntuación del sujeto en el test condicionada a un nivel de habilidad es:

$$P(U_j = u_j|\theta) = \prod_{j=1}^{n} P_j(\theta)^{u_j} Q_j(\theta)^{1-u_j}$$

Pero el principio de independencia local de los ítems será válido si y sólo si se verifica la unidimensionalidad de los ítems, y viceversa.

Como consecuencia de estos dos supuestos básicos, en la TRI se define una función de regresión no lineal de las puntuaciones del ítem sobre el rasgo o habilidad medida por el test y su representación gráfica se denomina Curva Característica del Item (CCI), función matemática que relaciona la probabilidad de éxito de un sujeto en un ítem y la habilidad medida por el mismo:

$$f(u_j|\theta) = P_j(\theta)^{u_j} Q_j(\theta)^{1-u_j}$$

En el caso de que el espacio latente sea multidimensional, la función de regresión no lineal se denomina Función Característica del Item (FCI).

Lord (1980), y Hambleton y Swaminathan (1985) definen la propiedad de *invarianza* de la CCI en una muestra de sujetos que han contestado a un ítem y se les ha ajustado un modelo de respuesta al ítem, por la cual se asegura que la probabilidad de éxito de un sujeto en un ítem es independiente de la distribución de la habilidad de la población a la que pertenece, i.e., que el éxito de un sujeto al responder a un ítem es independiente del número de sujetos situados en ese mismo nivel de habilidad. Por consiguiente, la probabilidad de éxito de un sujeto en un ítem sólo obedece a la forma de la CCI de ese ítem o, lo que es lo mismo, está subordinada a la invarianza de los parámetros de ese ítem, siempre y cuando éste se ajuste a un modelo de respuesta apropiado.

Cada modelo de respuesta describe una familia de curvas en función de los parámetros del ítem que contenga y que son: el parámetro de dificultad (b_j), de discriminación (a_j) y de pseudo-azar (c_j). De este modo, la CCI resulta ser un modelo empírico al que se aproxima un modelo teórico-matemático exponencial. Además y como consecuencia de lo expuesto, esta función monótona creciente verifica el supuesto de homocedasticidad en cada uno de los niveles de habilidad, para los cuales existe un conjunto de probabilidades de acierto al ítem que siguen una distribución normal y cuyos parámetros están condicionados al nivel de habilidad.

3.2. Modelos de respuesta al ítem (MRI)

Los Modelos de Respuesta al Item (MRI) son modelos teórico-matemáticos que establecen una relación entre el rasgo latente –variable no observable– y el rendimiento en el ítem –valor observable–. Una clasificación tradicional de los MRI, que aunque algo anticuada sigue aún vigente, fue propuesta por McDonald (1982; López, 1995):

- Modelos unidimensionales *vs.* multidimensionales en función del número

de rasgos subyacentes.

- Modelos lineales *vs.* no lineales según sea la relación entre el rasgo y la respuesta a los ítems.
- Modelos de respuesta dicotómica *vs.* politómica conforme a cómo se evalúe la respuesta al ítem.

El grueso de los modelos definidos a lo largo de los años de investigación en la TRI se han encuadrado en la categoría de modelos unidimensionales, no lineales y dicotómicos. Sin embargo, en las últimas décadas del siglo XX se prestó especial atención a los modelos que trataban ítems de respuesta politómica o de elección múltiple (Bock, 1972; Masters, 1982; Samejima, 1969) dentro del marco de la evaluación educativa. Hambleton y Swaminathan (1985) añadieron a la última categoría de la clasificación de McDonald (1982) la posibilidad de que el formato de un ítem fuera continuo como una extensión al formato politómico. Pero tanto unos como otros limitan o restringen su campo de aplicación a un espacio unidimensional del rasgo, sin que por ello estuvieran y estén exentos de investigación y nuevas aportaciones los distintos modelos que surgen de las combinaciones de la anterior clasificación. Los principales MRI y características más representativas son:

1. *Modelo de ogiva normal de 2-p* de Lord (1952, 1953a). La función matemática de este modelo es:

$$P_j(\theta) = \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2\pi}} \exp\left(-z_j^2/2\right) dz \tag{3.1}$$

donde $P_j(\theta)$ es la probabilidad de que un sujeto con un nivel de habilidad θ seleccionado aleatoriamente de una población acierte el ítem j; a_j y b_j son los parámetros de discriminación y dificultad del ítem; z_j es la puntuación de desviación o normal desviada de una distribución de media b_j y desviación típica $1/a_j$.

Lord y Novick (1968) justifican la aplicación de este modelo matemático a partir de la descripción de la regresión lineal de una variable aleatoria hipotética (Γ_j) que subyace al ítem sobre la escala de habilidad. La variable Γ_j representa la *propensión* de un sujeto de la población a responder correctamente el ítem y adopta valores en el intervalo $(-\infty, +\infty)$. Para cada valor de habilidad existe una distribución condicional normal de la variable aleatoria Γ_j con parámetros $N(\mu_{j|\theta}, \sigma^2_{j|\theta})$.

2. *Modelo logístico de 2-p* de Birnbaum (1968). La función logística desarrollada por Birnbaum es una buena aproximación a la función de ogiva normal, facilita su cálculo y su aplicación práctica. Para explicar el rendimiento de los sujetos en un test de respuesta dicotómica, este modelo incorpora un valor constante, $D = 1{,}702$, para ajustar los resultados de la función logística a los de la función de ogiva; mediante esta aproximación las diferencias entre ambos modelos son inferiores a 0'01. La expresión matemática del modelo logístico de 2-p es:

$$P_j(\theta) = \frac{\exp\left[Da_j(\theta - b_j)\right]}{1 + \exp\left[Da_j(\theta - b_j)\right]}$$

y tras ser factorizado:

$$P_j(\theta) = \frac{1}{1 + \exp\left[-Da_j(\theta - b_j)\right]} \tag{3.2}$$

Como se puede observar, al igual que el anterior, el modelo logístico de 2-p se define con los parámetros de dificultad (b_j) y de discriminación (a_j) del ítem; además, la probabilidad de que un sujeto con habilidad θ seleccionado aleatoriamente de una población acierte un ítem j depende también del parámetro de habilidad.

3. *Modelo de 3-p* de Birnbaum (1968). La ampliación que Birnbaum realizó sobre el modelo logístico de 2-p tenía como propósito incorporar un nuevo parámetro para dar constancia del posible efecto de la *adivinación* de la respuesta correcta por sujetos de baja habilidad. Con este modelo se podría controlar estadísticamente el acierto por azar, sobre todo en el caso de ítems politómicos. Incorporando el parámetro de *pseudo-azar*, denotado c_j –denominación que Hambleton y Swaminathan (1985) consideran más apropiada que la de adivinación– la expresión matemática resultante es:

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + \exp\left[-Da_j(\theta - b_j)\right]} \tag{3.3}$$

El parámetro c_j adopta valores entre 0 y 1 aunque, por lo general, depende del número de opciones de respuesta del ítem (k; $c_j = 1/k$). Siempre que se considere la posibilidad de responder por azar, la CCI será una función de regresión no lineal cuya asíntota horizontal por la izquierda es al menos el valor c_j.

El parámetro de pseudo-azar descalifica de logístico al modelo de 3-p debido a sus propiedades matemáticas y psicométricas; el nuevo parámetro cambia la definición del parámetro de dificultad de los modelos logísticos considerándose ahora como el punto del continuo de habilidad en donde:

$$P_j(\theta) = \frac{1 + c_j}{2} \tag{3.4}$$

No obstante, no todos los psicómetras están de acuerdo con esta afirmación.

4. *Modelo logístico de 1-p* o modelo de Rasch (1966). Retomando el modelo logístico de 2-p de Birnbaum (Ecuación 3.2) y suponiendo que el parámetro de discriminación es constante para todos los ítems, se obtiene la expresión matemática del modelo de Rasch:

$$P_j(\theta) = \frac{1}{1 + \exp\left[-D(\theta - b_j)\right]} \tag{3.5}$$

para el que la respuesta del sujeto al ítem sólo depende de la dificultad de éste.

5. *Modelo de 4-p* de Barton y Lord (1981) y McDonald (1967). Este modelo tiene más interés teórico que práctico debido a la escasez de investigación contrastada. El parámetro γ_j que se añade a la función del modelo de 3-p es un valor menor de 1 y representa la posibilidad de que sujetos con alta habilidad no respondan correctamente al ítem ya sea por descuido, por falta de información acerca del test o por otras causas (Hambleton y Swaminathan, 1985). Su expresión:

$$P_j(\theta) = c_j + \frac{\gamma_j - c_j}{1 + \exp\left[-Da_j(\theta - b_j)\right]} \tag{3.6}$$

6. *Modelos de ogiva normal para 1-p, 3-p* y *4-p* de Lord (1952). Aunque este autor se centró casi en exclusiva en el modelo de ogiva normal de 2-p, desde un punto de vista teórico es conveniente recordarlos. Sus interpretaciones se ajustan a las descritas para los modelos logísticos correspondientes por número de parámetros que les preceden:

   - Modelo de ogiva normal de 1-p:

$$P_j(\theta) = \int_{-\infty}^{\theta - b_j} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dx \tag{3.7}$$

- Modelo de ogiva normal de 3-p:

$$P_j(\theta) = c_j + (1 - c_j) \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dx \tag{3.8}$$

- Modelo de ogiva normal de 4-p:

$$P_j(\theta) = c_j + (\gamma_j - c_j) \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dx \tag{3.9}$$

Para estos modelos, y sobre todo para los modelos logísticos, los parámetros de los ítems y el parámetro de habilidad de los sujetos se caracterizan por:

- El parámetro de habilidad θ es un valor no observable directamente por el rendimiento en un ítem. Se define en una escala arbitraria a la que se puede ajustar una escala de intervalo y aproximar los datos al supuesto teórico de que se distribuyen según una ley normal de media μ y desviación típica σ (Camilli y Shepard, 1994). El rango de valores para el parámetro de habilidad establecido por consenso es $(-3, +3)$, evitando así la arbitrariedad en las conclusiones a las que se podría llegar al tomar como intervalo $(-\infty, +\infty)$.
- El parámetro de dificultad del ítem b_j está medido en la misma escala de intervalo que el parámetro de habilidad, $(-3, +3)$. El valor que adopte este parámetro del ítem será el punto de inflexión de la CCI en el cual la probabilidad de acertar el ítem es del 50 %. Un valor bajo de b_j indicaría baja dificultad del ítem, en caso contrario, si b_j toma valores altos, el ítem es difícil.
- Como θ y b_j están definidos en la misma escala, cuando uno de ellos se fija el otro también queda fijado.
- En principio, el parámetro de discriminación del ítem a_j se define dentro del amplio rango $(-\infty, +\infty)$. Pero poco sentido tiene hablar de un valor de discriminación negativo referido a la habilidad de ejecución de un ítem; por lo tanto, se cree más conveniente fijar su intervalo de variación en $(0, +3)$ o estrecharlo a $+2'5$ por su límite superior. Con respecto a su relación con la CCI, a_j es una magnitud proporcional a la pendiente de la CCI, así que altos a_j conllevan CCI más empinadas que bajos valores del mismo. La relación entre el punto de máxima pendiente (β) de la curva y a_j es:

$$\beta = 0'425\, a_j$$

- Para el caso del modelo de 3-p, el parámetro de pseudo-azar c_j toma valores entre 0 y 1; el parámetro de dificultad del ítem (b_j) se define como la proyección del punto de inflexión de la CCI sobre la habilidad cuando la probabilidad de responder con éxito el ítem se obtiene con la Ecuación 3.4. En este modelo, la pendiente de la CCI cuando $\theta = b_j$ es proporcional a los parámetros a_j y c_j:

$$\beta = 0'425\, a_j\, (1 - c_j)$$

7. Otros MRI son:

   a) *Modelo de respuesta graduada* de Samejima (1969). Permite estimar los parámetros de los ítems cuyas respuestas contienen dos o más categorías ordenadas, e.g., correcto, parcialmente correcto, falso. Este modelo define $P_{jl}(\theta)$ como la probabilidad de que el sujeto responda a una determinada categoría (l) en un nivel de habilidad dado. Su formulación es:

$$P_{jl}(\theta) = P^*_{jl}(\theta) - P^*_{j(l+1)}(\theta)$$

   donde $P^*_{jl}(\theta)$ es la probabilidad de que un sujeto de habilidad θ elija la categoría l o mayor como respuesta al ítem j y se obtiene aplicando el modelo logístico o de ogiva normal que mejor se ajuste a los datos. Cada categoría del ítem tiene un parámetro de dificultad (b^*_{jl}) y el parámetro de discriminación (a_j) es común para todas las categorías del mismo; en el caso del modelo logístico de 2-p, su cálculo es:

$$P^*_{jl}(\theta) = \frac{1}{1 + \exp\left[-Da_j(\theta - b^*_{jl})\right]} \tag{3.10}$$

   Como consecuencia, existe un Curva Característica para cada una de las Categorías del Item (CCCI) también denominada Función de Respuesta al Item (FRI).

   b) *Modelo de crédito parcial* de Masters (1982). Es un caso especial del modelo de respuesta graduada para ítems con más de dos categorías ordenadas y en donde la respuesta puede ser parcialmente correcta, es decir, está en función de los conocimientos del sujeto y por ello su respuesta aporta información de la habilidad del sujeto aunque no sea del todo correcta. La probabilidad de respuesta está expresada de modo similar al modelo de respuesta graduada:

$$P^*_{jl}(\theta) = \frac{1}{1 + \exp\left[-D(\theta - b^*_{jl})\right]} \tag{3.11}$$

donde b^*_{jl} es la dificultad de la categoría l del ítem j, con la particularidad de que $b^*_{j(l-1)} \leq b^*_{jl}$.

c) *Modelo de respuesta nominal* de Bock (1972). Es un modelo de respuesta categórico donde las categorías no están graduadas o no son ordinales. Para su estudio se ajusta el patrón de respuesta a un modelo logístico multivariante cuya expresión es:

$$P_{jl}(\theta) = \frac{\exp Z_l(\theta)}{\sum_{L=1}^{m} \exp Z_L(\theta)} \tag{3.12}$$

donde l es la categoría de respuesta al ítem elegida por el sujeto de las m categorías de respuesta posibles del ítem ($L = 1, 2, \ldots, m$) y $Z_l(\theta)$ es la transformación logit multivariante en esa categoría. Para cada categoría del ítem existe una curva característica de respuesta por lo que, para cada ítem, hay L FRI. Una especificación del modelo es que, para un nivel de habilidad θ, la suma de las probabilidades de elección de cada una de las m categorías de un ítem es igual a la unidad:

$$\sum_{L=1}^{m} P_{jL}(\theta) = 1$$

y en consecuencia:

$$\sum_{L=1}^{m} Z_L(\theta) = 0$$

Otra clasificación de los modelos de TRI la elaboraron Thissen y Steinberg (1986) en la que sólo se consideran los modelos paramétricos y unidimensionales, es decir, modelos que asumen que existe una única variable latente subyacente al modelo de respuesta al ítem y es aleatoria. Con estos supuestos, los parámetros de los modelos pueden ser estimados con procedimientos de máxima verosimilitud marginal. La clasificación de estos autores de los modelos que cumplen sus criterios es la siguiente:

1. *Modelos binarios*. Se incluyen los modelos de respuesta libre binaria:
   - Modelos de ogiva normal de Lord (1952, 1953a).
   - Modelo de Rasch (1960).
   - Modelo logístico de 2-p de Birnbaum (1968).
   - Funciones *spline* de Winsberg, Thissen y Wainer (1983).

2. *Modelos diferenciales*. Abarcan los modelos de respuesta politómica como el modelo de respuesta graduada de Samejima (1969), que define la probabilidad de una categoría en función de la diferencia entre las categorías superior e inferior a ella.

3. *Modelos divididos por el total*. Son modelos de respuesta politómica donde la probabilidad de responder a una categoría es un cociente de exponenciales:

   - Modelo de crédito parcial de Masters (1982).
   - Modelo de escalas de clasificación de Andrich (1978).
   - Modelo de respuesta nominal de Bock (1972).
   - Modelo de Masters y Wright (1984).

4. *Modelos acumulados por la izquierda*. Son modelos de respuesta binaria que contemplan el efecto del azar:

   - Modelo de 3-p de Birnbaum (1968).
   - Modelo de Choppin (1983).

5. *Modelos acumulados por la izquierda y divididos por el total*. Aquí se incluyen a aquellos modelos de elección múltiple definidos a partir del modelo de Samejima (1969) y el modelo de Thissen y Steinberg (1984).

Mellenbergh (1994) elaboró una clasificación de los modelos de respuesta al ítem a partir de la generalización de la TRI que el mismo autor ha denominado Teoría Lineal Generalizada de Respuesta al Item (*Generalized Linear Item Response Theory*, GLIRT). Los supuestos iniciales de la TRI se amplían para ser útiles cuando el formato de los ítems no es dicotómico; los modelos lineales generalizados de la TRI asumen que el rasgo latente puede ser continuo –modelos de rasgo latente– o nominal –modelos de clase latente–. En GLIRT, la ecuación de regresión que relaciona la respuesta del sujeto con el rasgo o los rasgos latentes está en función de la transformación de la respuesta esperada del sujeto a un ítem, denotada $g(\tau_{ij})$, y es una combinación lineal de las variables explicativas latentes y observables, continuas o nominales (z) y de una variable latente continua o nominal t utilizada en todos los modelos de TRI:

$$g(\tau_{ij}) = b_j + a_j t_i + c_{1j} z_{1i} + c_{2j} z_{2i} + \cdots + c_{pj} z_{pi} \tag{3.13}$$

En esta igualdad t_i y z_{ij} son las puntuaciones de los sujetos en las $(p+1)$ variables; b_j, a_j y c_{pj} son los parámetros del ítem j y τ_{ij} es la probabilidad que tiene un sujeto i de responder correctamente a un ítem j. La GLIRT mantiene

el supuesto de independencia local de los ítems. Mediante la Ecuación 3.13 se obtiene un modelo de respuesta al ítem generalizado que es una función lineal lograda al efectuar transformaciones logit, probit y la complementaria log-log sobre cada uno de los MRI.

3.3. La estimación de parámetros

Los modelos de la TRI permiten conocer la probabilidad que tiene un sujeto de habilidad θ de acertar un ítem j, definido éste a partir de uno, dos o tres parámetros. Un problema de estos modelos es que los parámetros de habilidad y de los ítems son desconocidos, lo único que se conoce es la respuesta del sujeto al ítem, con la cual se pueden hacer estimaciones para obtener los verdaderos valores de la habilidad y de los ítems, estimaciones que a su vez dependen del MRI que mejor explique los datos. Encontrar dichos parámetros no es un proceso directo, sino que requiere de sucesivas iteraciones –e.g., mediante el método de Newton-Raphson– con el objetivo de que los estimadores sean insesgados, suficientes, eficientes y consistentes.

Existen varios procedimientos de estimación de parámetros, de los cuales los más utilizados son: el Método de Mínimos Cuadrados (MMC) si el modelo es lineal, e independientemente del tipo de modelo, los métodos de Máxima Verosimilitud (MV) y de Estimación bayesiana (EB) o estimación Esperada a Posteriori (EAP), siendo la no-linealidad la característica principal de los MRI; de estos dos últimos métodos, el de MV y sus distintas variantes son los más empleados. El MMC también se puede aplicar a los modelos logísticos si éstos se linealizan (Baker, 1992).

3.3.1. El método de máxima verosimilitud (MV)

Sea una muestra de N sujetos a los que se les aplica un test formado por n ítems y cada uno de éstos está definido por parámetros de dificultad, discriminación y pseudo-azar; el vector de vectores de respuesta de los N sujetos es:

$$\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N)$$

constituido por los vectores de respuesta individuales a los n ítems del test:

$$\begin{aligned} \mathbf{U}_1 &= (u_1, u_2, \ldots, u_n) \\ \mathbf{U}_2 &= (u_1, u_2, \ldots, u_n) \end{aligned}$$

$$\begin{aligned} &\vdots \\ \mathbf{U}_N &= (u_1, u_2, \ldots, u_n) \end{aligned}$$

Si los ítems son de formato dicotómico ($u_{ij} = 1$ si es acierto y $u_{ij} = 0$ si es fallo), entonces el patrón de respuesta de un sujeto ($\mathbf{U}_i$) se aproxima a una distribución binomial, donde la probabilidad de acertar un ítem j por un sujeto de habilidad θ_i es $P_j(\theta_i)$ y la probabilidad de fallarlo es $Q_j(\theta_i) = 1 - P_j(\theta_i)$.

Para estimar los parámetros se recurre a la *función de verosimilitud*, función matemática que más se aproxima al verdadero valor de la probabilidad de acertar un ítem por un sujeto de habilidad θ, la cual es una matriz de respuestas $N \times n$:

$$L(\mathbf{U}|\mathbf{\Theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}) = L(u_1, u_2, \ldots, u_n|\mathbf{\Theta}, \mathbf{h}) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_j(\theta_i)^{u_{ij}} Q_j(\theta_i)^{1-u_{ij}}$$

Los mejores estimadores de los parámetros de los ítems y de la habilidad de los sujetos serán aquellos que maximicen esta función. Para facilitar los cálculos sin que se modifiquen los resultados, se opta por utilizar el logaritmo natural de la función de verosimilitud:

$$\ln L(\mathbf{U}|\mathbf{\Theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{i=1}^{N}\sum_{j=1}^{n}[u_{ij} \ln P_j(\theta_i) + (1 - u_{ij}) \ln Q_j(\theta_i)] \tag{3.14}$$

Los máximos son las raíces obtenidas de las derivadas primeras de la Ecuación 3.14 con respecto de cada uno de los parámetros de los ítems y de la habilidad por separado:

$$\frac{\delta \ln L}{\delta \mathbf{r}} = \sum_{i=1}^{N}\sum_{j=1}^{n} \frac{u_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \frac{\delta P_j(\theta_i)}{\delta \mathbf{r}} = 0 \tag{3.15}$$

donde $\mathbf{r}$ es el vector de vectores de parámetros a estimar definido como:

$$\mathbf{r} = \begin{bmatrix} \mathbf{\Theta} \\ \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{bmatrix}$$

vector con dimensión $(N+n)$ si la estimación se hace sobre el modelo logístico de 1-p, $(N+2n-2)$ si es el modelo logístico de 2-p y dimensión $(N+3n-2)$ si

es el modelo de 3-p. Resolver la Ecuación 3.15 es bastante complicado, por lo que se asumen ciertas restricciones para solucionar el problema (Baker, 1992; Birnbaum, 1968):

- Si los sujetos se eligen aleatoriamente de una población, entonces las respuestas de los sujetos a los ítems serán independientes unas de otras y, por lo tanto, la esperanza matemática de las derivadas cruzadas entre dos sujetos cualesquiera es igual a cero; esto reduce el proceso a la evaluación de la primera derivada sujeto-a-sujeto.
- Los sujetos y los ítems son independientes, por lo que la covariación entre ellos también es nula.
- Desde un punto de vista estadístico, la selección de ítems que componen el test es aleatoria y hace posible asumir que la covarianza entre los parámetros de los ítems es cero, de manera que los parámetros de un ítem se pueden estimar independientemente del resto de los ítems del test.

Es obvio que estas asunciones reducen bastante el problema de la estimación de parámetros, acotándolo a la resolución de las siguientes derivadas primeras parciales de la función de verosimilitud:

$$\frac{\delta \ln L}{\delta \theta_i} = \sum_{j=1}^{n} \frac{u_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \frac{\delta P_j(\theta_i)}{\delta \theta_i} = 0 \tag{3.16}$$

$$\frac{\delta \ln L}{\delta a_j} = \sum_{i=1}^{N} \frac{u_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \frac{\delta P_j(\theta_i)}{\delta a_j} = 0 \tag{3.17}$$

$$\frac{\delta \ln L}{\delta b_j} = \sum_{i=1}^{N} \frac{u_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \frac{\delta P_j(\theta_i)}{\delta b_j} = 0 \tag{3.18}$$

$$\frac{\delta \ln L}{\delta c_j} = \sum_{i=1}^{N} \frac{u_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \frac{\delta P_j(\theta_i)}{\delta c_j} = 0 \tag{3.19}$$

A pesar de la simplificación de cálculo para encontrar los valores máximos de la función de verosimilitud, un problema añadido es que se desconocen los parámetros de los ítems y de la habilidad, por lo que se ha de recurrir a algún método con el cual se puedan obtener buenos estimadores. Uno de los más utilizados es el algoritmo numérico iterativo o método de optimización de Newton-Raphson basado en la expansión de las series de Taylor (Baker, 1992; Birnbaum, 1968; Bunday, 1984; Hosking, Joyce y Turner, 1978; Isaacson y Keller, 1966; Kennedy y Gentle, 1980; Lord, 1980). Este algoritmo fija un

valor inicial de parámetro e itera hasta llegar al valor óptimo, cuya discrepancia con el valor inicial no excede de un error (ε) fijado a priori (los criterios de convergencia son, e.g., $\varepsilon = 0'001$ ó $\varepsilon = 0'0001$). En el caso de que el sistema de Ecuaciones 3.16, 3.17, 3.18 y 3.19 no se pueda separar, es decir, que se deban estimar conjuntamente todos los parámetros, entonces se implementaría la modalidad multivariante del método de Newton-Raphson, con la cual se buscará la solución a una matriz compuesta por las segundas derivadas de las anteriores hasta que se llegue a la convergencia. Sin embargo, las segundas derivadas pueden ser indefinidas, provocando la no-convergencia de la función de máxima verosimilitud; por lo tanto, para saldar este inconveniente se sustituye la matriz de segundas derivadas por la matriz con las Funciones de Información (FI) –i.e., grado de precisión de un ítem para medir la habilidad– de cada uno de los parámetros, alternativa que es conocida como método *scoring* de Fisher (Rao, 1965, p. 302).

El método MV tiene las siguientes variantes:

1. *El método de máxima verosimilitud conjunta* (MVC). Se emplea cuando se desconocen los parámetros de los ítems y de habilidad. Se procede de forma cíclica, en donde cada uno de los ciclos se subdivide en dos etapas: en la primera, se hacen estimaciones sobre los ítems suponiendo conocidos los parámetros de habilidad; en la segunda etapa, se estima la habilidad a partir de las estimaciones de los parámetros de los ítems obtenidos en la etapa anterior. Con el primer ciclo se llega a la primera aproximación de los parámetros de los ítems y de la habilidad. Para concluir el primer ciclo se fija la escala de medida de la habilidad (comúnmente, $\hat{\mu}_\theta = 0$ y $\hat{\sigma}_\theta = 1$) y se ajustan a esta transformación los parámetros de los ítems resultantes en la primera etapa; una vez que se han hecho las transformaciones oportunas, con estos valores se calcula el $\ln L_1$ y comienza un segundo ciclo en cuya primera etapa, estimación de los parámetros de los ítems, se toman los valores de habilidad estimados en el ciclo anterior; en la segunda etapa del segundo ciclo se estiman los parámetros de la habilidad tomando los valores estimados de los ítems en la primera etapa de este ciclo. Se vuelven a transformar los estimadores a una escala determinada, se computa el $\ln L_2$ para el segundo ciclo y se comparan los dos ciclos:

$$|\ln L_2 - \ln L_1| \leq \varepsilon \tag{3.20}$$

   Si se verifica esta desigualdad, el proceso iterativo concluye y los parámetros son los conseguidos en el segundo ciclo; en caso contrario, cuando $|\ln L_2 - \ln L_1| > \varepsilon$, comienza un tercer ciclo y así sucesivamente hasta que se cumpla la Ecuación 3.20.

2. *El método de máxima verosimilitud condicional* (MVCON). Este método se aplica para estimar el parámetro de dificultad de un ítem o el parámetro de la habilidad cuando el modelo que mejor explica los datos es el modelo de Rasch o el modelo de 1-p. En el caso de la estimación del parámetro de dificultad, se considera el total de aciertos del sujeto en el test como un estimador suficiente de su habilidad (Andersen, 1972, 1973). Entonces, conocidos los parámetros de habilidad y considerando que el parámetro de dificultad de un ítem es un *parámetro estructural*[1], la función de MVCON es (Hambleton y Swaminathan, 1985):

$$\begin{aligned} L(\mathbf{U}|t,\mathbf{b}) &= L(u_1,u_2,\ldots,u_n|t,\mathbf{b}) \\ &= \frac{P(\mathbf{U}|\theta,\mathbf{b})}{P(t|\theta,\mathbf{b})} \end{aligned} \tag{3.21}$$

donde $P(\mathbf{U}|\theta,\mathbf{b})$ es la probabilidad de obtener un patrón de respuestas $\mathbf{U}$ en un nivel de habilidad θ según el modelo de Rasch y $P(t|\theta,\mathbf{b})$ es la probabilidad de obtener una puntuación total t independientemente del patrón de respuesta.

Las primeras aproximaciones de b_j al parámetro verdadero de dificultad se consiguen con el mismo proceso descrito para el método MVC pero, a diferencia de éste, los valores de habilidad permanecen fijos a lo largo del proceso iterativo, centrado en estimar b_j. Para estimar el parámetro de habilidad por MVCON no es necesario que el modelo que mejor explique los datos sea el modelo de Rasch, basta con conocer los parámetros de los ítems y derivar el logaritmo natural de la función de máxima verosimilitud respecto de la habilidad:

$$\frac{\delta \ln L(\mathbf{U}|\theta)}{\delta\theta_i} = \sum_{j=1}^{n} \frac{u_{ij} - P_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)} \frac{\delta P_j(\theta_i)}{\delta\theta_i} = 0$$

El proceso sigue una secuencia iterativa comparando estimaciones de dos ciclos sucesivos como en el método MVC.

3. *El método de máxima verosimilitud marginal* (MVM). Esta variante intenta dar solución a los problemas que plantea el método MVC. Antes de explicar su argumentación, es necesario definir el concepto de *parámetro*

[1] El parámetro de dificultad de un ítem es un parámetro estructural porque las estimaciones sobre él tienden a converger hacia el verdadero valor paramétrico conforme se incrementa el número de sujetos que responden al ítem, independientemente del nivel de habilidad que tengan estos sujetos (Neyman y Scott, 1948).

incidental que hace referencia a la habilidad de los sujetos θ y con el que se enfatiza que, para un ítem dado, el valor de θ varía en función de los sujetos (Neyman y Scott, 1948).

Por MVC se obtienen estimaciones del parámetro de habilidad a partir de hipotéticos valores de los parámetros de los ítems y viceversa. Esto presenta el problema de que al incrementarse el número de sujetos (parámetros incidentales) se hace casi imposible encontrar convergencias de los parámetros de los ítems (parámetros estructurales). Entonces, lo que propone el método MVM es estimar los parámetros estructurales mediante la función de verosimilitud desligada de los parámetros de habilidad (Bock y Aitkin, 1981; Bock y Lieberman, 1970):

$$P(\mathbf{U}_i|\mathbf{a},\mathbf{b},\mathbf{c}) = \int_{-\infty}^{+\infty} \prod_{j=1}^{n} P_j(\theta)^{u_{ij}} Q_j(\theta)^{(1-u_{ij})} g(\theta) \partial\theta \equiv P(\mathbf{U}_i|\mathbf{U}) \tag{3.22}$$

donde $g(\theta)$ es la función de densidad de la distribución continua de la habilidad; $P(\mathbf{U}_i|\mathbf{a},\mathbf{b},\mathbf{c})$ es la probabilidad marginal de encontrar un patrón de respuesta $\mathbf{U}_i$ en un sujeto i extraído al azar de una población cuya habilidad es desconocida, pero en la que se asume que θ se distribuye según la función de densidad $g(\theta)$. La función $P(\mathbf{U}_i|\mathbf{a},\mathbf{b},\mathbf{c})$ sólo depende de los parámetros de los ítems, ya que ha sido integrada respecto de la habilidad y es equivalente a $P(\mathbf{U}_i|\mathbf{U})$.

En un test de n ítems hay 2^n patrones de respuesta posibles. Ante la posibilidad de encontrar s sujetos con un mismo patrón de respuestas $\mathbf{U}_i$, la función de verosimilitud vendrá dada por:

$$L(\mathbf{U}_i|\mathbf{a},\mathbf{b},\mathbf{c}) \propto \prod_{\mathbf{U}_i=1}^{2^n} P^s(\mathbf{U}_i|\mathbf{U})$$

cuyo logaritmo natural es:

$$\ln L(\mathbf{U}_i|\mathbf{a},\mathbf{b},\mathbf{c}) = s \sum_{\mathbf{U}_i=1}^{2^n} \ln P(\mathbf{U}_i|\mathbf{U}) \tag{3.23}$$

Los estimadores de los ítems son las soluciones de la Ecuación 3.23 al derivarla con respecto de a_j, b_j y c_j, proceso engorroso que ha sido sustituido por la aproximación que se obtiene con la cuadratura de Gauss-Hermite (Hambleton y Swaminathan, 1985). Otra solución a la integral (Ecuación 3.22) es la implementación del algoritmo EM propuesto por Bock y Aitkin (1981; Mislevy y Bock, 1990).

Con estos métodos se consiguen estimadores de los verdaderos parámetros de los ítems y de la habilidad de los examinados. Pero para precisar el valor de todos ellos es conveniente hacer intervalos de confianza que delimiten el rango del parámetro. Si existe un amplio grupo de sujetos, se puede suponer que el estimador $\hat{\theta}$ se aproxima a una distribución normal de media θ y de desviación típica o error típico de estimación es $\sigma_{\hat{\theta}|\theta} = 1/\sqrt{I(\hat{\theta})}$, donde $I(\theta)$ es la FI del estimador sobre el parámetro en ese nivel de habilidad. Fijado un nivel de confianza, el valor del verdadero parámetro de la habilidad se encontrará en el intervalo (Lord, 1980):

$$\hat{\theta} - |z_{\frac{\alpha}{2}}| \frac{1}{\sqrt{I(\hat{\theta})}} \leq \theta \leq \hat{\theta} + |z_{\frac{\alpha}{2}}| \frac{1}{\sqrt{I(\hat{\theta})}}$$

Cuanto mayor sea el número de ítems del test, mayor es la FI y, por lo tanto, menor es el error típico de estimación, con lo cual el intervalo confidencial estrechará sus límites y se logrará una mejor determinación del parámetro de habilidad. Sin embargo, si los ítems no son máximo-discriminantes en el nivel de habilidad que se estudia, la función de información del test no aumenta en ese nivel de habilidad, sino que tiende a disminuir; a este efecto se le conoce como *paradoja de Birnbaum* (Birnbaum, 1968).

Al igual que para el parámetro de habilidad, se pueden especificar intervalos de confianza para los parámetros de los ítems:

$$\begin{aligned} \hat{a} - |z_{\frac{\alpha}{2}}|\sigma_{\hat{a}|a} &\leq a \leq \hat{a} + |z_{\frac{\alpha}{2}}|\sigma_{\hat{a}|a} \\ \hat{b} - |z_{\frac{\alpha}{2}}|\sigma_{\hat{b}|b} &\leq b \leq \hat{b} + |z_{\frac{\alpha}{2}}|\sigma_{\hat{b}|b} \\ \hat{c} - |z_{\frac{\alpha}{2}}|\sigma_{\hat{c}|c} &\leq c \leq \hat{c} + |z_{\frac{\alpha}{2}}|\sigma_{\hat{c}|c} \end{aligned}$$

donde, en función del modelo, los errores típicos de los parámetros de los ítems son:

- si el modelo ajustado es el modelo de Rasch o de 1-p:

$$\sigma_{\hat{b}|b} = \frac{1}{\sqrt{I(\hat{b})}}$$

- para el modelo de 2-p:

$$\sigma_{\hat{b}|b} = \sqrt{\frac{I(\hat{a})}{I(\hat{a})I(\hat{b}) - I^2(\hat{a}\hat{b})}}$$

$$\sigma_{\hat{a}|a} = \sqrt{\frac{I(\hat{b})}{I(\hat{a})I(\hat{b}) - I^2(\hat{a}\hat{b})}}$$

- con respecto al modelo de 3-p, las estimaciones del parámetro de pseudo-azar por MV no son buenas, por lo que los valores de c_j se suelen escoger por algún método ad hoc y sólo se estiman los parámetros a_j y b_j (Lord, 1968, p. 1014).

3.3.2. La estimación bayesiana (EB) o esperada a posteriori (EAP)

Cuando un ítem es acertado o fallado por todos los sujetos, o cuando un sujeto acierta todos o ninguno de los ítems del test, la función MV no tiene máximo y, por lo tanto, el proceso de estimación no tiene solución porque la función no converge. El método de estimación bayesiana (EB; Mislevy, 1986; Swaminathan y Gifford, 1982, 1985, 1986) es una alternativa al método MVC cuando se da alguno de los casos citados. Para encontrar los estimadores de los parámetros de los ítems y de la habilidad de manera conjunta –si se desconocen todos los parámetros– o solamente de la habilidad –si se conocen los parámetros de los ítems– se recurre al teorema de Bayes, que añade a la función de verosimilitud una función de densidad a priori –real o hipotética– de cada uno de los parámetros a estimar a partir de una distribución conocida de los parámetros. Los estimadores son aquellos que maximizan las respectivas distribuciones de densidad a posteriori –de aquí la denominación de estimación esperada a posteriori (EAP)– para un patrón de respuestas $\mathbf{U}_i$; su expresión es:

$$f(\boldsymbol{\Theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}|\mathbf{U}) \propto L(\mathbf{U}|\boldsymbol{\Theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}) \left[\prod_{j=1}^{n} f(a_j) f(b_j) f(c_j)\right] \left[\prod_{i=1}^{N} f(\theta_i)\right] \tag{3.24}$$

donde $f(\boldsymbol{\Theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}|\mathbf{U})$ es la función de densidad de la distribución a posteriori de los parámetros de los ítems y de la habilidad de los sujetos; $L(\mathbf{U}|\boldsymbol{\Theta}, \mathbf{a}, \mathbf{b}, \mathbf{c})$ es la función de verosimilitud; $f(\theta_i)$ es la función de densidad a priori de la habilidad del sujeto i; y $f(a_j)$, $f(b_j)$ y $f(c_j)$ son las distribuciones a priori de los parámetros de los ítems que, por lo general, son:

- Una distribución χ^2 para a_j.
- Una distribución normal estandarizada para b_j porque este parámetro tiene la misma escala que la habilidad.
- Una distribución Beta para c_j.

El procedimiento a seguir es similar a la rutina de la estimación por MVC. La mayor ventaja de EB o EAP es que la estimación es directa y se logran estimadores próximos a los parámetros verdaderos, debido a que las desviaciones de los estimadores de éstos son controladas por las distribuciones de densidad a priori.

Cuando los parámetros de los ítems son conocidos, para la estimación EAP de la habilidad hay que encontrar el máximo estimador a posteriori de la desigualdad:

$$f(\theta_i|\mathbf{U}_i) \propto L(\mathbf{U}_i|\theta_i)f(\theta_i) \tag{3.25}$$

donde $f(\theta_i)$ es la función de densidad a priori de la habilidad de un sujeto i que se distribuye según una curva normal $N(0,1)$ (Owen, 1975; Swaminathan y Gifford; 1982), $L(\mathbf{U}_i|\theta_i)$ es la función de verosimilitud para la estimación del parámetro de habilidad y $f(\theta_i|\mathbf{U}_i)$ es la función de densidad a posteriori de la cual se obtiene el mejor estimador de la habilidad. Cuando hay N sujetos, la Ecuación 3.25 se generaliza:

$$f(\theta_1, \theta_2, \ldots, \theta_N|\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N) \propto L(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N|\theta_1, \theta_2, \ldots, \theta_N)f(\theta_1, \theta_2, \ldots, \theta_N)$$

La estimación bayesiana también sigue un proceso iterativo por el que se comparan las funciones de densidad a posteriori del último ciclo con las funciones de densidad a posteriori del ciclo inmediatamente anterior; cuando la diferencia entre ellas no sobrepasa un valor de error fijado de antemano, se termina el proceso.

3.4. Los estadísticos de medición apropiada

Los estudiosos e investigadores de la TRI que, desde hace poco más de 20 años, se han interesado por identificar a aquellos sujetos cuyos patrones de respuesta no se ajustan al modelo seleccionado, han desarrollado una serie de procedimientos y estadísticos para tal fin. A todos ellos se les ha incluido en el área de la *medición apropiada*, término acuñado por Levine y Rubin (1979).

En la literatura aparecen otras designaciones empleadas indistintamente tales como *índices de ajuste de persona*, *índices de escalabilidad*, *índices de respuesta atípica* o *índices de precaución*. Su objetivo es detectar los patrones que rompen la relación que debe existir entre ellos y el nivel de habilidad o rasgo del sujeto al que corresponde dicho patrón; en definitiva, evaluar el ajuste entre el MRI y el sujeto. Los índices de medición apropiada expuestos en la siguientes secciones son:

1. Los índices de precaución de Tatsuoka (1984), y Tatsuoka y Linn (1983).
2. El análisis de residuales de Rudner (1983), Smith (1985), Wright y Masters (1982), y Wright y Stone (1979).
3. La curva de respuesta de la persona de Trabin y Weiss (1983).
4. El método de comparación de las curvas características de los ítems de Rosenbaum (1987).
5. El estadístico de curvatura de la función de verosimilitud Drasgow, Levine y McLaughlin (1987).
6. Los índices de ajuste óptimo de Drasgow y Levine (1986), Drasgow, Levine y Zickar (1996), y Levine y Drasgow (1988).
7. Los estadísticos basados en la función de verosimilitud de Drasgow, Levine y McLaughlin (1991), Drasgow, Levine y Williams (1985), Levine y Rubin (1979), y Molenaar y Hoijtink (1990).
8. El estadístico ω de Wollack (1997).
9. El índices de medición apropiada para los tests adaptativos informatizados de Bradlow, Weiss y Cho (1998), McLeod y Lewis (1999), y van Krimpen-Stoop y Meijer (1999, 2000).

La última sección de este capítulo se describe un estadístico de ajuste de persona aplicado al Análisis de Estructura de Covarianza (AEC) propuesto por Reise y Widaman (1999).

3.4.1. Extensión de los índices de precaución a la TRI

Tatsuoka (1984), y Tatsuoka y Linn (1983) establecieron un nexo entre los estadísticos para detectar patrones atípicos por comparación con un grupo normativo y los supuestos de la TRI, buscando una correspondencia entre la curva-S, el índice de precaución C_i de Sato (1975) y las curvas características de la TRI.

En el modelo logístico de 1-p, la CCI se describe por la variable aleatoria de habilidad θ y el valor fijo del parámetro de dificultad de un ítem b_j. El resultado es una función monótona creciente a lo largo del continuo de habilidad:

$$P_{b_j}(\theta) = \frac{1}{1 + \exp[-D(\theta - b_j)]} \; \forall \, j \in n \tag{3.26}$$

Cuando lo que se quiere representar es la Curva Característica de la Persona (CCP) a partir de sus respuestas a cada uno de los n ítems del test, θ es un valor fijo y b_j es una variable aleatoria continua. En este caso, la función de respuesta de la persona es una función monótona decreciente a lo largo del continuo de dificultad:

$$S_{\theta_i}(b) = \frac{1}{1 + \exp[-D(\theta_i - b)]} \; \forall \, i \in N \tag{3.27}$$

Estas dos funciones son simétricas respecto del eje vertical $\theta = \theta_0$ para la CCI y respecto de $b = b_0$ para la CCP, pero en ambos casos cuando $\theta_0 = b_0$. Ambas curvas intersecan en el punto $(\theta_0 + b_0)/2$ si $\theta_0 \neq b_0$. Además, a partir de las Ecuaciones 3.26 y 3.27 se describen otras dos funciones: a) la Curva Característica del Test (CCT), la cual es el promedio de las n CCI:

$$T(\theta) = \frac{\sum_{j=1}^{n} P_{b_j}(\theta)}{n}$$

y b) la Curva Característica del Grupo (CCG), promedio de las N CCP:

$$G(b) = \frac{\sum_{i=1}^{N} S_{\theta_i}(b)}{N}$$

Las funciones $T(\theta)$ y $G(b)$ mantienen la monotonía de las funciones $P_{b_j}(\theta)$ y $S_{\theta_i}(b)$, respectivamente.

Para el modelo logístico de 2-p, que incluye el parámetro de discriminación del ítem (a_j), y para el modelo de 3-p, que aporta el parámetro de pseudo-azar (c_j), las cuatro funciones descritas se obtienen de igual modo.

Tatsuoka y Linn (1983) encontraron relación entre la curva-S y las CCT, y entre la curva-P y las CCG, si se sustituye la matriz de puntuaciones de la tabla S-P (Tabla 2.2) por una matriz de probabilidades en la que:

$$\sum_{j=1}^{n} P_{b_j}(\hat{\theta}_i) = \sum_{j=1}^{n} u_{ij} \sum_{i=1}^{N} S_{\theta_i}(\hat{b}_j) = \sum_{i=1}^{N} u_{ij}$$

y por lo tanto:

$$\begin{aligned} P(M_{i.}^{S}) &= \sum_{j=1}^{n} P_{b_j}(\hat{\theta}_i) = T(\theta) \\ P(M_{.j}^{P}) &= \sum_{i=1}^{N} S_{\theta_i}(\hat{b}_j) = G(b) \end{aligned}$$

donde $\hat{\theta}_i$ es el estimador del parámetro de habilidad y $\hat{b}_j$ es el estimador del parámetro de dificultad.

