INTRODUCCIÓN A LA TEORÍA DE RESPUESTA AL ITEM, UNA HERRAMIENTA PARA EL ANÁLISIS DE VARIABLES LATENTES: APLICACIÓN A LA MEDICIÓN DE LA CALIDAD DE VIDA DE LA INFANCIA

SÁNCHEZ RIVERO, Marcelino
Departamento de Economía Aplicada y Organización de Empresas
Universidad de Extremadura
correo-e: sanriver@unex.es

RESUMEN

La abundancia de conceptos en el ámbito de las ciencias sociales, en general, y de la ciencia económica aplicada, en particular, que no pueden ser observados de forma directa, y que requieren, por consiguiente, la utilización de indicadores indirectos para su cuantificación, exige que los mismos sean analizados mediante técnicas estadísticas que definen una variable latente a partir de un conjunto de ítems. Entre estas herramientas, la denominada Teoría de Respuesta al Item facilita al investigador aplicado una metodología que construye una variable latente continua a partir de múltiples variables observadas discretas (dicotómicas o policotómicas). En la presente comunicación, se abordan los aspectos esenciales de esta metodología y se realiza un ejercicio aplicado en el que, a modo ilustrativo, se analiza la calidad de vida de la infancia a nivel mundial. Haciendo uso de los datos recogidos en el informe "Estado Mundial de la Infancia 2001", elaborado por UNICEF, se cuantifica la calidad de vida de más de medio centenar de países, empleando para ello un total de 15 indicadores, que recogen aspectos sanitarios, educativos, demográficos, económicos, etc. de la infancia en dichos países. La estimación de los parámetros de dificultad y de discriminación de estos 15 indicadores hace posible identificar aquellos que causan las diferencias más importantes en calidad de vida entre países, mientras que a partir de la estimación del valor latente de calidad de vida de cada país se propone una clasificación de los mismos, agrupándolos en países con elevada calidad de vida de la infancia, países con calidad de vida media y países en los que la infancia disfruta de una escasa calidad de vida.

Palabras clave: Teoría de Respuesta al Item, dificultad, discriminación, calidad de vida.

1. Introducción

Los modelos y procedimientos de medición clásicos, conocidos bajo la denominación genérica de **Teoría clásica del test**, asumen que las diferencias en las respuestas de los individuos a un test determinado son debidas, única y exclusivamente, a la variación en la "capacidad" (actitud, percepción, opinión, disposición, etc.) de aquellos. Por consiguiente, todas las demás fuentes potenciales de variación (como las diferencias en el grado de dificultad de las preguntas que forman el test, las condiciones externas o las circunstancias personales del individuo) se consideran constantes o, a lo sumo, se presupone que tienen un efecto puramente aleatorio.

Esta imposibilidad de aislar, y analizar separadamente, las características del individuo y las características del test se ha convertido en uno de los principales inconvenientes de la Teoría clásica del test, iniciada con los estudios de Spearman (1904) y desarrollada con la formulación axiomática de Novick (1966). Además, esta teoría definía el concepto de capacidad en función de cada test particular, de forma que cuando un test es "difícil" el individuo tendrá baja capacidad, mientras que cuando un test es "fácil" el individuo tendrá elevada capacidad. Esta teoría clásica define la **dificultad** de un item como "la proporción de individuos **del grupo de interés** que responde correctamente al item". Es decir, la mayor o menor dificultad de un item depende de la capacidad de los individuos que están siendo medidos (*grupo-dependiente*) y, a la inversa, la capacidad de los individuos depende de si los ítems utilizados en el test son fáciles o difíciles (*test-dependiente*). Algo similar sucede con el grado de **discriminación** de cada item, que se define en términos de un grupo concreto de individuos. Por consiguiente, es muy difícil, desde esta óptica, comparar individuos que hayan realizado diferentes tests o comparar ítems cuyas características, de dificultad y discriminación, se hayan definido utilizando diferentes grupos de individuos.

Otro inconveniente de la teoría clásica del test es el relacionado con la fiabilidad del test. Según esta teoría, la fiabilidad de un test hace referencia a la correlación entre los valores del test, cuando éste es realizado varias veces de formas *paralelas*. Sin embargo, la construcción de *tests paralelos* es bastante difícil, si no imposible. Además, la fiabilidad de un test influye también sobre el error estándar de medida, que es considerado igual para todos los individuos que realizan el test, asunción que, según demuestran algunos autores (como Lord (1984)), no debe ser admitida.

Finalmente, la teoría clásica del test gira en torno al test en su conjunto, y no en torno a cada item particular que lo forma. Es decir, el valor alcanzado por el individuo en el test sólo permite valorar su capacidad global ante dicho test, pero impide descender a un análisis individual de cada item. Por tanto, no es posible, utilizando la teoría clásica del test, realizar predicciones sobre el comportamiento de los individuos ante un item concreto o responder a la pregunta de cuál es la probabilidad de que un individuo responda de forma correcta a un item determinado.

Este cúmulo de inconvenientes de la teoría clásica del test recomienda la utilización de una metodología de medición alternativa que permita, entre otras cosas, lo siguiente:

1º) Que las características de los ítems no sean dependientes de los individuos que responden al test.