La primera extensión de los índices de precaución de Sato a la TRI, es el índice $ECI1$ (*Extended Caution Index*):

$$ECI1_i = 1 - \frac{\sum_{j=1}^{n}(u_{ij} - p_{i.})(u_{.j} - p_{..})}{\sum_{j=1}^{n}[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)](u_{.j} - p_{..})} \tag{3.28}$$

Este índice es el resultado de sustituir en la Ecuación 2.3 de C_i los valores M_{ij}^{S} y $p_{.j}$ por sus equivalentes en la TRI, $S_{\theta_i}(\hat{b}_j)$ y $T(\hat{\theta}_i)$.

A partir de la estimación de parámetros por el procedimiento de MV se verifica que $u_{.j} = N\sum_{j=1}^{n} G(\hat{b}_j)$. Si se sustituye en la Ecuación 3.28 el factor $(u_{.j} - p_{..})$ por $[G(\hat{b}_j) - G]$, se consigue el segundo índice de precaución $ECI2$:

$$ECI2_i = 1 - \frac{\sum_{j=1}^{n}(u_{ij} - P_{i.})[G(\hat{b}_j) - G]}{\sum_{j=1}^{n}[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)][G(\hat{b}_j) - G]} \tag{3.29}$$

donde $G = \sum_{j=1}^{n} G(b_j)/n$ es el promedio de las funciones de respuesta de grupo.

El tercer índice es una razón de correlaciones:

$$ECI3_i = 1 - \frac{\frac{\sum_{j=1}^{n}(u_{ij} - P_{i.})[G(\hat{b}_j) - G]}{\hat{\sigma}_j(u_{ij})\hat{\sigma}_j[G(\hat{b}_j)]}}{\frac{\sum_{j=1}^{n}[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)][G(\hat{b}_j) - G]}{\hat{\sigma}_j[S_{\hat{\theta}_i}(\hat{b}_j)]\hat{\sigma}_j[G(\hat{b}_j)]}} \tag{3.30}$$

El cuarto índice, $ECI4$, es una medida de la relación entre la puntuación observada u_{ij} y el vector teórico $S_{\hat{\theta}_i}$ en el nivel de habilidad θ_i:

$$ECI4_i = 1 - \frac{\sum_{j=1}^{n}(u_{ij} - P_{i.})[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)]}{\sum_{j=1}^{n}[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)][G(\hat{b}_j) - G]} \tag{3.31}$$

donde $T = \sum_{i=1}^{N} T(\theta_i)/N$ es el promedio de las funciones de respuesta del test.

El quinto índice es una razón de correlaciones derivada de $ECI4$:

$$ECI5_i = 1 - \frac{\frac{\sum_{j=1}^{n}(u_{ij}-P_{i.})[S_{\hat{\theta}_i}(\hat{b}_j)-T(\hat{\theta}_i)]}{\hat{\sigma}_j(u_{ij})\hat{\sigma}_j[S_{\hat{\theta}_i}(\hat{b}_j)]}}{\frac{\sum_{j=1}^{n}[S_{\hat{\theta}_i}(\hat{b}_j)-T(\hat{\theta}_i)][G(\hat{b}_j)-G]}{\hat{\sigma}_j[S_{\hat{\theta}_i}(\hat{b}_j)]\hat{\sigma}_j[G(\hat{b}_j)]}} \tag{3.32}$$

Tatsuoka y Linn (1983) clasificaron estos índices en dos categorías. La primera incluye a $ECI2$ y $ECI3$ porque son medidas referidas al grupo y miden la relación entre el patrón de respuesta observado del sujeto i y la variable normalizada derivada del grupo al que pertenece; estos índices son similares al índice de conformidad de la norma (NCI) de Tatsuoka y Tatsuoka (1982) y, conceptualmente, al coeficiente de correlación biserial-personal (r_{bisper}) de Donlon y Fischer (1968). En la segunda categoría se encuentran $ECI4$ y $ECI5$, orientados al sujeto por comparar el patrón de respuestas observado con el patrón teórico que le corresponde según la CCP en un nivel θ fijado; ambos estadísticos están relacionados con el procedimiento de Trabis y Weiss (1979) que se describirá más adelante (sección 3.4.3).

Tatsuoka (1984, 1996) añadió un sexto índice de precaución, $ECI6$, y desarrolló la estandarización de algunos ECI anteriores debido a la dependencia que éstos tienen de los valores de habilidad. El índice $ECI6$ se obtiene a partir de $ECI4$ reemplazando los valores G del denominador por $P_{i.}$:

$$ECI6_i = 1 - \frac{\sum_{j=1}^{n}(u_{ij} - P_{i.})[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)]}{\sum_{j=1}^{n}[S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)][G(\hat{b}_j) - P_{i.}]} \tag{3.33}$$

Este nuevo índice da información acerca del sujeto, comparando el vector de respuesta observado con las probabilidades de respuesta $P_{i.}$ obtenidas con el MRI.

En cuanto a las versiones estandarizadas de los índices $ECI1$, $ECI2$, $ECI4$ y $ECI6$ se ha recurrido a la formulación equivalente y abreviada de los mismos, y desarrolladas para el modelo logístico de 1-p:

$$\begin{aligned} ECI1_i &= 1 - \frac{Cov(u_{i.}, u_{.j})}{Cov(P_{i.}, u_{.j})} \\ ECI2_i &= 1 - \frac{Cov(u_{i.}, G)}{Cov(P_{i.}, G)} \\ ECI4_i &= 1 - \frac{Cov(u_{i.}, P_{i.})}{Cov(P_{i.}, G)} \end{aligned}$$

$$ECI6_i = 1 - \frac{Cov(u_{i.}, P_{i.})}{Cov(P_{i.}, P_{i.})}$$

Las expresiones estandarizadas son las que siguen:

- Esperanza matemática y varianza del índice $ECI1$:

$$\begin{aligned} E(ECI1|\theta_i) &= 1 - \frac{Cov(P_{i.}, u_{i.})}{Cov(P_{i.}, u_{i.})} = 0 \\ Var(ECI1|\theta_i) &= \frac{\sum_{j=1}^{n} \sigma_{ij}^2 (u_{.j} - P_{..})^2}{n^2 Cov^2(P_{i.}, u_{i.})} \end{aligned}$$

  donde $\sigma_{ij}^2 = P_j(\theta_i)Q_j(\theta_i)$. Por lo tanto:

$$ECI1_z = \frac{n\, Cov(P_{i.} - u_{i.}, u_{..})}{\left[\sum_{j=1}^{n} \sigma_{ij}^2 (u_{.j} - P_{..})^2\right]^{1/2}} \tag{3.34}$$

- Esperanza matemática y varianza del índice $ECI2$:

$$\begin{aligned} E(ECI2|\theta_i) &= 1 - \frac{Cov(P_{i.}, G)}{Cov(P_{i.}, G)} = 0 \\ Var(ECI2|\theta_i) &= \frac{\sum_{j=1}^{n} \sigma_{ij}^2 (G_j - G)^2}{n^2 Cov^2(P_{i.}, G)} \end{aligned}$$

  Entonces:

$$ECI2_z = \frac{n\, Cov(P_{i.} - u_{i.}, G)}{\left[\sum_{j=1}^{n} \sigma_{ij}^2 (G_j - G)^2\right]^{1/2}} \tag{3.35}$$

- Esperanza matemática y varianza del índice $ECI4$:

$$\begin{aligned} E(ECI4|\theta_i) &= 1 - \frac{Var(P_{i.})}{Cov(P_{i.}, G)} \\ Var(ECI4|\theta_i) &= \frac{\sum_{j=1}^{n} \sigma_{ij}^2 (P_{ij} - T_i)^2}{n^2 Cov^2(P_{i.}, G)} \end{aligned}$$

  Por lo tanto:

$$ECI4_z = \frac{n\, Cov(P_{i.} - u_{i.}, P_{i.})}{\left[\sum_{j=1}^{n} \sigma_{ij}^2 (P_{ij} - T_i)^2\right]^{1/2}} \tag{3.36}$$

- Esperanza matemática y varianza del índice $ECI6$:

$$\begin{aligned} E(ECI6|\theta_i) &= 1 - \frac{Cov(P_{i.}, u_{i.})}{Cov(P_{i.}, u_{i.})} = 0 \\ Var(ECI6|\theta_i) &= \frac{\sum_{j=1}^{n} \sigma_{ij}^2 (P_{ij} - T_i)^2}{n^2 Cov^2(P_{i.}, P_{i.})} \end{aligned}$$

Por consiguiente:

$$ECI6_z = \frac{n\, Cov(P_{i.} - u_{i.}, P_{i.})}{\left[\sum_{j=1}^{n} \sigma_{ij}^2 (P_{ij} - T_i)^2\right]^{1/2}} \tag{3.37}$$

3.4.2. El análisis de residuales

Trabajando con el modelo de Rasch, Wright y Stone (1979) investigaron acerca de cómo este modelo podía explicar la factibilidad de un patrón de respuestas. Para ello, definieron el término *matriz esperada* como la matriz S-P de N sujetos que responden a n ítems de un test de la cual se han eliminado los ítems a los que todos los sujetos contestaron correcta e incorrectamente, y a los sujetos que han acertado y fallado todos los ítems. Siguiendo las mismas suposiciones de Sato (1975) pero aplicando los principios de la TRI y utilizando el modelo de Rasch, los patrones *esperados* serían los correspondientes o bien a sujetos más hábiles que tienen mayor probabilidad de acertar ítems fáciles que sujetos menos hábiles, o bien a sujetos que ante dos ítems de distinta dificultad tienen mayor probabilidad de acertar el ítem fácil que el ítem con mayor dificultad. Las desviaciones de lo que sería un patrón esperado, patrón *inesperado*, se pueden detectar mediante el cálculo de un índice de residuales. Para implementar éste, los autores introdujeron previamente el índice Z_{ij} para evaluar la respuesta dada por el sujeto i al ítem j:

$$Z_{ij} = \frac{u_{ij} - P_j(\hat{\theta})}{\{P_j(\hat{\theta})[1 - P_j(\hat{\theta})]\}^{1/2}}$$

donde u_{ij} es la respuesta del sujeto al ítem dicotómico (0 ó 1), $P_j(\hat{\theta}_i)$ es la probabilidad estimada de acertar el ítem según el modelo de Rasch por un sujeto de habilidad estimada $\hat{\theta}_i$, en función del porcentaje de aciertos en el test (recuérdese que en el modelo de Rasch la puntuación directa es un estimador suficiente del nivel de habilidad del sujeto). El índice Z_{ij} expresa los residuales estandarizados en términos de la habilidad del sujeto y de la dificultad del ítem,

y sigue una distribución normal estandarizada. Otra alternativa es calcular el cuadrado de Z_{ij} y contrastar su valor con una distribución χ^2 con un grado de libertad necesario para estimar el parámetro de habilidad del sujeto. Pero para justificar el patrón de respuestas, el índice más adecuado es la media cuadrática de los residuales estandarizados, resultado de sumar los valores Z_{ij}^2 del sujeto i en los j ítems del test:

$$U_i = \frac{\sum_{j=1}^{n} Z_{ij}^2}{n} = \sum_{j=1}^{n} \frac{[u_{ij} - P_j(\hat{\theta})]^2}{n\, P_j(\hat{\theta})[1 - P_j(\hat{\theta})]} \tag{3.38}$$

El índice de residuales U_i sigue una distribución χ_ν^2 con grados de libertad igual al número de ítems del test menos 1 ($\nu = n - 1$) y con él se sabría cómo de inesperado es el patrón de respuestas. Al probar este índice para detectar patrones atípicos, los autores apreciaron que se solía rechazar la hipótesis nula cuando la habilidad del sujeto y la dificultad del ítem eran dispares entre sí. Para solventar este inconveniente, Wright y Masters (1982) ponderaron el índice U_i y generalizaron su aplicabilidad a todos los MRI:

$$W_i = \frac{\sum_{j=1}^{n} Var_j(\hat{\theta})[u_{ij} - P_j(\hat{\theta})]^2}{\sum_{j=1}^{n} Var_j(\hat{\theta})} \tag{3.39}$$

El nuevo índice W_i es la media cuadrática ponderada de los residuales, en donde el factor de ponderación son las varianzas de cada uno de los cuadrados de los residuales $[Var_j(\hat{\theta})]$. Cuando los datos se ajustan al modelo, el índice W_i sigue aproximadamente una distribución media cuadrática de media 1 y varianza:

$$q_i^2 = \frac{\sum_{j=1}^{n} [C_{ij} - Var_j(\hat{\theta})^2]}{[\sum_{j=1}^{n} Var_j(\hat{\theta})]^2}$$

donde C_{ij} es el índice de curtosis de u_{ij}. Por conveniencia y para poder comparar distintos patrones de respuesta, Wright y Masters (1982), y Wright y Stone (1979) expresaron los índices de residuales U_i y W_i como estadísticos normalizados con media 0 y desviación típica 1:

$$ZU_i = (\ln U_i + U_i - 1)\left(\frac{\nu}{8}\right)^{1/2} \tag{3.40}$$

$$ZW_i = \frac{3(W_i^{1/3} - 1)}{q_i} + \frac{q_i}{3} \tag{3.41}$$

Rudner (1983) generalizó los estadísticos del análisis de residuales al modelo de 3-p. Los dos índices que propuso fueron el residual estandarizado de la media cuadrática (F_1):

$$F_1 = \frac{1}{n}\sum_{j=1}^{n}\frac{[u_{ij} - P_j(\hat{\theta}_i)]^2}{P_j(\hat{\theta}_i)Q_j(\hat{\theta}_i)} \tag{3.42}$$

y el otro estadístico, proporcional al W_i, está basado en la media cuadrática ponderada de ajuste para el modelo de 3-p cuya expresión es:

$$W_3 = \frac{\sum_{j=1}^{n}[u_{ij} - P_j(\hat{\theta}_i)]^2}{\sum_{j=1}^{n} P_j(\hat{\theta}_i)Q_j(\hat{\theta}_i)} \tag{3.43}$$

La media de W_3 es 1, por lo que valores superiores a éste indican que el patrón de respuesta no se ajusta al modelo y, por lo tanto, es un patrón atípico; si $W_3 < 1$ se considera al patrón observado en acuerdo con el esperado según el modelo. El estadístico W_3 ha sido empleado posteriormente por Drasgow, Levine y McLaughlin (1987), y Rudner, Bracey y Skaggs (1996).

Smith (1985) propuso dos estadísticos relacionados con los índices de residuales anteriores. Para evaluar el ajuste del patrón de respuestas al modelo, este autor trabajó con subtests. Si un test de n ítems es dividido en S subtests ($s = 1, 2, \ldots, S$) que contienen k ítems, el estadístico de residuales intrasubtests no ponderado de ajuste para un patrón es:

$$UB_i = \frac{1}{S-1}\sum_{s=1}^{S}\frac{\left\{\sum_{j=1}^{k}[u_{ij} - P_j(\theta)]\right\}^2}{\sum_{j=1}^{k} P_j(\theta)[1 - P_j(\theta)]} \tag{3.44}$$

El estadístico UB_i sigue, según los análisis posteriores de Kogut (1988), una distribución χ^2 con S grados de libertad si se emplea θ, o $S-1$ grados de libertad si se trabaja con $\hat{\theta}$ estimada por MV.

El estadístico de residuales intersubtests ponderado por el número de ítems del subtest en estudio es:

$$UW_i = \frac{1}{k}\sum_{j=1}^{k}\frac{[u_{ij} - P_j(\hat{\theta})]^2}{n\,P_j(\hat{\theta})[1 - P_j(\hat{\theta})]} \tag{3.45}$$

3.4.3. La Curva de Respuesta de Persona (CRP)

Para estudiar el ajuste de las respuestas de un sujeto al modelo de TRI, Trabin y Weiss (1979, 1983) apostaron por un procedimiento gráfico basado en varios estudios previos: los de Moiser (1940, 1942) en el ámbito de la Psicofísica, la derivación que Lumsden (1977, 1978) hizo de la ley de juicio categórico de Thurstone y en los referentes a los tests adaptativos estratificados de Weiss (1973). A este procedimiento le nombraron Curva de Respuesta de Persona (CRP), designación adaptada al estudio de la variabilidad de la respuesta del sujeto, al ajuste de ésta a un MRI y reformulado de anteriores denominaciones como las de curva característica de persona empleada por Lumsden y *trace line* por Weiss, en sus respectivos estudios.

Como su propio nombre indica, existe una CRP para cada sujeto, o mejor dicho, existen dos CRP por sujeto, una que se obtiene a partir de los datos empíricos –la CRP observada– y otra resultado del MRI al que se supone que se ajustan los datos –la CRP esperada–. Para graficar la CRP observada o empírica, primero hay que agrupar los ítems del test en S subtests representativos de cada nivel de dificultad y disponerlos en orden de dificultad creciente, con la particularidad de que todos los subtests contengan el mismo número de ítems (k). Dentro de cada subtest, los ítems se ordenan por su grado de discriminación desde el ítem más discriminativo al de menor discriminación. A continuación, se calcula la proporción de respuestas correctas en cada uno de los subtests de dificultad. La CRP aparece al unir estas proporciones como función de los niveles de dificultad de los ítems, es decir, a partir de pares ordenados representados en ejes de coordenadas en donde los niveles de dificultad de los ítems se sitúan en el eje de abscisas y la proporción de aciertos en el eje de ordenadas. La forma de la curva proporcionaría información sobre:

- Cómo el sujeto ha dado respuesta al test.
- El nivel de habilidad estimado del sujeto al proyectar el punto medio de la curva (50 % de aciertos en el test) sobre el eje de abscisas.
- Si el sujeto ha respondido al azar a los ítems cuya dificultad es superior al nivel de habilidad estimado para él.
- Si el sujeto ha descuidado sus respuestas a los ítems más fáciles con respecto a su nivel de habilidad.
- La dimensionalidad del patrón de respuestas.

Aunque con la representación gráfica de la CRP observada se puedan dilucidar algunos o cada uno de estos cinco puntos, con ella no se puede asegurar si dicha forma es producto de las características del sujeto, o bien es consecuencia

del mero azar. Para poder tomar decisiones acerca de lo típico o atípico del patrón de respuestas, la CRP observada debe ser comparada con una CRP esperada. Esta curva se obtiene a partir del modelo de 1-p, 2-p ó 3-p, y necesita de los valores estimados de los parámetros de habilidad y de los parámetros de los ítems conocidos a priori. Para cada subtest, si las respuestas del sujeto se rigen por el MRI elegido, se verifica que la habilidad del sujeto es invariante en todos los subtests ($\theta_1 = \theta_2 = \ldots = \theta_S$). Una vez decidido el MRI y estimada la habilidad, se calcula la probabilidad esperada de acertar cada uno de los ítems del test $[P_j(\hat{\theta})]$. Al representar estos valores de probabilidad en los mismos ejes de coordenadas que los de la CRP empírica, la comparación de ambas curvas podría ser un indicio del grado de ajuste del patrón de respuestas. Sin embargo, esto sería una prueba intuitiva carente de fundamento para catalogar a un patrón de típico o atípico. Por esto, los autores propusieron realizar una prueba χ^2 de bondad de ajuste sobre el índice $D(\hat{\theta})$ de las diferencias entre las curvas. El cálculo de $D(\hat{\theta})$ entre la CRP observada y esperada:

$$D(\hat{\theta}) = \sum_{s=1}^{S} D_s(\hat{\theta}) \tag{3.46}$$

donde $D_s(\hat{\theta})$ es la diferencia de la proporción media de aciertos en la CRP observada (U_s) y la CRP teórica ($\hat{p}_s$) en cada uno de los S subtests de dificultad ($s = 1, 2, \ldots, S$):

$$\begin{aligned} D_s(\hat{\theta}) &= U_s - \hat{p}_s \\ U_s &= \sum_{j=1}^{k} \frac{u_{ij}}{k} \\ \hat{p}_s &= \sum_{j=1}^{k} \frac{P_j(\hat{\theta})}{k} \end{aligned}$$

donde u_{ij} es la respuesta de acierto (1) o fallo (0) al ítem j, k es el número de ítems el cada subtest y $\hat{p}_s$ es la probabilidad media de acierto esperada por subtest.

Para saber si el patrón de respuestas se ajusta al modelo, se calcula el índice $D(\hat{\theta})$ con la Ecuación 3.46 y se contrasta con un valor χ^2_ν con grados de libertad $\nu = S - 1$, es decir, el número de subtests menos 1. Si la prueba es estadísticamente significativa, el patrón de respuesta es atípico y, evaluando las diferencias entre las proporciones medias de aciertos por subtests, el investigador podría describir el patrón de respuestas como consecuencia de

que el sujeto haya contestado al azar, haya descuidado sus respuestas o haya hecho trampa.

Sin embargo, la distribución condicional del índice $D(\hat{\theta})$ es desconocida y no se podría asegurar que fuera un índice asintóticamente estandarizado independiente de θ. Klauer y Rettig (1990) trabajaron sobre la idea de la CRP y el índice D, aportando tres pruebas estadísticas asintóticas estandarizadas para probar la hipótesis de invarianza del parámetro de habilidad. La primera de ellas, el estadístico χ^2_{SC} de bondad de ajuste por subtest es similar al propuesto por Trabin y Weiss (1979, 1983):

$$\chi^2_{SC} = \sum_{s=1}^{S} \frac{D'^2_s(\hat{\theta})}{I_s(\hat{\theta})} \tag{3.47}$$

donde

$$\begin{aligned} D'^2_s(\hat{\theta}) &= \sum_{j=1}^{k} [u_{ij} - P_j(\hat{\theta})]\omega_j(\hat{\theta}) \\ \omega_j(\hat{\theta}) &= \frac{\delta P_j(\hat{\theta})/\delta\hat{\theta}}{P_j(\hat{\theta})[1 - P_j(\hat{\theta})]} \end{aligned}$$

e $I_s(\hat{\theta})$ es la FI del subtest s. El estadístico χ^2_{SC} se contrasta con una χ^2_ν con $\nu = S - 1$ grados de libertad. Además de χ^2_{SC}, también propusieron el estadístico χ^2_W de Wald si el objetivo era comparar las habilidades estimadas en los subtests:

$$\chi^2_W = \sum_{s=1}^{S} I_s(\hat{\theta}_s) \left(\hat{\theta}_s - \sum_{s=1}^{S} \lambda_s \hat{\theta}_s \right)^2 \tag{3.48}$$

donde $\hat{\theta}_s$ es la habilidad estimada en el subtest s y $\lambda_s = I_s(\hat{\theta}_s) / \sum_{s=1}^{S} I_s(\hat{\theta}_s)$. Esta prueba χ^2_W se distribuye según χ^2_ν con grados de libertad $\nu = S - 1$.

El tercer índice de Klauer y Rettig (1990) es el criterio de razón de verosimilitud de Neyman-Pearson:

$$\chi^2_{LR} = 2 \sum_{s=1}^{S} \sum_{j=1}^{k} u_{ij} \lg \frac{P_j(\hat{\theta}_s)}{P_j(\hat{\theta})} + (1 - u_{ij}) \lg \frac{1 - P_j(\hat{\theta}_s)}{1 - P_j(\hat{\theta})} \tag{3.49}$$

De nuevo, la prueba estadística sobre el patrón de respuestas χ^2_{LR} se contrasta con una χ^2_ν cuyos grados de libertad son $\nu = S - 1$.

Del estudio de simulación Montecarlo que llevaron a cabo Klauer y Rettig (1990) se dedujo que: $\chi^2_{SC} \leq \chi^2_{LR} \leq \chi^2_W$, los resultados de χ^2_{SC} eran más estables y χ^2_W fue la menos robusta frente a una mala aproximación de la estimación de la habilidad a la distribución normal.

Sijtsma y Meijer (2001) han adaptado el procedimiento de identificación de patrones de respuesta atípicos mediante la CRP al ámbito de los modelos de respuesta no paramétricos.

3.4.4. El método de comparación de las CCI de Rosenbaum

Rosenbaum (1987) sugirió que a partir de las CCI se podrían identificar patrones atípicos siempre y cuando los ítems y las respuestas de los sujetos estén definidos en una escala latente, es decir, que estén ordenados en dificultad creciente. Esta ordenación es la misma que la del escalograma de Guttman (1944, 1950) con la diferencia de que éste es un modelo determinístico y Rosenbaum emplea modelos de variable latente –el modelo de Rasch, los modelos de doble monotonía de Mokken, el modelo logístico de 2-p, algunos modelos multidimensionales...– para la detección de los patrones atípicos.

Sea un sujeto i con un patrón de respuestas en n ítems dicotómicos $\mathbf{U}_i = (u_1, u_2, \ldots, u_n)$ y $P(\mathbf{U} = u)$ es la distribución de ese vector de respuestas en una población; entonces, un modelo de variable latente establece que para el vector $\mathbf{U}$ hay asociada una distribución condicional para cada uno de los sujetos que depende de θ, variable latente e inobservable característica de cada sujeto:

$$P(\mathbf{U} = u) = \int \prod_{j=1}^{n} P(u_j = 1|\theta = \theta_i)^{u_j} [1 - P(u_j = 1|\theta = \theta_i)]^{1-u_j} dF(\theta) \tag{3.50}$$

donde $F(\theta) = P(\theta \leq \theta)$ es la distribución acumulada de θ en la población y $P(u_j = 1|\theta)$ es la CCI de j. De acuerdo con este supuesto, el ítem j es de modo uniforme más difícil que el ítem k si $P(u_j = 1|\theta) \leq P(u_k = 1|\theta)\ \forall\theta$, i.e., la CCI del ítem j está por debajo de la CCI del ítem k. Gráficamente, si las CCI no se cruzan para un determinado θ, los patrones de respuesta son normales, pero si se cruzan serán un indicador de un patrón atípico.

3.4.5. Los estadísticos de curvatura de la función de verosimilitud

Drasgow, Levine y McLaughlin (1987) propusieron dos estadísticos para identificar patrones atípicos basados en la función de verosimilitud. Una respuesta atípica influye en la forma de la función de verosimilitud alejándola del punto máximo, achatando la curva debido a que no hay ningún valor de habilidad que permita al MRI seleccionado ajustarse al patrón de respuestas. El primer estadístico de curvatura es la estimación de la varianza *jackknife* normalizada, la cual se obtiene tras estimar dos parámetros de habilidad por MV con el modelo de 3-p: el primero de ellos, $\hat{\theta}$ para el test de longitud n y el segundo, $\hat{\theta}_{(j)}$ para el test con $n-1$ ítems por eliminación del ítem j. Los *pseudo-valores* son:

$$\hat{\theta}_j^* = n\hat{\theta} - (n-1)\hat{\theta}_{(j)}^*$$

para $j = 1, 2, \ldots, n$. La estimación jackknife del parámetro de habilidad y su varianza son:

$$\begin{aligned} \hat{\theta}^* &= \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_j^* \\ \sigma^2(\hat{\theta}^*) &= \frac{\sum_{j=1}^{n}(\hat{\theta}_j^*)^2 - \frac{1}{n}\left(\sum_{j=1}^{n}\hat{\theta}_j^*\right)^2}{n(n-1)} \end{aligned}$$

Pero $\sigma^2(\hat{\theta}^*)$ no es un índice de medición apropiada ya que depende de θ. Para saldar este problema, los autores recurrieron al empleo de la FI, ya que ésta es el recíproco de la varianza asintótica de $\hat{\theta}$ y así la estimación jackknife puede ser normalizada. Entonces, el índice de ajuste del patrón de respuestas es:

$$JK = \sigma^2(\hat{\theta}^*)\, I(\hat{\theta}) \tag{3.51}$$

Cuando el patrón de respuestas no se ajusta al modelo, la función de verosimilitud es relativamente plana y la estimación de la varianza es mayor que para un patrón de respuesta normal.

El segundo índice compara las curvaturas de las funciones de verosimilitud esperada y observada. Si la función de verosimilitud es más chata para el patrón de respuesta atípico que para el patrón de respuesta normal, entonces

la FI observada será menor que la FI esperada. En consecuencia, el índice de medición apropiada es una razón de funciones de información:

$$O/E = \frac{-\frac{\delta^2 l_0}{\delta\theta^2}\Big|_{\theta=\hat{\theta}}}{I(\hat{\theta})} \tag{3.52}$$

donde l_0 es el logaritmo de la función de verosimilitud del patrón de respuestas observado e $I(\hat{\theta})$ es la FI esperada.

En el estudio en el que Drasgow, Levine y McLaughlin (1987) pusieron a prueba los índices JK y O/E concluyeron que, a pesar de estar estandarizados, padecían de una pobre identificación de patrones atípicos.

3.4.6. Los estadísticos de ajuste de persona óptimos

Levine y Drasgow (1984), y Drasgow, Levine y Zickar (1996) consideraron que un índice de medición apropiada era *óptimo* cuando ningún otro índice obtenía tasas de identificaciones correctas de patrones atípicos mejores que él, i.e., cuando tiene efectividad absoluta para detectar un determinado tipo de respuesta atípica. El estadístico de ajuste óptimo es una razón de probabilidad basada en el lema de Neyman-Pearson, por el cual se contrasta un modelo de respuesta normal frente a un modelo de respuesta atípico. Un patrón de respuestas de un sujeto i a un test de n ítems es $\mathbf{U}_i = (u_1, u_2, \ldots, u_n)$, donde $u_j = 1$ si la respuesta es correcta y $u_j = 0$ si la respuesta es errónea; entonces, bajo la hipótesis nula de que el patrón es normal o típico según un MRI que se supone ser el adecuado, $P_i(u_j|\theta)$ es la probabilidad de acertar el ítem j y $Q_i(u_j|\theta)$ es la probabilidad de fallarlo. Por lo tanto, la probabilidad del patrón normal de respuestas $\mathbf{U}_i$ para un sujeto de habilidad θ es:

$$P_n(\mathbf{U}|\theta) = \prod_{j=1}^{n} P(u_j|\theta)^{u_j} Q(u_j|\theta)^{1-u_j}$$

Si los ítems del test son dicotómicos, habría un total de 2^n patrones de respuesta posibles para un sujeto y la probabilidad de que ocurra cada uno de ellos es:

$$P_n(\mathbf{U}) = \int P_n(u_j|\theta)\, f(\theta)\, d\theta$$

donde $f(\theta)$ es la función de densidad de θ. Para poder aplicar el lema de Neyman-Pearson hay que seleccionar otro modelo de respuesta que defina un

tipo de respuesta atípica a m de los n ítems del test (azar, copia, perseverancia...). Existen $S_k = \binom{n}{m}$ formas de escoger m ítems del test para generar un respuesta atípica y la probabilidad de ocurrencia de un patrón atípico es:

$$P_a(\mathbf{U}|\theta) = \sum_k P_a(\mathbf{U}|\theta, S_k)P(S_k)$$

en el que $P(S_k) = 1/\binom{n}{m}$, por lo que:

$$P_a(\mathbf{U}) = \int \left[\sum_k P_a(\mathbf{U}|\theta, S_k)P(S_k)\right] f(\theta)\, d\theta$$

La probabilidad $P_a(\mathbf{U})$ se calcularía mediante un algoritmo numérico de cuadratura. El índice de ajuste óptimo es el resultado de la razón de probabilidades:

$$\lambda(\mathbf{U}) = \frac{P_a(\mathbf{U})}{P_n(\mathbf{U})} \tag{3.53}$$

Si $\lambda(\mathbf{U})$ se aleja de 1 el patrón es atípico. Además, el lema de Neyman-Pearson aporta una prueba estadística en la que se fija la tasa de error tipo I para rechazar la hipótesis nula de que el patrón es normal en un valor:

$$\alpha = \text{cte}\, P_n(\mathbf{U})$$

rechazando la hipótesis nula si:

$$P_a(\mathbf{U}) \geq \text{cte}\, P_n(\mathbf{U})$$

El índice $\lambda(\mathbf{U})$ es óptimo si y sólo si $P_n(\mathbf{U})$ y $P_a(\mathbf{U})$ son correctas. Mientras que la elección del modelo para calcular $P_n(\mathbf{U})$ no tiene problema alguno, la descripción del modelo de respuestas atípicas sí es un obstáculo. En los estudios de Drasgow y Levine (1986), y Drasgow, Levine y McLaughlin (1987) se recurrío a la resolución de un algoritmo sobre funciones simétricas para calcular $\lambda(\mathbf{U})$. Un procedimiento de cómputo menos costoso fue ideado por Levine y Drasgow (1988), quienes hicieron una derivación intuitiva de dicho algoritmo para generar modelos de respuesta atípicos y manipular patrones pero sin emplear funciones simétricas.

Posteriormente, Drasgow, Levine y Zickar (1996) definieron cinco modelos de respuesta atípica:

1. *Trampa debida al conocimiento de las respuestas.* Este es el caso de la administración del mismo test con n_1 ítems en distintas ocasiones, de las cuales y tras la primera de ellas, algún o algunos sujetos se han quedado con el test y los sujetos a los que se les va a administrar a continuación lo consiguieron y prepararon las respuestas. Para descubrir este tipo de patrones atípicos, el examinador incluye nuevos ítems en el test (n_2).

2. *Falsificación de respuestas en escalas de personalidad y en inventarios biográficos.* El procedimiento que se sigue para detectar este tipo de patrones atípicos es dividir el test en dos subtests. En el primero de ellos aparecerían los ítems que son fáciles de falsificar (n_1) y el segundo subtest lo formarían los ítems protegidos contra el engaño (n_2). Para los autores, una forma de representar la falsificación de las respuestas es por un incremento del rasgo latente del sujeto en Δ unidades en los n_1 ítems susceptibles de falsificación ($\theta_1 = \theta + \Delta$), cuyo patrón de respuestas es $\mathbf{U}_1$. El rasgo latente se mantiene invariante en el patrón de respuestas de los ítems infalsificables.

3. *Inexperiencia con los ordenadores y con los tests informatizados.* Los tests informatizados requieren de ciertos conocimientos del sujeto para su adecuada realización, lo cual puede ser un problema para aquellos que no están familiarizados con el ordenador. En consecuencia, los sujetos con destrezas informáticas son de esperar que respondan consistentemente a todos los ítems y en concordancia con la habilidad o rasgo que el test está midiendo. Los sujetos con pocos o ningún conocimiento en informática, según Drasgow, Levine y Zickar (1996), se ajustan a un proceso bietápico para dar respuesta a los ítems: en la primera etapa se esfuerzan por comprender y adaptarse al instrumento de medida, por lo que las respuestas a los n_1 primeros ítems se pueden considerar como respuestas dadas al azar; en la segunda etapa, los sujetos se han adaptado al proceso de evaluación y sus respuestas a los n_2 ítems siguientes serían acordes con un patrón normal.

4. *Copia de respuestas en un test desconocido.* Es el patrón atípico más usual y más difícil de identificar. Los sujetos que han contestado a un test no informan del número de ítems que han copiado, por lo que el investigador para detectarlos debe tener en cuenta las $\binom{n}{m}$ formas de copiar m ítems

de un total de n, a las que les corresponden la misma probabilidad de ocurrencia.

5. *Modelo de respuestas para espurias bajas de Levine y Rubin.* Este modelo aporta una descripción razonable de anomalías en la medida. Si el test no puede ser contestado en el cuaderno de los ítems, sino que las respuestas deben marcarse en una hoja de codificación, el sujeto puede cometer el fallo de no contestar a un ítem y codificar la respuesta del ítem siguiente en la casilla asignada al ítem anterior. El resultado es que el examinador desconoce cuáles de las respuestas dadas se ajustan al modelo normal y cuáles a un modelo de respuesta atípico. Entonces, sería adecuado considerar series de m ítems y asignarles a todos la misma probabilidad de que sean atípicos.

Otro estadístico óptimo de ajuste que contempla un algoritmo para crear modelos para respuestas atípicas aplicado al modelo de Rasch fue elaborado por Klauer (1995). Este autor definió modelos alternativos a las generalizaciones biparamétricas del modelo de Rasch que contienen un parámetro específico de persona (η_i) añadido al parámetro de habilidad θ_i. El parámetro η_i describe la magnitud y dirección de las desviaciones del modelo original. La fórmula general del modelo alternativo es:

$$P(\mathbf{U}_i|\theta,\eta) = \mu(\theta,\eta)h(\mathbf{U}_i)\exp[\eta T(\mathbf{U}_i) + \theta R(\mathbf{U}_i)] \tag{3.54}$$

donde

$$\begin{aligned} \mu(\theta,\eta) &= \prod_{j=1}^{n}[1+\exp(\theta_i-\beta_j)]^{-1} \\ h(\mathbf{U}_i) &= \exp\left(-\sum_{j=1}^{n} x_j\beta_j\right) \end{aligned}$$

$\mathbf{U}_i$ es el patrón de respuestas del sujeto i a los n ítems;

$R(\mathbf{U}_i)$ la variable fila de puntuaciones;

$T(\mathbf{U}_i)$ un estadístico que depende del modelo alternativo escogido.

Al modelo alternativo se recurre cuando:

- La habilidad del sujeto no es invariante al dividir el test en subtests, i.e., hay tantos parámetros de habilidad como subtests. Si cada subtests

tiene $j = 1, 2, \ldots, k$ ítems y denotando al parámetro de habilidad de cada subtest θ_s $(s = 1, 2, \ldots, S)$, el modelo alternativo es:

$$P(\mathbf{U}_i|\theta, \eta) = \mu(\theta, \eta)h(\mathbf{U}_i)\exp[\eta R_1(\mathbf{U}_i) + \theta R(\mathbf{U}_i)]$$

Si $\eta = 0$ el modelo que resulta es el de Rasch.

- Hay un parámetro de discriminación del ítem específico para el sujeto:

$$\mu(\theta, \eta) = \prod_{j=1}^{k}\{1 + \exp[\eta(\theta - \beta_j)]\}^{-1}$$

Si $\eta = 1$ se obtiene el modelo de Rasch; si $0 < \eta < 1$, la dificultad del ítem decrece; si $\eta = 0$, todos los ítems tienen la misma dificultad; cuando $\eta > 1$ la dificultad del ítem se incrementa en relación con una escala Guttman; si $\eta < 0$, los ítems están colocados en orden inverso de dificultad. El modelo resultante es:

$$P(\mathbf{U}_i|\theta, \eta) = \mu(\theta, \eta)\exp[\eta M(\mathbf{U}_i) + \theta R(\mathbf{U}_i)]$$

donde $M(\mathbf{U}_i) = -\sum_{j=1}^{n}\beta_j x_j$.

- Si se viola el supuesto de independencia local, el modelo alternativo es:

$$P(\mathbf{U}_i|\theta, \eta) = \mu(\theta, \eta)h(\mathbf{U}_i)\exp\left[\eta\sum_{j=1}^{n-1}X_j(\mathbf{U}_i)X_{j+1}(\mathbf{U}_i) + \theta R(\mathbf{U}_i)\right]$$

Si $\eta = 0$ el modelo coincide con el de Rasch; si $\eta > 0$, el sujeto acierta el ítem j lo que le facilita el acertar al ítem $j + 1$; si por el contrario $\eta < 0$, el acertar al ítem j reduce la probabilidad de acertar el ítem $j + 1$.

Para probar la existencia del patrón atípico, Klauer (1995) define la hipótesis nula $H_0 : \eta = \eta_0$ para un cierto valor de η_0 frente a la alternativa $H_1 : \eta \neq \eta_0$. Además, especifica una función aleatorizada ϕ para contrastar la significación estadística de H_0, tal que:

$$P(\text{rechazar } H_0|\theta, \eta_0) = E(\phi|\theta, \eta_0) \leq \alpha$$

3.4.7. Los estadísticos basados en la función de verosimilitud

Levine y Rubin (1979) barajaron la posibilidad de que la existencia de más de un rasgo latente por sujeto en la ejecución del test fuera la causa de patrones de respuesta atípicos. Este sería el caso en el que un sujeto de baja habilidad copia las respuestas de un compañero cercano a él y con nivel de habilidad superior. Por lo tanto, el patrón de respuestas del sujeto de baja habilidad contendría respuestas representativas de su nivel (respuestas acertadas en los ítems más fáciles) y respuestas representativas del nivel de habilidad del sujeto del que se copió (respuestas acertadas a los ítems de dificultad mayor a su nivel de habilidad). Por la supuesta presencia de un parámetro de habilidad variable, los modelos de TRI clásicos –o como los autores denominaron *modelos estándar o constantes*– no serían válidos para evaluar el grado de acuerdo entre un patrón de respuestas y el nivel de habilidad del sujeto, ya que para ellos la habilidad de un sujeto es constante en todos los ítems del test (θ). Si se asume un modelo de respuesta que contemple valores de habilidad (θ_i) independientes para cada uno de los ítems del test, valores que en conjunto siguieran una distribución normal de media θ_0 y varianza σ^2, este modelo se ajustaría mejor al patrón de respuesta del sujeto que un modelo constante. A este modelo alternativo que permite valores de habilidad variables lo llamaron *modelo gaussiano*.

En el modelo gaussiano, la probabilidad de acertar el ítem j es $P_j(\theta_i)$ y no $P_j(\theta)$ como en los modelos clásicos. La relación entre ambos es que el modelo constante es el caso límite del gaussiano cuando $\sigma^2 = 0$. La probabilidad condicional de que un sujeto extraído aleatoriamente de la muestra y con habilidad θ_i tenga un patrón de respuesta $\mathbf{U} = (u_1, u_2, \ldots, u_n)$ según el modelo gaussiano es:

$$\begin{aligned} f(\mathbf{U}|\theta_0, \sigma^2) &= \int \ldots \int \prod_{j=1}^{n} P_j(\theta_i)^{u_j} Q_i(\theta_i)^{1-u_j} \sigma^{-1} \phi \left[\frac{\theta_i - \theta_0}{\sigma} \right] d\theta_1 \ldots d\theta_n \\ &= \prod_{j=1}^{n} \int P_j(t)^{u_i} Q_j(t)^{1-u_j} \sigma^{-1} \phi \left[\frac{t - \theta_0}{\sigma} \right] dt \end{aligned} \tag{3.55}$$

donde $\phi(x) = (2\pi)^{1/2} \exp^{-x^2/2}$ es la función de densidad gaussiana y $Q_j = (1 - P_j)$.

A partir de este modelo, Levine y Rubin (1979) propusieron tres estadísticos o índices de medición apropiada para detectar aquellos patrones de respuesta que no estaban en concordancia con los niveles de habilidad del sujeto. Estos índices serían una medida de la bondad de ajuste de un modelo psicométrico

al patrón de respuestas individual ítem-por-ítem. Si los valores de los índices son elevados es porque hay acuerdo entre las respuestas dadas y el nivel de habilidad del sujeto; si por el contrario los valores de los índices son bajos, el patrón de respuesta no está reflejando el nivel de habilidad del sujeto. Los tres índices de medición apropiada fueron:

1. *La probabilidad marginal de un patrón de respuesta.* Se obtiene calculando el promedio de la distribución de habilidad en la población. Si la probabilidad marginal de un patrón es baja, se debe a que el patrón es poco probable bien para sujetos que con alto nivel de habilidad fallan ítems fáciles, o bien para sujetos de baja habilidad que aciertan ítems difíciles. Sea $\mathbf{U}^*$ un particular vector de puntuaciones observadas. Si se especifica una función de densidad de habilidad $g(\theta)$, aunque por lo general es una función desconocida, para obtener la probabilidad marginal de dicho vector de respuestas hay que resolver la siguiente integral para el modelo constante:

$$\int_{-\infty}^{+\infty} f(\mathbf{U}^*|\theta)g(\theta)\,d\theta$$

La función de densidad $g(\theta)$ resume la información de la habilidad de la muestra antes de que ésta responda al test. Sin embargo, si se prescinde de dicha información y se tiene sólo en cuenta la estimación de la habilidad, se podría sustituir $g(\theta)$ por otra función de densidad, $\tilde{g}(\theta)$, centrada en torno a $\hat{\theta}$ y cuya varianza tiende a 0. El valor $\hat{\theta}$ es el estimador máximo-verosímil de la habilidad tras maximizar $f(\mathbf{U}^*|\theta)$. Entonces, la integral:

$$\int_{-\infty}^{+\infty} f(\mathbf{U}^*|\theta)\tilde{g}(\theta)\,d\theta$$

converge a $f(\mathbf{U}^*|\hat{\theta})$. El logaritmo del máximo de esta función es el primer índice de medición apropiada, denotado l_0:

$$l_0 = \ln f(\mathbf{U}^*|\hat{\theta}) \tag{3.56}$$

Un índice apropiado basado en la probabilidad marginal similar al de la Ecuación 3.56 es el estimador de la distribución de habilidad $g(\theta)$ ya sea a partir de la distribución observada $\hat{\theta}$, ya sea mediante métodos de observación verdadera. El camino más sencillo es suponer que la función de habilidad $\ln f(\mathbf{U}^*|\theta)$ es unimodal y simétrica en $\theta = \hat{\theta}$. Este índice es una aproximación a la derivada de segundo orden de $\ln f(\mathbf{U}^*|\theta)$:

$$l_0 + \frac{1}{2}(\theta - \hat{\theta})^2 l_2$$

donde l_2 es la derivada segunda del $\ln f(\mathbf{U}^*|\theta)$ en $\theta = \hat{\theta}$. Si la función de densidad de la habilidad se aproxima a la función de densidad normal tipificada, entonces la función la probabilidad marginal es:

$$\frac{1}{\sqrt{2\pi}} \int \exp^{l_0} \exp^{\frac{1}{2}(\theta-\hat{\theta})^2 l_2} \exp^{-\frac{1}{2}\theta^2} d\theta = \exp^{l_0} \exp^{\frac{1}{2}\hat{\theta}^2(\frac{l_2}{1-l_2})} (1 - l_2)^{-\frac{1}{2}}$$

o su equivalente:

$$l_0 + \frac{1}{2}\hat{\theta}^2(\frac{l_2}{1 - l_2}) - \frac{1}{2}\ln(1 - l_2)$$

2. *Razones de probabilidad*. Este índice se calcula mediante el logaritmo de la función de verosimilitud, con el cual se sabría el grado de ajuste logrado. Se supone que el modelo gaussiano es el que mejor se explica los datos porque permite que la habilidad varíe a través de los ítems. Para ello, compara el logaritmo de la función de probabilidad del modelo constante (l_0) y su generalización según el supuesto gaussiano (l_n):

$$LR = l_n - l_0 \tag{3.57}$$

donde

$$\begin{aligned} l_0 &= \ln \max_{\theta} P(\mathbf{U}^*|\theta) = \ln P(\mathbf{U}|\hat{\theta}) \\ l_n &= \ln \max_{\theta_i} P(\mathbf{U}^*|\theta_0, \sigma^2) \end{aligned}$$

Cuanto mayor sea LR mejor es el ajuste del patrón de respuestas al modelo gaussiano.