2º) Que los valores obtenidos por los individuos al realizar el test no sean dependientes de éste último.

3º) Que el análisis se realice a nivel de item, en lugar de hacerlo a nivel de test.

4º) Que no sean necesarios tests paralelos para verificar la fiabilidad del test.

Esta metodología alternativa, que consigue, entre otros, los objetivos anteriores, es la llamada **Teoría de Respuesta al Item**.

2. La Teoría de Respuesta al Item.

Como su propio nombre indica, la Teoría de Respuesta al Item (en adelante TRI) analiza el comportamiento de los tests, no a un nivel agregado del test en conjunto, sino al nivel desagregado de cada item. En el caso más simple, que será el que se desarrolle en el presente texto, las respuestas a cada item se pueden codificar como respuesta "correcta" o como respuesta "incorrecta", de forma que si se asocia una variable U_{ij} para representar la respuesta del individuo *j* al item *i*, esta variable tomará el valor 1 si la respuesta es correcta, y 0 si la respuesta es incorrecta. La probabilidad de responder correctamente al item *i* dependerá del parámetro θ, que es valor que toma la capacidad (latente) de cada individuo[1]. De esta forma, los individuos que posean escasa capacidad latente tendrán probabilidades más bajas de responder correctamente al item que aquellos otros individuos con una elevada capacidad latente. Por consiguiente, la probabilidad $P_i(\theta)$ es una función monótonamente creciente del parámetro θ. Esta función es conocida con el nombre de **función de respuesta al item** o **curva característica de item** (o CCI).

Además del parámetro θ, la probabilidad $P_i(\theta)$ va a depender también, en función del modelo que se proponga para su estimación, de otros dos parámetros, $a_i \in (-\infty;+\infty)$ y $b_i \in (-\infty;+\infty)$, conocidos con el nombre de parámetro de **discriminación** y parámetro de **dificultad**, respectivamente.

El parámetro de dificultad es aquel punto de la escala latente de capacidad para el cual la probabilidad de responder correctamente al item es, exactamente, del 50 %. En realidad, este parámetro es un parámetro de localización, puesto que indica la posición de la CCI en la escala latente. Así, cuanto mayor sea el valor de b_i, mayor será también la capacidad necesaria para que un individuo tenga una probabilidad de 0,5 de responder correctamente al item en cuestión, lo cual significa que los ítems "difíciles" tendrán una CCI situada en el extremo superior de la escala de capacidad, mientras que la CCI de los ítems "fáciles" se ubicará en el extremo inferior de dicha escala. Cuando los valores de capacidad se estandarizan, los valores de b_i suelen variar de -2 a +2, de manera que valores de b_i próximos a -2

[1] En principio, el parámetro θ puede tomar valores entre $-\infty$ y $+\infty$, aunque en la práctica los valores que suele tomar son los valores reales comprendidos entre – 3 y + 3.

indican una extrema facilidad del item, mientras que valores próximos a +2 denotan una elevada dificultad.

Por su parte, el parámetro de **discriminación** a_i está relacionado con la pendiente de la CCI en el punto b_i, de forma que cuanto mayor sea la pendiente de la curva, mayores serán las diferencias en las probabilidades $P_i(\theta)$ de valores latentes próximos, lo que indica que el item en cuestión tiene una gran capacidad de discriminación entre los individuos con escasa capacidad y los que tienen elevada capacidad latente. A pesar de que, teóricamente, el parámetro a_i está definido en la escala $(-\infty;+\infty)$, cuando un item tiene un parámetro de discriminación negativo, lo habitual es que dicho item sea eliminado del test puesto su definición es errónea y/o se ha producido una pérdida de información de los individuos con mayor capacidad latente (un parámetro a_i negativo significa que la probabilidad $P_i(\theta)$ es menor para los individuos de capacidad elevada que para los de baja capacidad). Por otro lado, no es habitual que el valor de a_i exceda de 2, por lo que en la práctica el rango habitual del parámetro de discriminación es $(0;2)$. Valores de a_i próximos a 2 denotan ítems con gran poder de discriminación, de la misma forma que valores de a_i próximos a 0 se asocian a ítems escasamente discriminatorios.

Los diferentes modelos que formalizan la probabilidad de responder correctamente a un item en función de los parámetros θ, a_i y b_i parten de una serie de hipótesis relativas a los datos a los que se van a aplicar dichos modelos. Entre estas hipótesis hay dos que son fundamentales: la hipótesis de **independencia local** y la hipótesis de **unidimensionalidad**.

La hipótesis de independencia local significa que, dado un valor determinado de capacidad latente (θ), la probabilidad de que un individuo responda correctamente a un item determinado es independiente (localmente) de los demás ítems utilizados en el test. Esto implica, por consiguiente, que la probabilidad de que el individuo *j* (con capacidad latente θ_j) responda (correctamente o incorrectamente) a un conjunto de varios ítems es igual al producto de las probabilidades marginales de respuesta a cada uno de esos ítems. De esta forma, si se considera un conjunto de *n* ítems, la probabilidad conjunta $P(U_{1j};U_{2j};U_{3j};\ldots;U_{nj}/\theta_j)$ del individuo *j* se calculará mediante la siguiente expresión:

$$P(U_{1j};U_{2j};U_{3j};\ldots;U_{nj}/\theta_j)=P(U_{1j}/\theta_j)\,P(U_{2j}/\theta_j)\,P(U_{3j}/\theta_j)\ldots P(U_{nj}/\theta_j)=$$

$$=\prod_{i=1}^{n}P(U_{ij}/\theta_j)$$

En definitiva, esta hipótesis establece que la relación estadística entre los ítems de un test desaparece cuando se fija un valor concreto de capacidad latente (θ_j), es decir, los ítems son *localmente independientes*.

Por su parte, la hipótesis de unidimensionalidad establece que el conjunto de ítems utilizados en el test pretende cuantificar una única dimensión o variable latente. En realidad, es muy difícil, en la práctica, que sólo exista un único factor que explique las respuestas de los individuos al test, puesto que sobre el mismo influyen factores de diversa naturaleza, como factores cognoscitivos, de personalidad, de

motivación, etc. Por tanto, lo que verdaderamente exige esta hipótesis de unidimensionalidad es la existencia de un factor "dominante" que determine la respuesta al test. Este factor dominante es la capacidad latente que se está intentando medir a través del test. Los modelos TRI en los que existe sólo una capacidad latente dominante se conocen con el nombre de modelos *unidimensionales*, en oposición a los modelos *multidimensionales*[2], en los que es necesario considerar más de una dimensión latente para explicar el comportamiento de los individuos ante el test.

En realidad, estas dos hipótesis están relacionadas. Así, si después de fijar un valor latente concreto (θ_j), los ítems no son localmente independientes, es decir, están correlacionados, ello es debido a que estos ítems tienen varias dimensiones latentes en común, de forma que si se considera una única dimensión y las demás dimensiones se mantienen constantes, entonces los ítems serán independientes. Por consiguiente, cuando se verifica la hipótesis de unidimensionalidad, se cumplirá también la independencia local, hasta el punto de que, como demuestra Lord (1980), estas dos hipótesis son equivalentes. Sin embargo, la hipótesis de independencia local se puede verificar, aun cuando el modelo TRI no sea unidimensional, para lo cual es necesario especificar el espacio latente completo, es decir, tener en cuenta todas las dimensiones latentes que influyen sobre el test en cuestión. Sólo en el caso de considerar todos los rasgos latentes será posible mantenerlos constantes y que se verifique la independencia local de los ítems para cada dimensión latente.

Existen diferentes tests para contrastar tanto la hipótesis de independencia local como la hipótesis de unidimensionalidad, cuyo tratamiento detallado excede de los objetivos del presente texto. En cualquier caso, entre estos tests cabe citar el test de razón de verosimilitud de Martin-Löf (1973), el test Q_2 de Van den Wollenberg (1982) y el test R_2 de Glas (1988).

3. El modelo 1PL (o modelo de Rasch).

La necesidad de utilizar una función monótonamente creciente con una asíntota inferior en el valor 0 y una asíntota superior en el valor 1 para representar las probabilidades $P_i(\theta)$ aconseja el uso de funciones de distribución acumuladas. Aunque el primer modelo TRI utilizó la función de distribución de la curva normal, modelizando las probabilidades $P_i(\theta)$ de la siguiente forma:

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/z}\, dz$$

los modelos TRI más utilizados son aquellos que modelizan las probabilidades anteriores a través de la función logística, debido a que las estimaciones de los parámetros utilizando la distribución normal y la

[2] Estos modelos han sido poco estudiados. Entre los investigadores que han profundizado en los modelos TRI multidimensionales destaca, principalmente, McDonald (1981).

logística no difieren sustancialmente y, sobre todo, porque ésta última distribución es matemáticamente más sencilla de utilizar y, por tanto, de interpretar.

El origen del modelo 1PL se encuentra en los trabajos sobre medición psicológica y educativa realizados por Rasch a finales de los años 40 y principios de los años 50 del pasado siglo. A partir de un modelo de Poisson, Rasch desarrolló en 1960 su modelo, al que bautizó como "un modelo estructural para los ítems de un test", y en el que la probabilidad de que el individuo *j* (con valor latente θ_j) responda correctamente al item *i*, con parámetro de dificultad b_i, se modeliza de la siguiente forma:

$$P\left(U_{ij}=1/\theta_j\right)=\frac{\frac{\theta_j}{b_i}}{1+\frac{\theta_j}{b_i}}=\frac{\theta_j}{\theta_j+b_i}$$

De esta forma, la probabilidad de que el individuo *j* responda correctamente al item *i* será tanto mayor cuanto mayor sea su valor de capacidad latente (θ_j) y menor sea la dificultad del item (b_i). Si los parámetros del anterior modelo se expresan en una escala logarítmica, dicho modelo es equivalente al llamado **modelo logístico de un parámetro** o modelo 1PL, cuya expresión es la siguiente:

$$P_i\left(\theta_j\right)=\frac{1}{1+\exp\left[-\left(\theta_j-b_i\right)\right]}$$