3. *Estimación del grado de anormalidad del patrón*, $\hat{\sigma}^2$. El estimador de la varianza de la distribución de habilidad θ_i en el modelo gaussiano es próximo a 0 si el patrón de respuesta del sujeto es consecuencia de un único nivel de habilidad constante para todos los ítems del test. Si existen otros factores intervinientes en las respuestas del mismo (e.g., respuestas al azar o mala comprensión de las instrucciones de respuesta al test) $\hat{\sigma}$ se aleja de 0 y cuanto mayor sea la desviación más atípico es el patrón de respuestas.

Drasgow (1982) amplió el trabajo de Levine y Rubin (1979) al cuestionarse cuál sería el MRI que se debería ajustar a los datos. Si bien el modelo de 3-p fue el primero que se empleó para describir la probabilidad de respuesta en tests de elección múltiple, dicho modelo ha sido criticado por las dificultades que conlleva la estimación de los parámetros de los ítems con MV, concretamente la del parámetro de pseudo-azar (Lord, 1968, p. 1014). Por esta razón, Drasgow plantea la posibilidad de poder emplear el modelo de Rasch en tests de elección múltiple con el fin de obtener una óptima estimación de los parámetros y, en consecuencia, una mejoría de la tasa de identificaciones de patrones atípicos. Junto con esta hipótesis sobre el modelo de respuesta, este autor modificó el estadístico l_0 para paliar los efectos de la longitud del tests y del número de ítems omitidos por los que se dejaba afectar. El nuevo índice es la media geométrica de l_0:

$$l_g = \exp^{l_0/m} \tag{3.58}$$

donde m es el número de ítems contestados en el test.

Hasta el momento, el índice de medición apropiada basado en el logaritmo de la función de verosimilitud más empleado era l_0, debido al alto porcentaje de identificación de patrones atípicos. Sin embargo, l_0 presentaba dos inconvenientes:

- No estaba estandarizado, lo cual implica que clasificar un patrón de respuesta como normal o atípico dependa de θ.
- La distribución de l_0 era desconocida, pero ésta es necesaria para poder probar la hipótesis nula de que el patrón de respuestas es normal.

Para solucionar estos dos problemas Drasgow, Levine y Williams (1985) propusieron la versión estandariza del estadístico l_0. El objetivo era poder comparar los valores del índice de medición apropiada de sujetos en distintos niveles de habilidad. Su estudio se centró de manera especial en aquellos patrones atípicos que contenían respuestas omitidas. La hipótesis planteada fue que un bajo índice de medición apropiada para un patrón con un sustancial porcentaje de respuestas omitidas, sería menos indicativo de la atipicidad del mismo que un valor más elevado del índice de otro sujeto con un diferente patrón con respuestas omitidas. En su trabajo, presentaron una generalización del modelo de 3-p al que denominaron *modelo del histograma* cuyos supuestos son:

- La unidimensionalidad del espacio latente.

- Es un modelo de habilidad constante con independencia local.
- Las respuestas a los ítems son condicionalmente independientes.

Sea $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n)$ el vector aleatorio de opciones de respuesta y $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ un vector de constantes que indica la opción elegida del ítem. Se asume que para una variable unidimensional aleatoria de habilidad (θ):

$$P(\mathbf{V}_1 = v_1 \bigcap \mathbf{V}_2 = v_2, \bigcap \ldots \bigcap \mathbf{V}_n = v_n | \theta = t) = \prod_{j=1}^{n} P(\mathbf{V}_j = v_j | \theta = t)$$

para todo $\mathbf{v}$ y t real. Además, si v_j^* es la opción de respuesta correcta del ítem j, entonces para algunos a_j, b_j y c_j y para todo t:

$$P(\mathbf{V}_j = v_j^* | \theta = t) = c_j + \frac{1 - c_j}{1 + \exp[-a_j(t - b_j)]} \tag{3.59}$$

Este modelo no es un modelo verosímil para describir el comportamiento de ejecución de un test, pero sí se puede considerar como un modelo descriptivo para los datos de un test que podría apoyar la extensión de las técnicas de medición apropiada a los ítems politómicos con un alto porcentaje de respuestas omitidas. Las funciones de la forma:

$$P_{jl}(t) = P(\text{la opción } l \text{ es elegida en el ítem } j | \theta = t)$$

para $l = 1, 2, \ldots, A + 1$, generalizan las FRI y son denominadas *funciones de opción de respuesta*. Entonces, según el modelo del histograma, la probabilidad de un patrón de respuesta $\mathbf{V} = \mathbf{v}$ en una muestra de sujetos con habilidad $\theta = t$ es:

$$P(\mathbf{V} = \mathbf{v} | \theta = t) = \prod_{j=1}^{n} \sum_{l=1}^{A+1} \delta_l(v_j) P_{jl}(t) \tag{3.60}$$

donde las primeras $A + 1$ integrales positivas son utilizadas como las puntuaciones para la opción elegida; $\delta(k) = 1$ si $k = l$ y $\delta(k) = 0$ en cualquier otro caso. La Ecuación 3.60 ha sido utilizada para estimar la habilidad por MV con ítems politómicos. La habilidad estimada según un modelo dicotómico $(\hat{\theta}_d)$ se obtiene al maximizar la función de verosimilitud:

$$L = \prod_{j=1}^{n} [u_j P_j(t) + (1 - u_j) Q_j(t)]$$

donde u_j es 1 ó 0 según sea v_j la opción correcta o la errónea, respectivamente, $P_j(t)$ es la probabilidad de acertar el ítem según la Ecuación 3.59 y $Q_j(t) = 1 - P_j(t)$. La función lineal de las puntuaciones a los ítems es:

$$l_0 = \sum_{j=1}^{n} [u_j \ln P_j(\hat{\theta}_d) + (1 - u_j) \ln Q_j(\hat{\theta}_d)] \tag{3.61}$$

siendo l_0 es el máximo del logaritmo de la función de verosimilitud en un modelo de ítems dicotómicos.

Para ítems politómicos, l_0 en el modelo del histograma es el logaritmo de la función de verosimilitud:

$$\max_{\theta} \sum_{j=1}^{n} \sum_{l=1}^{A+1} \delta(v_j) \ln P_{jl}(\theta)$$

Sin embargo, la primera derivada de esta función de verosimilitud no es continua, lo cual dificulta encontrar su máximo y a lo que Drasgow, Levine y Williams (1985) dieron solución al incluir la habilidad estimada por MV desde un modelo dicotómico ($\hat{\theta}_d$) en la función de verosimilitud al modelo del histograma, cuya expresión final es:

$$l_{0.h} = \sum_{j=1}^{n} \sum_{l=1}^{A+1} \delta(v_j) \ln P_{jl}(\hat{\theta}_d) \tag{3.62}$$

Para reducir la dependencia que poseen las Ecuaciones 3.61 y 3.62 de la habilidad, los estandarizaron y definieron dos nuevos estadísticos:

$$l_z = \frac{l_0 - E[l_0]}{\sigma[l_0]} \tag{3.63}$$

$$l_{z.h} = \frac{l_{0.h} - E_h[\hat{\theta}_d]}{\sigma_h[\hat{\theta}_d]} \tag{3.64}$$

donde $E[l_0]$, $\sigma[l_0]$, $E_h[\hat{\theta}_d]$ y $\sigma_h[\hat{\theta}_d]$ son las medias condicionales y las desviaciones típicas del modelo de 3-p y del modelo del histograma. La constante $E[l_0]$ es

el valor esperado condicional de la variable aleatoria $l_0(t)$ calculada según el modelo de 3-p para ítems dicotómicos con parámetro estimado de habilidad $\hat{\theta}_d = t$ y vector de respuestas de un sujeto $\mathbf{U}_j$:

$$l_0(t) = \sum_{j=1}^{n} [\mathbf{U}_j \ln P_j(t) + (1 - \mathbf{U}_j) \ln Q_j(t)] \tag{3.65}$$

Entonces,

$$\begin{aligned} E[l_0] &= E[l_0(t)|\hat{\theta}_d = t] = \sum_{j=1}^{n} [P_j(t) \ln P_j(t) + Q_j(t) \ln Q_j(t)] \\ \sigma^2[l_0] &= Var[l_0(t)|\hat{\theta}_d = t] = \sum_{j=1}^{n} P_j(t) Q_j(t) \left[\ln \frac{P_j(t)}{Q_j(t)}\right]^2 \end{aligned}$$

y para el modelo del histograma:

$$\begin{aligned} E_h(t) &= E\left[\sum_{j=1}^{n} \sum_{l=1}^{A+1} \delta_l(\mathbf{V}_j) \ln P_{jl}(t) | \hat{\theta}_d = t\right] \\ &= \sum_{j=1}^{n} \sum_{l=1}^{A+1} P_{jl}(t) \ln P_{jl}(t) \\ \sigma_h^2(t) &= Var\left[\sum_{j=1}^{n} \sum_{l=1}^{A+1} \delta_l(\mathbf{V}_j) \ln P_{jl}(t) | \hat{\theta}_d = t\right] \\ &= \sum_{j=1}^{n} \left[\sum_{l} \sum_{k} P_{jl}(t) P_{jk}(t) \ln P_{jl}(t) \ln \frac{P_{jl}(t)}{P_{jk}(t)}\right] \end{aligned}$$

La estandarización redujo sustancialmente la dependencia los índices l_0 y $l_{0.h}$ de la habilidad y las distribuciones de éstos eran más variables que las de l_z y $l_{z.h}$ a través de los niveles de habilidad.

Molenaar y Hoijtink (1990, 1996) cuestionaron la capacidad del estadístico l_z para detectar patrones de respuesta atípicos cuando se trabaja con el modelo de Rasch, con parámetros de habilidad estimados por MV y con tests cortos (50 ítems o menos) ya que, bajo estas condiciones, la distribución de l_z se aleja de la normalidad. Sea el patrón de respuestas de un sujeto en un test de n ítems de respuesta dicotómica $\mathbf{U} = (u_1, u_2, \ldots, u_n)$ ajustado al modelo de Rasch, entonces la puntuación total $r = \sum_{j=1}^{n} u_j$ es un estimador suficiente

de la habilidad. El estadístico propuesto por los autores para evaluar el ajuste del patrón de respuestas fue:

$$l(\mathbf{U}) = d_0 + M(\mathbf{U}) \tag{3.66}$$

cuyos sumandos son:

$$\begin{aligned} M(\mathbf{U}) &= -\sum_{j=1}^{n} b_j x_j \\ d_0 &= -\sum_{j=1}^{n} \ln[1+\exp(\hat{\theta}-b_j)] + r\hat{\theta} \end{aligned} \tag{3.67}$$

donde $\hat{\theta}$ es la estimación de la habilidad por MV para la puntuación total r; d_0 es independiente de $\mathbf{U}$; $l(\mathbf{U})$ y $M(\mathbf{U})$ tienen el mismo orden que $\mathbf{U}$ y sus medias, varianzas e índices de sesgo y curtosis son iguales, ya que sólo difieren en la constante aditiva d_0. Con el fin de simplificar los cálculos, Molenaar y Hoijtink (1990, 1996) escogieron $M(\mathbf{U})$ como detector de patrones de respuesta atípicos cuya prueba estadística para evaluar la hipótesis nula de ajuste del patrón cuando el test tiene pocos ítems ($n \leq 20$) requiere calcular la probabilidad de excedencia (*Probability of Exceedance*, PE):

$$PE(\mathbf{U}) \approx P\left(\chi^2_\nu > \frac{|M(\mathbf{U})+a|}{b}\right) \tag{3.68}$$

donde

$$\begin{aligned} \nu &= \frac{8}{\gamma^2} \\ \gamma &= \frac{E[M(\mathbf{U})^3] - 3E[M(\mathbf{U})^2]E[M(\mathbf{U})] + 2\{E[M(\mathbf{U})]\}^3}{\sigma^3} \\ \sigma^2 &= E[M(\mathbf{U})^2] - E[M(\mathbf{U})]^2 \\ E[M(\mathbf{U})] &= -\sum_{j=1}^{n} b_j P_j(\hat{\theta}) \\ E[M(\mathbf{U})^2] &= -\sum_{j=1}^{n} b_j^2 P_j(\hat{\theta}) + 2\sum_{j<k} b_j b_k P_{jk}(\hat{\theta}) \\ E[M(\mathbf{U})^3] &= -\sum_{j=1}^{n} b_j^3 P_j(\hat{\theta}) - 3\sum_{j\neq k} b_j^2 b_k P_{jk}(\hat{\theta}) - 6\sum_{h<j<k} b_h b_j b_k P_{hjk}(\hat{\theta}) \\ b &= \left\{\frac{E[M(\mathbf{U})^2]}{2\nu}\right\}^{1/2} \end{aligned}$$

$$a = -b\nu - E[M(\mathbf{U})]$$

Para tests largos se deben emplear los métodos Montecarlo, sobre todo cuando el rango de variabilidad del parámetro de dificultad es amplio, ya que para resolver la Ecuación 3.68 hay que recurrir a funciones simétricas. Bedrick (1997), y Liou y Chang (1992) desarrollaron sendos algoritmos para facilitar el cálculo de los momentos de la distribución de la dicha ecuación y obtener $PE(\mathbf{U})$. Liou y Chang emplearon el método de Mehta y Patel (1990) válido para tests muy largos; Bedrick recurrió a las aproximaciones Edgeworth y *saddlepoint* (Pierce y Peters, 1992) más eficaces que los algoritmos anteriores cuando tanto el parámetro de habilidad como los parámetros de dificultad de los ítems son desconocidos.

Snijders (2001), partiendo de los estudios y de la misma idea que Molenaar y Hoijtink (1990, 1996), ha propuesto un estadístico de medición apropiada para emplearlo cuando el parámetro de habilidad es desconocido y debe ser estimado, ampliando su utilidad a los modelos de 2-p, 3-p y de respuesta politómica:

$$l_z^* = \frac{W(\hat{\theta}_i)}{\sqrt{n}\,\tau(\hat{\theta}_i)} \tag{3.69}$$

donde n es el número de ítems del test y:

$$\begin{aligned}
W(\hat{\theta}_i) &= l_0 - E(l_0) + c(\hat{\theta}_i) r_0(\hat{\theta}_i) \\
c(\hat{\theta}_i) &= \frac{\sum_{j=1}^{n} a_j (\hat{\theta}_i - b_j) P_j'(\hat{\theta}_i)}{\sum_{j=1}^{n} a_j P_j'(\hat{\theta}_i)} \\
r_0(\hat{\theta}_i) &= \frac{J(\hat{\theta}_i)}{2I(\hat{\theta}_i)} \\
\tau_j^2 &= \frac{1}{n} \sum_{j=1}^{n} [a_j(\hat{\theta}_i - b_j) - a_j c(\hat{\theta}_i)] P_j(\hat{\theta}_i)[1 - P_j(\hat{\theta}_i)] \\
I(\hat{\theta}_i) &= \sum_{j=1}^{n} \frac{P_j'^2(\hat{\theta}_i)}{P_j(\hat{\theta}_i)[1 - P_j(\hat{\theta}_i)]} \\
J(\hat{\theta}_i) &= \sum_{j=1}^{n} \frac{P_j'(\hat{\theta}_i) P_j''(\hat{\theta}_i)}{P_j(\hat{\theta}_i)[1 - P_j(\hat{\theta}_i)]}
\end{aligned}$$

Para calcular l_z^*, $\hat{\theta}$ es el estimador máximo-verosímil ponderado de θ, para el cual:

$$\frac{J(\hat{\theta}_i)}{2I(\hat{\theta}_i)} + \sum_{j=1}^{n}[u_j - P_j(\hat{\theta}_i)]\frac{P'_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)[1 - P_j(\hat{\theta}_i)]} = 0$$

El estadístico l_z^* sigue una distribución próxima a la normal tipificada, ya que su varianza se aleja de 1 cuando aumenta el número de ítems del test y, en general, es sesgada negativa y leptocúrtica.

Drasgow, Levine y McLaughlin (1991) extendieron el uso del estadístico l_z al ámbito de los tests *multiunidimensionales*. Un test multiunidimensional es un test multidimensional cuyos ítems cumplen el supuesto de independencia local y está dividido en subtests unidimensionales. En base a esta definición, sea un test con S subtests; $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_S$ son las puntuaciones totales en cada subtest y $\mathbf{U}_s = (u_{s1}, u_{s2}, \ldots, u_{sn_s})$ es el vector de respuestas a los ítems en un subtest. Para cada subtest existe una determinada habilidad $(\theta_1, \theta_2, \ldots, \theta_S)$. El cálculo de la probabilidad de obtener un vector de vectores de respuestas $\mathbf{U} = (u_{11}, \ldots, u_{1n_1}, u_{21}, \ldots, u_{2n_2}, \ldots, u_{S1}, \ldots, u_{Sn_s})$ igual a alguno formado por 0s y 1s $(\mathbf{U}^*)$ dado que $\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2, \ldots, \theta_S = \hat{\theta}_S$, se lleva a cabo en dos fases: en la primera, por el supuesto de independencia local de los ítems se verifica que:

$$P(\mathbf{U} = \mathbf{U}^*|\theta = \hat{\theta}) = \prod_{s=1}^{S}\prod_{j=1}^{n_s} P(u_{sj} = u_{sj}^*|\theta = \hat{\theta})$$

En la segunda fase, por el supuesto de unidimensionalidad de cada uno de los subtests se cumple que:

$$P(\mathbf{U} = \mathbf{U}^*|\theta = \hat{\theta}) = \prod_{s=1}^{S}\prod_{j=1}^{n_s} P(u_{sj} = u_{sj}^*|\theta_s = \hat{\theta}_s)$$

Entonces, la multiunidimensionalidad implica que:

$$P[f_1(\mathbf{U}_1) = f_1(\mathbf{U}_1^*), f_2(\mathbf{U}_2) = f_2(\mathbf{U}_2^*), \ldots, f_S(\mathbf{U}_S) = f_S(\mathbf{U}_S^*)|\theta = \hat{\theta}] = \prod_{s=1}^{S} P[f_s(\mathbf{U}_s) = f_s(\mathbf{U}_s^*)|\theta_s = \hat{\theta}_s]$$

donde $f_s(\mathbf{U}_s)$ son funciones de variables independientes y, por lo tanto, las S funciones son también independientes. Cada una de ellas es un índice de medición apropiada unidimensional. El estadístico l_0 para tests multiunidimensionales queda definido en la expresión:

$$
\begin{aligned}
l_0(\mathbf{U}^*) &= \ln \max_t P(\mathbf{U} = \mathbf{U}^* | \theta = \mathbf{t}) \\
&= \ln \max_t \prod_{s=1}^{S} P(\mathbf{U} = \mathbf{U}^* | \theta_s = t_s) \\
&= \ln \prod_{s=1}^{S} \max_{t_s} P(\mathbf{U} = \mathbf{U}^* | \theta_s = t_s) = \sum_{s=1}^{S} l_0(\mathbf{U}_s^*)
\end{aligned} \tag{3.70}
$$

El estadístico l_0 estandarizado es:

$$
l_{zm} = \frac{l_0(\mathbf{U}^*) - E[l_0(\mathbf{U})|\theta = \hat{\theta}]}{\sigma[l_0(\mathbf{U})|\theta = \hat{\theta}]} \tag{3.71}
$$

La media condicionada y la desviación típica condicionada de l_0 para tests multiunidimensionales se expresan como funciones de medias y desviaciones típicas condicionadas unidimensionales:

$$
\begin{aligned}
E[l_0(\mathbf{U})|\theta = \hat{\theta}] &= \sum_{s=1}^{S} E[l_0(u_s)|\theta_s = \hat{\theta}_s] \\
\sigma^2[l_0(\mathbf{U})|\theta = \hat{\theta}] &= \sum_{s=1}^{S} \sigma^2[l_0(u_s)|\theta_s = \hat{\theta}_s]
\end{aligned}
$$

3.4.8. El estadístico ω de Wollack

Wollack (1997) retoma la idea sobre la que Frary *et al.* (1977) elaboraron el índice g_2 (sección 2.2). Cuando los datos son explicados por el modelo de respuesta nominal de Bock (1972), el estadístico ω de Wollack pretende detectar patrones con respuestas atípicas, en concreto, patrones con respuestas copiadas. Por el modelo de Bock, la probabilidad de que un sujeto de habilidad θ_i escoja la opción l del ítem j es:

$$
P_{jl}(\theta_i) = \frac{\exp(\zeta_{jl} + \lambda_{jl}\theta_i)}{\sum_{L=1}^{m} \exp(\zeta_{jl} + \lambda_{jl}\theta_i)}
$$

donde ζ_{jl} y λ_{jl} son los parámetros de dificultad y discriminación de la opción l del ítem j ($L = 1, 2, \ldots, m$).

A todos los sujetos que contestan el test se les considera posibles copiadores (C) y sus respuestas son contrastadas con los patrones de sujetos *fuente* (F)

según la disposición de asientos en el momento de la realización del test. Para cada par de sujetos cuyas respuestas al ítem j son u_{jC} la del sujeto C y u_{jF} la del sujeto F, el número de ítems contestados con la misma opción es:

$$h_{CF} = \sum_{j=1}^{n} I(u_{jC} = u_{iF})$$

donde $I(u_{jC} = u_{iF}) = 1$ si los dos sujetos eligen la misma opción e $I(u_{jC} = u_{iF}) = 0$ en cualquier otro caso. La probabilidad de que la respuesta al ítem sea la misma se obtiene calculando la probabilidad de que el sujeto C escoja la respuesta dada por el sujeto F. Este valor esperado es:

$$E(h_{CF}|\theta_C, \mathbf{U}_F, \xi) = E\left[\sum_{j=1}^{n} I(u_{jC} = u_{iF}|\theta_C, \mathbf{U}_F, \xi)\right] = \sum_{j=1}^{n} P(u_{jC} = u_{iF}|\theta_C, \mathbf{U}_F, \xi)$$

en donde ξ representa a la matriz que contiene los valores de los parámetros de los ítems. Entonces, el estadístico ω compara el número de ítems con la misma respuesta y el número esperado a consecuencia del azar:

$$\begin{aligned} \omega &= \frac{h_{CF} - E(h_{CF}|\theta_C, \mathbf{U}_F, \xi)}{\sigma_{h_{CF}} - E(h_{CF}|\theta_C, \mathbf{U}_F, \xi)} \\ &= \frac{h_{CF} - \sum_{j=1}^{n} P(h_{CF}|\theta_C, \mathbf{U}_F, \xi)}{\sqrt{\sum_{j=1}^{n} P(h_{CF}|\theta_C, \mathbf{U}_F, \xi)[1 - P(h_{CF}|\theta_C, \mathbf{U}_F, \xi)]}} \end{aligned} \tag{3.72}$$

El estadístico ω sigue una distribución normal estandarizada ya que, asumiendo que las respuestas a los ítems son localmente independientes, h_{CF} es la suma de n ensayos de Bernoulli independientes y, por el teorema central del límite, la distribución de h_{CF} se aproxima a la normal cuando el número de ítems tiende a infinito.

Wollack y Cohen (1998) hicieron un estudio sobre la potencia y las tasas de error tipo I de ω cuando el parámetro de habilidad es desconocido y para calcular el índice hay que estimar θ. A raíz de los resultados, los autores concluyeron que la tasa de error tipo I no estuvo afectada al sustituir los parámetros verdaderos por sus estimadores y la potencia era casi la misma, dependiendo ésta del número de sujetos de la muestra. Wollack, Cohen y Serlin (2001) han comparado las tasas de error tipo I y la potencia de g_2 de Frary *et al.* (1977) y ω con dos criterios de comparación: pares de sujetos y familia de sujetos (*familywise*). Una familia conlleva un conjunto de comparaciones realizadas fijando un sujeto C, considerado como copiador potencial, frente a cada uno de

los sujetos F. Hay tantas familias como sujetos y el número de comparaciones dentro de una de ellas depende del número de sujetos F sentados alrededor del sujeto C. Las tasas de error tipo I y la potencia de ω tomando como criterio de comparación una familia han sido más aceptables que comparando pares de sujetos.

3.4.9. Los estadísticos para tests adaptativos informatizados (TAI)

Debido al creciente uso de los TAI o CAT (*Computerized Adaptive Testing*), Bradlow, Weiss y Cho (1998) propusieron un estadístico *outlier* para identificar diferentes clases de patrones atípicos en este tipo de tests. Como su nombre indica, los TAI son tests adaptados a las respuestas del sujeto, i.e., se van construyendo en función de la respuesta que dé el sujeto a un ítem (acierto o fallo) procedente de un banco de ítems. Después de contestar a un ítem se presentará el siguiente con mayor o menor grado de dificultad al anterior e igualado, tanto como sea posible, al nivel de habilidad estimado del sujeto; para que el test cese, el investigador fija un criterio de terminación. En los TAI, los ítems son seleccionados para maximizar la información acerca de la habilidad del sujeto en cada ítem administrado, por lo que, en un diseño TAI perfecto siempre $P_j(\theta_i) = 0{,}5$, es decir, $b_j \approx \theta_i$ y al concluir el test el promedio de dificultad de los ítems es aproximadamente igual al nivel de habilidad del sujeto. Bradlow *et al.* estaban interesados en detectar sujetos con patrones outlier multivariantes a partir del estadístico para patrones de respuesta univariante basado en la suma acumulada w_h:

$$w_h = \sum_{j=1}^{h} u_j$$

donde $\mathbf{U} = (u_1, u_2, \ldots, u_n)$ es el vector de respuestas dicotómicas de un sujeto de habilidad θ y representa el número de respuestas correctas antes de un ítem h de los $j = 1, 2, \ldots, n$ administrados; w_h sigue una distribución binomial mixta con media:

$$\mu_h = \sum_{j=1}^{h} P_j$$

y varianza:

$$\sigma_h^2 = \sum_{j=1}^{h} P_j(1 - P_j)$$

Después del ítem h, se consideraría que w_h es un patrón outlier:

$$\begin{aligned} \text{si } w_h &\leq \mu_h - z_{(1-\frac{\alpha}{2})}\sigma_h \\ \text{o si } w_h &\geq \mu_h + z_{(1-\frac{\alpha}{2})}\sigma_h \end{aligned}$$

donde $0 < \alpha/2 < 0{,}5$ y $z_{(1-\alpha/2)}$ es la puntuación bajo la curva normal tipificada (los autores recurren a esta distribución por su aproximación a la distribución binomial mixta). Para cada h, sea $\alpha_0(h)$ el valor α más pequeño que identifica a w_h como outlier; se define *probabilidad de cola nominal* de la respuesta del sujeto en h como $\alpha_0 = \text{mín}_h[\alpha_0(h)]$, en contraste con la *probabilidad de cola real* de que w_h sobrepase los límites del intervalo. Entonces, el estadístico para identificar patrones outlier multivariantes de Bradlow *et al.* (1998) es:

$$K(\theta) = \max_{h\in[1,n]} \frac{|w_h - \mu_h|}{\sigma_h} \tag{3.73}$$

i.e., la mayor desviación en valor absoluto estandarizado del número de respuestas correctas w_h. Como θ es desconocido, los autores aconsejaron emplear el estadístico $K = E[K(\theta)|\mathbf{U}]$. Sin embargo, no se podía asegurar que $K \sim N(0,1)$, por lo que intentaron solventar el problema de cuál es la prueba estadística de K implementando la desigualdad de Kolmogorov (Feller, 1968) y varios procesos de simulación para obtener la función de densidad $g(K)$ con objeto de normalizar el estadístico.

McLeod y Lewis (1999) idearon un estadístico de medición apropiada adecuado para identificar patrones de respuesta atípicos en los TAI y al que denotaron Z_c. Para estos autores, el patrón de respuesta atípico característico en los TAI es el resultante de la memorización de los ítems más difíciles del test que tienen mayor frecuencia de aparición. El recuerdo de estos ítems conlleva el conocimiento de las respuestas a los mismos, lo cual provocará un incremento en la puntuación total del sujeto en la próxima aplicación del test.

El estadístico Z_c es una extensión del índice de precaución $ECI4_z$ de Tatsuoka y Linn (1983)(Ecuación 3.36). Para calcularlo hay que dividir el total de ítems aplicados en tres partes una vez se haya estimado el parámetro de dificultad de los mismos. La primera parte, contendría a los ítems más fáciles del test que representarían un tercio del total; en la segunda partición estarían los ítems más difíciles del test y serían otro tercio del total; por último, los ítems restantes formarían el grupo de los ítems con dificultad media. La expresión de Z_c es:

$$Z_c = \frac{\overline{F[P(\hat{\theta}) - u_j]} - \overline{D[P(\hat{\theta}) - u_j]}}{\left\{ \frac{\sum_{f=1}^{n_F} P(\hat{\theta})[1-P(\hat{\theta})]}{n_F^2} + \frac{\sum_{d=1}^{n_D} P(\hat{\theta})[1-P(\hat{\theta})]}{n_D^2} \right\}^{1/2}} \tag{3.74}$$

donde

$\overline{F[P(\hat{\theta}) - u_j]}$ es la media residual de los ítems fáciles administrados;

$\overline{D[P(\hat{\theta}) - u_j]}$ la media residual de los ítems difíciles administrados;

n_F el número de ítems fáciles administrados ($f = 1, 2, \ldots, n_F$);

n_D el número de ítems difíciles administrados ($d = 1, 2, \ldots, n_F$).

El índice Z_c es una función del promedio de la media de los residuales de los ítems fáciles y difíciles administrados. Si Z_c es positivo, el sujeto no acierta los ítems fáciles y sí a los difíciles, con lo que se concluiría que el patrón es atípico. Un inconveniente para identificar patrones atípicos en los TAI cuando se emplea el índice Z_c es que son necesarias dos series de ítems, ya que el sujeto debe recibir al menos un ítem fácil y otro difícil, si no el índice no puede ser calculado.

Van Krimpen-Stoop y Meijer (1999, 2000) simularon una distribución, a la que connotaron T, para detectar patrones atípicos de respuesta en los TAI basándose en el estadístico l_z de Drasgow, Levine y Williams (1985), y en el índice l_z^* de Snijders (2001). Describieron dos métodos para definir la distribución T:

- *Método 1.* La distribución T es simulada a partir de una muestra de valores de habilidad extraídos de una distribución normal $N(0, 1)$, donde cada valor θ representa el nivel de habilidad de un sujeto que contesta un test. Con estos valores de habilidad se simulan las respuestas a los ítems de un TAI de acuerdo con un MRI. Para cada patrón de respuestas $T = t$, donde t es el valor observado de un índice de medición apropiada –los autores calculaban l_z o l_z^*–, representa la distribución simulada basada en las características del banco de ítems. El nivel de significación para la prueba de contraste de la hipótesis nula de que el patrón de respuestas es normal, es el mismo para todas las distribuciones T simuladas.
- *Método 2.* Se simula una distribución T para cada valor θ, es decir, dado θ se simula un alto número de patrones de respuesta (diseño de test) y $T = t$ se calcula para cada patrón. En este caso, la distribución simulada

se origina con las características del banco de ítems y el valor θ dado a priori. La prueba estadística emplea un nivel de significación para cada nivel de θ; si las distribuciones T son las mismas para todos los niveles de habilidad, este método queda reducido al método anterior. En los TAI no aparecen siempre las mismas secuencias de ítems y, por lo tanto, existen diferentes diseños de test: uno para cada sujeto de la muestra. Entonces, para obtener la distribución T hay que usar diseños de test probabilísticos los cuales asuman el proceso probabilístico de elección del ítem. En consecuencia, hay que simular un gran número de patrones de respuesta para cada θ con distintos diseños de test. Sea $\mathbf{d}$ un diseño de test y $T(\mathbf{U})$ el estadístico del patrón de respuestas observado $[\mathbf{U} = (u_1, u_2, \ldots, u_n)]$. Como en los TAI la selección de ítems depende de la respuesta a los ítems ya contestados, $\mathbf{d}$ depende de θ y de $\mathbf{U}$, y cualquier función de $\mathbf{U}$ está condicionada por $\mathbf{d}$ y θ. Con este método, la distribución de probabilidad de T es $f(T|\mathbf{d}, \theta)$ y para que sea independiente de $\mathbf{d}$ se multiplica por la función de probabilidad de éste, resultando:

$$f(T\mathbf{d}|\theta) = f(\mathbf{d}|\theta)f(T|\mathbf{d}, \theta)$$

Comparar los valores del estadístico a través de todos los sujetos con igual θ es difícil porque cada sujeto responde a un TAI distinto al de otro sujeto, pero comparar los niveles de significación observados del estadístico sí es factible. Estos valores se pueden obtener simulando la distribución condicionada del estadístico en θ bien con un diseño de test fijo, o bien con un diseño de test probabilístico. En ambos casos, la distribución de T se lograría replicando m veces el test. Una vez definida dicha distribución, el catalogar a un patrón de normal o de atípico se llevaría a cabo con un contraste de hipótesis.

Van Krimpen-Stoop y Meijer (2001) han vuelto a retomar el estadístico l_z^* de Snijders (2001) para utilizar el procedimiento de suma acumulada (*Cumulative Sum*, CUSUM) de Page (1954) en este entorno de evaluación. Sea $l_{z,s}^*$ el valor del estadístico l_z^* de cada uno de los $s = 1, 2, \ldots, S$ subtests en los que se ha dividido el test; los $l_{z,s}^*$ están distribuidos independiente e idénticamente. Sea M un valor de referencia predeterminado. Entonces, el proceso CUSUM bilateral se define mediante dos límites:

$$\begin{aligned} C_s^+ &= \text{máx}[0, (l_{z,s}^* - M) + C_{s-1}^+] \\ C_s^- &= \text{mín}[0, (l_{z,s}^* + M) + C_{s-1}^-] \end{aligned}$$

proceso que comienza siendo $C_0^+ = C_0^- = 0$ y $|l_{z,s}^*| > M$. Sea h el umbral tal que cese la rutina cuando, o bien $C_s^+ > h$, o bien $C_s^- < -h$, lo cual será indicativo de que el patrón de respuestas es atípico. Tanto M como h se establecen a partir del supuesto de que $l_{z,s}^*$ sigue la ley normal; lo habitual es que M sea 1/2 de la magnitud de cambio en la media de $l_{z,s}^*$ que el investigador esté dispuesto a aceptar y h se calcula con la aproximación de Siegmund (1985), conocidos el nivel de significación de la prueba y el valor M:

$$\frac{2}{\alpha} = \frac{\exp[2M(h + 1{,}166)] = 2M(h + 1{,}166) - 1}{2M^2}$$

Para que las tasas de error tipo I sean conservadoras, el proceso CUSUM puede ser implementado cuando el número de ítems de cada subtest sea mayor o igual que 10 y el test se divida en 3, 4 ó 5 subtests.

Bradlow y Weiss (2001) han listado una serie de estadísticos para identificar patrones outliers en los TAI y han propuesto métodos de normalización de los mismos para estandarizarlos.

3.5. Un índice de ajuste de persona según el Análisis de Estructura de Covarianza (AEC)

El Análisis de Estructura de Covarianza (AEC; Bielby y Hauser, 1977; Bollen, 1988; Duncan, 1975; Goldberger y Duncan, 1973; Jöreskog, 1974, 1977) es un método que busca relaciones teóricas entre una serie de constructos o variables latentes y variables observables o medibles que son indicadores de aquellas. Para encontrar esta relación es necesario, primero, estimar de los parámetros que definen el modelo y, en segundo lugar, valorar la bondad de ajuste del modelo o los modelos posibles a los datos observados. El interés principal del AEC ha sido evaluar el grado en el cual los parámetros del modelo reproducen la matriz de covarianzas observadas entre las variables medidas. Reise y Widaman (1999) apostaron por la ampliación de la valoración del ajuste al modelo a nivel de sujeto y así se podría saber la proporción de éstos que no se ajustan al modelo de AEC estimado. Un modelo de AEC es un modelo lineal que pretende, o bien enlazar las variables latentes y las observadas, o bien especificar la relación entre varias variables latentes o entre varias variables observadas. La fórmula general del modelo estructural es:

$$S \approx \Lambda\Phi\Lambda' + \Psi = \Sigma^*$$

donde

S es la matriz de covarianzas observadas de p variables;

Σ^* la matriz de covarianzas reproducida;

Λ una matriz $g \times m$ de cargas de las $g = 1, 2, \ldots, p$ variables sobre $m = 1, 2, \ldots, r$ factores;

Φ una matriz de covarianzas $m \times m$ entre las variables latentes;

Ψ una matriz diagonal $g \times g$ de varianzas de factor único.

Para estimar los parámetros de este modelo se suele recurrir al método de MV de una función de ajuste asumiendo, previamente, la normalidad multivariante. Para evaluar el ajuste del sujeto al modelo, Reise y Widaman (1999) propusieron un índice basado en el logaritmo de la función de máxima verosimilitud, cuya argumentación es similar a la de l_0 de Levine y Rubin (1979):

$$P_L = -\frac{1}{2}[p \ln(2\pi) + \ln|\Sigma^*| + (\mathbf{U}_i - M)\Sigma^{*-1}(\mathbf{U}_i - M)] \tag{3.75}$$

En esta ecuación, $\mathbf{U}_i$ es el vector de respuestas del sujeto i a las p variables y M es el vector de las medias de las variables muestrales, $p \ln(2\pi) + \ln|\Sigma^*|$ es una constante para todos los sujetos y $(\mathbf{U}_i - M)\Sigma^{*-1}(\mathbf{U}_i - M)$ es la fórmula de la distancia cuadrática de Mahalanobis que varía a través de los sujetos. La función P_L es siempre negativa, por lo que si su valor absoluto es pequeño, el patrón de respuestas se ajustaría al modelo estimado de AEC y estaría en concordancia con la matriz de covarianzas reproducida Σ^*; por el contrario, si P_L arroja valores absolutos grandes, el patrón se desviaría del modelo AEC estimado y se catalogaría como patrón outlier (Bollen, 1987, 1990). Al igual que l_0, P_L depende del nivel de habilidad, por lo que para poder comparar sujetos entre sí es necesario desvincular los valores P_L de esta variable. Además, P_L carece de una prueba estadística con la que poder contrastar los resultados. Con el fin de solventar estos dos problemas, Reise y Widaman (1999) optaron por un nuevo índice de ajuste, IND_{χ^2}, definido a partir de la diferencia para cada sujeto entre los P_L calculados con un modelo sustantivo –modelo de un factor, $P_{L,\text{sust}}$– y con un modelo saturado –modelo que reproduce exactamente la matriz de covarianzas, $P_{L,\text{sat}}$–, cuya expresión es:

$$IND_{\chi^2} = -2[P_{L,\text{sust}} - P_{L,\text{sat}}] \tag{3.76}$$

El índice IND_{χ^2} es una razón de probabilidad que se contrasta con una prueba de ajuste χ^2 con grados de libertad igual a la diferencia en el número de parámetros estimados en el modelo saturado y en el modelo sustantivo.

Valores positivos y elevados de IND_{χ^2} aparecerían en sujetos con patrones atípicos de respuesta o en sujetos que no contribuyen al ajuste del modelo.

Capítulo 4

Estudio experimental

En las más recientes investigaciones aplicadas y en diversos estudios de simulación sobre la distribución del estadístico l_z se ha puesto en tela de juicio la normalidad de la misma bajo diversas condiciones experimentales y, por lo tanto, se ha cuestionado la potencia del estadístico para identificar patrones de respuesta atípicos. El objetivo de esta aplicación es comprobar bajo qué condiciones el estadístico l_z de Drasgow, Levine y Williams (1985) se distribuye según la ley normal, manipulando la distribución de habilidad, la longitud del test, el proceso de estimación de la habilidad, el modelo de respuesta dicotómica al que se ajustan los patrones de respuesta y el parámetro de discriminación.

La primera sección de este capítulo está dedicada a los trabajos que han promovido esta investigación; en las secciones siguientes se explica el proceso experimental realizado, los resultados y las conclusiones derivadas de éstos.

4.1. Estudios previos

Antes de comentar los trabajos dedicados a la evaluación de la distribución del estadístico l_z, se han resumido los concernientes al índice pionero, el índice l_0 de Levine y Rubin (1979), con el fin de justificar su estandarización de modo más detallado a como se hizo cuando se presentó en el capítulo anterior (sección 3.4.7).

Levine y Rubin (1979) analizaron la capacidad para identificar patrones atípicos de los tres índices que propusieron, l_0, LR y σ^2. Para ello simularon los patrones de respuesta de una muestra de habilidad normal con los 85 ítems del SAT-V (*Scholastic Aptitude Test, Verbal Section*), ajustados al modelo

de 3-p cuyos parámetros se tomaron del trabajo de Lord (1968). A partir de esta muestra se generaron dos más que contenían patrones atípicos, una con espurias altas y otra con espurias bajas. Para la submuestra con espurias altas se manipularon las respuestas del 20 % de los ítems, con el fin de crear patrones de sujetos que habrían copiado sus respuestas de otro sujeto con mayor nivel de habilidad. La transformación de las respuestas consistía en que los sujetos acertaran este 20 % de ítems, independientemente de si su respuesta inicial fue de acierto o fallo, desviándolas de la respuesta original en 4, 10, 20 y 40 %. Las espurias bajas también se concentraron en el 20 % de los ítems del test, con objeto de representar patrones de sujetos que habrían contestado al azar. Los ítems del SAT-V tienen cinco opciones de respuesta, por lo que para crear este tipo de espurias se tomó la probabilidad de acertar el ítem igual a 1/5 y la de fallarlo igual a 4/5. La magnitud de la desviación de la respuesta original fue también de 4, 10, 20 y 40 %.

Para evaluar cuál de los tres índices era más eficaz en la detección de patrones atípicos, recurrieron al método gráfico de las curvas ROC (*Receiver Operating Characteristic*) empleadas en la Teoría de Detección de Señales de Green y Swets (1966). Para calcular la curva empírica ROC de un índice, éste se calcula en un grupo de sujetos con patrones normales y en un grupo de sujetos con patrones atípicos. Una vez calculados, los sujetos se ordenan de menor a mayor según la magnitud del índice. Entonces, la curva ROC se construye a partir de pares ordenados $[x(t), y(t)]$, donde:

$$
\begin{aligned}
x(t) &= \text{proporción de examinados con patrones normales con el índice de medición apropiada menor o igual que } t \\
y(t) &= \text{proporción de examinados con patrones atípicos con el índice de medición apropiada menor o igual que } t
\end{aligned}
$$

Siendo t un valor aleatorio que indica, o bien el porcentaje máximo permitido de patrones normales incorrectamente clasificados como atípicos por un índice (e.g., un 5 %), o bien un valor muy bajo del índice de medición apropiada que con respecto a la curva ROC será próximo a la diagonal $x = y$. Si el índice detecta bien los patrones atípicos, la curva estará por encima de la diagonal y más alejada de ésta cuanto mejor poder de identificación tenga el índice. La curva empírica proporciona una estimación de la probabilidad de que sujetos con patrones normales sean mal clasificados por un criterio que, a su vez, debe ser lo bastante riguroso como para detectar a un porcentaje de sujetos con un determinado tipo de patrón atípico.

Para comprobar la eficacia de los tres índices, los autores trabajaron con una muestra de 3000 sujetos, de los cuales 200 tenían el mismo porcentaje de desviación en sus respuestas; los otros 2800 presentaban patrones normales. El proceso que se siguió fue, comparar la curva ROC de las submuestras de sujetos con patrones atípicos frente a la curva ROC de la submuestra de sujetos con patrones normales.

Con respecto a los patrones con espurias bajas (azar), el interés de los autores por estos gráficos se concentraba en la parte más bajas de las curvas ya que, el criterio t que clasificara equivocadamente a más del 30 % de los sujetos con patrones normales, no debería ser empleado al menos en tests de aptitudes. En función del porcentaje de desviación de las respuestas, los tres índices (l_0, LR y σ^2) detectaron bien a aquellos patrones con un 20 % o más de desviación o atipicidad.