La estimación de los parámetros de este modelo 1PL se realiza a través de métodos máximo-verosímiles. Así, si se considera una muestra de *N* individuos que realizan un test con *n* ítems, la función de verosimilitud **conjunta** (llamada así porque depende tanto de los parámetros de capacidad como de los parámetros de item) es la siguiente:

$$L\left(\boldsymbol{\theta},\mathbf{b};u\right)=\prod_{i=1}^{n}\prod_{j=1}^{N}\frac{\left(\theta_j/b_i\right)^{u_{ij}}}{1+\frac{\theta_j}{b_i}}=\frac{\prod_{i=1}^{n}b_i^{u_{i.}}\prod_{j=1}^{N}\theta_j^{u_{.j}}}{\prod_{i=1}^{n}\prod_{j=1}^{N}\left(1+\frac{\theta_j}{b_i}\right)}$$

donde $\boldsymbol{\theta}\equiv\left(\theta_1,\theta_2,\ldots,\theta_N\right)$; $\mathbf{b}\equiv\left(b_1,b_2,\ldots,b_n\right)$; $u\equiv\left(u_{ij}\right)$; $u_{i.}=\sum_{j=1}^{N}u_{ij}$; $u_{.j}=\sum_{i=1}^{n}u_{ij}$. Las sumas marginales $u_{i.}$ y $u_{.j}$ son estadísticos suficientes de los parámetros b_i y θ_j, respectivamente. Pues bien, igualando a cero las primeras derivadas parciales del logaritmo de $L\left(\boldsymbol{\theta},\mathbf{b};u\right)$ con respecto a $\boldsymbol{\theta}$ y con respecto a $\mathbf{b}$ se obtienen los estimadores máximo-verosímiles conjuntos (MVC) de los parámetros del modelo 1PL.

Sin embargo, los estimadores MVC presentan varios inconvenientes. Posiblemente, el inconveniente más importante es que estos estimadores no son consistentes ni asintóticamente insesgados cuando $N\to\infty$. Además, algunas de las propiedades deseables de los tests de hipótesis que se llevan a cabo con el modelo 1PL no se verifican cuando sus parámetros han sido estimados mediante estimadores MVC. Finalmente, la posible presencia de las llamadas **sumas nulas** o **sumas perfectas** (es decir,

cuando las sumas marginales son iguales a cero, o iguales al número total de individuos o de ítems: $u_{i.}=0$; $u_{.j}=0$; o bien $u_{i.}=n$; $u_{.j}=N$) aconseja eliminar aquellos ítems que todos los individuos hayan respondido incorrectamente (suma nula) o hayan respondido correctamente (suma perfecta), así como a todos aquellos individuos que hayan respondido incorrecta (suma nula) o correctamente (suma perfecta) a todos los ítems[3].

Para salvar los inconvenientes de la estimación por máxima verosimilitud conjunta, la función de verosimilitud puede ser condicionada a las sumas marginales de los ítems ($u_{i.}$), con lo que dicha función dependerá únicamente de los parámetros de capacidad (θ_j). La expresión de la función de verosimilitud **condicionada** es la siguiente:

$$L(\boldsymbol{\theta};u/u_{i.})=\frac{\prod_{j=1}^{N}\theta^{u_{.j}}}{\prod_{i=1}^{n}\gamma_{u_{i.}}}$$

donde $\gamma_{u_{i.}}$ son funciones combinatorias de los parámetros de capacidad. Como se puede observar, la anterior función de verosimilitud condicionada sólo contiene parámetros de capacidad, por lo que la maximización de su logaritmo permite obtener los estimadores máximo-verosímiles condicionados (o estimadores MVCO). De forma similar, si se condiciona la función de verosimilitud conjunta a las sumas marginales $u_{.j}$ es posible obtener los estimadores MVCO de los parámetros de dificultad de los ítems. Sin embargo, en la práctica, la estimación máximo-verosímil condicionada sólo se utiliza para estimar los parámetros b_i, los cuales, una vez estimados, son considerados conocidos y son incorporados a la función de verosimilitud al objeto de estimar los parámetros θ_j mediante procedimientos de estimación máximo-verosímiles convencionales.

Finalmente, una alternativa a la estimación por máxima verosimilitud condicionada consiste en marginalizar la función de verosimilitud conjunta. Para ello, es preciso integrar con respecto a la función de densidad común de los parámetros de capacidad θ_j, la cual se considera normal con media μ y desviación típica σ. Por consiguiente, la función de verosimilitud marginal viene dada por la siguiente expresión:

---

[3] La razón por la que se aconseja esta eliminación es evitar una interpretación incorrecta de los parámetros del modelo. Así, un item con parámetro de dificultad estimado igual a $-\infty$ (en el caso de suma perfecta) podría interpretarse como un item infinitamente fácil, por lo que, en teoría, cualquier otro individuo no incluido en la muestra respondería correctamente al mismo. Sin embargo, es posible que individuos no incluidos en la muestra respondan de forma incorrecta. Una justificación similar sería válida, aunque a la inversa, para el caso en que $b_i \to \infty$ (en el caso de suma nula).

$$L(\mathbf{b};u,\mu,\sigma) = \int_{-\infty}^{\infty} L(\boldsymbol{\theta},\mathbf{b};u)\ f(\theta;\mu,\sigma)\,d\theta$$

De esta forma, los parámetros b_i pueden ser estimados maximizando el logaritmo de la función $L(\mathbf{b};u,\mu,\sigma)$, para lo cual se puede utilizar, entre otros, el algoritmo EM. Los estimadores así obtenidos reciben el nombre de estimadores máximo-verosímiles marginales (MVM).