En el caso de las espurias altas (copia), los sujetos que presentaban este tipo de patrones eran mejor identificados que los anteriores. Los autores supusieron que esto era debido a que para generar las espurias bajas los sujetos eran forzados a responder por azar, mientras que los sujetos con espurias altas *saben* la respuesta correcta ya que la copian de sujetos con niveles de habilidad mayores.

Levine y Rubin (1979) en su artículo, además de parecer modestos respecto a la presentación y el apoyo al uso de los tres índices, pretendían impulsar investigaciones que sugirieran procedimientos y estadísticos de medición apropiada que fueran superiores a los suyos. De ahí que, tras la exposición de los resultados de la investigación, lanzaran los siguientes interrogantes:

1. El cálculo de los índices de medición apropiada implica un proceso de dos etapas. En la primera, se estiman los parámetros de los ítems en una muestra grande de sujetos cuyas respuestas son en principio normales; en la segunda, se calculan los índices. Entonces, si para calcular los índices de medición apropiada se utilizan parámetros estimados de los ítems, ¨cómo afectan los errores de estimación al cálculo de los índices?
2. Cuando se trabaja con datos reales suele ser difícil identificar algunos patrones atípicos, es decir, patrones que no podrían ser catalogados a priori por el investigador como patrones por azar o por copia y, por lo tanto, serían tratados como respuestas normales; ¨cómo afecta este desconocimiento de la existencia de patrones atípicos a la estimación de parámetros y, en consecuencia, al cálculo de los índices de medición apropiada?
3. Los ítems con respuestas omitidas, ¨deberían ser utilizados para incrementar la potencia de los índices de medición apropiada?

4. ¨Se deberían tener en cuenta las relaciones entre varios ítems y subtests en el MRI, y utilizarlas para identificar patrones atípicos?

5. Los modelos psicométricos desarrollados hasta ese momento estaban basados en la unidimensionalidad del rasgo latente subyacente a las respuestas al test, respuestas que eran dicotómicas o se dicotomizaban; ¨estos modelos son lo bastante válidos para apoyar la teoría de la medición apropiada con datos reales?

Para intentar dar respuesta a algunas de estas cuestiones, en concreto a las dos primeras y a la última, Levine y Drasgow (1982) utilizaron el índice l_0 para identificar patrones atípicos con datos reales y simulados. Escogieron una muestra normal inicial de 3200 sujetos y el mismo test que Levine y Rubin (1979), el SAT-V. Esta muestra la dividieron en las siguientes submuestras: a) 2800 primeros sujetos con patrones normales; b) 200 sujetos –desde el 3001 al 3200– cuyos patrones fueron modificados para provocar espurias bajas en el 20 % de sus respuestas originales, con valores de $P_j(\theta) = 1/5$ y $Q_j(\theta) = 4/5$; c) 102 sujetos de los 200 anteriores que tenían al menos el 10 % de sus respuestas originales modificadas. Con estas condiciones, llevaron a cabo cuatro estudios:

- *Estudio 1.* Con datos simulados se probó el efecto de la estimación de parámetros sobre la identificación de patrones atípicos. Para ello, se estimaron los parámetros de los ítems y de la habilidad por MV en la muestra de 2800 sujetos normales. Para esta muestra se calculó l_0 de cada uno de los patrones, con los parámetros simulados de los ítems y suponiendo $\theta = \hat{\theta}$. Los parámetros estimados de los ítems se utilizaron para calcular l_0 en la muestra de 102 sujetos con patrones atípicos. Los resultados mostraron que hubo bastante coincidencia en los valores de l_0 obtenidos con parámetros simulados y estimados aunque, pormenorizando, los l_0 calculados con parámetros estimados eran un poco más pequeños en la mayoría de los patrones atípicos de la submuestra de 102 sujetos con espurias bajas.

- *Estudio 2.* Recurriendo de nuevo a la simulación, se cuestionó el efecto de la presencia de patrones atípicos no identificados a priori en la muestra normativa, en el proceso de estimación de parámetros y en el cálculo del índice de medición apropiada. Se estimaron los parámetros de los ítems y de la habilidad en una muestra de 3000 sujetos, formada por las submuestras de los 2800 sujetos con patrones normales y de los 200 con patrones atípicos. Se calcularon los l_0 y se compararon con los obtenidos en la muestra de 102 sujetos con patrones atípicos hallados en el Estudio 1. La conclusión fue que la presencia de una submuestra grande de

patrones atípicos (200 sujetos) dentro de la muestra total, no devalúa los parámetros estimados de los ítems ni, por lo tanto, los valores de l_0.

- *Estudio 3*. En este análisis usaron datos reales con objeto de probar si el modelo de 3-p era lo suficientemente descriptivo para detectar patrones atípicos con espurias bajas. A partir de las respuestas de 3000 sujetos al SAT-V con una baja tasa de respuestas omitidas (los sujetos debían haber respondido al menos al 90 % de los ítems del test) se estimaron los parámetros de los ítems. De esta muestra se seleccionaron 200 sujetos –desde el 2801 al 3000– para modificarles las respuestas a un 20 % de los ítems y así provocar patrones con espurias bajas. Los índices de medición apropiada en este caso fueron el estadístico l_0 y el índice de razón de verosimilitud LR de Levine y Rubin (1979). En los resultados se apreció una alta tasa de identificaciones correctas de patrones atípicos con ambos índices. El índice LR era más efectivo cuanto más bajo era el nivel de significación, mientras que l_0 lo era cuando los niveles de significación resultaban más altos.

- *Estudio 4*. Por último se quiso confirmar la aplicabilidad de l_0 en cualquier test. A partir de una muestra de datos reales de 10000 sujetos que respondieron al GRE-V (*Graduate Record Examination, Verbal Section*), el cual consta de 95 ítems de opción múltiple, se estimaron los parámetros de los ítems con 3000 sujetos con un amplio rango de habilidad e ilimitado número de respuestas omitidas. De los 7000 sujetos restantes, 2470 fueron seleccionados por haber contestado al menos el 86 % de los ítems, de los cuales 2270 formaron la submuestra de sujetos con patrones normales y a 200 se les modificó su patrón de respuesta para provocar patrones con espurias bajas. Se calculó l_0 en estas dos últimas submuestras con los parámetros estimados en la muestra de los 3000 sujetos. Este estudio confirmó la eficacia de l_0 para detectar patrones atípicos independientemente del instrumento de medida.

Drasgow (1982) comparó los resultados de la modificación de l_0 que él mismo propuso, i.e., la media geométrica l_g, con LR, $\hat{\sigma}^2$ y el coeficiente de correlación biserial-personal r_{bisper} de Donlon y Fischer (1968). El método que siguió para manipular los patrones de respuesta atípica (espurias altas y bajas) fue el mismo que emplearon Levine y Rubin (1979), pero en este caso los patrones de partida eran de una muestra de datos reales de 10000 sujetos que contestaron al GRE-V. Se crearon tres submuestras con la mitad de los sujetos de la muestra real –del 5001 al 10000–; una submuestra estuvo formada por 115 sujetos con habilidad moderadamente baja, en la que se manipularían sus respuestas para

originar espurias altas; en la otra submuestra, a los patrones de respuesta de 200 sujetos con alto nivel de habilidad se les generarían espurias bajas.

Las respuestas atípicas se provocaron en 20 ítems escogidos aleatoriamente del total (95 ítems), de los cuales, si alguno había sido omitido en el patrón de partida, no se tenía en cuenta para la manipulación de la respuesta. Para comparar los resultados, se seleccionó la tercera submuestra con 255 sujetos, cuyos patrones de respuesta no fueron modificados.

Para valorar el efecto de la estimación de los parámetros de los ítems y la efectividad de los índices para identificar patrones atípicos, se realizaron los procesos de estimación con dos muestras de distinto tamaño, dada la dificultad de obtener una óptima estimación con el modelo de 3-p. En una submuestra de 3000 sujetos –del 1 al 3000–, tras el proceso de estimación para los modelos de 1-p y 3-p se calcularon los índices l_g, LR, $\hat{\sigma}^2$ y r_{bisper} (este último sólo en el modelo de 3-p) en las submuestras con las respuestas reales y con espurias. Los resultados que obtuvo Drasgow (1982) fueron que las espurias bajas se identificaron en un alto porcentaje en ambos modelos, siendo el modelo de 3-p algo más efectivo; además, los índices de medición apropiada de TRI detectaron patrones atípicos mejor que r_{bisper}. La identificación de espurias altas estuvo por debajo de la deseada y de manera similar en los dos MRI.

El segundo tamaño muestral para contrastar los efectos de la estimación fue de 500 sujetos, escogiendo entre el 1 y el 4991 aquellos a quienes correspondía la décima posición. De nuevo se estimaron los parámetros con los modelos de 1-p y 3-p para esta submuestra de patrones normales y para la que contenía patrones con espurias bajas (no se empleó la submuestra con espurias altas, debido a la baja tasa de identificaciones correctas aparecidas en el análisis anterior). La identificación de las espurias bajas fue bastante buena; no se apreció efecto del escaso tamaño muestral en la estimación de los parámetros, ya que, tanto en el modelo de 1-p como en el de 3-p, las tasas de identificaciones correctas fueron altas y similares, mostrándose algo superior el modelo de 3-p.

Las conclusiones de esta investigación de Drasgow (1982) fueron las siguientes:

- Los índices de medición apropiada no parecen estar muy afectados por las diferencias entre el modelo de Rasch y el de 3-p, aunque con éste se obtenga una mejor tasa de identificaciones correctas de los patrones con espurias bajas. Desde un punto de vista práctico, el modelo de 3-p no es mejor que el modelo de Rasch, aunque el primero conlleva un elevado coste computacional y temporal.
- La estimación de los parámetros en muestras pequeñas, no parece impac-

tar en la identificación de patrones atípicos en ninguno de los índices de medición apropiada empleados.

- Los índices de medición apropiada basados en los supuestos de la TRI dominan y funcionan mejor que el estadístico r_{bisper}. En el modelo de Rasch, el índice l_g tiene tasas de identificaciones correctas superiores a las de r_{bisper} en niveles de significación de 0.05 y 0.07.
- En tareas de medición apropiada, un factor importante que hay que controlar es la distribución de la habilidad, ya que las distribuciones de algunos de los índices de medición apropiada están estrechamente relacionadas con aquéllas, lo cual afecta en consideración a la identificación de patrones atípicos. En este estudio las distribuciones de la habilidad de los sujetos normales y la de sujetos con espurias bajas eran muy parecidas, lo cual haría sospechar que las elevadas tasas de identificaciones correctas de patrones atípicos en esta segunda muestra no fue causa de un artefacto estadístico. Esto se debería a que cuando $\hat{\theta} \cong \theta$, el test aporta bastante información sobre la habilidad y es posible obtener una expresión aproximada para la media y la varianza de l_0, cuyos valores en el SAT-V eran próximos a los de la distribución normal tipificada.

  Entre la distribución de la muestra de sujetos normales y la que contenía espurias altas sí hubo divergencias, lo cual perjudicó al logro de elevadas tasas de identificaciones correctas de este tipo de patrones atípicos. Este resultado contradecía el obtenido por Levine y Rubin (1979), en el cual los patrones atípicos con espurias altas eran mejor identificados que los que presentaban espurias bajas. La razón de esto residiría en que el test empleado por Drasgow (1982), el GRE-V, sólo tenía tres ítems con un valor del parámetro estimado de pseudo-azar inferior a 0.10 y siete ítems lo tenían inferior a 0.15. Por lo tanto, en este test fueron muy pocos los ítems en los que resultase muy probable una respuesta incorrecta por sujetos con un nivel de habilidad moderado. En consecuencia, el proceso de manipulación para generar patrones atípicos con espurias altas no produjo patrones especialmente inusuales. Por el contrario, el test empleado por Levine y Rubin (1979), el SAT-V, contenía bastantes ítems con valores muy bajos del parámetro estimado de pseudo-azar (20 ítems con $\hat{c_j} \leq 0{,}10$ y 42 ítems con $\hat{c_j} \leq 0{,}15$).

En resumen, parece más adecuado emplear el modelo de 3-p para identificar patrones atípicos, cuando casi todos los ítems del test tienen valores del parámetro de pseudo-azar iguales o superiores a 0.20. En tests en los que muchos de sus ítems $\hat{c_j} \approx 0$, el modelo de Rasch puede ser más efectivo en medición apropiada.

Dados los problemas que planteaba el estadístico l_0, como ya se señalaron en el capítulo anterior (índice no estandarizado y distribución desconocida), los trabajos que se realizaron en el campo de la medición apropiada se enfocaron hacia el análisis de la distribución de l_z y al uso de este índice para identificar patrones de respuesta atípicos. En el mismo artículo que se presenta la estandarización de l_0 por Drasgow, Levine y Williams (1985), se comprobó la distribución de los estadísticos l_z y $l_{z.h}$ –para el modelo del histograma– y sus capacidades para detectar patrones atípicos. Los autores consideraron la presencia de respuestas omitidas, algo que en las investigaciones precedentes no fue directamente estudiado, sino que, por el contrario, esta posibilidad de respuesta –o mejor dicho de no-respuesta a un ítem– era considerada como una respuesta fallida o se ignoraba.

Se realizaron tres estudios a partir de una muestra de datos reales de 75000 sujetos que contestaron al SAT-V. De la misma se extrajeron 3000 patrones de respuestas con los que se estimaron por MV los parámetros de los ítems. Con ellos se construyeron los histogramas necesarios para estimar los parámetros de habilidad ($\hat{\theta}_d$) por MV para ítems dicotómicos. En muestras con patrones de respuesta normales desarrollaron los siguientes ejercicios:

- *Estudio 1.* Con una muestra de 464 sujetos se examinaron las distribuciones de l_0, $l_{0.h}$, l_z y $l_{z.h}$. Graficando los valores de l_0 y $l_{0.h}$ frente a $\hat{\theta}_d$, se demostró la dependencia de los estadísticos de $\hat{\theta}_d$, sobre todo la de $l_{0.h}$. Con el mismo procedimiento gráfico, se cuestionó si la estandarización de ambos estadísticos eliminaría la dependencia de la habilidad si $\hat{\theta}_d = \theta$; en caso contrario, si $\hat{\theta}_d \neq \theta$, la estandarización sólo sería un procedimiento heurístico que reduciría la dependencia. Se confirmó la ausencia de relación entre l_z y $l_{z.h}$.

  Algo que llamó la atención de Drasgow, Levine y Williams (1985) fue el elevado número de patrones con valores de $l_{z.h}$ muy por debajo del esperado, lo que les llevó a inspeccionar dichos patrones de respuestas. La peculiaridad de éstos era el alto porcentaje de respuestas omitidas (el 35 %, i.e., 30 ó más ítems) y de sujetos que no habían contestado al menos el 77 % del test. La dependencia de $l_{z.h}$ del número de respuestas omitidas y de patrones de sujetos que no finalizaban el test, se confirmó en una segunda muestra de 456 sujetos distintos a los anteriores. Lo que realmente afectaba a esta relación era la omisión de respuesta a los ítems más fáciles del test y a los que tenían distractores muy efectivos.

- *Estudio 2.* Teniendo en cuenta los resultados del Estudio 1, se eliminaron de las muestras de los análisis sucesivos a los sujetos con más de un 35 % de respuestas omitidas y a los que no contestaban al menos al 77 % de

los ítems, ya que eran catalogados como patrones con espurias bajas, no siendo válidos para el objetivo del trabajo. Por lo tanto, en este segundo estudio se limita el número de respuestas omitidas. Se representaron las distribuciones de los estadísticos l_z y $l_{z.h}$ en una muestra de 3478 sujetos que cumplían ambos requisitos, y además sus valores de habilidad estimada con el modelo de 3-p para ítems dicotómicos se encontraban en el intervalo (-2.05,+2.05). Las distribuciones se perfilaron asimétricas respecto a la normal estandarizada, hecho que contrastó y aseguró la prueba de normalidad de Kolmogorov-Smirnov.

Sin embargo, Drasgow, Levine y Williams (1985) afirman que esta desviación de la distribución normal no era esencial para sus propósitos; lo importante era la invarianza de la distribución a lo largo del continuo de $\hat{\theta}_d$. La distribución condicional de $l_{z.h}$ presentó más variabilidad, lo que podría deberse a que en la muestra existieran verdaderos patrones de respuesta atípica; esto les condujo a plantear la hipótesis del tercer estudio.

- *Estudio 3.* Con los parámetros de los ítems del SAT-V se simularon los patrones de respuesta de 4000 sujetos con el modelo de 3-p y el modelo del histograma, cuyos niveles de habilidad pertenecían a una distribución normal y $\theta \in$(-2.05,+2.05). Con parámetros de habilidad estimados ($\hat{\theta}_d$) se calcularon l_z y $l_{z.h}$. En ausencia de posibles patrones con respuestas atípicas, de nuevo la distribución de $l_{z.h}$ fue más variable que la de l_z.

Con estos tres estudios los autores concluyeron que, a pesar de la no absoluta independencia de l_z y $l_{z.h}$ del nivel de habilidad, la estandarización no es tan estadísticamente significativa como para que tales índices no puedan ser utilizados para identificar patrones de respuesta atípicos.

A continuación probaron la efectividad de los índices estandarizados para detectar patrones atípicos. De una muestra real de 3478 sujetos con patrones normales se escogieron 300 por tener menos del 35 % de respuestas omitidas, o bien más del 77 % del test contestado. A estos patrones se les manipularon las respuestas para provocar patrones atípicos con espurias bajas y con espurias altas. Se modificaron las respuestas al 10, 20 y 30 % de los ítems. Los patrones con espurias altas se generaron cambiando la respuesta inicial del sujeto a ser correcta, independientemente de la respuesta original. La creación de patrones con espurias bajas fue algo más compleja. Primero se determinó el porcentaje de respuestas omitidas de cada vector de respuestas, q; a continuación, cada ítem fue puntuado como omitido con probabilidad q y a cada una de las cinco opciones de respuesta del ítem le correspondió una probabilidad $(1 - q)/5$.

Con este procedimiento de creación de espurias bajas pretendían reflejar la tendencia de los sujetos a omitir respuestas en un test.

Tras la modificación de las respuestas a los ítems, se estimaron los parámetros de habilidad de los patrones con respuestas espurias usando el modelo de 3-p, valores que se usaron para calcular l_z y $l_{z.h}$ de cada uno de estos patrones.

Al igual que Levine y Rubin (1979), Drasgow, Levine y Williams (1985) emplearon el procedimiento de las curvas ROC para visualizar la efectividad de los índices para detectar patrones atípicos. Los resultados que obtuvieron fueron los siguientes:

- Sin tener en cuenta el tipo de respuesta espuria generada, cuanto mayor es el número de ítems con respuesta atípica (30 %), más alta es la tasa de identificaciones correctas de los dos índices.
- Cuando los patrones de respuesta son atípicos por espurias bajas, la tasa de identificaciones correctas aumenta con el incremento de los valores de habilidad. El índice $l_{z.h}$ identifica mejor este tipo de respuestas atípicas, lo cual podría ser debido a que, como los ítems que se manipularon fueron escogidos al azar, algunos de éstos eran ítems demasiado fáciles para fallarlos o para no ser entendidos por el sujeto. El modelo del histograma no se deja afectar por este hecho, ya que contempla en su formulación la función de opción de respuesta y es sensible a la identificación de sujetos que escogen una opción poco común.
- Cuando los patrones de respuesta son atípicos con espurias altas, la tasa de identificaciones correctas disminuye conforme aumentan los valores de habilidad. En esta ocasión, el índice l_z identifica mejor esta clase de respuestas atípicas que el índice $l_{z.h}$, ya que el modelo de 3-p no diferencia entre las distintas opciones de respuesta, sino que valora si la respuesta al ítem es correcta o incorrecta.

Las conclusiones a las que se llegó tras este estudio fueron: a) la estandarización reduce sustancialmente la dependencia de los índices l_0 y $l_{0.h}$ respecto de la habilidad; b) las distribuciones de l_0 y $l_{0.h}$ son más variables que las de l_z y $l_{z.h}$ a través de los niveles de habilidad; c) l_z identifica correctamente patrones atípicos que presentan muy bajo número de respuestas omitidas, por lo que en presencia de éstas en altos porcentajes sería conveniente emplear el índice $l_{z.h}$. Esta misma decisión se debería adoptar si se quiere mejorar la identificación de patrones atípicos con espurias bajas cuando los ítems son politómicos. Sin embargo, l_z es más efectivo que $l_{z.h}$ para detectar patrones atípicos con espurias altas.

Drasgow, Levine y McLaughlin (1987) analizaron la estandarización y la potencia de nueve índices denominados *índices de medición apropiada prácticos*, que fueron el estadístico l_z, los estadísticos de ajuste F_1 y W_3 de Rudner (1983), los estadísticos de curvatura de la función de verosimilitud JK y O/E de estos mismos autores, el índice de precaución C_i de Sato (1975), los índices $ECI2_z$ y $ECI4_z$ de Tatsuoka (1984), y la varianza de la opción del ítem IOV (*Item-Option Variance*). El estadístico IOV es un índice fácil de calcular. Sea N_{jk} el número de sujetos que escogen la opción k del ítem j y $\bar{X}_{jk}$ la media de respuestas correctas de estos sujetos. Se identifica la opción más seleccionada por los sujetos de alta habilidad, que por lo general será la opción correcta, y la opción escogida por la mayoría de los sujetos de baja habilidad. Para los patrones de respuesta atípicos se esperan valores de $\bar{X}_{jk}$ inconsistentes y el empleo de la varianza correspondiente a la opción de respuesta puede ser usada como un índice de medición apropiada:

$$IOV = Var(\bar{X}_{jk})$$

Valiéndose de las respuestas al SAT-V de 49470 sujetos de una muestra real y empleando el modelo de 3-p, estimaron los valores de habilidad. Se extrajo una submuestra de 1000 sujetos con $\hat{\theta} \sim N(0,1)$ y se simularon los patrones de respuesta con el modelo del histograma, para así incluir en la experimentación el caso de respuestas omitidas y para calcular la probabilidad condicional de las posibles categorías de respuesta dado $\hat{\theta}$.

Los resultados de la estandarización con patones normales se describieron a través del análisis de las curvas ROC de cada uno de los nueve índices. Las distribuciones de IOV y C_i no siguieron la ley normal, algo que en principio se esperaba ya que no están estandarizados; sin embargo, F_1, $ECI2_z$ y $ECI4_z$ son índices estandarizados, pero no produjeron resultados acordes con la normalidad. Los que sí se ajustaron a la distribución normal fueron l_z, W_3, JK y O/E si la distribución de habilidad era normal. Por lo tanto, las tasas de identificaciones de patrones atípicos de IOV, C_i y F_1 debían ser interpretadas con cautela.

Para estudiar el poder de identificación de los índices de medición apropiada, se generó una muestra de 4000 patrones de respuesta normal y otra de 24000 patrones con respuestas atípicas. En esta última, 12000 patrones contenían espurias altas, provocadas al cambiar las respuestas originales del 15 y 30 % de los ítems a ser respuestas correctas; los otros 12000 patrones presentaban espurias bajas de sujetos con alto nivel de habilidad, al sustituir la respuesta inicial en el 15 y 30 % de los ítems por una respuesta aleatoria a una de las opciones, teniendo cada una de éstas una probabilidad de 0.2.

Se calcularon los índices de ajuste en la muestra de patrones normales y en las de patrones atípicos, así como el índice LR de ajuste óptimo de Levine y Drasgow (1984). Los parámetros de los ítems y de la habilidad se obtuvieron de la muestra de patrones normales aplicando el método MV. Las tasas más altas de identificaciones correctas y más próximas a las de LR, tanto de espurias altas como de espurias bajas, las tuvieron l_z, W_3 y $ECI2_z$, y tasas moderadas las de $ECI4_z$, JK y O/E. Los índices IOV, C_i y F_1 no identificaron adecuadamente patrones atípicos.

Drasgow, Levine y McLaughlin (1987) indagaron más acerca de las distribuciones empíricas de los tres índices que mejor identificaban la atipicidad de los patrones de respuesta (l_z, W_3, $ECI2_z$), incluyendo la prueba de normalidad de Kolmogorov-Smirnov con 1000 patrones de respuestas normales y 24000 patrones con espurias. No se encontraron diferencias estadísticamente significativas, aunque el resultado debería tomarse con precaución, ya que la prueba de Kolmogorov-Smirnov es conservadora cuando los momentos empíricos son sustituidos en la distribución teórica.

Reise (1995) aportó más información acerca de la estandarización, la distribución y la potencia para detectar patrones atípicos de l_z en el ámbito de los tests de personalidad, para lo cual empleó cuatro escalas del MPQ (*Multidimensional Personality Questionnaire* de Tellegen, 1982). Estas escalas están compuestas por ítems dicotómicos ajustados al modelo logístico de 2-p y la longitud de las seleccionadas es 24, 27, 28 y 34 ítems. Reise consideró como parámetros verdaderos de los ítems aquellos que se estimaron por MV a partir de los datos reales de una muestra de 2000 sujetos. Con los parámetros estimados de los ítems se simularon 10000 patrones de respuesta para cada escala, estando repartidos en cinco grupos de habilidad con 2000 sujetos (θ= -2.0,-1.0,0.0,+1.0,+2.0) que se definieron como los valores verdaderos de habilidad. Con tales patrones se desarrollaron dos estudios:

- *Estudio 1.* Análisis de l_z con parámetros verdaderos de la habilidad y de los ítems. Las distribuciones de l_z en las cuatro escalas eran sesgadas negativas en el continuo de habilidad. Este comportamiento del estadístico fue justificado por Reise (1995) recurriendo a la investigación de Drasgow (1982), en la que se afirmaba que l_z tenía una buena aproximación a la distribución normal cuando la función de información del test era alta, hecho que no ocurría en las escalas del MPQ para valores altos de habilidad.
- *Estudio 2.* Análisis de l_z con parámetros estimados de la habilidad y de los ítems. Conocidos los parámetros de los ítems se estimaron los paráme-

tros de habilidad por MV. Los resultados sobre la variabilidad y el sesgo del estadístico coincidieron con los obtenidos mediante los parámetros verdaderos. La dispersión de l_z estaba relacionada con la función de información del test: a menor información, menor varianza. También estaba relacionada la variabilidad de l_z con el número de ítems de las escalas cuyos parámetros de dificultad eran cercanos a $\hat{\theta}$ y tenían parámetros de discriminación elevados, de modo que en ese nivel de habilidad la escala aportaba mucha información y, por lo tanto, la varianza de l_z se aproximaba a 1. Las distribuciones del estudio con parámetros estimados diferían de la normalidad más que las obtenidas con parámetros verdaderos.

En este mismo trabajo de Reise (1995) se evaluó cómo afectaban a la potencia de l_z para identificar patrones atípicos factores tales como el parámetro de discriminación, la distribución del parámetro de dificultad, el número de ítems del test, el método de estimación y el tipo de atipicidad. En los tests de personalidad, los patrones atípicos característicos son los derivados de: la pérdida de atención en determinados ítems, la respuesta azarosa o el engaño deliberado. Ante la carencia de modelos de respuesta para generar este tipo de patrones y dejando de lado la etiología de los mismos, Reise conceptuó que un patrón de respuesta atípica sería aquel que no deriva de un modelo de TRI unidimensional, por lo que estaría compuesto por respuestas aleatorias que no aportarían información sobre la habilidad del sujeto. Esta definición es una adaptación de la Teoría de Respuesta de Persona (TRP) de Strandmark y Linn (1987), según la cual el modelo de respuestas atípicas generalizado incorporaría un parámetro de discriminación de persona, a_p, constante para todos los ítems del test. La formulación de este modelo es:

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp[-(a_p a_j)(\theta_i - b_j)]}$$

El parámetro a_p implica que la FI del test para este modelo sea:

$$I(\theta_i) = \sum_{j=1}^{n} (a_p a_j)^2 Q_j(\theta_i) \frac{[P_j(\theta_i) - c_j]^2}{(1 - c_j)^2}$$

por lo que la información del test puede ser manipulada mediante un proceso de simulación, que otorga diferentes valores del parámetro a_p a los sujetos de una muestra. Cuando $a_p = 1$ el modelo resultante es el de 3-p; cuando $a_p = 0$, el ítem no aporta información psicométrica con respecto a $\hat{\theta}_i$. Desde la TRP, Reise y Due (1991) desarrollaron la hipótesis de que un patrón de

respuestas atípico sería aquel que aporta menos información psicométrica para estimar la habilidad del sujeto que la que predicen los parámetros de un MRI especificado. Entonces, retomando el trabajo de Reise (1995), se especificó que un patrón sería atípico si sus respuestas no fueran generadas por el modelo unidimensional de 2-p. Los patrones atípicos se crearon siguiendo varios pasos: primero, se describieron los parámetros de los ítems y se manipularon los parámetros de discriminación, reduciendo su valor inicial a $a'_j = 0{,}0$ (aleatoriedad y ausencia de información de las respuestas); segundo, se simularon 1000 patrones de respuesta con ningún ítem con $a'_j = 0{,}0$, con un ítem con $a'_j = 0{,}0$, con dos ítems con $a'_j = 0{,}0$, y así sucesivamente hasta que todos los ítems de las escalas tuvieron $a'_j = 0{,}0$; tercero, cálculo de l_z de esos patrones pero usando los parámetros de discriminación originales y los parámetros de habilidad verdaderos y estimados por tres procedimientos, MV, EAP y estimación biponderada (*biweight estimation*).

Reise (1995) comenzó la presentación de los resultados señalando que, según los dos estudios anteriores, la distribución de l_z se alejaba de la normal estandarizada y que, por lo tanto, las tasas de identificaciones correctas de patrones atípicos estarían sesgadas. En general, cuanto mayor era el número de ítems con respuesta atípica y mayor diferencia existía entre b_j y la habilidad, mayores eran la tasas de identificaciones correctas de l_z con parámetros de habilidad estimados. En cuanto al método de estimación, para pocas respuestas atípicas las diferencias entre MV, EAP y biponderada fueron mínimas, aumentando las divergencias entre ellos con el incremento de dichas respuestas y obteniéndose las mejores tasas en los l_z calculados con $\hat{\theta}$ mediante el proceso biponderado. Entre la estimación máximo-verosímil y la bayesiana, el porcentaje de patrones atípicos fue muy similar.

La mayoría de investigaciones precedentes a la que a continuación se va a exponer, justificaban la aproximación a la normalidad de la distribución de los índices de medición apropiada l_z, $ECI4_z$ y W_3 en ausencia de patrones de respuesta atípicos. Noonan *et al.* (1992) estudiaron el efecto que podrían tener el MRI y el número de ítems sobre la distribución de estos tres estadísticos, así como la estabilidad de los mismos ante diferentes tasas de falsos positivos. Elaboraron un estudio de simulación Montecarlo con parámetros conocidos de habilidad e ítems. Para el fin que perseguían, emplearon los modelos de 2-p y 3-p, y dos longitudes de test, $n = 40$ y $n = 80$ ítems. Los parámetros de los ítems se generaron con distribuciones uniformes y los parámetros de habilidad ajustados a una distribución $N(0, 1)$. Cada combinación $MRI \times n$ contenía 2000 patrones de respuesta simulados y replicados 50 veces. Los índices l_z, $ECI4_z$ y W_3 se calcularon junto con sus medias, desviaciones típicas e índices de sesgo y curtosis; para el análisis de la estabilidad de los índices, las tasas

de falsos positivos fueron 0.01, 0.05 y 0.10.

Mediante la correlación de Pearson se comprobó la ausencia de relación lineal entre θ y los tres índices. En ausencia de patrones atípicos, $ECI4_z$ fue el índice cuya distribución más se aproximó a la normal estandarizada y también el más estable a través de todas las réplicas, independientemente de la longitud del test y del modelo de respuesta. La distribución de W_3 era la más alejada de la normalidad y sus valores los menos estables; a su vez, las tasas de falsos positivos de este índice fueron las más afectadas por la interacción $MRI \times n$. La distribución de l_z, en ausencia de patrones atípicos, estuvo distorsionada por elevados índices de sesgo negativos y de curtosis positivos, siendo más acusada cuando el número de ítems era menor y el MRI empleado fue el de 2-p. Entre l_z y W_3 hubo altos coeficientes de correlación de Pearson, algo que sorprendió a Noonan *et al.* (1992), ya que el fundamento de ambos estadísticos es muy dispar.

La investigación de Nering (1995) barajó distintas fuentes de influencia sobre l_z, como fueron la distribución de θ [$\theta = 0{,}0$, $\theta \sim N(0, 0{,}5)$ y $\theta \sim N(0, 1)$], la amplitud del parámetro de dificultad [(-1.0,+1.0), (-2.0,+2.0) y (-3.0,+3.0)], el método de estimación del parámetro de habilidad (MV y EAP) y el parámetro de pseudo-azar ($c_j = 0{,}20$ y $c_j = 0{,}0$). El parámetro de discriminación fue constante para todas las condiciones ($a_j = 1{,}50$). Se simularon 1000 patrones de respuesta a un test de 25 ítems. La normalidad de l_z se evaluó con los estadísticos descriptivos: media, desviación típica, sesgo y curtosis, incluyendo la prueba de Kolmogorov-Smirnov. También se calcularon las tasas de falsos positivos y se realizó un estudio de recubrimiento sobre el parámetro de habilidad con tres medidas: la correlación de Pearson, la exponencial de la raíz del error medio cuadrático (*Root Mean Square Error*, $RMSE$):

$$RMSE = \exp \sqrt{\frac{\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i)^2}{N}} \tag{4.1}$$

y el error medio con signo (*Average Signed Bias*, ASB):

$$ASB = \frac{\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i)}{N} \tag{4.2}$$

En cuanto a la estimación de θ, ASB detectó sobrestimación del parámetro; las tres medidas coincidieron en que cuanto más estrecho era el rango de variación del parámetro de dificultad, peores estimaciones se obtenían. Los mejores resultados se daban en las condiciones en que se empleó el método de

estimación EAP y en las que $c_j = 0{,}0$. En consecuencia, la mala estimación de θ podría provocar diferentes valores de l_z.

Las distribuciones del estadístico de Drasgow, Levine y Williams (1985) tanto con valores verdaderos de θ como estimados, eran asimétricas negativas y leptocúrticas, aunque con el método EAP los resultados fueron menos acusados. La prueba de Kolmogorov-Smirnov confirmó que estas distribuciones se alejaban de la normal estandarizada. El parámetro de pseudo-azar no tuvo influencia sobre la media y la desviación típica de l_z, pero sí que cuando $c_j = 0{,}2$ aparecieron diferencias en el índice de sesgo y de curtosis, originando distribuciones más negativamente sesgadas y más leptocúrticas. Comprobado el recubrimiento del parámetro de habilidad, Nering (1995) sugirió que una mejor estimación conllevaría la mejoría de los valores de l_z. En esta aplicación se observó que con el procedimiento EAP, tanto los parámetros de habilidad como l_z mejoraban en comparación con el uso de parámetros verdaderos y estimadores máximo-verosímiles. Incluso las medias y las desviaciones típicas de l_z tras la estimación bayesiana, fueron más próximas a las estandarizadas si $c_j = 0{,}0$.

Nering (1997) amplió los datos de su anterior investigación al ámbito de los TAI. Trabajó con dos bancos de 250 y 500 ítems, y para cada uno de ellos se fijó como valor crítico para finalizar el test el error de estimación asociado a $\hat{\theta}$ cuando éste era 0.30, 0.25 y 0.20. Con el fin de escoger los parámetros del modelo de 3-p, se emplearon tres distribuciones normales para el parámetro de discriminación, una distribución normal para el parámetro de pseudo-azar y una distribución uniforme para el parámetro de dificultad. Se simularon 10000 patrones de respuesta dicotómica para cada banco de ítems bajo la supuesta distribución de habilidad $N(0, 1)$. El análisis de la distribución normal los índices l_z y $ECI4_z$ se llevó a cabo calculando los cuatro primeros momentos y se especificó el número medio de ítems que se logró en cada uno de los dos bancos.

Los hallazgos de Nering (1997) en el contexto de los TAI coincidieron con los encontrados en el ámbito de los tests convencionales, es decir, las distribuciones de l_z y $ECI4_z$ tenían índices de sesgo negativo y positivo, respectivamente, y diseñaban curvas leptocúrticas. Estas desviaciones de la normalidad eran más marcadas cuando, en vez de los valores reales de θ, se empleaban las estimadores máximo-verosímiles de la habilidad. Un tanto de lo mismo ocurría a mayor poder discriminativo de los ítems, a menor restricción del valor crítico para finalizar el test, con menor número medio de ítems contestados y en el banco de 500 ítems. De hecho, estos cuatro factores de influencia sobre la no-normalidad de los estadísticos forman una cadena, ya que para que se satisfaga

la teoría asintótica debe de haber un elevado número de ítems conseguidos, lo cual se logra cuando, bien los ítems son poco discriminativos, bien el valor crítico para finalizar el test es más restrictivo, o bien el banco tiene pocos ítems. El estadístico l_z tuvo índices de sesgo y curtosis mayores en valor absoluto que los de $ECI4_z$. Comparando los resultados del uso de parámetros verdaderos de habilidad frente a los estimados por MV, la distribución de l_z se aproxima más a la normal usando parámetros verdaderos, siendo indiferente para $ECI4_z$.

La correlación entre los dos índices resultó elevada, al igual que con los tests convencionales, incrementándose con ítems menos discriminativos y un criterio de terminación del test más restrictivo. La relación lineal que mantuvieron l_z y $ECI4_z$ con $\hat{\theta}$ fue nula.

Las tasas de falsos positivos estuvieron afectadas por el número de ítems y se acercaron a las esperadas cuantos más ítems se administraban, algo que favorecería a su vez la normalidad de los índices de medición apropiada; sin embargo, aunque se contestara a más de 50 ítems, las distribuciones de ambos discrepaban de la normal.

Nering (1997) concluyó su indagación con un aviso de precaución en el uso de l_z y $ECI4_z$ en los TAI, ya que el proceso implícito de evaluación de los sujetos por tales tests afecta al proceso de estimación de la habilidad y, por tanto, al cálculo de dichos índices de ajuste de persona, por lo que tendrían una utilidad limitada en este contexto de medida.

McLeod y Lewis (1999) también evaluaron la distribución de los estadísticos l_z, $ECI4_z$ y Z_c en los TAI, y su poder para identificar patrones con respuestas atípicas producto de la memorización a un porcentaje de ítems de mayor frecuencia de aparición. Utilizaron el test GRE-Q (*Graduate Record Examination, Quantitative Section*), compuesto de un banco de 348 ítems. Antes de iniciar su proceso experimental, definieron dos grupos de patrones de respuesta, cada uno con 1650 sujetos, simulándolos con el modelo de 3-p. El primero de ellos era el grupo de sujetos con respuestas memorizadas (GM), para el que se manipuló las respuestas de los 50 ítems de mayor frecuencia de aparición al 5 % de los sujetos con puntuaciones más altas. La modificación de las respuestas a estos ítems consistió en darlas por acertadas al margen de su respuesta original (en estos ítems $\bar{x}_{b_j} = 1{,}26$, $\bar{x}_{a_j} = 1{,}27$ y $\bar{x}_{c_j} = 0{,}13$). Con los parámetros estimados de los ítems, se generaron los patrones de respuesta a los 298 ítems restantes. En el segundo grupo, al que llamaron grupo *nulo* (GN), no hubo ningún ítem memorizado. El parámetro de habilidad seguía una distribución uniforme discreta; en la estimación de los parámetros de los ítems y de la habilidad se implementó el procedimiento MV.

Tras las simulaciones se seleccionaron 28 ítems del banco en ambos grupos y se calcularon los tres índices de medición apropiada. El estudio de correlación entre los índices apoyó las investigaciones previas acerca de la alta covariación entre l_z y $ECI4_z$, pero demostró muy poca relación de éstos con Z_c. En cuanto a las distribuciones de los tres índices, ninguna resultó normal estandarizada ni para el GN, en el cual las medias y las desviaciones típicas fueron infravaloradas, ni para el GM, en donde las medias estuvieron por encima y las desviaciones típicas por debajo de las esperadas. Por lo tanto, se debería cuestionar la potencia de los mismos para identificar patrones con respuestas memorizadas en los TAI; de hecho, el estadístico l_z fue el menos eficaz de los tres. El índice $ECI4_z$ tuvo las mejores tasas de identificaciones correctas, con valores $\alpha = 0{,}025$ y $\alpha = 0{,}10$, y Z_c las logró cuando $\alpha < 0{,}025$.

A las deducciones anteriores sobre la distribución de l_z en los TAI se sumó el trabajo de van Krimpen-Stoop y Meijer (1999). Estos autores realizaron tres estudios en donde se comparó la distribución de l_z con la de l_z^* de Snijders (2001), en el contexto de los tests convencionales (lápiz y papel) y de los TAI:

- *Estudio 1.* Se trabajó con tres tests convencionales de longitud 20, 50 y 80 ítems, y un TAI con un banco de 400 ítems, todos ellos ajustados al modelo logístico de 2-p. Un total de 10 muestras de 10000 sujetos se simularon para nueve valores fijos de habilidad (θ= -2.0,-1.5,-1.0,-0.5,0.0,+0.5,+1.0,+1.5,+2.0) y una muestra cuya habilidad seguía una distribución normal tipificada. Se calcularon las medias, las desviaciones típicas y los índices de sesgo y curtosis de ambos índices de ajuste de persona, con valores de habilidad verdaderos y estimados por máxima verosimilitud ponderada (MVP) –el estadístico l_z^* sólo se puede calcular con parámetros estimados de habilidad–.

  Usando $\hat{\theta}$ en los tests convencionales, las medias y las desviaciones típicas de l_z y de l_z^* eran próximas a las normalizadas, pero las distribuciones resultaron negativamente sesgadas y leptocúrticas. En el TAI ambos estadísticos obtuvieron medias y desviaciones típicas alejadas de las estándar, que también se dibujaron asimétricas negativas y leptocúrticas. La distribución de l_z en los tests de lápiz y papel y en el TAI siguió el mismo patrón cuando se calculaba en función de θ, aunque los valores de los cuatro momentos se asemejaban más a los de la normal que empleando $\hat{\theta}$.

- *Estudio 2.* En la siguiente aplicación realizaron un estudio de remuestreo con técnicas *bootstrapping* paramétricas para derivar las distribuciones de l_0, l_z y l_z^*. En este caso, siguiendo con el modelo de 2-p, se empleó una muestra de 400 sujetos para los mismos cuatro tests del Estudio 1. Para

los tests convencionales se generaron 1000 réplicas por cada sujeto con $\hat{\theta}$; para el TAI, fueron 500 patrones por sujeto los que se remuestrearon. De nuevo, evaluaron el efecto de trabajar con parámetros verdaderos y estimados de habilidad.

Para comprobar si las distribuciones teóricas y empíricas de los tres estadísticos se asemejaban, van Krimpen-Stoop y Meijer (1999) aplicaron la prueba de correspondencia de criterio χ^2 de Pearson sobre el nivel de significación $p^*(\alpha)$. El proceso seguido para obtener $p^*(\alpha)$ fue calcular los l_0, l_z y l_z^* de los patrones originales de respuesta; diez intervalos de longitud 0.10 caracterizaron la distribución de $p^*(\alpha)$, bajo la restricción de que la proporción esperada de patrones simulados con un $p^*(\alpha)$ en un determinado intervalo fuera 0.10. La prueba χ^2 de Pearson evaluó si los valores $p^*(\alpha)$ seguían una distribución uniforme con $E[\chi^2] = 9$.

En los tres tests convencionales, la distribución de $p^*(\alpha)$ fue aproximadamente uniforme tanto para θ como para $\hat{\theta}$. En el TAI, con θ las distribuciones de $p^*(\alpha)$ resultaron uniformes para l_0 y l_z; con $\hat{\theta}$, los $p^*(\alpha)$ se desviaron de manera significativa de la uniformidad en los tres índices de medición apropiada.

- *Estudio 3*. El objetivo fue examinar la capacidad de identificación de varios tipos de patrones atípicos en los tests convencionales usando la distribución teórica, es decir, asumiendo la distribución normal de l_z y l_z^*. Recurrieron a tres tests con 10, 20 y 50 ítems de opción múltiple, y generaron 200 patrones de respuesta atípica de tres tipos: consecuencia del azar, producto de la bidimensionalidad de θ y por violación del supuesto de independencia local de los ítems. También manipularon la media de la distribución del parámetro de discriminación, $a_j \sim N(1{,}5, 0{,}2)$ y $a_j \sim N(1{,}0, 0{,}2)$.

  El estadístico l_z^* reconoció a estos patrones mejor que l_z, sobre todo en tests cortos con parámetros de discriminación moderados. En tests largos e ítems más discriminativos, ambos índices identificaron la atipicidad de forma similar. Los patrones en los que se simularon respuestas por azar eran mejor detectados que los otros dos tipos (bidimensionalidad y no-independencia local).