Para más detalles sobre los diferentes métodos de estimación máximo-verosímiles del modelo 1PL y sobre otros métodos de estimación alternativos, puede consultarse Fischer y Molenaar (1995, pp. 39-51).

Finalmente, una amplia variedad de contrastes estadísticos se han desarrollado en la literatura para verificar varios aspectos del modelo 1PL. De hecho, han sido varios los estadísticos que se han propuesto para contrastar globalmente el modelo, pero también para determinar la posible violación de algunas de las asunciones del mismo, como la propiedad de suficiencia de las sumas marginales, la unidimensionalidad de la variable latente o el paralelismo y el crecimiento estrictamente monótono de las curvas características. Dentro de esta batería de tests destacan aquellos que comparan las frecuencias observadas con los correspondientes valores esperados (tests basados en el contraste de Pearson), los estadísticos de razón de verosimilitud y los estadísticos de Wald. Un tratamiento detallado de todos estos tests queda fuera de objetivo general de este texto, por lo que se remite al lector interesado a Fischer y Molenaar (1995, pp.69-95).

4. El modelo 2PL (o modelo de Birnbaum).

El modelo 1PL anteriormente presentado supone que todos los ítems discriminan de igual forma entre individuos con distinta capacidad latente. Pero esta hipótesis es poco realista, ya que, en la práctica, lo habitual es que la discriminación entre los individuos de reducida capacidad latente y aquellos otros de elevada capacidad difiera de unos ítems a otros, lo que obliga a introducir en el modelo 1PL un nuevo parámetro, conocido como **parámetro de discriminación** (a_i), que permite identificar qué ítems tienen mayor poder de discriminación de la capacidad latente de los individuos. Aunque el primero en formular un modelo TRI con dos parámetros basado en la función de distribución normal fue Lord (1952), el padre del modelo 2PL es Birnbaum (1968), quien sustituyó la distribución normal por una distribución logística. De esta forma, la expresión matemática del modelo 2PL es la siguiente:

$$P_i(\theta_j) = \frac{1}{1+\exp\left[-a_i(\theta_j - b_i)\right]}$$

Para este modelo, la función de verosimilitud conjunta toma la siguiente forma:

$$L(\boldsymbol{\theta},\mathbf{a},\mathbf{b};u) = \prod_{i=1}^{n}\prod_{j=1}^{N} P_i(\theta_j;a_i,b_i)^{u_{ij}}\ \left[1-P_i(\theta_j;a_i,b_i)\right]^{1-u_{ij}}$$

donde $\mathbf{a} \equiv (a_1, a_2, \ldots, a_n)$. Si se toman logaritmos en la anterior función de verosimilitud conjunta, se calculan las correspondientes derivadas parciales y se igualan a cero se obtendrá un sistema de ecuaciones de estimación, que puede ser resuelto de forma conjunta a través de un proceso iterativo, consistente en fijar unos valores iniciales para los parámetros θ_j, en resolver las ecuaciones para los parámetros de los ítems (a_i y b_i), en fijar los parámetros de los ítems, y en resolver las ecuaciones con estimaciones mejoradas de los parámetros de capacidad θ_j, y así sucesivamente. Los estimadores así obtenidos se conocen con el nombre, al igual que en el modelo 1PL, de estimadores máximo-verosímiles conjuntos (o MVC). Sin embargo, algunos problemas de convergencia y ciertas propiedades no deseables de los estimadores MVC han motivado que el método de estimación de los parámetros del modelo 2PL más comúnmente utilizado sea el de máxima verosimilitud marginal, en el cual, considerando que la distribución de la variable latente viene dada por la función de densidad $f(\theta)$, la probabilidad marginal asociada al vector de respuestas $\mathbf{u} \equiv (u_{1.}, u_{2.}, \ldots, u_{n.})$ se calcula de la siguiente forma:

$$P(\mathbf{u}/\mathbf{a},\mathbf{b}) = \int_{-\infty}^{\infty} P_i(\theta;a_i,b_i)^{u_{i.}} \left[1 - P_i(\theta;a_i,b_i)\right]^{1-u_{i.}} f(\theta)\, d\theta$$

A partir de la expresión anterior, la función de verosimilitud marginal viene dada por:

$$L(\mathbf{a},\mathbf{b};\mathbf{u}) = \prod_{u=1}^{2^n} \pi_u^{r_u}$$

siendo π_u la probabilidad asociada al vector de respuestas $\mathbf{u}$ y r_u la frecuencia observada de dicho vector. La maximización del logaritmo de la anterior función de verosimilitud, que como puede observarse sólo depende de los parámetros de item, conduce a los estimadores MVM de los parámetros a_i y b_i. El principal inconveniente de este método, desarrollado por Bock y Lieberman (1970) es que el elevado número de posibles vectores de respuesta (2^n) hace que el proceso de estimación sea muy lento cuando el número de ítems (n) es muy elevado. Para salvar este inconveniente, Bock y Aitkin (1981) modificaron este método de estimación, introduciendo el algoritmo EM, haciéndolo plenamente operativo para cualquier número de ítems, pero especialmente para cuando $n > 10$. Los estimadores MVM de los parámetros de item son consistentes y asintóticamente eficientes, como demuestra Andersen (1980).