Por último, un estudio de simulación de Li y Olejnik (1997) indagó en la influencia que podría tener la dimensionalidad del rasgo latente sobre la distribución de cinco índices de medición apropiada, así como en la identificación de patrones atípicos. Los índices con los que trabajaron fueron l_z, los residuales estandarizados ZW y UZ de Wright y Masters (1982) y Wright y Stone

(1979), y los índices de precaución $ECI2_z$ y $ECI4_z$. Con dos longitudes de test, 30 y 60 ítems, simularon 50000 patrones de respuesta unidimensionales y bidimensionales para cada longitud, con una forma reducida del modelo multidimensional no compensatorio de Sympson (1978), y ajustados al modelo de Rasch. La segunda dimensión se originó con el método de Hoffman (1959); suponiendo que los tests eran de logro y que se construyeron para medir un rasgo dominante, los tests tenían ítems de similar dificultad, pero era más reducida en aquellos que contemplaban la segunda dimensión y, por lo tanto, esta dimensión debía ser menos difícil que la primera. Las distribuciones de los parámetros de habilidad y de dificultad eran normales tipificadas. Para el cálculo de los índices de ajuste de cada sujeto se utilizaron los estimadores máximo-verosímiles.

De cara a estudiar las distribuciones de los índices se calcularon sus medias, desviaciones típicas, índices de sesgo y curtosis, junto con la prueba de normalidad de Kolmogorov-Smirnov. Los resultados indicaron que las medias y las desviaciones típicas se aproximaban a las de la distribución normal, pero con índices de sesgo y curtosis positivos, lo cual contradice en parte a las investigaciones preliminares. La prueba de Kolmogorov-Smirnov fue estadísticamente significativa en todas las condiciones experimentales, lo que según los autores se podría justificar por el tamaño muestral empleado. La conclusión de este cometido fue que los cinco índices de medición apropiada se desviaban de la normalidad, al margen de la dimensionalidad y de las longitudes de test.

Como consecuencia de la ausencia de normalidad en las distribuciones de los índices, Li y Olejnik (1997) buscaron los niveles críticos de las distribuciones empíricas, fijando en el percentil 95 la tasa de identificaciones correctas de las muestras. Los cinco índices se mostraron conservadores y los niveles críticos apenas variaron por la longitud del test o por la dimensionalidad.

Con el mismo procedimiento anterior de generación de patrones, crearon respuestas espurias altas y bajas en los cuatro tests ($n \times Dimensión$) para muestras de 500 sujetos. En estas muestras se seleccionaron 50 sujetos a los que provocaron espurias altas en el 20 % de los ítems, mediante la modificación aleatoria de respuestas incorrectas por respuestas correctas; en otra submuestra de 50 sujetos se incluyeron las espurias bajas por la modificación inversa. La estimación de los parámetros de habilidad y dificultad se realizó con las muestras que contenían patrones normales y atípicos. Se evaluaron las tasas de falsos positivos de los cinco índices a través de los niveles críticos de las distribuciones empíricas de cada uno de ellos obtenidos anteriormente.

El número de falsos positivos para muestras con patrones atípicos fue similar al obtenido en muestras ausentes de ellos. Por tanto, el que hubiera

en una muestra respuestas espurias no supuso una amenaza contra el buen funcionamiento de los índices de medición empleados. En cuanto a las tasas de identificaciones correctas, ZU fue el que más bajas las obtuvo. Las espurias altas en el test bidimensional de 60 ítems fueron las mejor detectadas por todos los índices; las que peor se identificaron fueron las espurias bajas en el test bidimensional de 30 ítems. Con un ANOVA los autores testimoniaron que la potencia de l_z, $ECI2_z$, $ECI4_z$, ZU y ZW depende de la dimensionalidad y del tipo de atipicidad de las respuestas. En general, todos ellos identificaron bien las espurias bajas en tests unidimensionales y las espurias altas en tests bidimensionales, lo cual resulta contradictorio con estudios anteriores a éste. El índice más efectivo fue l_z.

4.2. Procedimiento

Se ha generado una matriz de 1000 patrones de respuesta por el algoritmo de Hambleton y Cook (1983) según el cual, una vez elegido el MRI con el que se va a trabajar, requiere las siguientes especificaciones: el número de sujetos de la muestra (N), la distribución de la habilidad, el número de ítems (n) y los valores de los parámetros de habilidad y de los ítems. El proceso de generación de matrices de respuesta comienza con la sustitución de los parámetros de los ítems en el modelo para obtener la probabilidad de acierto del sujeto i en el ítem j (P_{ij}). Estas probabilidades de acierto se ordenan en una matriz $N \times n$ que es transfomada en otra matriz de igual dimensión, pero cuyos componentes son las puntuaciones en los ítems, es decir, 1s y 0s en función de si el sujeto acierta o falla el ítem. Esta conversión se consigue comparando el valor P_{ij} con un número aleatorio escogido de una distribución uniforme $U(0,1)$ de modo que, si este número aleatorio es menor o igual que P_{ij} entonces $P_{ij} = 1$ y en caso contrario $P_{ij} = 0$. El tamaño muestral es $N = 1000$ sujetos y las longitudes de test son n= 10, 25, 50 y 75 ítems. Con objeto de valorar si la distribución de habilidad afecta a la distribución de l_z se ha trabajado con tres distribuciones de habilidad normales $N(0,1)$: no sesgada, asimétrica positiva con índice de sesgo $g_1 = +1$ y asimétrica negativa con índice de sesgo $g_1 = -1$. También se examinó la influencia del MRI, por lo que se contemplan los resultados obtenidos con los modelos logísticos de 1-p y 2-p, y con el modelo de 3-p. Los parámetros de los ítems proceden del trabajo de Narayanan y Swaminathan (1996) en el que el test original estaba formado por 40 ítems ajustados al modelo de 3-p (Tabla 4.1). Las medias y desviaciones típicas de los parámetros de dificultad (b_j) y discriminación (a_j) de las cuatro longitudes de test se describen en la Tabla 4.2; para el modelo de

3-p, el parámetro de pseudo-azar es constante para todos los ítems ($c_j = 0{,}20$). La posible influencia del parámetro de discriminación se valoró incrementando los parámetros originales en magnitudes de 0.30 y 0.60.

Tabla 4.1. Parámetros de los ítems de Narayanan y Swaminathan (1996)

Item	a_j	b_j	c_j	**Item**	a_j	b_j	c_j
1	0.44	-0.30	0.20	21	0.92	1.13	0.20
2	0.55	-1.06	0.20	22	0.64	-1.55	0.20
3	0.82	1.02	0.20	23	1.01	0.81	0.20
4	0.52	-1.96	0.20	24	0.61	-0.53	0.20
5	1.02	1.28	0.20	25	0.70	1.05	0.20
6	0.82	0.61	0.20	26	1.02	0.64	0.20
7	0.92	0.42	0.20	27	0.48	2.12	0.20
8	0.65	1.68	0.20	28	1.01	0.91	0.20
9	0.56	-2.70	0.20	29	0.53	0.87	0.20
10	0.29	-1.39	0.20	30	0.36	-2.63	0.20
11	0.35	-1.12	0.20	31	1.12	-1.21	0.20
12	0.31	-1.37	0.20	32	0.86	-0.57	0.20
13	1.05	0.10	0.20	33	0.59	-1.29	0.20
14	0.51	-0.09	0.20	34	0.56	0.40	0.20
15	0.73	0.61	0.20	35	1.09	1.11	0.20
16	0.88	0.95	0.20	36	0.88	-0.93	0.20
17	1.11	-0.35	0.20	37	0.96	-1.21	0.20
18	1.32	0.57	0.20	38	1.06	2.11	0.20
19	0.55	1.09	0.20	39	0.92	0.62	0.20
20	1.40	1.64	0.20	40	0.75	-1.01	0.20

Tabla 4.2. Estadísticos descriptivos de b_j y a_j

	$n = 10$	$n = 25$	$n = 50$	$n = 75$
$\bar{x}_{b_j}$	-0.178	0.022	-0.030	-0.013
s_{b_j}	1.551	1.204	1.231	1.214
$\bar{x}_{a_j}$	0.717	0.747	0.793	0.778
s_{a_j}	0.327	0.298	0.266	0.276

Para analizar el efecto del método de estimación de los parámetros de habilidad y de los ítems sobre la distribución de l_z se han empleado dos procedimientos de estimación: MVM y EAP, llevadas a cabo con el programa BILOG v. 3.04 (Mislevy y Bock, 1990). El análisis del método de estimación se ha

acompañado de un estudio de recubrimiento de los parámetros de habilidad mediante los tres mismos índices empleados por Nering (1995): el coeficiente de correlación de Pearson (ρ), la exponencial de la raíz del error medio cuadrático $RMSE$ (Ecuación 4.1) y el error medio con signo ASB (Ecuación 4.2).

La distribución de l_z se ha descrito con los estadísticos media, desviación típica, sesgo y curtosis. Se han completado estos resultados con la prueba de normalidad de Lilliefors (1967; Marascuilo y McSweeney, 1977), una variante de la prueba de Kolmogorov-Smirnov útil cuando la media y la desviación típica de la distribución son desconocidas. La variable bajo estudio (l_z) está estandarizada respecto a su media y desviación típica empírica. Los datos normalizados Z_i se comparan con los de la distribución normal $N(0,1)$ mediante el estadístico de Kolmogorov-Smirnov de máxima diferencia (M.D.) definido como:

$$M.D. = \text{máx}\,|P(Z < Z_i) - \hat{P}(Z < Z_i)| \tag{4.3}$$

donde $P(Z < Z_i)$ es la probabilidad de frecuencias acumuladas bajo la curva normal correspondiente a Z_i y $\hat{P}(Z < Z_i)$ son las frecuencias relativas acumuladas de la variable normalizada. La hipótesis nula sobre la normalidad de la distribución de l_z se rechaza si $M.D. > L_{N,(1-\alpha)}$, valor crítico que depende del tamaño muestral y del nivel de significación. Esta prueba no está afectada ni por la localización ni por la escala de l_z y el contraste lo realiza según la forma de la distribución por aproximación no lineal a la tabla de Lilliefors. Tanto los estadísticos descriptivos como la prueba no paramétrica de Lilliefors se han calculado con el programa SySTAT v. 10.0 (2000).

También se han evaluado las tasas de falsos positivos (FP) del estadístico l_z para un contraste bilateral en dos niveles de significación: 5 % y 1 %; en el caso de que la distribución de dicho estadístico no sea la normal tipificada, la tasas de FP bien serán infraestimadas o bien sobrestimadas.

Los análisis se han agrupado en tres bloques:

- *Bloque 1.* Modelo logístico de 1-p. En este modelo el parámetro de discriminación es constante e igual a 1 para todos los ítems. El objetivo es obtener información sobre la distribución de l_z manipulando la longitud del test (10, 25, 50 y 75 ítems), la distribución de habilidad:
  - $\theta \sim N(0,1)$ centrada y no sesgada;
  - $\theta \sim N(0,1)$ centrada y sesgada en magnitud $g_1 = +1$;
  - $\theta \sim N(0,1)$ centrada y sesgada en magnitud $g_1 = -1$;

comparando dos métodos de estimación de parámetros (MVM y EAP) frente a parámetros verdaderos. En total 36 estudios de la distribución de l_z en el modelo de 1-p.

- *Bloque 2*. Modelo logístico de 2-p. En este apartado, además de las tres variables descritas en el Bloque 1 (número de ítems, distribución de habilidad y método de estimación), se examina la repercusión del parámetro de discriminación:

  - Condición 1 (C1): valores de a_j verdaderos.
  - Condición 2 (C2): valores de a_j verdaderos con incremento de 0.30.
  - Condición 3 (C3): valores de a_j verdaderos con incremento de 0.60.

  Con todo, son en total 108 distribuciones de l_z.

- *Bloque 3*. Modelo de 3-p. Para completar los resultados de los dos bloques anteriores se incluye este modelo con el fin de poder valorar si existe influencia del parámetro de pseudo-azar en la distribución del estadístico l_z. Para ello, a todas las condiciones experimentales del Bloque 2 se les han añadido a los ítems de los tests el parámetro c_j con una magnitud de 0.20. De nuevo son 108 análisis con los que se pretende perfilar esta investigación sobre la distribución de l_z.

En la Tabla 4.3 se resumen las condiciones experimentales más relevantes de los trabajos expuestos en la sección 4.1 y que han suscitado la presente tesis. En la segunda columna de esta tabla se ha denotado con V las aplicaciones que trabajaron con datos reales y con S las que simularon los patrones de respuestas.

4.3. Resultados

Los resultados se han recogido en tablas en las que la información se agrupa de la siguiente manera: en la primera tabla aparecen los datos referentes al estudio de recubrimiento de los parámetros de habilidad; las tres tablas siguientes recogen los valores de los estadísticos descriptivos de l_z, la prueba de normalidad de Lilliefors y el error tipo I empírico, ordenadas de modo que la primera de ellas contiene los resultados sobre l_z empleando parámetros verdaderos, la segunda empleando los estimadores máximo-verosímiles y la tercera con los estimadores obtenidos por EAP. La nomenclatura utilizada en cada tabla y común a todas ellas es: n para la longitud del test, C la abreviatura de la

condición experimental en la que se trabaja con el parámetro de discriminación y N el tamaño muestral, esta última sólo en los datos referidos a los estimadores máximo-verosímiles ya que la función de verosimilitud no converge cuando en el patrón todas las respuestas son 1s ó 0s y, por lo tanto, estos sujetos se eliminaron de la muestra para el cálculo de los estadísticos descriptivos de l_z. En las tablas correspondientes a la distribución de l_z, $\hat{\mu}_{l_z}$ es la media, $\hat{\sigma}_{l_z}$ la desviación típica, g_1 el coeficiente de sesgo o asimetría de Fisher, g_2 el índice de apuntamiento o curtosis, $M.D.$ la prueba estadística para la normalidad de Lilliefors y $p(\alpha)$ el error tipo I asociado a ella.

Tabla 4.3. Resumen de las condiciones experimentales de los estudios citados y las del presente trabajo sobre la distribución normal de l_z

Variables / **Estudio**	Datos	Modelo	n	θ	$\hat{\theta}$	$\ddot{}l_z \sim N$?
Drasgow *et al.* (1985)	V	3-p	SAT-V 85		MV	Sí
Drasgow *et al.* (1987)	S	3-p	SAT-V 85	Normal $g_1 = 0$	MV	Sí
Reise (1995)	S	2-p	MPQ 24, 27 28, 34	Constante Normal $g_1 = 0$	MV	No $g_1 < 0$
Noonan *et al.* (1992)	S	2-p 3-p	40 80	Normal $g_1 = 0$	-	No $g_1 < 0$ $g_2 > 0$
Nering (1995)	S	3-p	25	Constante Normal $g_1 = 0$	MV EAP	No $g_1 < 0$ $g_2 > 0$
Nering (1997)	S	3-p	TAI 250 500	Normal $g_1 = 0$	MV	No $g_1 < 0$ $g_2 > 0$
McLeod y Lewis (1999)	S	3-p	GRE-Q 28	Uniforme	MV	No $\hat{\mu} < 0$ $\hat{\sigma} < 0$
van Krimpen-Stoop y Meijer (1999)	S	2-p	20,50 80; TAI:400	Constante Normal $g_1 = 0$	MVP	No $g_1 < 0$ $g_2 > 0$
Li y Olejnik (1997)	S	1-p	30 60	Normal $g_1 = 0$	MV	No $g_1 > 0$ $g_2 > 0$
Este estudio	S	1-p 2-p 3-p	10, 25 50, 75	Normal $g_1 = -1, 0, +1$	MVM EAP	??

En el Apéndice se encuentran los gráficos de cada una de las distribuciones de l_z que han resultado de este estudio experimental.

4.3.1. Bloque 1: Modelo logístico de 1-p

Distribución de habilidad no sesgada

Estudio de recubrimiento (Tabla 4.4)

Tabla 4.4. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$

	MVM				EAP		
n	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	0.765	2.373	-0.031	1000	0.768	1.867	-0.044
25	0.903	1.660	-0.032	1000	0.905	1.514	-0.039
50	0.943	1.467	-0.029	999	0.943	1.391	-0.032
75	0.958	1.372	-0.038	1000	0.959	1.327	-0.040

El coeficiente de correlación de Pearson como indicador del grado de recubrimiento del parámetro de habilidad es el casi el mismo en ambos procedimientos de estimación, apareciendo las diferencias en el tercer decimal. Con MVM $\rho_{\theta,\hat{\theta}} = 0{,}765$ y con EAP $\rho_{\theta,\hat{\theta}} = 0{,}768$ en el test de 10 ítems, en el test de 25 ítems $\rho_{\theta,\hat{\theta}} = 0{,}903$ con MVM y $\rho_{\theta,\hat{\theta}} = 0{,}905$ con EAP, si hay 50 ítems $\rho_{\theta,\hat{\theta}} = 0{,}943$ en ambos procedimientos de estimación, y con 75 ítems $\rho_{\theta,\hat{\theta}} = 0{,}958$ con MVM y $\rho_{\theta,\hat{\theta}} = 0{,}959$ con EAP. Según este coeficiente de correlación, cuanto mayor es el número de ítems del test la estimación del parámetro de habilidad mejora independientemente del procedimiento de estimación que se escoja.

Al examinar el índice $RMSE$ se puede apreciar que la estimación con el procedimiento bayesiano es más acertada que con el de máxima verosimilitud marginal, aunque las diferencias entre ellos no sean muy grandes. En el test de 10 ítems es donde las discrepancias son mayores: con MVM, $RMSE = 2{,}373$, mientras que con EAP, $RMSE = 1{,}867$. Conforme aumenta el número de ítems las diferencias en la estimación de θ se van reduciendo y en el test de 75 ítems empleando MVM, $RMSE = 1{,}372$, y con EAP, $RMSE = 1{,}327$. Al igual que el coeficiente de correlación, $RMSE$ confirma que conforme aumenta la longitud del test mejor es la estimación del parámetro y utilizando el método EAP se logran mejores resultados.

En cuanto al índice ASB, ambos procedimientos sobrestiman el parámetro verdadero de habilidad. Con el procedimiento de estimación esperada a posteriori se sobrestima el parámetro de habilidad más que con máxima verosimilitud marginal. Por ejemplo, en el test de 10 ítems $ASB = -0{,}031$ y $ASB = -0{,}044$ con MVM y EAP, respectivamente, con 50 ítems por MVM es $-0{,}029$ y por EAP es $-0{,}032$. No aparece un efecto definido de la longitud del test en ASB.

Distribución de l_z con parámetros verdaderos (Tabla 4.5)

Tabla 4.5. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

							α nominal	
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0,05	0,01
10	0.042	0.949	-0.764**	0.088	0.084**	0.000	0.046	0.014
25	-0.021	0.986	-0.430**	0.234	0.027	0.080	0.047	0.010
50	-0.004	0.997	-0.421**	-0.162	0.052**	0.000	0.048	0.009
75	-0.030	0.999	-0.401**	0.123	0.046**	0.000	0.050	0.009

* p¡0.05

** p¡0.01

Empleando parámetros verdaderos, el estadístico de tendencia central de l_z es infravalorado en tres de las cuatro longitudes de test. El test de 10 ítems es el único que da un valor central mayor a 0, $\hat{\mu}_{l_z} = 0{,}042$, mientras que con 25, 50 y 75 ítems las medias son menores a la estándar, siendo la más próxima a este valor la obtenida en el test de 50 ítems, $\hat{\mu}_{l_z} = -0{,}004$. Sobre la media, el aumento del número de ítems en el test no tiene relevancia, algo que sí repercute en la desviación típica de l_z: en el test de 10 ítems $\hat{\sigma}_{l_z} = 0{,}949$, magnitud que se va incrementando hasta $\hat{\sigma}_{l_z} = 0{,}999$ en el test de 75 ítems. En todos los casos, el estadístico de dispersión es inferior al esperado.

El índice de sesgo de la distribución es menor a 0 en los cuatro tests, pero acercándose a este valor conforme aumenta la longitud del test y, por lo tanto, corrigiéndose en cierta medida la asimetría negativa de la distribución. Cuando $n = 10$ $g_1 = -0{,}764$, el valor más bajo y alrededor de -0.415 en los tests de 25, 50 y 75 ítems. En cualquier caso, todas presentan diferencias estadísticas con la simetría al 1 % de nivel de significación.

En cuanto al achatamiento de la curva, el índice g_2 no se aleja de 0 significativamente, siendo el test de 10 ítems el que presenta la curva más mesocúrtica de todos, $g_2 = 0{,}088$. El único test con valores negativos de forma es el de 50 ítems, $g_2 = -0{,}162$, y el valor más alto es $g_2 = 0{,}234$ en el test de 25 ítems, poniendo de manifiesto la ausencia de relación del tamaño del test sobre este índice de curtosis.

Según la prueba de normalidad de Lilliefors, en el único test que l_z sigue una distribución normal es el de 25 ítems, donde $M.D. = 0{,}027$ con $p(\alpha) = 0{,}080$.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.6)

Tabla 4.6. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM									
							α nominal		
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	0.213	0.897	-0.873**	0.451**	0.112**	0.000	0.048	0.019	1000
25	0.005	0.878	-0.497**	0.870**	0.031*	0.022	0.048	0.014	1000
50	-0.130	0.941	-0.448**	0.229	0.049**	0.000	0.052	0.012	999
75	-0.220	0.911	-0.268**	0.221	0.025	0.156	0.047	0.012	1000

* p¡0.05

** p¡0.01

Empleando estimadores máximo-verosímiles de la habilidad y la dificultad de los ítems, las medias de l_z en los tests de 10 y 25 ítems son mayores que 0 sobre todo en el primero de ellos, $\hat{\mu}_{l_z} = 0{,}213$, y casi el valor de referencia en el segundo, $\hat{\mu}_{l_z} = 0{,}005$. En los tests de 50 y 75 ítems, la media de l_z es infravalorada, $\hat{\mu}_{l_z} = -0{,}130$ y $\hat{\mu}_{l_z} = -0{,}220$.

La desviación típica de la distribución también es inferior a la estándar, siendo el valor más bajo $\hat{\sigma}_{l_z} = 0{,}878$ del test de 25 ítems y el más cercano a 1, $\hat{\sigma}_{l_z} = 0{,}941$ del test de 50 ítems.

Las distribuciones son asimétricas negativas con $p(\alpha) < 0{,}01$, restableciéndose el sesgo conforme aumenta el número de ítems: con 10 ítems $g_1 = -0{,}873$, con 25 ítems $g_1 = -0{,}497$, con 50 $g_1 = -0{,}448$ y con 75 ítems $g_1 = -0{,}268$.

Las curvas de los tests de 10 y 25 ítems son leptocúrticas, $g_2 = 0{,}451$ y $g_2 = 0{,}870$, mientras que las de los tests de 50 y 75 ítems son mesocúrticas, $g_2 = 0{,}229$ y $g_2 = 0{,}221$.

El número de ítems del test no repercute en los resultados de las medias pero sí de manera positiva en las desviaciones típicas, en los índices de asimetría y en los de curtosis.

Sólo en el test de 75 ítems l_z sigue una distribución normal, $M.D. = 0{,}025$ con $p(\alpha) = 0{,}156$. De las otras tres, si $n = 25$ la prueba de Lilliefors es significativa al 5 %, y si $n = 10$ y $n = 50$ lo es al 1 %.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.7)

Tabla 4.7. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

							α nominal	
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	0.307	0.900	-0.956**	0.494**	0.109**	0.000	0.050	0.019
25	0.177	0.916	-0.614**	0.718**	0.036**	0.004	0.045	0.014
50	0.020	0.960	-0.545**	0.114	0.061**	0.000	0.047	0.011
75	-0.097	0.924	-0.355**	0.158	0.034**	0.010	0.045	0.013

* p¡0.05

** p¡0.01

La media de la distribución de l_z con estimadores bayesianos se acerca a 0 conforme aumenta el número de ítems, aunque el test de 75 ítems no confirma esta tendencia. Los valores medios son $\hat{\mu}_{l_z} = 0{,}307$ con 10 ítems, $\hat{\mu}_{l_z} = 0{,}177$ con 25 ítems, $\hat{\mu}_{l_z} = 0{,}020$ con 50 ítems y $\hat{\mu}_{l_z} = -0{,}097$ con 75 ítems. Esto mismo ocurre con los valores de las desviaciones típicas, las cuales se aproximan a 1 conforme aumenta el número de ítems excepto en el de 75 ítems: $\hat{\sigma}_{l_z} = 0{,}900$ si $n = 10$, $\hat{\sigma}_{l_z} = 0{,}916$ si $n = 25$, $\hat{\sigma}_{l_z} = 0{,}960$ si $n = 50$, pero $\hat{\sigma}_{l_z} = 0{,}924$ si $n = 75$.

A mayor número de ítems la distribución es menos asimétrica negativa, con 10 ítems $g_1 = -0{,}956$ y con 75 ítems $g_1 = -0{,}355$; aun así, en todas ellas existen diferencias significativas respecto de la curva simétrica con $p(\alpha) < 0{,}01$.

También hay cierto efecto de la longitud del test sobre el estadístico de curtosis, el cual diseña curvas leptocúrticas en los tests de 10 y 25 ítems, $g_2 = 0{,}494$ y $g_2 = 0{,}718$, y mesocúrticas en los dos restantes, $g_2 = 0{,}114$ con $n = 50$ y $g_2 = 0{,}158$ con $n = 75$.

Con el procedimiento EAP, el estadístico l_z no sigue una distribución normal según la prueba de no paramétrica de Lilliefors con $p(\alpha) < 0{,}01$.

Las tasas de error tipo I se mantienen próximas al valor α nominal tanto con parámetros verdaderos como con parámetros estimados; son más consistentes al nivel de 0.05 y en la mayoría de los casos del nivel $\alpha = 0{,}01$ son sobrestimados.

Distribución de habilidad sesgada positiva

Estudio de recubrimiento (Tabla 4.8)

Tabla 4.8. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$

	MVM				EAP		
n	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	0.782	2.436	-0.007	1000	0.779	1.873	0.000
25	0.901	1.663	-0.008	1000	0.898	1.552	0.001
50	0.942	1.465	-0.023	996	0.942	1.404	-0.010
75	0.965	1.332	-0.021	994	0.966	1.298	-0.012

El aumento del número de ítems en el test mejora la estimación del parámetro de habilidad según el coeficiente de correlación de Pearson y el índice $RMSE$, tanto si la habilidad es estimada por MVM como por EAP. Con el procedimiento MVM en el test de 10 ítems $\rho_{\theta,\hat{\theta}} = 0{,}782$ y $RMSE = 2{,}436$, y en el test de 75 ítems $\rho_{\theta,\hat{\theta}} = 0{,}965$ y $RMSE = 1{,}332$. Tras la estimación EAP $\rho_{\theta,\hat{\theta}} = 0{,}779$ y $RMSE = 1{,}873$ si $n = 10$, y $\rho_{\theta,\hat{\theta}} = 0{,}966$ y $RMSE = 1{,}298$ si $n = 75$. Los coeficientes de correlación obtenidos con ambos procedimientos son similares y están igualados en el test de 50 ítems, $\rho_{\theta,\hat{\theta}} = 0{,}942$. Sin embargo, en el índice $RMSE$ hay diferencias más claras entre las dos estimaciones, logrando los mejores resultados el procedimiento EAP ya que los valores están más próximos a 1.

En general, ambos métodos tienden a sobrestimar el parámetro de sujeto, pero comparando los valores de ASB de uno y otro, las más altas sobrestimaciones de θ se obtienen con el método MVM y a mayor tamaño de test, con $n = 10$ $ASB = -0{,}007$ y con $n = 75$ $ASB = -0{,}021$, y con EAP si $n = 10$ $ASB = -0{,}000$ y si $n = 75$ $ASB = -0{,}012$.

Distribución de l_z con parámetros verdaderos (Tabla 4.9)

Los valores medios de la distribución con parámetros verdaderos son muy cercanos a 0, especialmente en los tests con menor número de ítems: en el de 10 ítems $\hat{\mu}_{l_z} = -0{,}002$ y en el de 25 ítems es $\hat{\mu}_{l_z} = 0{,}006$. El valor más alejado es el del test de 50 ítems, $\hat{\mu}_{l_z} = 0{,}055$.

En cuanto a la dispersión de l_z, hay una buena aproximación al valor estándar sin que se aprecie influencia alguna del tamaño del test. El resultado más aproximado a 1 es el del test de 25 ítems, $\hat{\sigma}_{l_z} = 1{,}011$, y el más distante el del test de 10 ítems, $\hat{\sigma}_{l_z} = 0{,}958$.

Las distribuciones son asimétricas negativas [$p(\alpha) < 0{,}01$] y más cuanto más corto es el test; con 10 ítems $g_1 = -0{,}789$, con 25 $g_1 = -0{,}511$, con 50 $g_1 = -0{,}530$ y con 75 ítems $g_1 = -0{,}297$.

Las cuatro longitudes de test delinean tanto curvas mesocúrticas con 25 y 75 ítems, $g_2 = 0{,}021$ y $g_2 = -0{,}024$, como leptocúrticas en los de 10 y 50 ítems, $g_2 = 0{,}404$ y $g_2 = 0{,}522$. El aumento de ítems en el test no influye en la forma de las curvas.

La prueba de normalidad de Lilliefors considera que la distribución de l_z no es normal en los tests de 10, 25 y 50 ítems al nivel de significación del 1 %, pero sí en el test de 75 ítems, $M.D. = 0{,}028$ con $p(\alpha) = 0{,}059$.

Tabla 4.9. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

							α nominal	
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	-0.002	0.958	-0.789**	0.404**	0.074**	0.000	0.049	0.011
25	0.006	1.011	-0.511**	0.021	0.056**	0.000	0.047	0.013
50	0.055	0.967	-0.530**	0.522**	0.049**	0.000	0.050	0.013
75	-0.034	1.037	-0.297**	-0.024	0.028	0.059	0.051	0.009

* p¡0.05

** p¡0.01

Distribución de l_z con parámetros estimados con MVM (Tabla 4.10)

Tabla 4.10. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

							α nominal		
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	-0.023	0.991	-0.797**	0.267	0.096**	0.000	0.054	0.013	1000
25	0.043	0.898	-0.471**	0.155	0.043**	0.000	0.049	0.014	1000
50	-0.086	0.900	-0.400**	0.875**	0.029*	0.044	0.056	0.014	996
75	-0.141	0.965	-0.241**	0.160	0.034**	0.008	0.043	0.012	994

* p¡0.05

** p¡0.01

La media de l_z con parámetros máximo-verosímiles es infravalorada en los tests de 10, $\hat{\mu}_{l_z} = -0{,}023$, de 50, $\hat{\mu}_{l_z} = -0{,}086$, y de 75 ítems, $\hat{\mu}_{l_z} = -0{,}141$, y sobrevalorada en el test de 25 ítems, $\hat{\mu}_{l_z} = 0{,}043$. La cercanía al valor esperado se obtiene en los tests con menor número de ítems.

La desviación típica es inferior a 1 en los cuatro tests sin que se manifieste intervención del tamaño de los mismos en los resultados. La desviación típica

más próxima a 1 es la del test de 10 ítems, $\hat{\sigma}_{l_z} = 0{,}991$, y la más alejada es la del test de 25 ítems, $\hat{\sigma}_{l_z} = 0{,}898$.

Las distribuciones son sesgadas negativas con $p(\alpha) < 0{,}01$, de modo más marcado en los tests con menor número de ítems, por ejemplo, con $n = 10$ $g_1 = -0{,}797$, logrando cierta mejoría con $n = 75$ ítems para el que $g_1 = -0{,}241$.

Aparece una sola curva leptocúrtica de l_z en el test de 50 ítems, $g_2 = 0{,}875$ con $p(\alpha) < 0{,}01$; las demás son mesocúrticas, siendo las más semejantes en forma a la de la normal las de los tests de 25 ítems, $g_2 = 0{,}155$, y de 75 ítems, $g_2 = 0{,}160$.

El estadístico de medición apropiada l_z no sigue una distribución normal cuando se calcula con parámetros estimados por MVM; la significación estadística del test de 50 ítems es al 5 %, $M.D. = 0{,}029$ con $p(\alpha) = 0{,}044$, y el resto de los tests la presentan al 1 %.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.11)

Tabla 4.11. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

							α nominal	
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	0.060	0.982	-0.888**	0.379*	0.100**	0.000	0.059	0.014
25	0.211	0.932	-0.592**	0.066	0.066**	0.000	0.044	0.013
50	0.059	0.913	-0.394**	0.291	0.039**	0.001	0.052	0.013
75	-0.015	0.978	-0.337**	0.142	0.038**	0.002	0.047	0.012

* p¡0.05

** p¡0.01

Con parámetros estimados por el procedimiento bayesiano los valores medios no son sensibles al tamaño del test; en el test de 75 ítems $\hat{\mu}_{l_z} = -0{,}015$, la más próxima a 0, en los de 10 y 50 ítems $\hat{\mu}_{l_z} = 0{,}060$ y $\hat{\mu}_{l_z} = 0{,}059$, mientras que con 25 ítems el valor medio se aleja en magnitud 0.211.

Las desviaciones típicas están infravaloradas con estos estimadores y su rango ha sido de 0.913, del test de 50 ítems, a 0.982, del test de 10 ítems. Este estadístico tampoco es muy influenciable por el número de ítems.

La asimetría de la curvas es negativa [$p(\alpha) < 0{,}01$], mejorando los resultados de g_1 conforme aumenta la longitud del test: con 10 ítems $g_1 = -0{,}888$ y con 75 ítems $g_1 = -0{,}337$.

Las curvas se dibujan mesocúrticas, excluyendo la del test de 10 ítems, $g_2 = 0{,}379$ con $p(\alpha) < 0{,}05$, que es leptocúrtica. La curva menos apuntada es la del test de 25 ítems, $g_2 = 0{,}066$. Con esto también se aprecia la ausencia de

relación entre el tamaño del test y g_2.

La prueba de normalidad es estadísticamente significativa en los cuatro tests con $p(\alpha) < 0{,}01$, es decir, se rechaza la hipótesis nula de normalidad de la distribución de l_z.

Las tasas de error tipo I están cerca del valor nominal en todos los casos estudiados, siendo sobrestimados si α nominal es 0.01.

Distribución de habilidad sesgada negativa

Estudio de recubrimiento (Tabla 4.12)

Tabla 4.12. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$

	MVM				EAP		
n	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	0.759	2.473	0.010	1000	0.756	1.922	-0.000
25	0.894	1.649	0.020	989	0.899	1.548	0.002
50	0.938	1.471	0.018	995	0.939	1.412	0.006
75	0.960	1.357	0.025	997	0.959	1.329	0.014

Los coeficientes de correlación de Pearson son casi los mismos, salvo errores de redondeo, con ambos procedimientos de estimación del parámetro de habilidad: con 10 ítems $\rho_{\theta,\hat{\theta}} = 0{,}759$ en el procedimiento MVM y $\rho_{\theta,\hat{\theta}} = 0{,}756$ con EAP, en el test de 25 ítems para MVM $\rho_{\theta,\hat{\theta}} = 0{,}894$ y para EAP $\rho_{\theta,\hat{\theta}} = 0{,}899$, llegando a la mínima diferencia en el test de 75 ítems en donde $\rho_{\theta,\hat{\theta}} = 0{,}960$ con MVM y $\rho_{\theta,\hat{\theta}} = 0{,}959$ con EAP. Los resultados ponen de manifiesto la mejoría en la estimación de este parámetro conforme aumenta la longitud del test.

El índice $RMSE$ también arroja resultados muy similares en los dos procedimientos, aunque se hallan mayores diferencias que con $\rho_{\theta,\hat{\theta}}$. Por el método de estimación MVM en los tests de 10 y 25 ítems se obtienen los valores más altos de $RMSE$ y, por lo tanto, los que indican una peor estimación de la habilidad, $RMSE = 2{,}473$ y $RMSE = 1{,}649$, respectivamente. Un valor intermedio a estos dos es el del test de 10 ítems con EAP, $RMSE = 1{,}922$. Los valores más próximos a 1 y con menos diferencias entre los dos métodos de estimación son los del test de 75 ítems, para el que $RMSE = 1{,}357$ con MVM y $RMSE = 1{,}329$ con EAP.

El índice ASB es positivo en todos los casos estudiados, lo que representa la infraestimación del parámetro al margen del método de estimación empleado;

esta tendencia es menos acusada en el procedimiento EAP. Sin embargo, conforme aumenta el número de ítems, el estimador se aleja del valor paramétrico y en el test de 75 ítems estimando con MVM resulta $ASB = 0{,}025$ y con EAP $ASB = 0{,}014$.

Distribución de l_z con parámetros verdaderos (Tabla 4.13)

Tabla 4.13. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

							α nominal	
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	-0.023	0.994	-1.050**	1.464**	0.094**	0.000	0.048	0.022
25	-0.037	1.046	-0.808**	0.897**	0.073**	0.000	0.047	0.018
50	-0.021	1.016	-0.335**	0.027	0.032*	0.019	0.042	0.006
75	-0.055	1.038	-0.429**	0.372*	0.038**	0.002	0.049	0.015

* p¡0.05

** p¡0.01

Si se calculan los estadísticos descriptivos de l_z con los parámetros verdaderos de habilidad y de los ítems, la media es infravalorada sin relación con la longitud del test en los resultados obtenidos. En el test de 75 ítems está el valor medio más bajo, $\hat{\mu}_{l_z} = -0{,}055$, y en el de 50 ítems el más cercano a 0, $\hat{\mu}_{l_z} = -0{,}021$.

Los estadísticos de dispersión son mayores a 1, excepto en el test de 10 ítems en el que es menor, $\hat{\sigma}_{l_z} = 0{,}994$. Con $n = 50$ se consiguen los resultados de media y desviación típica más cercanos a los estándar, $\hat{\sigma}_{l_z} = 1{,}016$, aunque en general, en los cuatro tests se obtienen estadísticos descriptivos próximos a los de la distribución normal tipificada.

La distribución es muy asimétrica negativa en el test con 10 ítems, $g_1 = -1{,}050$ con $p(\alpha) < 0{,}01$, sesgo que se suaviza conforme aumenta el número de ítems: con 25 ítems $g_1 = -0{,}808$, con 50 ítems $g_1 = -0{,}335$ y con 75 ítems $g_1 = -0{,}429$, sin que por ello se llegue a la no-significación estadística.

El estadístico de curtosis indica que las curvas son en su mayoría leptocúrticas, con 10 ítems $g_2 = 1{,}464$, con 25 ítems $g_2 = 0{,}897$, ambas al nivel de significación del 1 %, y con 75 ítems $g_2 = 0{,}372$ con $p(\alpha) < 0{,}05$. Sólo en el test de 50 ítems la distribución tiene una altura media, $g_2 = 0{,}027$. Existe un efecto corrector del tamaño del test sobre el estadístico de forma en el sentido de que cuanto mayor es el número de ítems menor es el apuntamiento de la curva.

La prueba de Lilliefors es significativa al nivel del 5 % en el test de 50 ítems, $M.D. = 0{,}032$ con $p(\alpha) = 0{,}019$, y con $p(\alpha) < 0{,}01$ en los otros tres, por lo que se descartaría la normalidad de la distribución de l_z.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.14)

Tabla 4.14. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

							α nominal		
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	0.046	0.976	-0.932**	0.638**	0.095**	0.000	0.046	0.020	1000
25	-0.418	1.250	-1.160**	2.852**	0.080**	0.000	0.046	0.018	989
50	-0.480	1.240	-0.596**	0.540**	0.042**	0.000	0.041	0.012	995
75	-0.404	1.229	-0.583**	0.655**	0.061**	0.000	0.055	0.014	997

* p¡0.05

** p¡0.01

Las medias y las desviaciones típicas obtenidas con estadísticos estimados por MVM se alejan de los valores esperados, infravalorando el primero de ellos y sobrevalorando el segundo salvo en el test de 10 ítems. Con esta longitud $\hat{\mu}_{l_z} = 0{,}046$ y $\hat{\sigma}_{l_z} = 0{,}976$, los cuales son los mejores que se han obtenido con estadísticos máximo-verosímiles. La media más alejada de 0 es la del test de 50 ítems, $\hat{\mu}_{l_z} = -0{,}480$, y la desviación más apartada de 1 es la del test de 25 ítems, $\hat{\sigma}_{l_z} = 1{,}250$.

Las distribuciones son asimétricas negativas con $p(\alpha) < 0{,}01$, sobre todo en los tests de 10 y 25 ítems, $g_1 = -0{,}932$ y $g_1 = -1{,}160$, restableciéndose con el incremento del número de ítems en el test.

La curva más parecida a la normal en cuanto a la curtosis es la del test de 50 ítems en donde $g_2 = 0{,}540$, pero todas son significativamente leptocúrticas con $p(\alpha) < 0{,}01$, e.g., en el test de 25 ítems $g_2 = 2{,}852$. Los tests de 10 y 75 ítems tienen índices de curtosis muy similares, $g_2 = 0{,}638$ en el primero de ellos y $g_2 = 0{,}655$ en el segundo, por lo que no se podría confirmar una mejoría de este índice por medio del tamaño del test.

Las distribuciones de l_z no siguen una ley normal con $p(\alpha) < 0{,}01$, decisión tomada a partir de los resultados de la prueba no paramétrica empleada.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.15)

Tabla 4.15. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

							α nominal	
n	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	0.131	0.971	-1.012**	0.671**	0.111**	0.000	0.048	0.020
25	0.123	0.963	-0.857**	1.098**	0.057**	0.000	0.043	0.019
50	-0.030	0.974	-0.387**	-0.097	0.043**	0.000	0.052	0.008
75	-0.039	0.950	-0.423**	0.188	0.041**	0.000	0.053	0.015

* p¡0.05

** p¡0.01

Tras la estimación EAP los resultados de las medias de l_z más cercanos al esperado son los de los tests con 50 y 75 ítems, que también son en los que se infravalora dicho estadístico, $\hat{\mu}_{l_z} = -0{,}030$ y $\hat{\mu}_{l_z} = -0{,}039$. La media más alejada es la del test con 10 ítems, $\hat{\mu}_{l_z} = 0{,}131$.

Las desviaciones típicas se encuentran en un intervalo de 0.950 del test de 75 ítems y 0.974 del test de 50 ítems, disminuyendo con el aumento del número de ítems.

La asimetría de las distribuciones está corregida a consecuencia de la longitud del test: con 10 ítems $g_1 = -1{,}012$ y con 50 ítems $g_1 = -0{,}387$, sin que por ello se alcance la simetría. Existen diferencias significativas relativas a la curva simétrica con $p(\alpha) < 0{,}01$.

Las curvas mesocúrticas aparecen en los tests de 50 y 75 ítems, $g_2 = -0{,}097$ y $g_2 = 0{,}188$, siendo la primera de éstas la única que se sitúa por debajo de la normal. Sin embargo, en los tests de 10 y 25 ítems se sobrepasa la altura media y las curvas son leptocúrticas con $p(\alpha) < 0{,}01$, $g_2 = 0{,}671$ y $g_2 = 1{,}098$.

La prueba no paramétrica de Lilliefors es estadísticamente significativa en todos los casos con $p(\alpha) < 0{,}01$, por lo tanto, se concluye con la no-normalidad de la distribución de l_z cuando se emplean estadísticos estimados por EAP.

Las tasas de FP se aproximan a los valores nominales con los tres tipos de parámetros empleados; son más consistentes e infraestimados si $\alpha = 0{,}05$, y menos consistentes y sobrestimados si $\alpha = 0{,}01$.

4.3.2. Bloque 2: Modelo logístico de 2-p

Distribución de habilidad no sesgada

Estudio de recubrimiento (Tabla 4.16)

Tabla 4.16. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$								
		MVM				EAP		
n	C	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	1	0.663	2.927	0.019	975	0.685	2.034	-0.044
	2	0.740	2.601	0.047	992	0.760	1.884	-0.044
	3	0.809	2.147	0.020	1000	0.812	1.765	-0.044
25	1	0.841	1.896	-0.009	996	0.849	1.674	-0.043
	2	0.894	1.684	-0.021	997	0.901	1.530	-0.046
	3	0.923	1.560	-0.014	1000	0.927	1.445	-0.036
50	1	0.919	1.520	-0.041	997	0.920	1.467	-0.047
	2	0.948	1.455	-0.019	1000	0.949	1.368	-0.030
	3	0.959	1.417	-0.046	999	0.960	1.337	-0.050
75	1	0.943	1.537	-0.174	999	0.944	1.446	-0.169
	2	0.963	1.479	-0.153	996	0.964	1.389	-0.153
	3	0.971	1.463	-0.138	996	0.972	1.373	-0.137

El estudio de recubrimiento cuando los parámetros de habilidad son estimados pone de manifiesto la tendencia a la mejoría en dichas estimaciones conforme aumenta el número de ítems y cuanto mayor es el parámetro de discriminación (C3), lo cual está reflejado en la conformidad de los índices $\rho_{\theta,\hat{\theta}}$ y $RMSE$, y en ambos métodos de estimación. Cuando se estima θ por MVM si el test tiene 10 ítems en C1 $\rho_{\theta,\hat{\theta}} = 0{,}663$ y $RMSE = 2{,}927$, mientras que en C3 $\rho_{\theta,\hat{\theta}} = 0{,}809$ y $RMSE = 2{,}147$, i.e., hay una mejoría notable de ambos índices cuanto mayor es el parámetro de discriminación. Esto mismo ocurre cuando se emplea EAP con el que, si el test tiene 10 ítems $\rho_{\theta,\hat{\theta}} = 0{,}685$ y $RMSE = 2{,}034$ en C1, y $\rho_{\theta,\hat{\theta}} = 0{,}812$ y $RMSE = 1{,}765$ en C3. Cuando el test tiene 75 ítems en C1, si θ se estima por MVM $\rho_{\theta,\hat{\theta}} = 0{,}943$ y $RMSE = 1{,}537$, y en C3 $\rho_{\theta,\hat{\theta}} = 0{,}971$ y $RMSE = 1{,}463$; si se estima con EAP en C1 $\rho_{\theta,\hat{\theta}} = 0{,}944$ y $RMSE = 1{,}446$, y en C3 $\rho_{\theta,\hat{\theta}} = 0{,}972$ y $RMSE = 1{,}373$. Estos resultados también muestran la optimización de las estimaciones del parámetro de habilidad; pero conforme aumenta la longitud del test, las diferencias entre las condiciones C1 y C3 no son tan dispares como en el caso del test de 10 ítems. Por lo tanto, cuanto menor es el número de ítems la magnitud del parámetro de discriminación tiene mayor consecuencia sobre la estimación de los parámetros de habilidad; además, cuanto mayor es el número de ítems y más discriminativos son éstos, mejor es la estimación realizada por ambos métodos. La diferencia entre ellos es muy pequeña, llegando a reducirse de modo notable cuantos más ítems tiene el test –diferencias de 0.001 en los tests de 50 y 75 ítems–.

En cuanto al análisis del tercer índice, ASB, en las estimaciones de θ por MVM cuando el test tiene 10 ítems el error cometido es positivo, es decir,

que el método MVM infraestima los parámetros de habilidad, lo cual discrepa del resto de condiciones experimentales y de EAP, para las que se sobrestima la habilidad ($ASB < 0$). Lo que sí es una tendencia general, analizando el valor absoluto de este índice, es que conforme disminuye el número de ítems se produce menor sesgo en la estimación ya que ASB se aproxima más a 0. Así, por MVM con $n = 10$ en C1 $ASB = 0{,}019$, en C2 $ASB = 0{,}047$ y en C3 $ASB = 0{,}020$, mientras que cuando $n = 75$ en C1 $ASB = -0{,}174$, en C2 $ASB = -0{,}153$ y en C3 $ASB = -0{,}138$; con EAP en las tres condiciones del test de 10 ítems $ASB = -0{,}044$ y con $n = 75$ en C1 $ASB = -0{,}169$, en C2 $ASB = -0{,}153$ y en C3 $ASB = -0{,}137$. En cuanto al parámetro de discriminación, éste no parece tener ningún impacto sobre ASB.

Distribución de l_z con parámetros verdaderos (Tabla 4.17)

Tabla 4.17. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.037	1.002	-0.858**	0.482**	0.099**	0.000	0.039	0.013
	2	0.013	0.992	-1.243**	2.751**	0.081**	0.000	0.044	0.017
	3	-0.014	1.048	-1.281**	2.018**	0.104**	0.000	0.041	0.022
25	1	-0.057	1.033	-0.337**	-0.171	0.037**	0.002	0.039	0.006
	2	-0.022	0.998	-0.402**	-0.106	0.039**	0.001	0.041	0.008
	3	-0.038	1.026	-0.654**	0.506**	0.064**	0.000	0.037	0.011
50	1	-0.040	1.026	-0.311**	-0.071	0.035**	0.006	0.041	0.009
	2	0.010	1.019	-0.318**	0.084	0.035**	0.006	0.055	0.007
	3	-0.063	1.061	-0.498**	0.324*	0.055**	0.000	0.046	0.015
75	1	0.070	0.993	-0.422**	0.678**	0.041**	0.001	0.056	0.010
	2	0.060	0.985	-0.318**	0.147	0.033*	0.014	0.054	0.012
	3	0.015	1.015	-0.381**	0.218	0.025	0.140	0.042	0.010

* p¡0.05

** p¡0.01

Cuando los parámetros empleados en el cálculo de l_z son los verdaderos, la media y la desviación típica del índice de medición apropiada se aproximan a los valores estándar. Las medias se acercan a 0 cuanto mayor es el parámetro de discriminación, de modo que con $n = 10$ en C1 $\hat{\mu}_{l_z} = 0{,}037$ y en C3 $\hat{\mu}_{l_z} = -0{,}014$; cuando $n = 75$ en C1 $\hat{\mu}_{l_z} = 0{,}070$ y en C3 es $\hat{\mu}_{l_z} = 0{,}015$.

Respecto a la desviación típica, se aproxima bastante a 1 en todas las condiciones experimentales, siendo el valor más bajo 0.985, aparecido en C2 del test de 75 ítems, y el valor más alto 1.061 de C3 del test de 50 ítems.

Los índices de sesgo son negativos y estadísticamente significativos al 1 % en todos los casos, lo que pone de manifiesto la asimetría de las distribuciones de

l_z. Al observar los resultados en cada longitud de test, la asimetría o sesgo es mayor cuanto más discriminativos son los ítems, e.g., en el test de 25 ítems en C1 $g_1 = -0{,}337$, en C2 es $g_1 = -0{,}402$ y en C3 es $g_1 = -0{,}654$. Sin embargo, esta tendencia no se verifica en el test de 75 ítems en el que cuanto más discriminativos son los ítems menos sesgada está la distribución: $g_1 = -0{,}422$ en C1, $g_1 = -0{,}318$ en C2 y $g_1 = -0{,}381$ en C3. El índice de sesgo incrementa su valor absoluto a medida que se reduce el número de ítems, así en C3 para 10, 25, 50 y 75 ítems el índice de asimetría es -1.281, -0.654, -0.498 y -0.381.

Ante los valores de curtosis, se observa que hay curvas mesocúrticas definidas por índices próximos y mayores a 0, exceptuando la C1 de los tests con 25 y 50 ítems, $g_2 = -0{,}171$ y $g_2 = -0{,}071$, y en C2 del test con 25 ítems, $g_2 = -0{,}106$; el valor más cercano y mayor a 0 está en C2 del test de 50 ítems, $g_2 = 0{,}084$. Sin embargo, las curvas obtenidas en las tres condiciones del test con 10 ítems son leptocúrticas, $g_2 = 0{,}482$ en C1, $g_2 = 2{,}751$ en C2 y $g_2 = 2{,}018$ en C3 [$p(\alpha) < 0{,}01$], junto con C3 de los tests de 25 y 50 ítems, $g_2 = 0{,}506$ con $p(\alpha) < 0{,}01$ y $g_2 = 0{,}324$ con $p(\alpha) < 0{,}05$, y C1 del test de 75 ítems, $g_2 = 0{,}678$ con $p(\alpha) < 0{,}01$. La distribución de l_z tiende a la altura media conforme los ítems son menos discriminativos, tendencia contraria a la obtenida en el test con 75 ítems en donde las curvas mesocúrticas se corresponden con las condiciones más discriminativas, $g_2 = 0{,}147$ en C2 y $g_2 = 0{,}218$ en C3.

La prueba de normalidad de Lilliefors aporta la existencia de diferencias significativas en la mayoría de las condiciones al 1 % y en C2 si $n = 75$ al 5 %. La distribución de l_z no sigue la ley normal, salvo en el test de 75 ítems en C3 donde $M.D. = 0{,}025$ con $p(\alpha) = 0{,}140$.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.18)

Cuando se calcula l_z con los parámetros estimados por MVM, la media de l_z se aleja de 0 tanto como las halladas en el test de 10 ítems en C1 y C2, en las que $\hat{\mu}_{l_z} = -0{,}545$ y $\hat{\mu}_{l_z} = -0{,}482$. Estos valores se corresponden a su vez con los valores más altos de desviaciones típicas, 1.468 en C1 y 1.429 en C2. Los valores medios más próximos a 0 aparecen en C1 y C2 del test de 75 ítems, $\hat{\mu}_{l_z} = 0{,}022$ y $\hat{\mu}_{l_z} = 0{,}034$, y sus desviaciones típicas no se alejan demasiado de 1, $\hat{\sigma}_{l_z} = 0{,}917$ en C1 y $\hat{\sigma}_{l_z} = 0{,}908$ en C2. A mayor número de ítems y parámetros de discriminación, mayores son las semejanzas con la media y la desviación típica esperadas.

Tabla 4.18. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM										
								α nominal		
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	1	-0.545	1.468	-1.340**	2.624**	0.090**	0.000	0.042	0.025	975
	2	-0.482	1.429	-1.451**	2.906**	0.097**	0.000	0.050	0.029	992
	3	0.240	0.943	-1.263**	1.729**	0.138**	0.000	0.050	0.022	1000
25	1	-0.388	1.322	-0.958**	1.948**	0.060**	0.000	0.045	0.014	996
	2	-0.282	1.271	-1.049**	2.012**	0.065**	0.000	0.037	0.021	997
	3	0.103	0.885	-0.625**	0.781**	0.043**	0.000	0.039	0.012	1000
50	1	-0.289	1.206	-0.615**	0.958**	0.048**	0.000	0.044	0.013	997
	2	0.072	0.906	-0.238**	0.225	0.030*	0.032	0.052	0.009	1000
	3	0.092	0.966	-0.473**	0.623**	0.046**	0.000	0.044	0.010	999
75	1	0.022	0.917	-0.498**	1.249**	0.059**	0.000	0.049	0.022	999
	2	0.034	0.908	-0.226**	0.136	0.030*	0.031	0.051	0.011	996
	3	-0.192	1.147	-0.747**	1.396**	0.045**	0.000	0.046	0.017	996

* p¡0.05

** p¡0.01

Los índices de sesgo vuelven a resultar negativos en todas las condiciones experimentales, siendo los más altos en valor absoluto y, por lo tanto, los que marcan una distribución más asimétrica en comparación con el resto, los tests de 10 ítems para los que en C1 $g_1 = -1{,}340$, en C2 $g_1 = -1{,}451$ y en la última condición $g_1 = -1{,}263$; esto mismo también ocurre en las dos primeras condiciones del test de 25 ítems en las que el índice de sesgo es $g_1 = -0{,}958$ en C1 y $g_1 = -1{,}049$ en C2. Todos los índices de asimetría son estadísticamente significativos con $p(\alpha) < 0{,}01$.

Casi la totalidad de las curvas son leptocúrticas con $p(\alpha) < 0{,}01$, mostrando apuntamientos de relevancia como en los tests de 10 y 25 ítems: si $n = 10$ $g_2 = 2{,}624$ en C1, $g_2 = 2{,}906$ en C2 y $g_2 = 1{,}729$ en C3, y en C1 si $n = 25$ $g_2 = 1{,}948$ y en C2 $g_2 = 2{,}012$; así como con $n = 75$, $g_2 = 1{,}249$ en C1 y $g_2 = 1{,}396$ en C3. Sólo hay dos curvas mesocúrticas, las de C2 de los tests con 50 y 75 ítems, $g_2 = 0{,}225$ y $g_2 = 0{,}136$, que a su vez son las distribuciones de l_z menos asimétricas negativas, $g_1 = -0{,}238$ y $g_1 = -0{,}226$.

En los índices de sesgo y curtosis se produce un efecto corrector de la asimetría y de la forma conforme aumenta el número de ítems y el índice de discriminación, algo a lo que parece no ajustarse el test de 75 ítems.

En cuanto a la prueba de normalidad, se pone de manifiesto la desviación de la ley normal de la distribución de l_z con $p(\alpha) < 0{,}01$ y en C2 de los tests de 50 y 75 ítems con $p(\alpha) < 0{,}05$, $M.D. = 0{,}030$ con $p(\alpha) = 0{,}032$ y $p(\alpha) = 0{,}031$.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.19)

Tabla 4.19. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP									
								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.256	0.918	-0.908**	0.670**	0.089**	0.000	0.052	0.017
	2	0.248	0.898	-1.047**	1.212**	0.091**	0.000	0.045	0.022
	3	0.231	0.924	-1.266**	1.760**	0.149**	0.000	0.051	0.023
25	1	0.244	0.883	-0.436**	-0.046	0.045**	0.000	0.049	0.011
	2	0.235	0.866	-0.493**	-0.037	0.049**	0.000	0.042	0.011
	3	0.226	0.906	-0.719**	0.738**	0.057**	0.000	0.039	0.011
50	1	0.197	0.929	-0.361**	-0.075	0.037**	0.002	0.049	0.012
	2	0.198	0.926	-0.336**	0.162	0.026	0.100	0.046	0.009
	3	0.200	0.957	-0.491**	0.402**	0.051**	0.000	0.050	0.010
75	1	0.161	0.931	-0.603**	1.015**	0.051**	0.000	0.047	0.016
	2	0.155	0.925	-0.307**	0.090	0.036**	0.003	0.052	0.008
	3	0.146	0.924	-0.310**	-0.047	0.029**	0.050	0.043	0.012

* p¡0.05

** p¡0.01

Las medias y desviaciones típicas de l_z cuando los parámetros es estiman por EAP son cercanas a las estandarizadas. La media tiende a 0 cuanto mayor es el número de ítems y mayor es el parámetro de discriminación; en C1 con $n = 10$ $\hat{\mu}_{l_z} = 0{,}256$, $n = 25$ $\hat{\mu}_{l_z} = 0{,}244$, $n = 50$ $\hat{\mu}_{l_z} = 0{,}197$ y con $n = 75$ $\hat{\mu}_{l_z} = 0{,}161$, disminuyendo dicho distanciamiento hasta $\hat{\mu}_{l_z} = 0{,}146$ en C3 del último test. El valor de desviación típica más próximo a 1 es 0.957 en C3 del test de 50 ítems y el más alejado es 0.866 en C2 del test con 25 ítems. A diferencia de la media, la desviación típica no está afectada por la longitud del test. Una ligera mejoría del estadístico de dispersión es conseguida por el incremento de la discriminación de los ítems.

En lo referente al índice de sesgo se confirma lo descrito anteriormente, es decir, la presencia de índices g_1 negativos que señalan la asimetría de las distribuciones de l_z con $p(\alpha) < 0{,}01$. Los índices más altos en valor absoluto aparecen en C2 y C3 del test de 10 ítems, $g_1 = -1{,}047$ y $g_1 = -1{,}266$, y los más pequeños en C2 y C3 del test de 75 ítems, $g_1 = -0{,}307$ y $g_1 = -0{,}310$. La longitud del test repara la asimetría y el poder discriminativo de los ítems la estropea.

La distribución de l_z es leptocúrtica al 1 % de nivel de significación en los tests de 10 ítems –en C1 $g_2 = 0{,}670$, C2 $g_2 = 1{,}212$ y C3 $g_2 = 1{,}760$–, en C3 de los tests con 25 y 50 ítems, $g_2 = 0{,}738$ y $g_2 = 0{,}402$, y en C1 del test de 75 ítems, $g_2 = 1{,}015$. Las restantes curvas son mesocúrticas, en especial las de C1 y C2 del test de 25 ítems, $g_2 = -0{,}046$ y $g_2 = -0{,}037$; con $n = 50$ en C1 $g_2 = -0{,}075$, la curva más chata, y en C2 $g_2 = 0{,}162$ la más alta. Los tests

más largos dibujan curvas con alturas medias, así como los que contienen ítems menos discriminativos aunque esto no se verifique del todo si $n = 75$.

Los resultados aportados por la prueba de Lilliefors llevan a que la distribución de l_z no sigue la ley normal en la mayoría de los casos con $p(\alpha) < 0{,}01$, con excepción del test de 75 ítems en C3, $M.D. = 0{,}029$ con $p(\alpha) = 0{,}050$ que sería significativa al nivel más conservador, y en el test de 50 ítems en C2, $M.D. = 0{,}026$ con $p(\alpha) = 0{,}100$, donde la distribución es normal.

En el nivel α nominal igual a 0.05, el error tipo I es más consistente que al nivel 0.01, siendo en éste con parámetros verdaderos donde se producen mayores fluctuaciones en la estimación de dicho error.

Distribución de habilidad sesgada positiva

Estudio de recubrimiento (Tabla 4.20)

Tabla 4.20. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$

		MVM				EAP		
n	C	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	1	0.683	2.748	0.073	980	0.717	2.007	0.000
	2	0.758	2.466	0.098	993	0.790	1.846	0.000
	3	0.775	2.329	0.096	999	0.806	1.808	0.000
25	1	0.835	1.861	0.020	989	0.859	1.670	-0.001
	2	0.898	1.677	0.025	1000	0.906	1.527	0.005
	3	0.915	1.593	0.033	1000	0.919	1.484	0.012
50	1	0.916	1.530	0.002	997	0.920	1.481	0.005
	2	0.942	1.450	0.003	995	0.945	1.389	0.004
	3	0.960	1.412	0.043	1000	0.957	1.350	0.045
75	1	0.939	1.612	-0.188	988	0.945	1.473	-0.170
	2	0.958	1.578	-0.161	985	0.962	1.442	-0.153
	3	0.971	1.532	-0.139	991	0.970	1.412	-0.128

Con el método MVM para estimar el parámetro de habilidad, el coeficiente de correlación de Pearson es bajo en las tres condiciones del test de 10 ítems ($\rho_{\theta,\hat{\theta}} = 0{,}683$ en C1, $\rho_{\theta,\hat{\theta}} = 0{,}758$ en C2 y $\rho_{\theta,\hat{\theta}} = 0{,}775$ en C3) indicando una notable discrepancia entre θ y $\hat{\theta}$. Esto también lo testimonia $RMSE$, en donde con $n = 10$ se obtienen los valores más altos, $RMSE = 2{,}748$ en C1, $RMSE = 2{,}466$ en C2 y $RMSE = 2{,}329$ en C3. Conforme aumenta el número de ítems y el parámetro de discriminación, tanto el coeficiente de correlación

como el índice $RMSE$ se aproximan a los valores esperados, el valor 1 en ambos casos. Esto se constata observando los resultados obtenidos, por ejemplo, en C1 del test de 25 ítems en donde $\rho_{\theta,\hat{\theta}} = 0{,}835$ y $RMSE = 1{,}861$, que llegan hasta $\rho_{\theta,\hat{\theta}} = 0{,}958$ y $RMSE = 1{,}578$ en C2 del test de 75 ítems, y $\rho_{\theta,\hat{\theta}} = 0{,}971$ y $RMSE = 1{,}532$ en C3 para esta misma longitud de test.

El índice ASB manifiesta que MVM infraestima el parámetro de habilidad, ya que en todas las condiciones dicho índice es positivo, excepto en el test de 75 ítems en el que se sobrestima dicho parámetro, $ASB = -0{,}188$ en C1, $ASB = -0{,}161$ en C2 y $ASB = -0{,}139$ en C3. A diferencia de $\rho_{\theta,\hat{\theta}}$ y $RMSE$, ASB se revela indiferente a la longitud del test y a la magnitud del parámetro de discriminación. Los valores positivos más alejados de 0 se hallan en el test de 10 ítems, $ASB = 0{,}098$ en C2 y $ASB = 0{,}096$ en C3; los más próximos a 0 son $ASB = 0{,}002$ y $ASB = 0{,}003$ en C1 y C2 con 50 ítems.

Al ser estimada la habilidad con el proceso EAP, a mayor número de ítems del test y mayor parámetro de discriminación mejor es la estimación realizada, tal y como muestran $\rho_{\theta,\hat{\theta}}$ y $RMSE$. En el test de 10 ítems, desde C1 a C3, ambas referencias mejoran desde valores $\rho_{\theta,\hat{\theta}} = 0{,}717$ y $RMSE = 2{,}007$ en C1, y $\rho_{\theta,\hat{\theta}} = 0{,}790$ y $RMSE = 1{,}846$ en C2 hasta $\rho_{\theta,\hat{\theta}} = 0{,}806$ y $RMSE = 1{,}808$ en la última condición del parámetro de discriminación. Dicho esto, el valor más próximo a 1 que cabría esperar tanto de $\rho_{\theta,\hat{\theta}}$ como de $RMSE$ debería ser en C3 del test con 75 ítems, algo que solamente corrobora el primero de estos índices ($\rho_{\theta,\hat{\theta}} = 0{,}970$), ya que los valores de $RMSE$ más cercanos a 1 se han obtenido en C2 y C3 del test de 50 ítems, $RMSE = 1{,}389$ y $RMSE = 1{,}350$.

El índice ASB muestra una buena estimación del parámetro de habilidad en el test con 10 ítems, en C1 y C2 de los tests de 25, $ASB = -0{,}001$ y $ASB = 0{,}005$, y 50 ítems, $ASB = 0{,}005$ y $ASB = 0{,}004$. En el caso del test de 75 ítems, los resultados difieren del resto de tests estudiados porque, primero, ASB indica que se sobrestima el parámetro de habilidad y, segundo, los valores absolutos son muy elevados en comparación con los demás resultados obtenidos, $ASB = -0{,}170$ en C1, $ASB = -0{,}153$ en C2 y $ASB = -0{,}128$ en C3.

Comparando los dos procesos de estimación, EAP consigue mejores aproximaciones al parámetro que MVM. Las diferencias entre ambas se reducen conforme aumenta el número de ítems del test y coinciden en la sobrestimación de la habilidad detectada por ASB en el test con 75 ítems.

Distribución de l_z con parámetros verdaderos (Tabla 4.21)

Tabla 4.21. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.012	0.995	-0.913**	1.086**	0.067**	0.000	0.036	0.020
	2	-0.080	1.018	-0.879**	0.623**	0.075**	0.000	0.051	0.016
	3	-0.091	1.082	-1.314**	2.705**	0.099**	0.000	0.045	0.022
25	1	0.015	1.020	-0.500**	-0.004	0.051**	0.000	0.048	0.011
	2	-0.012	0.963	-0.481**	0.265	0.035**	0.005	0.048	0.013
	3	-0.055	1.014	-0.582**	0.292	0.057**	0.000	0.043	0.010
50	1	-0.020	1.046	-0.412**	0.258	0.037**	0.003	0.040	0.011
	2	0.004	0.960	-0.370**	0.024	0.050**	0.000	0.053	0.012
	3	0.035	1.018	-0.778**	1.600**	0.057**	0.000	0.052	0.013
75	1	-0.050	1.005	-0.215**	-0.121	0.022	0.298	0.053	0.007
	2	0.012	0.992	-0.367**	0.434**	0.035**	0.006	0.049	0.012
	3	0.014	1.001	-0.431**	0.192	0.038**	0.002	0.043	0.015

* p¡0.05

** p¡0.01

Con los parámetros verdaderos, el cálculo de los estadísticos descriptivos de media y desviación típica de l_z son cercanos a los de la distribución normal tipificada. Los valores de la media más alejados de 0 se obtienen en C2 y C3 del test con 10 ítems, $\hat{\mu}_{l_z} = -0{,}080$ y $\hat{\mu}_{l_z} = -0{,}090$, mientras que el valor más próximo es el conseguido en C2 del test de 50 ítems, $\hat{\mu}_{l_z} = 0{,}004$. En cuanto a la desviación típica, los valores oscilan entre $\hat{\sigma}_{l_z} = 0{,}960$ de C2 del test de 50 ítems y $\hat{\sigma}_{l_z} = 1{,}082$ de C3 del test de 10 ítems. La desviación típica más cercana a la estándar es lograda en C3 del test con 75 ítems, $\hat{\sigma}_{l_z} = 1{,}001$, y se aproximan a ésta las C1 de los tests de 10 y 75 ítems con valores $\hat{\sigma}_{l_z} = 0{,}995$ y $\hat{\sigma}_{l_z} = 1{,}005$. En tanto que sobre la media no hay efecto ni de la longitud del test ni del parámetro de discriminación, este último sí favorece a los valores de la desviación típica.

El sesgo de las distribuciones es negativo y todas presentan diferencias significativas con la simetría, $p(\alpha) < 0{,}01$. Cuantos menos ítems tiene el test más asimétrica negativa es la curva: en los tests con 10 ítems $g_1 = -0{,}913$, $g_1 = -0{,}879$ y $g_1 = -1{,}314$, en los que tienen 75 ítems $g_1 = -0{,}215$, $g_1 = -0{,}367$ y $g_1 = -0{,}431$ en orden de condiciones de C1 a C3. Sobre el estadístico de simetría también interfieren los parámetros de discriminación, acusando el sesgo cuando son más elevados.

Las distribuciones son leptocúrticas con $p(\alpha) < 0{,}01$ en el test de 10 ítems, $g_2 = 1{,}086$ en C1, $g_2 = 0{,}623$ en C2 y en C3 $g_2 = 2{,}705$, al igual que las de C3 del test de 50 ítems, $g_2 = 1{,}600$, y C2 del test de 75 ítems, $g_2 = 0{,}434$. Los índices que dibujan curtosis media-baja aparecen en C1 de los tests con 25,

$g_2 = -0{,}004$, y 75 ítems, $g_2 = -0{,}121$, y la curtosis media más alta está en C3 del test de 25 ítems, $g_2 = 0{,}292$. Los tests más largos son los que muestran las distribuciones de altura media más satisfactoria.

La única condición experimental en la que la prueba de Lilliefors no rechaza la hipótesis nula de normalidad es en C1 del test de 75 ítems, $M.D. = 0{,}022$ con $p(\alpha) = 0{,}298$, las demás la rechazan con $p(\alpha) < 0{,}01$.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.22)

Tabla 4.22. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

								α nominal		
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	1	-0.494	1.474	-1.504**	3.324**	0.096**	0.000	0.049	0.025	980
	2	-0.454	1.338	-1.393**	3.308**	0.080**	0.000	0.043	0.021	993
	3	-0.404	1.439	-1.777**	4.633**	0.119**	0.000	0.049	0.027	999
25	1	-0.345	1.299	-0.857**	1.232**	0.069**	0.000	0.048	0.015	989
	2	0.069	0.798	-0.433**	0.410**	0.031*	0.024	0.055	0.016	1000
	3	0.093	0.866	-0.530**	0.337*	0.049**	0.000	0.041	0.008	1000
50	1	0.051	0.923	-0.359**	0.563**	0.030*	0.037	0.045	0.012	997
	2	0.067	0.899	-0.362**	0.050	0.044**	0.000	0.049	0.016	995
	3	0.091	0.939	-0.671**	1.230**	0.052**	0.000	0.052	0.014	1000
75	1	0.020	0.872	-0.055	0.252	0.034**	0.009	0.053	0.011	988
	2	0.032	0.898	-0.299**	0.619**	0.027	0.084	0.048	0.013	985
	3	0.060	0.934	-0.398**	0.347*	0.034**	0.008	0.052	0.011	991

* pi0.05

** pi0.01

La media de la distribución de l_z es infravalorada si $n = 10$ y en la primera de las condiciones si $n = 25$, tests en los que también se hallan las medias más alejadas de 0, $\hat{\mu}_{l_z} = -0{,}494$, $\hat{\mu}_{l_z} = -0{,}454$ y $\hat{\mu}_{l_z} = -0{,}404$ en C1, C2 y C3 con 10 ítems, y $\hat{\mu}_{l_z} = -0{,}345$ en C1 con 25 ítems. Conjuntamente, en estos cuatro tests es en donde se obtienen las desviaciones típicas más altas y que son, en el mismo orden de descripción anterior, $\hat{\sigma}_{l_z} = 1{,}474$, $\hat{\sigma}_{l_z} = 1{,}338$, $\hat{\sigma}_{l_z} = 1{,}439$ y $\hat{\sigma}_{l_z} = 1{,}299$. En el resto de condiciones experimentales los valores medios están próximos a 0 y oscilan entre 0.093, de C3 del test de 25 ítems, y 0.020, de C1 del test de 75 ítems. A mayor número de ítems y menor parámetro de discriminación, más proximidad tienen las medias a la estándar, exceptuando los casos citados de los tests con 10 y 25 ítems. Las desviaciones típicas de C2 y C3 del test de 25 ítems y los tests de 50 y 75 ítems son inferiores a 1 en un rango de $\hat{\sigma}_{l_z} = 0{,}798$, en C2 del test de 25 ítems, a $\hat{\sigma}_{l_z} = 0{,}939$, en C3 del test de 50 ítems. A mayor parámetro de discriminación se logran mejores valores de desviaciones típicas.

La asimetría de las distribuciones es negativa y presentan diferencias significativas con la simetría [$p(\alpha) < 0{,}01$], de modo más acusado cuanto menor es el número de ítems y a mayor parámetro de discriminación. Por ejemplo, en C3 con $n = 10$ $g_1 = -1{,}777$ y con $n = 75$ $g_1 = -0{,}398$; para esta longitud de test pero en C1 $g_1 = -0{,}055$, la única curva simétrica que aparece.

En cuanto a la forma, las curvas serían leptocúrticas en la generalidad, destacando las tres condiciones del test de 10 ítems: $g_2 = 3{,}324$ en C1, $g_2 = 3{,}308$ en C2 y $g_2 = 4{,}633$ en C3 con $p(\alpha) < 0{,}01$. En C3 de los tests de 25 y 75 ítems, $g_2 = 0{,}337$ y $g_2 = 0{,}347$, la significación ocurre al nivel del 5 %. Las distribuciones mesocúrticas de l_z son las del test de 50 ítems en C2, $g_2 = 0{,}050$, y C1 del test con 75 ítems, $g_2 = 0{,}252$. Sobre el estadístico de forma hay ciertas mejorías por el tamaño del test, pero no parece notable el parámetro de discriminación.

La prueba estadística no rechaza la hipótesis nula de normalidad de la distribución de l_z en el test de 75 ítems en C2, donde $M.D. = 0{,}027$ con $p(\alpha) = 0{,}084$. La ley normal no se verifica con $p(\alpha) < 0{,}05$ en C2 si $n = 25$, $M.D. = 0{,}031$ con $p(\alpha) = 0{,}024$, y C1 si $n = 50$, $M.D. = 0{,}030$ con $p(\alpha) = 0{,}037$; el resto de distribuciones no la verifican al nivel de significación del 1 %.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.23)

Conforme aumenta el número de ítems del test y el parámetro de discriminación, las medias y desviaciones típicas de l_z calculado con parámetros estimados por EAP tienden a los de una distribución normal tipificada. En los tests de 10 y 25 ítems las medias no son inferiores a 0.200, en decrecimiento desde C1 del test con 10 ítems, donde $\hat{\mu}_{l_z} = 0{,}240$, a C3 del test con 25, $\hat{\mu}_{l_z} = 0{,}207$. En los tests de 50 y 75 ítems los valores están alrededor de 0.190 y el más próximo a 0 se encuentra en C3 del test con 75 ítems, $\hat{\mu}_{l_z} = 0{,}041$.

En el estadístico de dispersión también afecta, aunque en menor grado que sobre la media, el parámetro de discriminación. Las desviaciones típicas oscilan entre $\hat{\sigma}_{l_z} = 0{,}835$, de C2 del test con 25 ítems, y $\hat{\sigma}_{l_z} = 0{,}946$, de C1 y C3 del test con 50 ítems.

Con este procedimiento de estimación la media de la distribución es sobrevalorada y la desviación típica infravalorada.

El índice de sesgo es negativo y significativo con $p(\alpha) < 0{,}01$, manifestando que la curva de l_z es siempre asimétrica. Esto se corrige moderadamente conforme aumenta el número de ítems y a menor parámetro de discriminación; así, en el test de 10 ítems $g_1 = -1{,}151$ en C1, $g_1 = -1{,}010$ en C2 y $g_1 = -1{,}146$ en C3; y en el test de 75 ítems, $g_1 = -0{,}236$ en C1, $g_1 = -0{,}379$ en C2 y $g_1 = -0{,}331$ en C3.

Por el índice de forma se describen varias distribuciones leptocúrticas tanto con $p(\alpha) < 0{,}05$ como con $p(\alpha) < 0{,}01$, sin efecto aparente del parámetro de discriminación y sin transcendencia de la longitud del test, con valores tan altos como los del test con 10 ítems en C1, $g_2 = 1{,}724$, C2, $g_2 = 0{,}966$ y C3, $g_2 = 1{,}349$, y en C3 del test de 50 ítems, $g_2 = 1{,}233$. Otras curvas son mesocúrticas: las de C1 si $n = 25$, $g_2 = 0{,}218$, C2 si $n = 50$, $g_2 = 0{,}100$, C1 y C3 del test de 75 ítems, $g_2 = 0{,}121$ y $g_2 = -0{,}061$.

Según la prueba de Lilliefors, las distribuciones de l_z en este apartado no son normales al nivel de significación del 5 %, pero si se examinaran al 1 %, sólo C1 y C2 del test de 75 ítems seguirían una distribución normal, $M.D. = 0{,}030$ con $p(\alpha) = 0{,}037$ y $M.D. = 0{,}031$ con $p(\alpha) = 0{,}024$.

Tabla 4.23. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.240	0.909	-1.151**	1.724**	0.103**	0.000	0.044	0.018
	2	0.243	0.879	-1.010**	0.966**	0.103**	0.000	0.050	0.019
	3	0.236	0.887	-1.146**	1.349**	0.129**	0.000	0.049	0.017
25	1	0.237	0.891	-0.551**	0.218	0.052**	0.000	0.046	0.012
	2	0.223	0.835	-0.566**	0.312*	0.048**	0.000	0.042	0.015
	3	0.207	0.880	-0.625**	0.307*	0.059**	0.000	0.041	0.008
50	1	0.197	0.946	-0.491**	0.558**	0.044**	0.000	0.045	0.010
	2	0.191	0.898	-0.441**	0.100	0.044**	0.000	0.052	0.014
	3	0.193	0.946	-0.746**	1.233**	0.054**	0.000	0.050	0.015
75	1	0.182	0.905	-0.236**	0.121	0.030*	0.037	0.052	0.010
	2	0.171	0.911	-0.379**	0.612**	0.031*	0.024	0.053	0.014
	3	0.041	0.933	-0.331**	-0.061	0.038**	0.002	0.045	0.010

* p¡0.05

** p¡0.01

Las tasas de FP más elevadas están en el nivel α nominal 0.01 cuando $n = 10$ tanto con parámetros verdaderos como con parámetros estimados. En el resto de longitudes de tests dichas tasas son consistentes, y muestran cierta tendencia a la infraestimación si α nominal es 0.05 y a la sobrestimación si α nominal es 0.01.

Distribución de habilidad sesgada negativa

Estudio de recubrimiento (Tabla 4.24)

Tabla 4.24. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$								
		MVM				EAP		
n	C	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	1	0.628	3.056	0.043	966	0.660	2.119	-0.000
	2	0.748	2.544	0.004	999	0.749	1.938	-0.000
	3	0.794	2.256	0.004	1000	0.795	1.834	-0.000
25	1	0.832	1.928	0.038	993	0.831	1.745	0.001
	2	0.889	1.701	0.035	994	0.888	1.583	0.002
	3	0.922	1.583	0.033	1000	0.912	1.508	0.003
50	1	0.910	1.548	0.011	993	0.914	1.501	-0.003
	2	0.947	1.446	-0.017	996	0.946	1.389	-0.032
	3	0.955	1.432	0.003	998	0.953	1.368	-0.009
75	1	0.937	1.458	-0.081	996	0.938	1.430	-0.086
	2	0.964	1.357	-0.068	995	0.963	1.325	-0.073
	3	0.973	1.332	-0.045	996	0.972	1.288	-0.055

Los coeficientes de correlación de Pearson en ambos procedimientos de estimación tienen valores muy similares cuando la distribución de habilidad es normal y sesgada negativa. La menor diferencia entre los dos procedimientos es de 0.001 en C2 y C3 del test de 10 ítems, en C1 y C2 del test de 25, en C2 del test de 50 y en todas las condiciones del test de 75 ítems, favoreciendo indistintamente a un procedimiento de estimación o al otro. La mayor diferencia es de magnitud 0.032 a favor de EAP aparecida en C1 del test de 10 ítems ($\rho_{\theta,\hat{\theta}} = 0{,}628$ con MVM y $\rho_{\theta,\hat{\theta}} = 0{,}660$ con EAP). Los dos procedimientos tienden a aproximar el estimador al valor verdadero de habilidad cuanto mayor es el número de ítems del test y el parámetro de discriminación.

El índice $RMSE$ establece mayores diferencias entre las dos estimaciones tal y como se puede observar comparando el test de 10 ítems, que con MVM tiene valores $RMSE = 3{,}056$ en C1, $RMSE = 2{,}544$ en C2 y $RMSE = 2{,}256$ en C3, mientras que con EAP $RMSE = 2{,}119$, $RMSE = 1{,}938$ y $RMSE = 1{,}834$ en C1, C2 y C3. Estas diferencias entre MVM y EAP van siendo menores cuando aumenta la longitud del test, a la vez que mejora la estimación del parámetro; por ejemplo, en el test de 75 ítems en C1, C2 y C3 con MVM $RMSE = 1{,}458$, $RMSE = 1{,}357$ y $RMSE = 1{,}332$, y con EAP $RMSE = 1{,}430$, $RMSE = 1{,}325$ y $RMSE = 1{,}288$.

Los valores de ASB en el procedimiento MVM indican que la habilidad es infraestimada en los tests de 10, 25 y 50 ítems, excluyendo en este último a C2 ($ASB = -0{,}017$). Los índices ASB más divergentes infraestimando el parámetro verdadero son los del test de 25 ítems, en donde $ASB = 0{,}038$ en

C1, $ASB = 0{,}035$ en C2 y $ASB = 0{,}033$ en C3, así como en la primera de las condiciones del test de 10 ítems, $ASB = 0{,}043$. Con 75 ítems se sobrestima el parámetro de habilidad con MVM: $ASB = -0{,}081$, $ASB = -0{,}068$ y $ASB = -0{,}045$ de C1 a C3; esto mismo ocurre con EAP, $ASB = -0{,}086$, $ASB = -0{,}073$ y $ASB = -0{,}055$ de C1 a C3, que además y a diferencia de $\rho_{\theta,\hat{\theta}}$ y $RMSE$, da estimadores de habilidad más alejados del parámetro que MVM. Disconforme con esta afirmación, según ASB el proceso EAP estima bien el parámetro en los tests de 10, $ASB = -0{,}000$, y 25 ítems, $ASB = 0{,}001$ en C1, $ASB = 0{,}002$ en C2 y $ASB = 0{,}003$ en C3; con 50 ítems se sobrestiman pero se acercan bastante al parámetro a excepción de la segunda de las condiciones ($ASB = -0{,}032$). La tendencia de las estimaciones es a mejorar conforme aumenta el parámetro de discriminación.

Distribución de l_z con parámetros verdaderos (Tabla 4.25)

Independientemente de la longitud del test y del parámetro de discriminación las medias de la distribución de l_z con parámetros verdaderos varían en un rango de $\hat{\mu}_{l_z} = -0{,}082$, obtenido en C1 del test de 75 ítems, y $\hat{\mu}_{l_z} = 0{,}030$ del test de 25 ítems en esa misma condición. Mientras que las medias infravaloradas se alejan de 0 en cantidades mínimas como $\hat{\mu}_{l_z} = -0{,}017$ de C3 del test de 25 ítems, la media sobrevalorada más próxima a 0 es $\hat{\mu}_{l_z} = 0{,}001$ del test de 75 ítems también en la última condición.

Al igual que el estadístico de tendencia central, el de dispersión tampoco se deja influir ni por el test ni por la discriminación. Se pueden observar valores próximos a 1 tanto con 10 ítems, $\hat{\sigma}_{l_z} = 1{,}009$ en C1, como con 50 ítems, $\hat{\sigma}_{l_z} = 0{,}995$ en C2. El intervalo de valores de la desviación típica con parámetros verdaderos es de $\hat{\sigma}_{l_z} = 0{,}950$, en C3 del test de 75 ítems, a $\hat{\sigma}_{l_z} = 1{,}049$ del test de 10 ítems en esta misma condición.

Las distribuciones de l_z son asimétricas negativas con $p(\alpha) < 0{,}01$ sobre todo cuanto mayor es el parámetro de discriminación; e.g., en el test de 25 ítems en C1 $g_1 = -0{,}438$, en C2 $g_1 = -0{,}465$ y en C3 $g_1 = -0{,}669$. Este estadístico también es influenciable por la longitud del test, intentando restablecer la simetría de la curva a medida que se crece en ítems; comparando con el test de 25 ítems, con $n = 75$ en C1 $g_1 = -0{,}378$, en C2 $g_1 = -0{,}235$ y en C3 $g_1 = -0{,}404$. Las distribuciones más sesgadas son las de C2 y C3 si $n = 10$, $g_1 = -1{,}008$ y $g_1 = -1{,}107$.

Tabla 4.25. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.002	1.009	-0.807**	0.615**	0.067**	0.000	0.044	0.018
	2	0.018	1.018	-1.008**	0.985**	0.095**	0.000	0.051	0.019
	3	-0.053	1.049	-1.107**	1.198**	0.098**	0.000	0.048	0.021
25	1	0.030	0.982	-0.438**	0.117	0.036**	0.003	0.051	0.011
	2	-0.055	0.986	-0.465**	0.183	0.035**	0.005	0.046	0.012
	3	-0.017	0.961	-0.669**	0.603**	0.050**	0.000	0.044	0.019
50	1	-0.024	1.018	-0.320**	-0.007	0.032*	0.020	0.049	0.012
	2	0.023	0.995	-0.501**	0.409**	0.037**	0.002	0.044	0.011
	3	0.006	0.973	-0.513**	0.233	0.045**	0.000	0.046	0.015
75	1	-0.082	1.048	-0.378**	0.135	0.039**	0.001	0.054	0.014
	2	0.003	0.951	-0.235**	-0.093	0.038**	0.002	0.047	0.010
	3	0.001	0.950	-0.404**	0.109	0.040**	0.001	0.042	0.011

* p¡0.05

** p¡0.01

Algunas de las curvas que son mesocúrticas se dibujan por debajo de la normal: en C1 del test de 50 ítems, $g_2 = -0{,}007$, y en C2 del test de 75 ítems, $g_2 = -0{,}093$. Son curvas leptocúrticas estadísticamente significativas con $p(\alpha) < 0{,}01$ las del test de 10 ítems en sus tres condiciones ($g_2 = 0{,}615$, $g_2 = 0{,}985$ y $g_2 = 1{,}198$), la de C3 si $n = 25$, $g_2 = 0{,}603$, y la de C2 si $n = 50$, $g_2 = 0{,}409$. Dicho esto, se puede deducir que sobre el estadístico de forma g_2 el parámetro de discriminación no ejerce un efecto evidente.

La prueba de normalidad de Lilliefors muestra diferencias significativas en todas las condiciones estudiadas con parámetros verdaderos al 1 % y en C1 del test de 25 ítems al 5 %, $M.D. = 0{,}032$ con $p(\alpha) = 0{,}020$. Por lo tanto, la distribución de l_z no sigue la ley normal.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.26)

Cuando se recurre a los parámetros estimados con MVM para estudiar la distribución de l_z, la media es infravalorada en la mayoría de las condiciones y sólo es sobrevalorada en C2 y C3 del test de 10 ítems, $\hat{\mu}_{l_z} = 0{,}176$ y $\hat{\mu}_{l_z} = 0{,}207$, y en C3 del test de 25 ítems, $\hat{\mu}_{l_z} = 0{,}116$. El incremento tanto en el número de ítems como en el parámetro de discriminación provocan una mejoría de los valores medios de l_z; comparando los tests de 50 y 75 ítems en orden de C1 a C3, en el primero de ellos $\hat{\mu}_{l_z} = -0{,}291$, $\hat{\mu}_{l_z} = -0{,}211$ y $\hat{\mu}_{l_z} = -0{,}197$, en el de 75 ítems son $\hat{\mu}_{l_z} = -0{,}261$, $\hat{\mu}_{l_z} = -0{,}182$ y $\hat{\mu}_{l_z} = -0{,}166$.

En cuanto a las desviaciones típicas están afectadas de la misma manera que las medias por la discriminación de los ítems, así en C1 de los tests de 10

y 50 ítems son $\hat{\sigma}_{l_z} = 1{,}398$ y $\hat{\sigma}_{l_z} = 1{,}241$, en C2 de estos tests son $\hat{\sigma}_{l_z} = 0{,}928$ y $\hat{\sigma}_{l_z} = 1{,}190$, y en la última de sus condiciones son $\hat{\sigma}_{l_z} = 0{,}931$ y $\hat{\sigma}_{l_z} = 1{,}147$.

Tabla 4.26. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

								α nominal		
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	1	-0.519	1.398	-1.266**	2.548**	0.080**	0.000	0.041	0.023	966
	2	0.176	0.928	-1.087**	1.820**	0.109**	0.000	0.046	0.018	999
	3	0.207	0.931	-1.123**	1.085**	0.135**	0.000	0.046	0.022	1000
25	1	-0.399	1.291	-0.842**	1.115**	0.064**	0.000	0.050	0.019	993
	2	-0.320	1.211	-0.863**	1.359**	0.063**	0.000	0.042	0.014	994
	3	0.116	0.847	-0.504**	0.042	0.056**	0.000	0.049	0.012	1000
50	1	-0.291	1.241	-0.633**	0.343*	0.056**	0.000	0.055	0.018	993
	2	-0.211	1.190	-1.018**	3.133**	0.056**	0.000	0.036	0.015	996
	3	-0.197	1.147	-0.799**	1.070**	0.052**	0.000	0.044	0.021	998
75	1	-0.261	1.219	-0.582**	0.440**	0.045**	0.000	0.047	0.012	996
	2	-0.182	1.082	-0.519**	0.481**	0.046**	0.000	0.045	0.010	995
	3	-0.166	1.099	-0.674**	0.724**	0.069**	0.000	0.045	0.014	996

* p¡0.05

** p¡0.01

El sesgo de la distribución es negativo y presenta diferencias significativas con la simetría al nivel del 1 %, definiendo curvas asimétricas y más cuanto menor es el número de ítems del test; con 10 ítems $g_1 = -1{,}266$ en C1, $g_1 = -1{,}087$ en C2 y $g_1 = -1{,}123$ en C3, y se reduce la asimetría en el test de 75 ítems a niveles de sesgo que son: en C1 $g_1 = -0{,}582$, en C2 $g_1 = -0{,}519$ y en C3 es $g_1 = -0{,}674$. Una excepción a esta tendencia del índice de asimetría está en C2 del test de 25 ítems, donde $g_1 = -1{,}018$, una distribución tan asimétrica como las de los tests con 10 ítems.