Finalmente, y a diferencia de lo que ocurre con el modelo 1PL, en el modelo 2PL no es tan importante contrastar la bondad global de ajuste del modelo a los datos como analizar otros aspectos relacionados con el mismo, como pueden ser la verificación de las hipótesis del modelo, la invariabilidad de los parámetros o la validez de las predicciones. Entre los tests para verificar la validez de las hipótesis del modelo destaca, una vez más, el utilizado para contrastar la unidimensionalidad de la variable latente, a cuyo fin se remite al lector interesado a Hattie (1985). La contrastación de la invariabilidad de los parámetros del modelo es una forma indirecta de probar la validez global del mismo, puesto que, tal y como demuestran Hambleton, Swaminathan y Rogers (1991), si el ajuste del modelo es aceptable, las estimaciones de los parámetros de capacidad deberían ser las mismas cuando se utilicen un conjunto diferente de ítems, y las estimaciones de los parámetros de dificultad y discriminación no deberían

cambiar sustancialmente al emplear diferentes muestras de individuos. Finalmente, el análisis de los residuos, o de los residuos estandarizados, del modelo puede ser útil para desarrollar tests estadísticos sobre las desviaciones entre los datos reales y las predicciones basadas en el modelo. Una exposición detallada de estos tests puede encontrarse en Yen (1981).

5. Construcción de una escala latente de la calidad de vida de la infancia.

El concepto "calidad de vida" es un concepto complejo que tiene un claro carácter multidisciplinar. De las diferentes disciplinas desde las que se puede abordar el estudio de la calidad de vida (biología, medicina, psicología, sociología, …), nuestro interés se centra en el enfoque de las ciencias sociales, en general, y de la economía, en particular, que relaciona la calidad de vida con el bienestar de la población y utiliza medidas económicas (como el Producto Interior Bruto o la renta *per cápita*) o sociales (como la tasa de analfabetismo o el índice de criminalidad).

Otro problema añadido en el concepto de calidad de vida es que la misma puede tener tanto una naturaleza subjetiva como objetiva. Es decir, la calidad de vida depende, en gran medida, de la percepción subjetiva que el individuo hace de determinadas condiciones de vida, pero depende también, y no en menor grado, de una serie de condiciones objetivas, las cuales se dejan al margen de la valoración subjetiva del individuo.

Además, la posibilidad de utilizar diferentes indicadores en función del aspecto concreto de la calidad de vida que se esté estudiando convierte a ésta última en un concepto multidimensional, en cuyo análisis hay que tener en cuenta también que no todos los indicadores tienen la misma importancia a la hora de cuantificar la calidad de vida de un individuo. Así, por ejemplo, el estado de salud es un indicador indiscutible de la calidad de vida, pero mientras que para las personas mayores este índice cobra una especial relevancia, para la población juvenil es un factor prácticamente irrelevante.

El carácter directamente inobservable de la calidad de vida (que permite considerarla como una variable latente) y la necesidad de utilizar diferentes indicadores para su cuantificación convierten a las técnicas TRI en un instrumento estadístico adecuado para el estudio de la misma. El objetivo que nos planteamos es jerarquizar un conjunto de países según la calidad de vida de la infancia y determinar cuáles de los indicadores empleados para su análisis discriminan mejor entre los países considerados en el análisis comparativo.

Para ello, se ha utilizado el Informe Mundial de la Infancia en 2001, realizado por UNICEF, en el que se recogen diversas estadísticas económicas y sociales de diferentes países del mundo referidas al bienestar de la infancia. Entre los indicadores de calidad de vida de la infancia, este informe recoge, además de algunos índices básicos (como esperanza de vida, tasa de mortalidad infantil, tasa de alfabetización, número de nacimientos anuales o PNB per cápita), indicadores de nutrición, sanitarios, educativos, demográficos, económicos, etc.

De este amplio conjunto se han seleccionado un total de 15 indicadores con el objeto de medir la calidad de vida de la infancia de 60 países. La descripción de estos 15 indicadores se muestra en la Tabla 5.1., recogiéndose en el Anexo 1 las respuestas de estos 60 países a los mismos.

Tabla 5.1.

Indicadores de calidad de vida de la infancia utilizados para cuantificar la variable latente[4]

Indicadores básicos

Item 1: ¿ Es la tasa de mortalidad de los menores de 5 años del país en el año 1999 inferior al 82 ‰ ?.

Item 2: ¿ Es el PNB per cápita del país en el año 1999 superior a 4.884 dólares ?.

Item 3: ¿ Supera la esperanza de vida al nacer en el país, en 1999, los 64 años ?.