La curtosis se corrige, al igual que el índice de sesgo, a mayor longitud del test sin que por ello se evite la significación estadística. Las curvas más leptocúrticas con $p(\alpha) < 0{,}01$ son las de C1 del test con 10 ítems, $g_2 = 2{,}548$, y C2 del test con 50 ítems, $g_2 = 3{,}133$. En este mismo test en C1 $g_2 = 0{,}343$, significativa al 5 %. Sólo una curva es mesocúrtica: la del test de 25 ítems en la última condición, $g_2 = 0{,}042$.

La distribución del l_z no es normal a raíz de los resultados de la prueba no paramétrica de Lilliefors con $p(\alpha) < 0{,}01$.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.27)

Tabla 4.27. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP									
								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.246	0.953	-1.002**	1.153**	0.081**	0.000	0.048	0.022
	2	0.243	0.925	-1.144**	1.925**	0.115**	0.000	0.049	0.018
	3	0.241	0.914	-1.153**	1.133**	0.131**	0.000	0.052	0.020
25	1	0.251	0.908	-0.530**	0.260	0.058**	0.000	0.047	0.015
	2	0.244	0.890	-0.540**	0.213	0.049**	0.000	0.046	0.014
	3	0.240	0.868	-0.595**	-0.032	0.059**	0.000	0.046	0.013
50	1	0.198	0.903	-0.382**	0.103	0.034**	0.007	0.052	0.011
	2	0.196	0.938	-0.600**	0.787**	0.049**	0.000	0.039	0.013
	3	0.203	0.903	-0.497**	-0.018	0.051**	0.000	0.053	0.012
75	1	0.135	0.947	-0.427**	-0.022	0.043**	0.000	0.047	0.014
	2	0.141	0.912	-0.369**	0.129	0.050**	0.000	0.047	0.016
	3	0.132	0.932	-0.456**	0.050	0.041**	0.001	0.042	0.011

* pj0.05

** pj0.01

Tras el procedimiento EAP, las medias de l_z más cercanas al valor estándar de la distribución normal se obtienen en el test de 75 ítems, $\hat{\mu}_{l_z} = 0{,}135$, $\hat{\mu}_{l_z} = 0{,}141$ y $\hat{\mu}_{l_z} = 0{,}132$ en C1, C2 y C3, respectivamente. La aproximación a 0 también ocurre dentro de cada longitud de test cuanto mayor es el parámetro de discriminación; por ejemplo, en el test con 25 ítems de C1 a C3, $\hat{\mu}_{l_z} = 0{,}251$, $\hat{\mu}_{l_z} = 0{,}244$ y $\hat{\mu}_{l_z} = 0{,}240$.

Las desviaciones típicas de l_z con parámetros estimados con EAP son infravaloradas, sin que haya mejoría por el número de ítems, pero sí por el parámetro de discriminación en sentido contrario a como lo hace en la media, es decir, a mayor discriminación las desviaciones típicas se alejan de la deseada. El rango en el que oscilan los valores de dispersión es de $\hat{\sigma}_{l_z} = 0{,}868$, en C3 del test de 25 ítems, a $\hat{\sigma}_{l_z} = 0{,}953$ en C1 del test de 10 ítems.

Las distribuciones más asimétricas negativas son las de los tests con 10 ítems, $g_1 = -1{,}002$ en C1, $g_1 = -1{,}144$ en C2 y $g_1 = -1{,}153$ en C3, y se reduce la asimetría conforme aumenta el número de ítems, llegando a $g_1 = -0{,}427$, $g_1 = -0{,}369$ y $g_1 = -0{,}456$ en C1, C2 y C3 si $n = 75$, aún siendo estadísticamente significativas. Cuanto menor es el parámetro de discriminación también son menos asimétricas las curvas, pero siempre significativas con $p(\alpha) < 0{,}01$.

La altura de las distribuciones es media, hallándose valores tales como en C3 de los tests con 25 y 50 ítems, $g_2 = -0{,}032$ y $g_2 = -0{,}018$, o en C1 con 75 ítems, $g_2 = -0{,}022$, que dibujan curvas algo por debajo de la normal. Las distribuciones que por su apuntamiento son leptocúrticas con $p(\alpha) < 0{,}01$ se delinean en los tests de 10 ítems, $g_2 = 1{,}153$ en C1, $g_2 = 1{,}925$ en C2 y

$g_2 = 1{,}133$ en C3, y en C2 del test de 50 ítems, $g_2 = 0{,}787$. El aumento tanto de ítems como del parámetro de discriminación favorece la similitud de la forma de estas curvas con la de altura media.

El estadístico l_z para detectar patrones atípicos no se ajusta a la ley normal según la prueba de Lilliefors con $p(\alpha) < 0{,}01$.

Las tasas de error tipo I son más consistentes al nivel α nominal 0.05 que al de 0.01, siendo a este nivel donde se sobrestima la tasa de error sobre todo en los tests de 10 ítems y con parámetros estimados por MVM.

4.3.3. Bloque 3: Modelo de 3-p

Distribución de habilidad no sesgada

Estudio de recubrimiento (Tabla 4.28)

Los valores del coeficiente de correlación de Pearson cuando el parámetro de habilidad es estimado por MVM y por EAP van aumentando conforme aumenta el número de ítems en el test, a la vez que cuanto mayor es el índice de discriminación. Las diferencias en este coeficiente entre ambos procedimientos son pequeñas y se reducen cuanto mayor es el número de ítems y el parámetro de discriminación. Así, cuando $n = 10$ en C1 por MVM $\rho_{\theta,\hat{\theta}} = 0{,}541$ y por EAP $\rho_{\theta,\hat{\theta}} = 0{,}572$, en C3 por MVM $\rho_{\theta,\hat{\theta}} = 0{,}725$ y por EAP $\rho_{\theta,\hat{\theta}} = 0{,}732$; cuando $n = 75$ en C1 por MVM $\rho_{\theta,\hat{\theta}} = 0{,}905$ y por EAP $\rho_{\theta,\hat{\theta}} = 0{,}911$, en C3 por MVM $\rho_{\theta,\hat{\theta}} = 0{,}954$ y por EAP $\rho_{\theta,\hat{\theta}} = 0{,}956$.

Esta tendencia a la similitud entre las dos estimaciones también la muestra $RMSE$ aunque en menor grado: en C1 del test de 10 ítems por MVM es 3.741, y por EAP, 2.221, y en el test de 75 ítems por MVM, 1.574 y por EAP, 1.497; en C3 del test de 10 ítems por MVM, $RMSE = 2{,}793$, y por EAP, $RMSE = 1{,}941$, y en el test de 75 ítems por MVM es 1.424, y por EAP, 1.345.

En general, con respecto a $\rho_{\theta,\hat{\theta}}$ y a $RMSE$, los valores más próximos a 1 se obtienen con el método EAP.

Tabla 4.28. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$								
		MVM				EAP		
n	C	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	1	0.541	3.741	-0.091	955	0.572	2.221	-0.044
	2	0.606	3.355	-0.202	972	0.627	2.136	-0.044
	3	0.725	2.793	-0.176	990	0.732	1.941	-0.044
25	1	0.747	2.316	-0.035	986	0.770	1.861	-0.044
	2	0.818	1.983	-0.059	985	0.833	1.717	-0.047
	3	0.862	1.819	-0.055	992	0.871	1.617	-0.047
50	1	0.872	1.710	-0.056	994	0.879	1.595	-0.049
	2	0.917	1.548	-0.054	991	0.924	1.456	-0.045
	3	0.937	1.499	-0.049	998	0.939	1.406	-0.042
75	1	0.905	1.574	-0.050	998	0.911	1.497	-0.054
	2	0.939	1.463	-0.057	997	0.942	1.397	-0.055
	3	0.954	1.424	-0.059	1000	0.956	1.345	-0.058

El índice ASB en los dos procedimientos detecta sobrestimación del parámetro de habilidad, especialmente en MVM: cuando el test tiene 10 ítems en C1 $ASB = -0{,}091$, en C2 $ASB = -0{,}202$ y en C3 $ASB = -0{,}176$. Estos resultados resaltan por su magnitud en comparación con EAP, que en ese mismo test en las tres condiciones de discriminación $ASB = -0{,}044$. Los valores de ASB indican que la sobrestimación del parámetro de habilidad en función de la longitud del test y del parámetro de discriminación es más homogénea con EAP, en donde el rango de ASB está en el intervalo $(-0{,}058, -0{,}042)$, mientras que con MVM es más variable, $(-0{,}202, -0{,}035)$.

Distribución de l_z con parámetros verdaderos (Tabla 4.29)

La media y la desviación típica empleando los parámetros verdaderos resultan aceptables como para afirmar que la distribución de l_z es normal. El valor medio más alejado de 0 en valor absoluto lo tiene C2 del test de 75 ítems, $\hat{\mu}_{l_z} = 0{,}071$, seguido de C1 del test de 10 ítems, $\hat{\mu}_{l_z} = 0{,}053$. Dentro de cada longitud de test, la condición de mayor discriminación (C3) es la que obtiene valores más cercanos al estándar, $\hat{\mu}_{l_z} = -0{,}008$, $\hat{\mu}_{l_z} = 0{,}002$, $\hat{\mu}_{l_z} = 0{,}013$ y $\hat{\mu}_{l_z} = -0{,}023$ con 10, 25, 50 y 75 ítems. Esta secuencia también evidencia que a mayor número de ítems más se aleja la media de l_z de 0 en el modelo de 3-p con distribución de habilidad no sesgada, aunque el efecto del tamaño del test no está claramente definido.

La desviación típica del estadístico en estudio oscila entre el valor $\hat{\sigma}_{l_z} = 1{,}053$ de C2 del test de 10 ítems y $\hat{\sigma}_{l_z} = 0{,}975$ de C1 del test de 25 ítems. Sobre este estadístico descriptivo no parece que haya efecto ni de la longitud

del test ni del parámetro de discriminación.

Tabla 4.29. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.053	0.978	-0.546**	0.030	0.054**	0.000	0.042	0.011
	2	-0.037	1.053	-0.952**	1.473**	0.061**	0.000	0.040	0.020
	3	-0.008	0.992	-0.777**	0.429**	0.080**	0.000	0.037	0.016
25	1	0.012	0.975	-0.368**	0.188	0.040**	0.001	0.047	0.011
	2	-0.007	1.026	-0.554**	0.211	0.050**	0.000	0.045	0.016
	3	0.002	0.988	-0.407**	0.081	0.043**	0.000	0.036	0.006
50	1	-0.017	1.003	-0.268**	0.027	0.033*	0.014	0.045	0.014
	2	0.027	0.983	-0.212**	0.113	0.019	0.456	0.055	0.009
	3	0.013	1.012	-0.453**	0.077	0.051**	0.000	0.038	0.014
75	1	0.027	1.032	-0.259**	-0.049	0.024	0.159	0.044	0.010
	2	0.071	0.987	-0.307**	0.137	0.043**	0.000	0.049	0.012
	3	-0.023	1.023	-0.347**	0.140	0.034**	0.009	0.055	0.010

* p¡0.05

** p¡0.01

Las curvas son asimétricas negativas y más a menor longitud del test, en cualquier caso significativas con $p(\alpha) < 0{,}01$. Con 10 ítems en C1 $g_1 = -0{,}546$, en C2 $g_1 = -0{,}952$ –la más asimétrica– y en C3 $g_1 = -0{,}777$. Pero a pesar de aumentar el número de ítems no se reducen las diferencias con el índice de sesgo estándar ($g_1 = 0$) y la prueba de asimetría sigue resultando significativa, así, en el test de 75 ítems C1 obtiene un índice de sesgo de $g_1 = -0{,}259$, $g_1 = -0{,}307$ en C2 y $g_1 = -0{,}347$ en C3. La curva menos asimétrica es la de C2 con $n = 50$, $g_1 = -0{,}212$.

El estadístico de forma g_2 señala que la distribución es mesocúrtica con independencia de la condición experimental, exceptuando las condiciones C1 y C2 del test de 10 ítems que las dibuja leptocúrticas con $p(\alpha) < 0{,}01$, $g_2 = 1{,}473$ y $g_2 = 0{,}429$. Los valores más cercanos al tipificado son: $g_2 = 0{,}030$ en C1 del test de 10 ítems, en C1 con 50 ítems, $g_2 = 0{,}027$ y con 75 ítems, $g_2 = -0{,}049$.

La prueba no paramétrica de Lilliefors es estadísticamente significativa en casi la totalidad de las condiciones al 1 % y en C1 del test de 25 ítems al 5 %, $M.D. = 0{,}033$ con $p(\alpha) = 0{,}014$. Para C2 del test de 50 ítems y C1 del test de 75 ítems, $M.D. = 0{,}019$ con $p(\alpha) = 0{,}456$ y $M.D. = 0{,}024$ con $p(\alpha) = 0{,}159$, la distribución de l_z sigue la ley normal.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.30)

Tabla 4.30. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

								α nominal		
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	1	-0.513	1.320	-0.837**	0.709**	0.074**	0.000	0.040	0.015	955
	2	-0.562	1.523	-1.312**	2.427**	0.100**	0.000	0.045	0.023	972
	3	-0.450	1.526	-1.785**	4.815**	0.118**	0.000	0.044	0.025	990
25	1	-0.470	1.276	-0.694**	0.834**	0.052**	0.000	0.046	0.013	986
	2	-0.406	1.249	-0.896**	2.076**	0.050**	0.000	0.041	0.018	985
	3	-0.374	1.293	-0.970**	1.786**	0.057**	0.000	0.037	0.021	992
50	1	-0.282	1.216	-0.607**	0.496**	0.047**	0.000	0.045	0.014	994
	2	-0.286	1.221	-0.714**	1.244**	0.045**	0.000	0.048	0.017	991
	3	-0.241	1.239	-0.863**	1.406**	0.062**	0.000	0.041	0.018	998
75	1	-0.263	1.208	-0.408**	0.309*	0.026	0.106	0.044	0.012	998
	2	-0.239	1.182	-0.565**	0.720**	0.040**	0.001	0.057	0.016	997
	3	0.074	0.885	-0.242**	0.376*	0.037**	0.003	0.061	0.015	1000

* p¡0.05

** p¡0.01

La media del estadístico l_z es infravalorada cuando se han utilizado los parámetros estimados por MVM. Las medias más alejadas de la estándar aparecen en los tests con 10 ítems, diferencias que se acortan conforme aumenta la longitud del test. Dentro de cada tamaño de test, cuanto menor es el parámetro de discriminación mayor es la diferencia con la media normalizada. Con 25 ítems $\hat{\mu}_{l_z} = -0{,}470$, $\hat{\mu}_{l_z} = -0{,}406$ y $\hat{\mu}_{l_z} = -0{,}374$ en C1, C2 y C3, y con 75 ítems son $\hat{\mu}_{l_z} = -0{,}263$, $\hat{\mu}_{l_z} = -0{,}239$ y $\hat{\mu}_{l_z} = 0{,}074$ en C1, C2 y C3 –esta última condición es en la única que aparece sobrestimada la media del estadístico–.

En cuanto a las desviaciones típicas, son mayores a 1 exceptuando la última condición del test de 75 ítems, $\hat{\sigma}_{l_z} = 0{,}885$, y más se aproxima a dicho valor a mayor número de ítems; por ejemplo, en C2 y C3 con 10 ítems $\hat{\sigma}_{l_z} = 1{,}523$ y $\hat{\sigma}_{l_z} = 1{,}526$, mientras que con 75 ítems $\hat{\sigma}_{l_z} = 1{,}182$ y $\hat{\sigma}_{l_z} = 0{,}885$ para las mismas condiciones. El parámetro de discriminación no parece repercutir en la mejoría de las desviaciones típicas de l_z.

De nuevo, al igual que con los parámetros verdaderos, las distribuciones son asimétricas negativas y más asimétricas a menor número de ítems y a mayor parámetro de discriminación con $p(\alpha) < 0{,}01$; en el test de 10 ítems $g_1 = -0{,}837$ en C1, $g_1 = -1{,}312$ en C2 y $g_1 = -1{,}785$ en C3. Sin embargo, el test de 75 ítems no se ajusta a lo descrito para el índice de sesgo, ya que en C1 $g_1 = -0{,}408$, en C2 $g_1 = -0{,}565$ y en C3 $g_1 = -0{,}242$, es decir, que con $n = 75$ la curva con parámetro de discriminación más alto es la menos asimétrica.

Resultan dispares los índices de curtosis obtenidos con los estimadores

máximo-verosímiles, aunque todos ellos son significativos y las curvas son leptocúrticas. Hay valores de g_2 indicativos, sin lugar a dudas, de una distribución leptocúrtica como los de C2 y C3 si $n = 10$ en los que $g_2 = 2{,}427$ y $g_2 = 4{,}815$, y C2 si $n = 25$, $g_2 = 2{,}076$; otros como en C1 y C3 del test de 75 ítems, $g_2 = 0{,}309$ y $g_2 = 0{,}376$ con $p(\alpha) < 0{,}05$, que de no recurrir a la prueba estadística, se podría decir que dibujan curvas mesocúrticas. En general, este índice es sensible a la longitud del test –disminuyendo su apuntamiento– pero no al parámetro de discriminación.

La única condición en la que la prueba de Lilliefors no es estadísticamente significativa aparece en C1 del test de 75 ítems, $M.D. = 0{,}026$ con $p(\alpha) = 0{,}106$, y por lo tanto, se acepta la hipótesis nula de distribución normal de l_z.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.31)

Tabla 4.31. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.221	0.935	-0.660**	0.113	0.081**	0.000	0.044	0.011
	2	0.202	0.914	-0.885**	0.865**	0.084**	0.000	0.042	0.015
	3	0.222	0.897	-0.891**	0.825**	0.096**	0.000	0.046	0.019
25	1	0.268	0.904	-0.486**	0.311*	0.049**	0.000	0.041	0.013
	2	0.258	0.932	-0.548**	0.192	0.058**	0.000	0.051	0.015
	3	0.251	0.892	-0.532**	0.250	0.051**	0.000	0.049	0.013
50	1	0.209	0.852	-0.303**	0.480**	0.028	0.062	0.049	0.018
	2	0.200	0.867	-0.266**	0.218	0.033*	0.011	0.054	0.017
	3	0.194	0.904	-0.412**	0.048	0.045**	0.000	0.045	0.008
75	1	0.186	0.915	-0.265**	0.215	0.030**	0.029	0.046	0.008
	2	0.173	0.877	-0.253**	0.080	0.038**	0.001	0.054	0.012
	3	0.168	0.895	-0.362**	0.379*	0.041**	0.000	0.059	0.015

* p¡0.05

** p¡0.01

Los estadísticos de tendencia central de l_z calculados con parámetros estimados con EAP tienen valores muy similares entre todas las condiciones experimentales, oscilando entre el valor más alto obtenido en el test de 25 ítems en C1, $\hat{\mu}_{l_z} = 0{,}268$, y el más cercano a 0 en el test de 75 ítems en C3, $\hat{\mu}_{l_z} = 0{,}168$. El número de ítems y el incremento del parámetro de discriminación mejoran la estimación de las medias, apareciendo las magnitudes más próximas a 0 en el test con 75 ítems.

Mientras que empleando los parámetros bayesianos la media se sobrevalora, la desviación típica es infravalorada. En sentido contrario a lo que ocurre con la media, en la desviación típica parecen ser malos influyentes la longitud del

test y el parámetro de discriminación, así, en tres de los cuatro tamaños de test los valores más bajos se obtienen en la condición de mayor discriminación, $\hat{\sigma}_{l_z} = 0{,}897$ con 10 ítems, $\hat{\sigma}_{l_z} = 0{,}892$ con 25 y $\hat{\sigma}_{l_z} = 0{,}895$ con 75 ítems. En el test de 50 ítems en C1 $\hat{\sigma}_{l_z} = 0{,}852$, el valor más bajo de desviación típica hallado.

Las distribuciones son asimétricas negativas y significativas con $p(\alpha) < 0{,}01$, sesgo que se corrige conforme aumenta el número de ítems. En el test de 10 ítems se logran valores tan bajos como los de C2 y C3, $g_1 = -0{,}885$ y $g_1 = -0{,}891$, y los más altos en el test de 50 ítems para C2, $g_1 = -0{,}266$, y en el test de 75 ítems C1 tiene un índice de sesgo $g_1 = -0{,}265$, C2 $g_1 = -0{,}253$ y C3 $g_1 = -0{,}362$. Dentro de cada longitud de test, el sesgo es más acusado en C3, $g_1 = -0{,}891$ con 10 ítems, $g_1 = -0{,}532$ con 25, $g_1 = -0{,}412$ con 50 y $g_1 = -0{,}362$ con 75 ítems.

El índice de curtosis es próximo a 0 en todos los casos, siendo significativos al 1 % los del test de 10 ítems en C2 y C3, $g_2 = 0{,}865$ y $g_2 = 0{,}825$, y C1 si $n = 50$, $g_2 = 0{,}480$, y al 5 % los de C1 si $n = 25$, $g_2 = 0{,}311$, y C3 si $n = 75$, $g_2 = 0{,}379$, manifestando que las curvas son leptocúrticas. Los valores más bajos de g_2 aparecen en C3 del test de 50 ítems, $g_2 = 0{,}048$, y en C2 del test de 75 ítems, $g_2 = 0{,}080$. El parámetro de discriminación no influye sobre la forma de la distribución de l_z calculado con parámetros estimados por EAP, y de modo poco eficiente lo hace el número de ítems del test.

La prueba de normalidad es estadísticamente significativa en todas las condiciones, aunque lo contrario ocurre en C1 si $n = 50$, $M.D. = 0{,}028$ con $p(\alpha) = 0{,}062$. Con $p(\alpha) < 0{,}05$, existen diferencias con la curva normal en C2 del test de 50 ítems, $M.D. = 0{,}033$ con $p(\alpha) = 0{,}011$, y en C1 si $n = 75$, $M.D. = 0{,}030$ con $p(\alpha) = 0{,}029$.

Las tasas de error tipo I son infraestimadas, consistentes y próximas al nivel α nominal de 0.05, y sobrestimadas y menos consistentes al nivel α nominal de 0.01 con mayor notoriedad tras el empleo de parámetros estimados por MVM.

Distribución de habilidad sesgada positiva

Estudio de recubrimiento (Tabla 4.32)

Tabla 4.32. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$								
		MVM				EAP		
n	C	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	1	0.550	3.688	-0.066	951	0.598	2.228	-0.000
	2	0.660	3.152	-0.029	982	0.680	2.082	0.000
	3	0.676	2.969	-0.024	981	0.705	2.032	0.000
25	1	0.755	2.263	0.017	986	0.802	1.818	-0.001
	2	0.805	2.058	0.002	987	0.850	1.694	-0.001
	3	0.852	1.876	-0.016	994	0.880	1.609	-0.016
50	1	0.864	1.737	0.001	996	0.885	1.595	-0.001
	2	0.899	1.605	-0.024	996	0.913	1.503	-0.020
	3	0.927	1.557	-0.029	1000	0.934	1.434	-0.023
75	1	0.900	1.593	-0.016	999	0.913	1.503	-0.012
	2	0.929	1.507	-0.020	1000	0.935	1.428	-0.013
	3	0.950	1.426	-0.021	999	0.953	1.359	-0.013

La estimación del parámetro de habilidad por MVM en el modelo de 3-p y distribución de habilidad sesgada positiva no consigue muy buenas aproximaciones al parámetro verdadero cuanto menor es el número de ítems y el parámetro de discriminación, tal y como lo confirman los resultados del coeficiente de correlación de Pearson y el índice de error $RMSE$. Correlaciones tan bajas como las obtenidas en el test de 10 ítems en las tres condiciones, $\rho_{\theta,\hat{\theta}} = 0{,}550$ en C1, $\rho_{\theta,\hat{\theta}} = 0{,}660$ en C2 y $\rho_{\theta,\hat{\theta}} = 0{,}676$ en C3, se incrementan hasta llegar al test de 75 ítems, alcanzando magnitudes de 0.900 en C1, 0.929 en C2, y 0.950 en C3. En el caso del índice $RMSE$, los valores extremos del test de 10 y 25 ítems, $RMSE = 3{,}688$ en C1 si $n = 10$ ó $RMSE = 2{,}058$ de C2 si $n = 25$, se reducen a $RMSE = 1{,}593$, $RMSE = 1{,}507$ y $RMSE = 1{,}426$ en C1, C2 y C3 del test de 75 ítems.

Estos dos indicadores del grado de estimación pero con EAP declaran los mismos resultados que con MVM. La diferencia entre los dos métodos es mínima para $\rho_{\theta,\hat{\theta}}$ y algo más considerables en el índice $RMSE$, aunque siempre es favorecida EAP. Recogiendo los mismos datos que los descritos en el párrafo precedente, en el test de 10 ítems $\rho_{\theta,\hat{\theta}} = 0{,}598$ en C1, $\rho_{\theta,\hat{\theta}} = 0{,}680$ en C2 y $\rho_{\theta,\hat{\theta}} = 0{,}705$ en C3; en el test de 75 ítems $\rho_{\theta,\hat{\theta}} = 0{,}913$, $\rho_{\theta,\hat{\theta}} = 0{,}935$ y $\rho_{\theta,\hat{\theta}} = 0{,}953$ en C1, C2 y C3. Para el índice $RMSE$ en el test de 10 ítems $RMSE = 2{,}228$, $RMSE = 2{,}082$ y $RMSE = 2{,}032$, y en el de 75 ítems $RMSE = 1{,}503$, $RMSE = 1{,}428$ y $RMSE = 1{,}359$ en C1, C2 y C3.

El índice de sesgo medio con signo ASB señala que hay una buena estimación del parámetro θ con valores próximos a 0, infraestimándolo y sobres-

timándolo, en su mayoría, tanto con MVM como con EAP. El valor más alejado del crítico con MVM es −0,066 en C1 del test de 10 ítems y los más cercanos a 0 son 0.002 en C2 del test de 25 ítems y 0.001 en C1 del test de 50. Una congruencia mejor entre parámetro y estimador es la que este índice muestra tras el procedimiento EAP ya que, como por ejemplo ocurre en los tests 10 ítems, $ASB = -0{,}000$ en C1 y $ASB = 0{,}000$ en C2 y C3, ó como en C1 y C2 del test de 25 ítems donde $ASB = -0{,}001$. El valor más alejado de 0 con EAP es −0,023 del test de 50 ítems con mayor discriminación.

Distribución de l_z con parámetros verdaderos (Tabla 4.33)

Tabla 4.33. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.033	0.946	-0.551**	0.373*	0.042**	0.000	0.043	0.013
	2	-0.049	0.982	-0.664**	0.255	0.060**	0.000	0.042	0.015
	3	-0.077	1.024	-0.827**	0.679**	0.066**	0.000	0.048	0.015
25	1	-0.011	0.964	-0.181*	0.010	0.020	0.412	0.051	0.009
	2	0.003	0.986	-0.449**	0.091	0.039**	0.001	0.050	0.012
	3	-0.048	1.012	-0.527**	0.455**	0.042**	0.000	0.042	0.011
50	1	-0.004	1.065	-0.277**	0.120	0.032**	0.020	0.046	0.011
	2	-0.031	0.944	-0.448**	0.261	0.042**	0.000	0.058	0.014
	3	-0.009	0.987	-0.346**	-0.014	0.043**	0.000	0.045	0.009
75	1	0.035	1.003	-0.114	-0.218	0.029*	0.040	0.041	0.008
	2	-0.007	0.999	-0.420**	0.562**	0.027	0.082	0.048	0.011
	3	0.020	0.988	-0.440**	0.495**	0.033*	0.014	0.052	0.011

* p¡0.05

** p¡0.01

Con los parámetros verdaderos la media y la desviación típica son muy próximos a los estándar. Dejándose afectar por la longitud del test pero no por la magnitud del parámetro de discriminación, los valores medios varían entre -0.077 de C3 del test con 10 ítems y 0.035 de la primera de las condiciones del test con 75 ítems. Los valores más cercanos a 0 son 0.003 de C2 con 25 ítems y −0,004 de C1 con 50 ítems.

La desviación típica consigue una mejoría cuando se incrementa el parámetro de discriminación en los tests de 10 ítems, $\hat{\sigma}_{l_z} = 0{,}946$, $\hat{\sigma}_{l_z} = 0{,}982$ y $\hat{\sigma}_{l_z} = 1{,}024$ de C1 a C3, y también si $n = 25$, $\hat{\sigma}_{l_z} = 0{,}964$, $\hat{\sigma}_{l_z} = 0{,}986$ y $\hat{\sigma}_{l_z} = 1{,}012$ en C1, C2 y C3. En el test de 50 ítems este parámetro no interviene y en el de 75 ítems el incremento en discriminación rebaja la magnitud de la dispersión, $\hat{\sigma}_{l_z} = 1{,}003$, $\hat{\sigma}_{l_z} = 0{,}999$ y $\hat{\sigma}_{l_z} = 0{,}988$ en C1, C2 y C3, respectivamente. No se manifiesta relación entre el tamaño del test y la desviación típica.

Donde sí aparece efecto del número de ítems es en el estadístico de sesgo, el cual indica la asimetría negativa, con $p(\alpha) < 0{,}01$, de las distribuciones de l_z de modo más acentuado en aquellas condiciones de mayor discriminación (C3). Así, a mayor discriminación en el test de 10 ítems, $g_1 = -0{,}827$, está la curva más sesgada; en el test de 25 ítems $g_1 = -0{,}527$, de 50 $g_1 = -0{,}346$ y en el de 75 ítems $g_1 = -0{,}440$. La curva menos asimétrica es la de C1 del test de 25 ítems, $g_1 = -0{,}181$ con $p(\alpha) < 0{,}05$, y en esta misma condición con 75 ítems la curva es simétrica, $g_1 = -0{,}114$.

Algunas curvas son mesocúrticas con valores de g_2 muy cercanos a 0, como en C1 si $n = 25$ que $g_2 = 0{,}010$ ó en C3 si $n = 50$ que $g_2 = -0{,}014$; pero también se ha obtenido significación estadística en C1 y C3 del test 10 ítems, $g_2 = 0{,}373$ con $p(\alpha) < 0{,}05$ y $g_2 = 0{,}679$ con $p(\alpha) < 0{,}01$, en C2 con 25 ítems, $g_2 = 0{,}455$, así como en C2 y C3 con 75 ítems, $g_2 = 0{,}562$ y $g_2 = 0{,}495$, todas éstas al 1 % de nivel de significación. En contraste, C1 si $n = 75$ muestra una curva más baja que la normal, $g_2 = -0{,}218$, pero mesocúrtica. En el estadístico de forma influye en su imperfección el parámetro de discriminación, pero no el número de ítems del test.

La distribución de l_z, según la prueba de Lilliefors, solamente es normal en dos casos: con 25 ítems en C1, $M.D. = 0{,}020$ con $p(\alpha) = 0{,}412$, y el otro en C2 del test con 75 ítems, $M.D. = 0{,}027$ con $p(\alpha) = 0{,}082$. Si se realizara el contraste al nivel menos conservador de los dos establecidos, tampoco serían significativas las curvas de C1 si $n = 50$, $M.D. = 0{,}032$ con $p(\alpha) = 0{,}020$, ni las de C1 y C3 si $n = 75$, $M.D. = 0{,}029$ con $p(\alpha) = 0{,}040$ y $M.D. = 0{,}033$ con $p(\alpha) = 0{,}014$.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.34)

Con parámetros estimados con MVM, la media del estadístico l_z es infravalorada, ya que en la totalidad de las condiciones experimentales tiene valores inferiores a 0. En los tests de 10 y 25 ítems, a menor longitud de test y menor discriminación, más se aleja de la media tipificada: con 10 ítems de C1 a C3, $\hat{\mu}_{l_z} = -0{,}481$, $\hat{\mu}_{l_z} = -0{,}472$ y $\hat{\mu}_{l_z} = -0{,}429$; con 25 ítems, $\hat{\mu}_{l_z} = -0{,}434$, $\hat{\mu}_{l_z} = -0{,}393$ y $\hat{\mu}_{l_z} = -0{,}352$. Esta tendencia de los valores medios contrasta con los obtenidos en los tests de 50 y 75 ítems, los cuales son indiferentes al parámetro de discriminación.

Las desviaciones típicas sólo se dejan intervenir por la discriminación de los ítems en los tests cortos (10 y 25 ítems), encontrándose los valores más próximos a 1 en aquellos casos en los que el parámetro de discriminación es menor; por citar uno de los dos, en el test de 25 ítems $\hat{\sigma}_{l_z} = 1{,}217$ en C1 y $\hat{\sigma}_{l_z} = 1{,}229$ en C2 y C3. Sin embargo, con 50 y 75 ítems, las desviaciones

típicas tienden al valor estándar conforme más discriminativos son los ítems, por ejemplo, si $n = 75$ $\hat{\sigma}_{l_z} = 1{,}223$ en C1, $\hat{\sigma}_{l_z} = 0{,}886$ en C2 y $\hat{\sigma}_{l_z} = 1{,}107$ en C3.

Tabla 4.34. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

								α nominal		
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	1	-0.481	1.271	-1.072**	2.399**	0.069**	0.000	0.035	0.016	951
	2	-0.472	1.346	-1.234**	2.605**	0.089**	0.000	0.046	0.023	982
	3	-0.429	1.348	-1.469**	3.933**	0.087**	0.000	0.043	0.019	981
25	1	-0.434	1.217	-0.664**	1.202**	0.050**	0.000	0.049	0.018	986
	2	-0.393	1.229	-1.090**	2.560**	0.076**	0.000	0.049	0.017	987
	3	-0.352	1.229	-0.860**	1.162**	0.060**	0.000	0.032	0.019	994
50	1	-0.155	1.191	-0.437**	0.592**	0.028	0.062	0.040	0.011	996
	2	0.079	0.843	-0.366**	0.453**	0.050**	0.000	0.056	0.014	996
	3	0.090	0.894	-0.329**	0.204	0.046**	0.000	0.050	0.015	1000
75	1	-0.267	1.223	-0.562**	1.276**	0.039**	0.001	0.046	0.010	999
	2	0.059	0.886	-0.309**	0.491**	0.026	0.121	0.047	0.013	1000
	3	-0.177	1.107	-0.810**	2.732**	0.037**	0.003	0.041	0.013	999

* p¡0.05

** p¡0.01

La asimetría se dibuja negativa y se corrige conforme aumenta el número de ítems, pero todas presentan diferencias significativas con $p(\alpha) < 0{,}01$. En el test de 10 ítems el índice de asimetría alcanza magnitudes muy pequeñas, $g_1 = -1{,}072$, $g_1 = -1{,}234$ y $g_1 = -1{,}469$ en C1, C2 y C3, al igual que en el de 25 ítems en la segunda condición, $g_1 = -1{,}090$; en el test de 50 ítems $g_1 = -0{,}437$ en C1, en C2 $g_1 = -0{,}366$ y $g_1 = -0{,}329$ en C3, a los que se asimila C2 en el de 75 ítems, $g_1 = -0{,}309$.

Con $p(\alpha) < 0{,}01$, los índices de curtosis diseñan curvas tan leptocúrticas como las de los tests de 10 ítems en C1 $g_2 = 2{,}399$, en C2, $g_2 = 2{,}605$ y $g_2 = 3{,}933$ en C3, de 25 ítems en C2, $g_2 = 2{,}560$, y C3 del test de 75 ítems, $g_2 = 2{,}732$. La única curva con altura media es la de C3 cuando $n = 50$, con índice de curtosis 0.204.

El parámetro de discriminación no afecta al estadístico de forma y sí al de simetría acusando el sesgo de las curvas.

La prueba de normalidad de Lilliefors concluye que l_z, en el modelo de 3-p con distribución de habilidad sesgada positiva, se distribuye normalmente en los tests de 50 ítems con la menor magnitud discriminativa, $M.D. = 0{,}028$ con $p(\alpha) = 0{,}062$, y de 75 ítems en la segunda de sus condiciones experimentales, $M.D. = 0{,}026$ con $p(\alpha) = 0{,}121$.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.35)

Tabla 4.35. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.055	0.01
10	1	0.200	0.877	-0.488**	-0.085	0.056**	0.000	0.032	0.013
	2	0.226	0.894	-0.837**	0.518**	0.096**	0.000	0.041	0.018
	3	0.225	0.936	-0.956**	0.762**	0.124**	0.000	0.048	0.019
25	1	0.250	0.894	-0.379**	0.306	0.046**	0.000	0.043	0.014
	2	0.220	0.905	-0.618**	0.621**	0.047**	0.000	0.044	0.011
	3	0.223	0.894	-0.596**	0.673**	0.060**	0.000	0.042	0.009
50	1	0.184	0.962	-0.357**	0.242	0.033*	0.013	0.047	0.013
	2	0.175	0.864	-0.502**	0.266	0.060**	0.000	0.055	0.012
	3	0.182	0.912	-0.443**	0.097	0.053**	0.000	0.048	0.013
75	1	0.166	0.928	-0.183*	-0.170	0.025	0.150	0.045	0.007
	2	0.155	0.906	-0.418**	0.416**	0.037**	0.003	0.041	0.012
	3	0.154	0.917	-0.467**	0.296	0.045**	0.000	0.036	0.012

* pj0.05

** pj0.01

Las medias de l_z calculadas con parámetros estimados con EAP oscilan entre $\hat{\mu}_{l_z} = 0{,}250$ de C1 del test con 25 ítems y $\hat{\mu}_{l_z} = 0{,}154$ de C3 del test con 75 ítems. Sin que se encuentre efecto del parámetro de discriminación sobre la mejoría de los valores medios, el tamaño del test sí que provoca la aproximación de los mismos a 0. Los valores más altos se concentran en los tests de 10 y 25 ítems, alrededor de 0.225, y los más bajos en los de 50 y 75 ítems, con un promedio de 0.170. Este estadístico es sobrevalorado a diferencia de la desviación típica que es infravalorada, mejorando su valor conforme aumenta la longitud del test: e.g., con 10 ítems en las tres condiciones son 0.877, 0.894 y 0.936; en el test de 75 ítems se logran dispersiones muy cercanas a 1, $\hat{\sigma}_{l_z} = 0{,}928$ en C1, $\hat{\sigma}_{l_z} = 0{,}906$ en C2 y $\hat{\sigma}_{l_z} = 0{,}917$ en C3.

Al igual que con los parámetros máximo-verosímiles antes analizados, las distribuciones de l_z vuelven a ser asimétricas negativas, pero a diferencia de aquellos, los índices de sesgo son muy similares entre todas las condiciones del parámetro de discriminación, con una moderada mejoría cuando aumenta la longitud del test. El índice g_1 tiene un rango de -0.956 obtenido en C3 del test de 10 ítems y -0.183 de C1 del test con 75 ítems, ésta última estadísticamente significativa al 5 %, el resto lo son al 1 %.

Las curvas parecerían mesocúrticas ya que el índice g_2 no llega a superar el valor de 1, es más, en C1 de los tests de 10 y 75 ítems adopta valores más bajos que el crítico, $g_2 = -0{,}085$ y $g_2 = -0{,}170$. Sin embargo, el índice de

curtosis muestra diferencias significativas respecto de la curva de altura media con $p(\alpha) < 0{,}01$ en C2 y C3 de los tests con 10, $g_2 = 0{,}518$ y $g_2 = 0{,}762$, y 25 ítems, $g_2 = 0{,}621$ y $g_2 = 0{,}673$, así como en C2 con 75 ítems, $g_2 = 0{,}416$. Con ello, se puede decir que la longitud del test apenas sí favorece la forma de la distribución de l_z y el parámetro de discriminación eleva, con salvedades, el empinamiento de las curvas.

Sólo en C1 del test de 75 ítems no es estadísticamente significativa la prueba de Lilliefors, $M.D. = 0{,}025$ con $p(\alpha) = 0{,}150$ y, por lo tanto, l_z con estas condiciones se distribuye según la ley normal. La primera condición del parámetro de discriminación en el test de 50 ítems es significativa al 5 %, $M.D. = 0{,}033$ con $p(\alpha) = 0{,}013$.

Las tasas de FP sufren más fluctuaciones al emplear parámetros máximo-verosímiles. Al nivel α nominal de 0.05 los FP son infraestimados y al 0.01 sobrestimados, siendo más estables en el primero de estos dos niveles nominales.

Distribución de habilidad sesgada negativa

Estudio de recubrimiento (Tabla 4.36)

Los resultados aportados por el estudio de recubrimiento sobre el procedimiento MVM señalan que la estimación del parámetro de habilidad no ha sido muy adecuada en el test de 10 ítems como así lo ponen de manifiesto los tres índices empleados, que en C1 son $\rho_{\theta,\hat{\theta}} = 0{,}472$, $RMSE = 4{,}261$ y $ASB = -0{,}128$, en C2 $\rho_{\theta,\hat{\theta}} = 0{,}618$, $RMSE = 3{,}333$ y $ASB = -0{,}114$, y en C3 $\rho_{\theta,\hat{\theta}} = 0{,}667$, $RMSE = 2{,}993$ y $ASB = -0{,}124$. Estos valores van aumentando para el coeficiente de correlación y disminuyendo para $RMSE$ y ASB conforme aumenta el tamaño del test y, de este modo, en el test de 75 ítems $\rho_{\theta,\hat{\theta}} = 0{,}904$, $RMSE = 1{,}578$ y $ASB = 0{,}008$ en C1, $\rho_{\theta,\hat{\theta}} = 0{,}928$, $RMSE = 1{,}484$ y $ASB = 0{,}002$ en C2, y $\rho_{\theta,\hat{\theta}} = 0{,}947$, $RMSE = 1{,}416$ y $ASB = -0{,}007$ en C3.

Tanto $\rho_{\theta,\hat{\theta}}$ como $RMSE$ muestran también mejoría de las estimaciones de la habilidad cuando la discriminación es mayor. El índice ASB, por el contrario, sobrestima más este parámetro cuando la discriminación es la mayor. La excepción en el signo de este índice está en C1 del test de 50 ítems, $ASB = 0{,}001$, y en C1 y C2 del test de 75 ítems, $ASB = 0{,}008$ y $ASB = 0{,}002$, con valores muy cercanos a 0, en donde también se han realizado mejores estimaciones según ASB, junto con C1 del test de 25 ítems y C3 del test de 75 ítems, $ASB = -0{,}007$.

Tabla 4.36. $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB entre θ y $\hat{\theta}$								
		MVM				EAP		
n	C	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	N	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
10	1	0.472	4.261	-0.128	954	0.516	2.354	-0.000
	2	0.618	3.333	-0.114	988	0.645	2.147	-0.000
	3	0.667	2.993	-0.124	989	0.701	2.040	-0.000
25	1	0.726	2.324	-0.007	980	0.750	1.937	-0.000
	2	0.810	1.988	-0.010	983	0.823	1.765	-0.001
	3	0.855	1.828	-0.012	990	0.860	1.667	0.000
50	1	0.863	1.732	0.001	991	0.868	1.642	-0.003
	2	0.898	1.622	-0.015	994	0.900	1.546	-0.009
	3	0.920	1.545	-0.021	996	0.921	1.481	-0.015
75	1	0.904	1.578	0.008	998	0.900	1.546	-0.005
	2	0.928	1.484	0.002	994	0.928	1.452	-0.006
	3	0.947	1.416	-0.007	995	0.944	1.393	-0.008

En lo referente a la estimación de la habilidad por el proceso bayesiano, el efecto de la longitud del test y del parámetro de discriminación son los mismos que los descritos para MVM. Sin embargo, en este caso, los valores de $\rho_{\theta,\hat{\theta}}$ y $RMSE$ no son tan extremos: en el test de 10 ítems $\rho_{\theta,\hat{\theta}} = 0{,}516$ y $RMSE = 2{,}354$ en C1, $\rho_{\theta,\hat{\theta}} = 0{,}645$ y $RMSE = 2{,}147$ en C2, y $\rho_{\theta,\hat{\theta}} = 0{,}701$ y $RMSE = 2{,}040$ en C3. Las diferencias en las estimaciones de ambos procedimientos son menores conforme aumenta el número de ítems. El índice ASB detecta sobrestimación del parámetro de habilidad, aunque es mínima en los tests de 10 y 25 ítems, $ASB = \pm 0{,}000$; el valor absoluto más alto aparece en C3 del test de 50 ítems, $ASB = -0{,}015$.

Distribución de l_z con parámetros verdaderos (Tabla 4.37)

La media de l_z con parámetros verdaderos varía entre $-0{,}049$ de C1 del test de 25 ítems y 0.052 de la tercera condición del test de 75 ítems. Con 10, 25 y 50 ítems la media se aproxima a la de la distribución normal en la condición de mayor discriminación, mientras que en el test de 75 ítems sucede la tendencia inversa, $\hat{\mu}_{l_z} = -0{,}012$ en C1, $\hat{\mu}_{l_z} = -0{,}047$ en C2 y $\hat{\mu}_{l_z} = 0{,}052$ en C3. En la mayoría de los casos el estadístico central es infravalorado.