Indicadores de salud

Item 4: ¿ Disponía en el año 1999 más de un 81 % de la población de fuentes mejoradas de agua potable ?.

Item 5: ¿ Son financiadas, en 1999, por el gobierno nacional al menos un 83 % de las vacunas administradas de forma periódica para proteger a los niños ?.

Item 6: ¿ Es el porcentaje de la población del país con inmunización completa contra la poliomelitis en el período 1997-1999 superior al 76 % ?.

Indicadores educativos

Item 7: ¿ Es la tasa de escolarización en la enseñanza primaria femenina en el período 1995-1999 superior al 79 % ?.

Item 8: ¿ Superaba en el año 1999 el 75 % la proporción de alumnos de enseñanza primaria que alcanzan el quinto grado ?.

Indicadores demográficos

Item 9: ¿ Es la tasa bruta de natalidad en 1999 inferior al 22 ‰ ?.

Item 10: ¿ Es la tasa global de fecundidad en el año 1999 inferior a 2,6 hijos por mujer ?.

Item 11: ¿ Vivía en 1999 más del 57 % de la población del país en núcleos urbanos ?.

Indicadores económicos

Item 12: ¿ Es la tasa media anual de inflación en el período 1990-1998 inferior al 19 % ?.

Item 13: ¿ Superaba en 1999 el 12 % el porcentaje del gasto del gobierno en salud ?.

Situación de las mujeres

Item 14: ¿ Es en 1999 el porcentaje de partos atendidos por personal especializado (médicos, enfermeras, parteras y agentes de atención primaria) superior al 56 % ?.

Ritmo de progreso

Item 15: ¿ Es el valor medio anual de reducción de la tasa de mortalidad en el período 1990-1999 superior al 1,3 % ?.

5.1. Análisis de los indicadores.

Dada la importancia de determinar el grado de dificultad de los 15 indicadores utilizados, pero también de cuantificar el grado de discriminación de los mismos, se ha optado por el modelo 2PL como el más idóneo para analizar estadísticamente la validez de los citados ítems como índices de la calidad de vida de la infancia a nivel mundial.

Tras una primera estimación de los parámetros de discriminación a través del software estadístico Winsteps, se aprecia que existen tres indicadores con discriminación negativa. Se trata del item 12 ($\hat{a}_{12} = -0{,}67$), del item 13 ($\hat{a}_{13} = -0{,}07$) y del item 15 ($\hat{a}_{15} = -0{,}17$). La presencia de ítems que discriminan negativamente puede distorsionar todo el análisis (especialmente, la propuesta de una jerarquización de los individuos según su nivel de capacidad) puesto que son indicadores para los que la probabilidad de respuesta correcta de los individuos con mayor capacidad latente es (o puede ser)

[4] En todos los casos, 1 significa “sí” (respuesta correcta) y 0 significa “no” (respuesta incorrecta).

menor que la de los individuos con menor capacidad. Por consiguiente, es conveniente eliminar estos tres ítems al objeto de evitar interpretaciones erróneas.

Con los 12 indicadores restantes se procede a estimar de nuevo el modelo 2PL y a analizar el signo de sus parámetros de discriminación, observándose que los ítems 4 y 6 poseen coeficientes estimados de discriminación negativos ($\hat{a}_4 = -0{,}19$; $\hat{a}_6 = -0{,}06$) por lo que, por las mismas razones que las anteriormente apuntadas, es conveniente su eliminación del análisis.

En una tercera estimación del modelo 2PL con los restantes 10 indicadores se obtienen los parámetros de dificultad y discriminación estimados que se muestran en la Tabla 5.2. Se puede apreciar en la citada tabla que sólo 4 de los 10 items analizados tienen un parámetro de dificultad positivo, siendo el item 2 el que mayor grado de dificultad presenta ($\hat{b}_2 = 5{,}42$), puesto que sólo aquellos países con valor latente de calidad de vida superior a 5,42 tienen una probabilidad superior al 50 % de tener un PNB *per cápita* superior a 4.884 dólares en el año 1999. También los indicadores 9 y 10 poseen un elevado grado de dificultad. Por su parte, los indicadores 1 y 3 son los menos difíciles de todos aquellos (6 en total) que tienen un parámetro de dificultad negativo, con un valor estimado de dicho parámetro de $\hat{b}_i = -2{,}50$. También son relativamente fáciles los ítems 5 y 7 ($\hat{b}_i = -2{,}15$ en ambos casos).

Sin embargo, los indicadores más difíciles no tienen por qué ser también los que mejor discriminen. De hecho, el item 2, que es el que mayor parámetro de dificultad arroja, es uno de los que menor capacidad de discriminación tienen ($\hat{a}_2 = 0{,}71$), situándose en niveles de discriminación similares a los de los indicadores 5, 7 y 8. No obstante, el item que peor discrimina es el 11, con un parámetro estimado igual a $\hat{a}_{11} = 0{,}13$. Por el contrario, uno de los dos indicadores más fáciles, el 3, es el segundo en poder de discriminación ($\hat{a}_3 = 1{,}49$), sólo superado por el item 9 ($\hat{a}_9 = 1{,}59$).