Las desviaciones típicas se definen en un intervalo de valores entre 0.969, de C1 si $n = 10$, y 1.033, de C3 si $n = 50$. Sobre el estadístico de dispersión no afecta ni la longitud del test ni el parámetro de discriminación.

Tabla 4.37. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros verdaderos									
								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.026	0.969	-0.674**	0.445**	0.060**	0.000	0.048	0.013
	2	-0.019	1.009	-0.856**	0.762**	0.067**	0.000	0.042	0.018
	3	-0.002	1.018	-1.022**	1.619**	0.082**	0.000	0.045	0.016
25	1	-0.049	1.012	-0.304**	-0.101	0.045**	0.000	0.038	0.009
	2	0.024	0.978	-0.459**	-0.010	0.045**	0.000	0.033	0.012
	3	-0.017	1.009	-0.661**	0.806**	0.050**	0.000	0.046	0.013
50	1	-0.016	0.977	-0.309**	0.085	0.041**	0.000	0.053	0.011
	2	-0.010	1.004	-0.323**	0.126	0.038**	0.002	0.048	0.012
	3	0.018	1.033	-0.439**	0.341*	0.039**	0.001	0.046	0.019
75	1	-0.012	1.006	-0.137	0.005	0.026	0.102	0.063	0.009
	2	-0.047	1.032	-0.342**	0.139	0.027	0.072	0.039	0.011
	3	0.052	0.977	-0.296**	0.019	0.032*	0.020	0.115	0.012

* p¡0.05

** p¡0.01

Las curvas son asimétricas negativas y significativas con $p(\alpha) < 0{,}01$, excepto la del test de 75 ítems en C1, $g_1 = -0{,}137$. El sesgo es corregido con el aumento en el número de ítems y por la menor magnitud del parámetro de discriminación, como se puede apreciar comparando los tests de 25 y 50 ítems, $g_1 = -0{,}304$, $g_1 = -0{,}459$ y $g_1 = -0{,}661$ en C1, C2 y C3 para el primero de ellos, y si $n = 50$ $g_1 = -0{,}309$, $g_1 = -0{,}323$ y $g_1 = -0{,}439$. La curva más sesgada es la de C3 del test con 10 ítems, $g_1 = -1{,}022$. Sin embargo, si $n = 75$ esta tendencia del índice de simetría no se verifica.

La distribución es significativamente leptocúrtica al 1 % en todas las condiciones del test de 10 ítems, $g_2 = 0{,}445$ en C1, $g_2 = 0{,}762$ en C2 y $g_2 = 1{,}619$ en C3; en C3 con 25, $g_2 = 0{,}806$, y 50 ítems, $g_2 = 0{,}341$ con $p(\alpha) < 0{,}05$. El resto de curvas son mesocúrticas, algunas con alturas medias-bajas como las de C1 y C2 con 25 ítems, $g_2 = -0{,}101$ y $g_2 = -0{,}010$. Las curvas más semejantes a la normal son las del test de 25 ítems en C2 y las del test de 75 ítems en C1 y C3, $g_2 = 0{,}005$ y $g_2 = 0{,}019$. El estadístico de forma es afectado por el parámetro de discriminación, cuanto más discriminativos son los ítems la curva es más alta, y por la longitud del test con efecto corrector al incrementarse el número de ítems.

La prueba de normalidad de Lilliefors es estadísticamente significativa en todas las condiciones experimentales, excepto C1 y C2 del test de 75 ítems, en donde $M.D. = 0{,}026$ con $p(\alpha) = 0{,}102$ y $M.D. = 0{,}027$ con $p(\alpha) = 0{,}072$; sólo al 1 % de nivel de significación no lo sería C3 en este mismo test, $M.D. = 0{,}032$ con $p(\alpha) = 0{,}020$.

Distribución de l_z con parámetros estimados con MVM (Tabla 4.38)

Tabla 4.38. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con MVM

								α nominal		
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01	N
10	1	-0.518	1.404	-1.055**	1.713**	0.074**	0.000	0.038	0.021	954
	2	-0.547	1.556	-1.536**	3.464**	0.115**	0.000	0.049	0.029	988
	3	-0.478	1.511	-1.563**	3.412**	0.124**	0.000	0.043	0.026	989
25	1	-0.399	1.271	-0.723**	0.903**	0.054**	0.000	0.040	0.015	980
	2	-0.392	1.310	-1.004**	2.365**	0.055**	0.000	0.042	0.017	983
	3	-0.387	1.371	-1.317**	3.196**	0.096**	0.000	0.046	0.020	990
50	1	-0.307	1.248	-0.653**	1.004**	0.066**	0.000	0.047	0.049	991
	2	-0.242	1.226	-0.570**	0.593**	0.041**	0.001	0.046	0.012	994
	3	-0.251	1.250	-0.942**	2.193**	0.064**	0.000	0.045	0.017	996
75	1	-0.225	1.219	-0.517**	0.588**	0.050**	0.000	0.053	0.013	998
	2	-0.212	1.221	-0.501**	0.164	0.049**	0.000	0.045	0.016	994
	3	-0.212	1.215	-0.783**	1.382**	0.044**	0.000	0.044	0.019	995

* p¡0.05

** p¡0.01

Estudiando la distribución de l_z con parámetros estimados por MVM, la media es infravalorada alcanzando los valores más bajos en el test de 10 ítems, $\hat{\mu}_{l_z} = -0{,}518$, $\hat{\mu}_{l_z} = -0{,}547$ y $\hat{\mu}_{l_z} = -0{,}478$ de C1 a C3. Estos valores se van incrementando y, por lo tanto, aproximándose a 0 conforme aumenta la longitud del test, logrando con 75 ítems $\hat{\mu}_{l_z} = -0{,}225$ en C1 y $\hat{\mu}_{l_z} = -0{,}212$ en C2 y C3.

La misma repercusión de aproximación al valor tipificado a consecuencia de la longitud del test lo padecen las desviaciones típicas que reducen sus magnitudes desde las más elevadas en los tests de 10 ítems, $\hat{\sigma}_{l_z} = 1{,}404$ en C1, $\hat{\sigma}_{l_z} = 1{,}556$ en C2 y $\hat{\sigma}_{l_z} = 1{,}511$ en C3, a las más bajas en los de 75 ítems, $\hat{\sigma}_{l_z} = 1{,}219$ en C1, $\hat{\sigma}_{l_z} = 1{,}221$ en C2 y $\hat{\sigma}_{l_z} = 1{,}215$ en C3.

El parámetro de discriminación sí repercute en el estadístico de tendencia central mejorándolo y no en el de dispersión.

El sesgo de la distribución es negativo dibujando las distribuciones más asimétricas con $p(\alpha) < 0{,}01$ en los tests de 10 y 25 ítems, $g_1 = -1{,}055$, $g_1 = -1{,}536$ y $g_1 = -1{,}563$ en C1, C2 y C3 si $n = 10$, $g_1 = -1{,}004$ y $g_1 = -1{,}317$ en C2 y C3 si $n = 25$. La asimetría es más acusada, además de con menor número de ítems, cuanto más discriminativos son éstos como se puede deducir de los datos descritos en estas líneas y que también se verifican en los tests de 50 y 75 ítems. La curva menos asimétrica es la C2 si $n = 75$, $g_1 = -0{,}501$.

El poder discriminativo de los ítems es una variable que perjudica a la curtosis de la distribución, ya que g_2 dibuja curvas leptocúrticas con $p(\alpha) <$ 0,01 en todas las condiciones experimentales, con exclusión de C2 del test de 75, $g_2 = 0{,}164$. Los demás datos aportados por el estadístico de forma oscilan entre 0.588, de C1 del test con 75 ítems, y 3.464, de C2 del test con 10 ítems. Con estos resultados se deriva que el incremento de la longitud del test corrige el apuntamiento de las curvas y el del parámetro de discriminación lo eleva.

Los resultados de la prueba de normalidad son todos estadísticamente significativos en el nivel de significación del 1 %, rechazándose la hipótesis nula de que la distribución de l_z es normal con parámetros estimados por MVM.

Distribución de l_z con parámetros estimados con EAP (Tabla 4.39)

Tabla 4.39. Estadísticos descriptivos de l_z, prueba de normalidad y tasas de error tipo I con parámetros estimados con EAP

								α nominal	
n	C	$\hat{\mu}_{l_z}$	$\hat{\sigma}_{l_z}$	g_1	g_2	M.D.	$p(\alpha)$	0.05	0.01
10	1	0.174	0.900	-0.828**	0.849**	0.076**	0.000	0.046	0.020
	2	0.212	0.899	-0.965**	1.050**	0.095**	0.000	0.044	0.014
	3	-0.096	1.122	-1.159**	1.824**	0.099**	0.000	0.044	0.023
25	1	0.284	0.882	-0.374**	0.223	0.049**	0.000	0.048	0.008
	2	0.265	0.861	-0.412**	0.090	0.049**	0.000	0.042	0.012
	3	0.267	0.872	-0.604**	0.421**	0.055**	0.000	0.055	0.015
50	1	0.214	0.816	-0.204**	0.148	0.028	0.065	0.054	0.014
	2	0.200	0.830	-0.340**	0.557**	0.034**	0.009	0.054	0.016
	3	0.196	0.851	-0.308**	0.434**	0.031*	0.023	0.057	0.015
75	1	0.192	0.833	-0.132	0.215	0.021	0.350	0.053	0.012
	2	0.183	0.880	-0.327**	0.304	0.029*	0.046	0.043	0.013
	3	0.173	0.859	-0.215**	-0.218	0.022	0.274	0.047	0.008

* p¡0.05

** p¡0.01

Los estadísticos de tendencia central de l_z calculados con parámetros obtenidos por EAP son sobrevalorados, apareciendo las medias más cercanas a 0 en el test de 75 ítems, $\hat{\mu}_{l_z} = 0{,}192$, $\hat{\mu}_{l_z} = 0{,}183$ y $\hat{\mu}_{l_z} = 0{,}173$ en C1, C2 y C3. A mayor parámetro de discriminación y longitud del test mejor es la tendencia de estos resultados a la media estándar. El único valor central que contrasta con el resto es el de la tercera condición del test de 10 ítems por ser inferior a 0, $\hat{\mu}_{l_z} = -0{,}096$.

Las desviaciones típicas varían entre 0.816 de C1 del test con 50 ítems y 1.122 de C3 del test con 10 ítems; este valor es el único mayor a 1, ya que la mayoría se encuentran en torno a 0.861 sin que se aprecie mejoría por el número de ítems y apenas por el parámetro de discriminación.

Cuanto menor es la longitud del test y mayor el parámetro de discriminación la curva es más asimétrica negativa con $p(\alpha) < 0{,}01$. En el test de 10 ítems $g_1 = -0{,}828$ en C1, $g_1 = -0{,}965$ en C2 y $g_1 = -1{,}159$ en C3, mientras que con 25 ítems $g_1 = -0{,}374$, $g_1 = -0{,}412$ y $g_1 = -0{,}604$. Esta característica de empeoramiento del sesgo no la verifican los tests de 50 y 75 ítems, en los que el índice más bajo aparece en la segunda de las condiciones, $g_1 = -0{,}340$ y $g_1 = -0{,}327$. Sólo la curva dibujada por C1 si $n = 75$ es simétrica, $g_1 = -0{,}132$.

La mitad de las curvas son mesocúrticas, e.g., en C1 y C2 del test de 25 ítems, $g_2 = 0{,}223$ y $g_2 = 0{,}090$, ó en C1 del test de 50 ítems, $g_2 = 0{,}148$; la otra mitad son leptocúrticas al 1 %, sobre todo las de los tests con 10 ítems: $g_2 = 0{,}849$, $g_2 = 1{,}050$ y $g_2 = 1{,}824$. En C3 con 75 ítems, $g_2 = -0{,}218$, se delinea una curva más baja que la normal pero no es significativamente discrepante de la curva media. Se podría afirmar que la altura de las curvas se rectifica conforme aumenta el tamaño del test y el apuntamiento es más acusado por el mayor poder discriminativo de los ítems.

La prueba de Lilliefors sostiene que las siguientes distribuciones de l_z calculado con estimadores bayesianos son normales: C1 del test con 50 ítems, $M.D. = 0{,}028$ con $p(\alpha) = 0{,}065$, y con 75 ítems C1 y C3, $M.D. = 0{,}021$ con $p(\alpha) = 0{,}350$ y $M.D. = 0{,}022$ con $p(\alpha) = 0{,}274$. El resto de las pruebas de normalidad muestran diferencias significativas con $p(\alpha) < 0{,}01$ ó con $p(\alpha) < 0{,}05$, como lo son C3 del test con 50 ítems, $M.D. = 0{,}031$ con $p(\alpha) = 0{,}023$, y C2 con 75 ítems, $M.D. = 0{,}029$ con $p(\alpha) = 0{,}046$.

El error tipo I es más consistente al nivel α nominal de 0.05 e infraestimado en la generalidad de condiciones, siendo llamativo el resultante en C3 si $n = 75$ calculando l_z con parámetros verdaderos, $\alpha = 0{,}115$, el cual puede ser consecuencia de las diversas interacciones de los factores: distribución de habilidad, MRI, longitud del test y magnitud del parámetro de discriminación. Un caso similar a éste por lo discrepante con el resto es el acontecido al nivel α nominal igual a 0.01, con parámetros máximo-verosímiles en C1 si $n = 50$, en donde $\alpha = 0{,}049$. Con respecto a este último nivel, el error tipo I es sobrestimado y fluctúa más que al de 0.05; los valores más altos han ocurrido en los tests de 10 ítems, indiferentemente del tipo de parámetros empleados.

4.4. Conclusiones

El estadístico de medición apropiada l_z de Drasgow, Levine y Williams (1985) es uno de los más utilizados para detectar patrones de respuesta atípicos debido a las altas tasas de identificaciones correctas y a las bajas tasas de falsos

positivos que ha obtenido en muy diversas investigaciones, tras los contrastes de los datos observados con la distribución normal tipificada a la que se ajusta l_z. Sin embargo, también han sido muchos los estudios que han puesto en entredicho la normalidad de este índice de ajuste de personas (sección 4.1) por la modificación de factores como el parámetro de discriminación, el MRI, la longitud del test... Todas ellas han derivado en la presente aplicación con el fin de ampliar el conocimiento sobre la distribución de l_z y así poder ser más objetivo a la hora de escogerlo e interpretarlo.

Si bien es cierto que el estadístico l_z está estandarizado y, por lo tanto, es independiente de los niveles de habilidad en los que se calcule, la distribución de habilidad y el empleo de valores de habilidad verdaderos o estimados no deberían repercutir en la distribución del estadístico. Desde los estudios de recubrimiento, los cuales han sido resumidos en la Tabla 4.40 mediante los valores medios de los índices con los que se han evaluado, se obtiene que los estimadores de habilidad más próximos a los parámetros verdaderos son los del modelo de 1-p resultantes del procedimiento de estimación EAP. Comparando las correlaciones de Pearson y el índice $RMSE$ en los tres MRI sin tener en cuenta el proceso de estimación, en el modelo de 3-p los estimadores se alejan más del valor verdadero de habilidad, atribuyendo al parámetro de pseudo-azar ser el causante de este distanciamiento. Los estimadores de la habilidad cuando los datos se ajustan al modelo de 2-p son de magnitud intermedia entre el modelo de 1-p y el de 3-p, por lo que, en dicho modelo la inexactitud parámetro-estimador estaría provocada por el parámetro de discriminación. Con respecto a éste, las estimaciones de θ en las condiciones experimentales que contenían ítems más discriminativos en los modelos de 2-p y 3-p mejoraban tanto con MVM como con EAP según $\rho_{\theta,\hat{\theta}}$ y $RMSE$; el poder discriminativo de los ítems tuvo mayor influencia en los tests más cortos (10 y 25 ítems), donde $\rho_{\theta,\hat{\theta}}$ y $RMSE$ se incrementaron notablemente desde la condición de menor discriminación (C1) a la de mayor discriminación (C3). El otro factor relacionado con la mejoría en la estimación de la habilidad es la longitud del test, por la cual, con el aumento en número de ítems se ganó en grado de acuerdo con los parámetros de sujeto.

El índice ASB muestra que ambos procedimientos de estimación tienden a sobrestimar la habilidad real sobre todo en el test más largo (75 ítems) y estimando con MVM. El parámetro de discriminación no tiene un efecto definido. El MRI que más sobrevalora θ es el 3-p trabajando con MVM y el de 2-p con EAP; cuando la distribución de habilidad no está sesgada también se produce este fenómeno en todos los modelos y en ambos procedimientos de estimación (Tabla 4.40).

Tabla 4.40. Promedios de $\rho_{\theta,\hat{\theta}}$, $RMSE$ y ASB							
		MVM			**EAP**		
MRI	$\theta \sim N$	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB	$\rho_{\theta,\hat{\theta}}$	$RMSE$	ASB
1-p	$g_1 = 0$	0.892	1.718	-0.033	0.894	1.525	-0.039
	$g_1 = +1$	0.898	1.724	-0.015	0.896	1.532	-0.005
	$g_1 = -1$	0.888	1.738	0.018	0.888	1.553	0.006
2-p	$g_1 = 0$	0.881	1.807	-0.044	0.887	1.559	-0.070
	$g_1 = +1$	0.879	1.816	-0.008	0.891	1.574	-0.032
	$g_1 = -1$	0.873	1.803	-0.003	0.877	1.586	-0.021
3-p	$g_1 = 0$	0.819	2.102	-0.079	0.830	1.682	-0.048
	$g_1 = +1$	0.814	2.119	-0.019	0.837	1.690	-0.008
	$g_1 = -1$	0.801	2.118	-0.036	0.813	1.748	-0.004

Lo cierto es que las diferencias entre MVM y EAP no son muy acusadas y, por lo tanto, la elección de un procedimiento u otro no debería contaminar el cálculo de l_z. Estas mismas conclusiones referentes al recubrimiento del parámetro de habilidad coinciden con las que obtuvieron Meijer y Nering (1997) y Nering (1995), aunque ellos no emplearon el modelo de 1-p. En la práctica es difícil encontrar una muestra de sujetos cuya habilidad o rasgo que se va a evaluar con un test sea exactamente normal y no sesgada, pero esto, estableciendo las pertinentes limitaciones de este estudio de simulación, no parece ser un aspecto del todo relevante en lo que se refiere a la estimación del parámetro de habilidad; en cuanto a su importancia para con la distribución de l_z será analizado en las líneas siguientes. El aumento del número de ítems del test y el incremento del parámetro de discriminación sí establecen pautas de mejorías en las estimaciones, lo que también concuerda con el estudio de Hambleton y Cook (1983).

Si se toma como referencia la prueba de normalidad de Lilliefors (1967; Marascuilo y McSweeney, 1977) la distribución de l_z no sigue la ley normal bajo las condiciones experimentales aquí planteabas salvo en contadas ocasiones, ya que existe significación estadística casi en la totalidad de las mismas. Sin embargo, estos resultados deben interpretarse con cautela porque, como toda prueba estadística, se deja afectar por el tamaño muestral; además, al igual que la prueba de Kolmogorov-Smirnov, puede rechazar la hipótesis nula de normalidad por divergencias con la distribución normal estandarizada bien en el parámetro de dificultad, en la escala de medida de los parámetros, en el índice de asimetría, o bien en el índice de curtosis; estos dos últimos podrían ser los causantes de la significación en este estudio junto con el tamaño muestral ($N = 1000$).

Si se analizan los valores de los estadísticos que se han elegido para la descripción de las curvas del índice de medición apropiada de Drasgow, Levine y Williams (1985) –media, desviación típica, sesgo y curtosis–, tal vez las discrepancias que demuestra la prueba de Lilliefors no sean tan acusadas. Con respecto al estadístico de tendencia central de l_z, los valores más cercanos a 0 siempre se han obtenido empleando los parámetros verdaderos de la habilidad y de los ítems, así como cuando la distribución de habilidad es centrada y no sesgada. Trabajando con parámetros verdaderos no parece que el MRI al que se ajustan los datos impacte en el comportamiento del índice de ajuste de personas; no obstante, si se ha recurrido a los estimadores de la habilidad y de los ítems para su cálculo, en el modelo de 1-p y 2-p las medias más centradas se han conseguido con los estimadores máximo-verosímiles, mientras que ajustando el modelo de 3-p, los estimadores bayesianos consiguen aproximar la media de l_z al valor esperado. El aumento del número de ítems y del parámetro de discriminación mejoran los resultados de las medias, las cuales son infravaloradas con parámetros verdaderos y estimadores máximo-verosímiles, y sobrevaloradas con estimadores bayesianos. Con parámetros verdaderos la presencia de pseudo-azar no influye, como tampoco lo hace el parámetro de discriminación en comparación con el modelo de 1-p.

El estadístico de dispersión es infravalorado con el uso de parámetros estimados por EAP y sobrevalorado con parámetros verdaderos y estimados por MVM. Los mejores resultados se han conseguido al calcular l_z con parámetros verdaderos y con estimadores esperados a posteriori al incrementarse el número de ítems y en las condiciones con mayor discriminación de éstos. El cómo sea el sesgo de la distribución de habilidad no parece que sea una traba para que la desviación típica de l_z obtenga buenas aproximaciones, lo que no se puede asegurar cuando se contrastan los MRI, ya que con el modelo logístico de 1-p la dispersión de l_z es más acertada que ajustando el modelo de 2-p y con éste, a su vez, mejor que con el modelo de 3-p.

Si sólo se sopesan las medias y las desviaciones típicas para catalogar de normal la distribución de índice de medición apropiada l_z se podría confirmar que no se desvirtúa la normalidad de este estadísdico y se podría recurrir a él para identificar patrones atípicos de respuesta, manteniendo l_z el estatus del mejor estadístico para detectar este tipo de respuesta indeseadas.

No cabe lugar a dudas que la distribución de l_z es asimétrica negativa sin que se haya mostrado corrección alguna por el empleo de parámetros verdaderos o de sus estimadores ni por el grado de discriminación de los ítems. En el modelo de 2-p y 3-p con estimadores máximo-verosímiles aparecieron las curvas más negativamente sesgadas, junto con las descritas cuando la distribución de

habilidad era centrada pero también asimétrica negativa. Con independencia de los parámetros utilizados, el modelo de 2-p produce índices de sesgo mayores. La asimetría de la curva es menos acusada en tests largos y con parámetros de discriminación bajos.

El índice de curtosis delinea curvas leptocúrticas sobre todo cuando l_z se ha calculado con los estimadores producto del proceso de MVM. La longitud del test es un factor importante en el estadístico de forma, ya que en los tests más cortos es en los que las curvas eran más empinadas. No hubo un efecto definido de la distribución de habilidad sobre la altura de las curvas, mientras que el MRI sí repercutió en la forma de las mismas provocando curvas más altas con el modelo de 2-p, aunque implementando el modelo de 3-p con parámetros estimados por MVM las alturas fueron superiores a las de aquel. El poder discriminativo de los ítems bien impidió que los valores del índice de curtosis se acercaran a los del apuntamiento medio –sobre todo en los tests con menos ítems–, bien lo favoreció, bien no tuvo relación con él.

Las tasas de falsos positivos se mantienen próximas a los niveles nominales, siendo más consistentes si α nominal es 0.05. Cuando $\alpha = 0{,}01$ hubo tendencia a la sobrestimación de dichas tasas, sobre todo en el test más corto (10 ítems) y con parámetros estimados por MVM. El parámetro de discriminación no tiene una influencia definida, pero en caso de que la haya, el incremento del poder discriminativo de los ítems provoca la mejoría de las identificaciones correctas. El parámetro de pseudo-azar no ha originado diferencias entre los FP del modelo de 2-p y los de 3-p. Comparando el uso de parámetros verdaderos frente a parámetros estimados, los valores de α más semejantes se han encontrado entre parámetros verdaderos y máximo-verosímiles si α nominal es 0.05, y entre los verdaderos y los bayesianos si el valor nominal es 0.01. El tipo de distribución de la habilidad no ha afectado en el error tipo I y sí lo ha hecho el MRI, de los cuales el de 1-p es el que más lo ha controlado. En definitiva, la prueba estadística de l_z para identificar patrones atípicos es conservadora y consistente en el nivel de significación nominal de 0.05, a pesar del sesgo y la curtosis de su distribución.

Los resultados de este estudio respecto a la distribución de l_z coinciden en su mayoría con los de Li y Olejnik (1997) –aunque en desacuerdo con los referentes al sesgo, ya que en su investigación la distribución de l_z era asimétrica positiva–, Nering (1995, 1997), Noonan *et al.* (1992), Reise (1995), Reise y Due (1991), y van Krimpen-Stoop y Meijer (1999, 2000). De conformidad con ellos, la distribución de l_z es centrada, asimétrica negativa y moderadamente leptocúrtica. Ante la dificultad de poder trabajar con los parámetros de habilidad verdaderos es conveniente recurrir a la estimación de dicho parámetro

por el procedimiento bayesiano o estimación esperada a posteriori y con tests de longitud próxima a 50 ítems o más, ya que con tests con menos ítems no se satisface la teoría asintótica; en cuanto al modelo de respuesta, las diferencias aparecidas entre los tres aquí empleados no llevan a desechar ninguno y, por lo tanto, el modelo más recomendable siempre será el que mejor se ajuste a los datos. Tampoco ha repercutido en gran medida el sesgo de la distribución de habilidad, pero es un factor importante que se debe tener en cuenta en el momento de interpretar los resultados.

Resumiendo, el estadístico de medición apropiada l_z de Drasgow, Levine y Williams (1985) sigue una distribución normal sesgada, por lo que los mejores límites para la regla de decisión del contraste de hipótesis se deberían definir en función de las características del sesgo o asimetría que el índice l_z presenta para una determinada muestra de sujetos. Dada la aplicabilidad social que tiene la identificación de patrones no representativos del rasgo subyacente al sujeto –como ya se apuntó en la introducción de esta tesis–, dicho índice podría ser empleado para tal fin avalado en las investigaciones que le han considerado como el mejor estadístico de detección de patrones atípicos. Parece tentador la creación programas informáticos que no sólo implementen l_z como detector de atipicidad en las respuestas, sino también que aporten información acerca de su distribución. En la bibliografía recogida hasta el momento sólo se sabe de los siguientes programas enfocados a la identificación de patrones atípicos:

- FORTRAN TCC3 de Baillie y Tatsuoka (1982) para $ECI1_z$, $ECI2_z$, $ECI4_z$ de Tatsuoka (1984, 1996) y l_z, utilizado por Birenbaum (1986).
- Un programa elaborado por Drasgow (1985) que calcula l_z, $ECI4_z$ y W_3 de Rudner (1983), e implementado por Noonan *et al.* (1992).
- RASCH/ECIZ de Nelson y Chatman (1985) para hallar los índices U_i de Wright y Stone (1979), $ECI2$, $ECI4$ de Tatsuoka y Linn (1983), $ECI2_z$ y $ECI4_z$.
- IPARM de Smith (1991) para el análisis de residuales con UB y UW de Smith (1985).
- RSP de Glas y Ellis (1994) calcula los índices de precaución de Tatsuoka (1984) adaptados al modelo de Rasch.
- WPERFIT de Ferrando y Lorenzo (2000) para el cómputo de l_z, $ECI4_z$ y la prueba χ^2_{SC} de bondad de ajuste para la CRP de Trabin y Weiss (1979, 1983; Klauer y Rettig, 1990).

- X-PAT de Doval, Núñez, Renom y Solanas (2001) analiza patrones de respuesta atípicos mediante el número de errores de Loevinger (1947, 1948), r_{bisper} de Donlon y Fischer (1968), C_i de Sato (1975), U_1^* de van der Flier (1977), C_i^* de Harnisch y Linn (1981) y NCI de Tatsuoka y Tatsuoka (1982).

La valoración de este estudio experimental queda limitado al ámbito del formato de respuesta dicotómico, por lo que futuras líneas de investigación podrían contemplar el análisis de la medición apropiada con modelos de respuesta politómica. Otro campo a explorar es el de la relación existente entre los patrones de respuesta atípica y el funcionamiento diferencial de los ítems, ya que la causa de aquellos puede estar en la presencia de diferencias estadísticas en los parámetros de los ítems cuando éstos son estimados en distintos grupos previamente igualados en un nivel de habilidad, y viceversa, la presencia de patrones atípicos de respuesta puede ocasionar que aparezcan diferencias estadísticas en los parámetros de los ítems cuando son dichos patrones los que están desacreditando el supuesto de invarianza de los parámetros. y, por supuesto, profundizar en las peculiaridades de un patrón de respuesta atípico en función de la variable a evaluar (rendimiento, personalidad, patología...) y del tipo de test, e.g., lápiz y papel o TAI.

Referencias

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34*, 42-54.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias. En R.A. Berk (Ed.), *Handbook of methods for detecting item bias* (pp. 96-116). Baltimore, MD: Johns Hopkins University Press.

Ansley, T.N. y Forsyth, R.A. (1985). An examination of the characteristic of unidimensional IRT parameters estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Baillie, R. y Tatsuoka, K.K. (1982). *TCC3* [Computer program]. Urbana, IL: University of Illinois at Urbana-Champaign, Computer-based Education Research Laboratory.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques*. New york: Marcel Dekker.

Barton, M.A. y Lord, F.M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Bulletin No. 81-20). Princeton, NJ: Educational Testing Service.

Bartram, D. (2001). The development of international guidelines on tests use: The International Test Commission Proyect. *International Journal of Testing, 1*, 33-53.

Bedrick, E.J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika, 62*, 191-199.

Bielby, W.T. y Hauser, R.M. (1977). Structural equation models. *Annual Review of Sociology, 3*, 137-163.

Binet, A. y Simon, T. (1973). *The development of intelligence in children*. New york: Arno. (Trabajo original publicado en 1916).

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10*, 167-174.

Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. En F.M. Lord y M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Reading, MA: Addison-Wesley.

Blanco, A. (1989). Fiabilidad y generalizabilidad de la observación conductual. *Anuario de Psicología, 43*, 5-32.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Bock, R.D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of AM algorithm. *Psychometrika, 46*, 443-459.

Bock, R.D., Gibbons, R. y Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bock, R.D. y Lieberman, M. (1970). Fitting a response model for n dicotomously scored items. *Psychometrika, 35*, 179-197.

Bollen, K.A. (1987). Outliers and improper solutions. *Sociological Methods and Research, 15*, 375-384.

Bollen, K.A. (1988). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 7*, 303-316.

Bollen, K.A. (1990). Outlier screening and a distribution-free test vanishing tetrads. *Sociological Methods and Research, 19*, 80-92.

Bradlow, E.T. y Weiss, R.E. (2001). Outlier measures and norming methods for computerized adaptive tests. *Journal of Educational and Behavioral Statistics, 26*, 85-104.

Bradlow, E.T., Weiss, R.E. y Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910-919.

Brennan, R.L. y Kane, M.T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277-289.

Bunday, B.D. (1984). *Basic optimisation methods*. London: Edward Arnold.

Camilli, G. y Shepard, L.A. (1994). *Methods of identifying biased test items*. Newbury Park, CA: Sage.

Choppin, B. (1983). *A two parameter latent trait model* (CSE Report No. 197). Los Angeles: University of California, Center for the Study of Evaluation.

Cronbach, L.J., Gleser, G.C., Nanda, H. y Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New york: Wiley.

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (Coord.), *Psicometría* (pp. 239-292). Madrid: Universitas.

Donlon, T.F. y Fischer, F.E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement, 28*, 105-113.

Doval, E., Núñez, M.I., Renom, J. y Solanas, A. (2001). *X-PAT: Un explorador de patrones de respuestas*. Comunicación presentada en el VII Congreso de Metodología de las Ciencias Sociales y de la Salud, Madrid.

Drasgow, F. (1982). Choice of test models for appropriateness measurement. *Applied Psychological Measurement, 6*, 297-308.

Drasgow, F. (1985). *A computer program to compute three appropriateness indices*. Unpublished computer software.

Drasgow, F. y Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology, 72*, 10-18.

Drasgow, F. y Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Drasgow, F., Levine, M.V. y McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M.V. y McLaughlin, M.E. (1991). Appropriateness for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

Drasgow, F., Levine, M.V. y Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Drasgow, F., Levine, M.V. y Zickar, M.J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47-64.

Duncan, O.D. (1975). *Introduction to structural equation models*. New york: Academic Press.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). New york: Wiley.

Ferrando, P.J. y Lorenzo, U. (2000). WPERFIT: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement, 60*, 479-487.

Frary, R.B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6*, 153-165.

Frary, R.B., Tideman, T.N. y Watts, T.M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2*, 235-256.

Fraser, C. y McDonald, R.P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Ghiselli, E.E. (1960). The prediction of predictability. *Educational and Psychological Measurement, 20*, 3-8.

Glas, C.A.W. y Ellis, J. (1994). Computer programs: RSP. *Rasch Measurement Transactions, 8*, 339-340.

Goldberger, A.S. y Duncan, O.D. (1973). *Structural equation models in the social sciences*. New york: Academic Press.

Goodman, L.A. y Kruskal, W.H. (1954). Measures of association for classifications. *Journal of the American Statistical Association, 49*, 732-764.

Green, D.M. y Swets, J.A. (1966). *Signal detection theory and psychophysics*. New york: Wiley.

Guttman L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman L. (1950). The basis for scalogram analysis. En S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star y J.A. Clausen (Eds.), *Measurement and prediction. Studies in social psychology in World War II* (Vol. 4) (pp. 60-90). Princeton, NJ: Princeton University Press.

Hambleton, R.K. y Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. En D.J. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). New york: Academic Press.

Hambleton, R.K. y Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R.K. y Traub, R.E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26*, 195-211.

Harnisch, D.L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*, 191-206.

Harnisch, D.L. y Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133-146.

Harnisch, D.L. y Tatsuoka, K.K. (1983). A comparison of appropriateness indices based on item response theory. En R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 104-122). Vancouver, Canada: Kluwer-Nijhoff Publishing.

Hattie, J.A. (1985). Methodological review: Assessing unidimensionality of test and items. *Applied Psychological Measurement, 9*, 139-164.

Hoffman, P.J. (1959). Generating variables with arbitrary properties. *Psychometrika, 24*, 265-267.

Holland, P.W. y Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523-1543.

Hosking, R.J., Joyce, D.C. y Turner, J.C. (1978). *First steps in numerical analysis*. London: Houder and Stoughton.

Hulin, C.L., Drasgow, F. y Parsons, C.K. (1983). *Item response theory. Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

International Tests Commission (2000). *Directrices internacionales para el uso de los tests* [On-line]. Disponible en: http://www.cop.es/tests/

Isaacson, E. y Keller, H.B. (1966). *Analysis of numerical methods*. New york: Wiley.

Jacobs, P.I. (1963). *A study of large score changes on the Scholastic Aptitude Test* (Research Bulletin No. 63-20). Princeton, NJ: Educational Testing Service.

Jöreskog, K.G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R.C. Atkinson, D.H. Krantz, R.D. Luce y P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2) (pp. 1-56). San Francisco: Freeman.

Jöreskog, K.G. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In P.R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265-287). Amsterdam: North-Holland.

Kane, M.T. y Brennan, R.L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 105-126.

Kennedy, W.J. y Gentle, J.E. (1980). *Statistical computing*. New york: Marcel Dekker.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*, 535-547.

Klauer, K.C. (1995). The assessment of person fit. En G.H. Fischer y I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 97-110). New york: Springer-Verlag.

Klauer, K.C. y Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43*, 193-206.

Kogut, J. (1988). *Asymptotic distribution of a person-fit statistic* (Research Report No. 88-13). Enschede, The Netherlands: University of Twente.

Kok, F.G. (1988). Item bias and test multidimensionality. En R. Langeheine y J. Rost (Eds.), *Latent trait and latent class models* (pp. 263-275). New york: Plenum.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61A*, 273-287.

Lawley, D.N. (1944). The factorial analysis of multiple item test. *Proceedings of the Royal Society of Edinburgh, 62A*, 74-82.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. En S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star y J.A. Clausen (Eds.), *Measurement and prediction. Studies in social psychology in World War II* (Vol. 4) (pp. 362-412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P.F. y Henry, N.W. (1968). *Latent structure analysis*. New york: Hougton Mifflin.

Levine, M.V. y Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M.V. y Drasgow, F. (1983a). Appropriateness measurement: Validating studies and variable ability models. En D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109-131). New york: Academic Press.

Levine, M.V. y Drasgow, F. (1983b). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.

Levine, M.V. y Drasgow, F. (1984). *Performance envelops and optimal appropriateness measurement* (Report No. 84-5). Champaign, IL: University of Illinois, Department of Educational Psychology, Model-based Measurement Laboratory. (ERIC Document Reproduction Service No. ED 263 126).

Levine, M.V. y Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.

Levine, M.V. y Rubin, B.D. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Li, M.F. y Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215-231.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*, 399-402.

Liou, M. y Chang, C.H. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika, 57*, 169-181.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph, 61* (No. 4).

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin, 45*, 507-530.

López, J.A. (1995). *Teoría de la respuesta al item: Fundamentos*. Barcelona: DM, PPU.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph* (No. 7).

Lord, F.M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 13*, 57-75.

Lord, F.M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Lumsden, J. (1977). Person reability. *Applied Psychological Measurement, 1*, 477-482.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31*, 19-26.

Marascuilo, L.A. y McSweeney, M. (1977). *Nonparametric and distribution--free methods for social sciences*. Monterey, CA: Cole Publishing Company.

Martínez, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G.N. y Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*, 529-544.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monographs* (No. 15).

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

McDonald, R.P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: LEA.

McDonald, R.P. y Ahlawat, K.S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-89.

McLeod, L.D. y Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147-160.

Mehta, C.R. y Patel, N.R. (1990). *Exact nonparametric inference: Introducing StatXact*. Cambridge, MA: Cytel Software.

Meijer, R.R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311-314.

Meijer, R.R. (1995). A supplement to "The number of Guttman errors as a simple and powerful person-fit statistic". *Applied Psychological Measurement, 19*, 166.

Meijer, R.R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3-8.

Meijer R.R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina y Whitney study. *Applied Psychological Measurement, 21*, 99-113.

Meijer R.R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147-160.

Meijer, R.R., Muijtjens, M.M. y van der Vleuten, C.P.M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9*, 77-89.

Meijer, R.R. y Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321-336.

Meijer, R.R. y Sijtsma, K. (1999). *A review of methods for evaluating the fit of item score patterns on a test* (Research Report No. 99-01). Twente, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

Meijer, R.R. y Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R.R., Sijtsma, K. y Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*, 283-298.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-307.

Miller, M.D. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement, 23*, 147-156.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Mislevy, R.J. y Bock R.D. (1990). *PC-BILOG 3.04: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Moiser, C.I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review, 47*, 355-366.

Moiser, C.I. (1942). Psychophysics and mental test theory II: The constant process. *Psychological Review, 48*, 235-249.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Netherlands: Mounton.

Molenaar, I.W. y Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75-106.

Molenaar, I.W. y Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9*, 27-45.

Mulaik, S.A. (1972). *The foundations of factor analysis*. New york: McGraw-Hill.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 1-19.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

Nandakumar, R. y Stout, W.F. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18*, 41-68.

Narayanan, P. y Swaminathan, H. (1996). Identification of items that show non-uniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Nelson, R.B. y Chatman, S.P. (1985). RASCH/ECIZ: A SAS PROC MATRIX program for Rasch analysis and person-fit statistics. *Applied Psychological Measurement, 9*, 325.

Nering, M.L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Nering, M.L. y Meijer, R.R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22*, 53-69.

Neyman, J. y Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 1-32.

Noonan, B.W., Boss, M.W. y Gessaroli, M.E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16*, 345-352.

Owen, R.J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100-115.

Parsons, C.K. (1983). The identification of people for whom job descriptive index scores are inappropriate. *Organizational Behaviour and Human Performance, 33*, 365-393.

Pierce, D.A. y Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *Journal of the Royal Statistical Society, Series B, 54*, 701-737.

Rao, C.R. (1965). *Linear statistical inference and its application*. New york: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Reckase, M.D., Carlson, J.E., Ackerman, T.A. y Spray, J.A. (1986). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the annual meeting of the Psychological Society, Toronto, Canada.

Reise, S.P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.

Reise, S.P. y Due, A.M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217-226.

Reise, S.P. y Flannery, Wm. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*, 9-26.

Reise, S.P. y Waller, N.G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143-151.

Reise, S.P. y Widaman, K.F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*, 3-21.

Rogers, H.J. y Hattie, J.A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

Rosenbaum, P.R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40*, 157-168.

Roznowsky, M.A., Tucker, L.R. y Humphreys, L.G. (1991). Three approaches to determining the dimensionality of binary data. *Applied Psychological Measurement, 14*, 127-137.

Rudner, L.M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207-219.

Rudner, L.M., Bracey, G. y Skaggs, G. (1996). The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education, 9*, 91-109.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement* (No. 17).

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tosho. (En japonés).

Schmitt, N., Chan, D., Sacco, J.M., McFarland, L.A. y Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41-53.

Schmitt, N., Cortina, J.M. y Whitney, D.J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17*, 143-150.

Shealy, R.T. y Stout, W.F. (1993). An item response theory model for test bias and differential test functioning. En P.W. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: LEA.

Siegmund, D. (1985). *Sequential analysis: Tests and confidence intervals*. New york: Springer-Verlag.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7*, 131-145.

Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory*. Amsterdam: Free University Press.

Sijtsma, K. y Meijer, R.R. (1992). A method for investigating the intersection of the item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

Sijtsma, K. y Meijer, R.R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191-208.

Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433-444.

Smith, R.M. (1991). *IPARM: Item and person analysis with the Rasch model*. Chicago: MESA Press.

Snijders, T.A.B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*, 331-342.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W.F. (1990). A new item response theory modeling approach with applications to multidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Strandmark, N.L. y Linn, R.L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement, 11*, 355-370.

Swaminathan, H. y Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 589-601.

Swaminathan, H. y Gifford, J.A. (1985). Bayesian estimation in two-parameter logistic model. *Psychometrika, 50*, 349-364.

Swaminathan, H. y Gifford, J.A. (1986). Bayesian estimation in three-parameters logistic model. *Psychometrika, 50*, 589-601.

Sympson, J.B. (1978). A model for testing with multidimensional items. En D.J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-88). Minneapolis, MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

SySTAT v. 10.0. [Computer software]. (2000). Chicago: SPSS, Inc.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

Tatsuoka, K.K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education, 9*, 65-75.

Tatsuoka, K.K. y Linn, R.L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*, 81-96.

Tatsuoka, K.K. y Tatsuoka, M.M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.

Tatsuoka, K.K. y Tatsuoka, M.M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 7*, 215-231.

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire*. Unpublished manuscript.

Thissen, D. y Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Thissen, D. y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Trabin, T.E. y Weiss, D.J. (1979). *The person response curve: Fit of individuals to item characteristic curve models* (Research Report No. 79-7). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Trabin, T.E. y Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. En D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83-108). New york: Academic Press.

van der Flier, H. (1977). Environmental factors and deviant response patterns. En y.H. Poortinga (Ed.), *Basic problems in Cross-Cultural Psychology*. Amsterdam: Swets and Zeitlinger.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse, The Netherlands: Swets and Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. En W.J. van der Linden y C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston: Kluwer-Nijhoff Publishing.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-218.

Weiss, D.J. (1973). *The stratified adaptive computerized ability test* (Research Report No. 73-3). Minneapolis, MN: University of Minnesota, Department of Psychology.

Winsberg, S., Thissen, D. y Wainer, H. (1983). *Estimation of the form of the item characteristic curve using monotone splines*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Wollack, J.A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*, 307-320.

Wollack, J.A. y Cohen, A.S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144-152.

Wollack, J.A., Cohen, A.S. y Serlin, R.C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement, 25*, 385-404.

Wright, B.D. y Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D. y Stone, M.H. (1979). *Best test design. Rasch measurement*. Chicago: MESA Press.

yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

yen, W.M. (1984). Effect of item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Zickar, M.J. y Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Indice de abreviaturas

CAT *Computerized Adaptive Testing*.

CCG Curva Característica del Grupo.

CCP Curva Característica de Persona.

CCT Curva Característica del Test.

CRP Curva de Respuesta de Persona.

EAP Estimación Esperada a Posteriori.

EB Estimación Bayesiana.

FI Función de Información.

FRI Función de Respuesta al Item.

GLIRT *Generalized Linear Item Response Theory*.

GRE-Q *Graduate Record Examination, Quantitative Section*.

GRE-V *Graduate Record Examination, Verbal Section*.

ITC *International Test Commission*.

MHM Modelo de Homogeneidad Monótona.

MPQ *Multidimensional Personality Questionnaire*.

MRI Modelo de Respuesta al Item.

MRL Modelo de Rasgo Latente.

MV Estimación por Máxima Verosimilitud.

MVC Estimación por Máxima Verosimilitud Conjunta.

MVCON Estimación por Máxima Verosimilitud Condicional.

MVM Estimación por Máxima Verosimilitud Marginal.

ROC *Receiver Operating Characteristic*.

SAT-V *Scholastic Aptitude Test, Verbal Section*.

TAI Test Adaptativos Informatizados.

TCT Teoría Clásica de Test.

TG Teoría de la Generalizabilidad.

TRI Teoría de Respuesta al Item.

TRP Teoría de Respuesta de Persona.

Apéndice: Gráficos de la distribución del estadístico de medición apropiada