Si se considera dificultad alta cuando el parámetro b_i es positivo y dificultad baja cuando $b_i \leq 0$, y discriminación alta cuando el parámetro a_i es superior a 1 y discriminación baja cuando $0 \leq a_i \leq 1$, es posible clasificar los 10 items analizados de la forma que se muestra en la Tabla 5.3.

Dado que en la definición de la variable latente no serán igualmente importantes los indicadores con escasa dificultad y discriminación que aquellos otros que sean más difíciles y que discriminen claramente entre los individuos, parece más adecuado ponderar más estos últimos que los primeros. Por tanto, al objeto de poder diferenciar entre unos países y otros y evitar la posible igualdad en sus valores latentes de calidad de vida, se ha procedido a estimar de nuevo el modelo 2PL, pero ponderando ahora los ítems de la siguiente forma:

- Ponderación para los indicadores con dificultad y discriminación bajas: 1

- Ponderación para los indicadores con dificultad alta y discriminación baja: 1,05

- Ponderación para los indicadores con dificultad baja y discriminación alta: 1,10

- Ponderación para los indicadores con dificultad y discriminación altas: 1,15

Tabla 5.2.
Estimación de los parámetros de dificultad y de discriminación en el modelo 2PL no ponderado

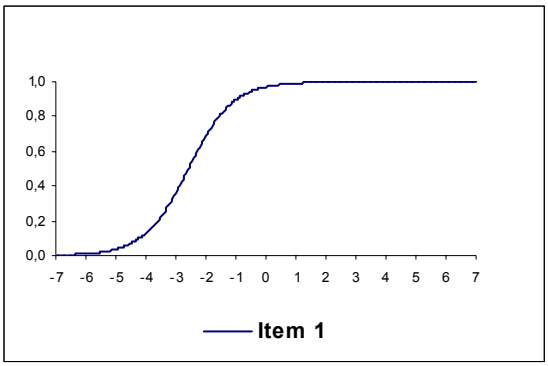
	Dificultad ($\hat{b}_i$)	Discriminación ($\hat{a}_i$)
Item 1	- 2,50	1,35
Item 2	5,42	0,71
Item 3	- 2,50	1,49
Item 5	- 2,15	0,94
Item 7	- 2,15	0,67
Item 8	- 0,55	0,68
Item 9	2,25	1,59
Item 10	2,25	1,38
Item 11	1,42	0,13
Item 14	- 1,48	1,46

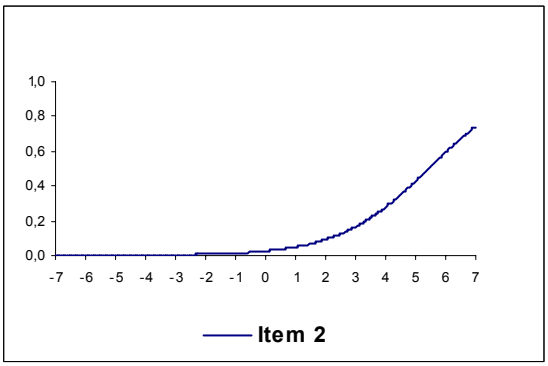
Tabla 5.3.
Clasificación de los ítems analizados en función de su dificultad y de su discriminación

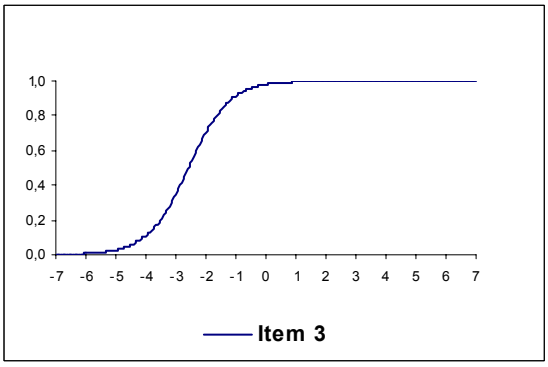
		Discriminación	
		Discriminación alta $a_i > 1$	Discriminación baja $0 \le a_i \le 1$
Dificultad	Dificultad alta $b_i > 0$	Item 9 Item 10	Item 2 Item 11
	Dificultad baja $b_i \le 0$	Item 1 Item 3 Item 14	Item 5 Item 7 Item 8

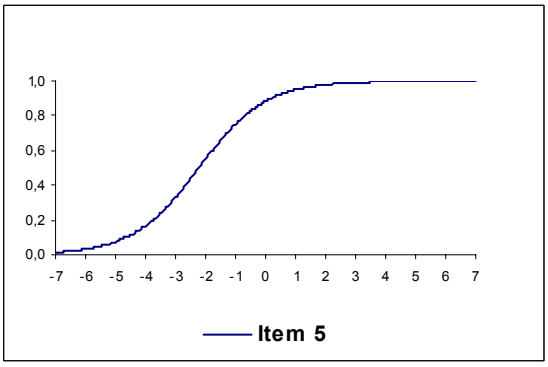
Con este sistema de ponderaciones, la mínima puntuación que puede obtener un país no cambia respecto al modelo no ponderado (0 en ambos casos), pero sí la puntuación máxima (10 puntos en el modelo sin ponderar; 10,7 en el modelo ponderado).

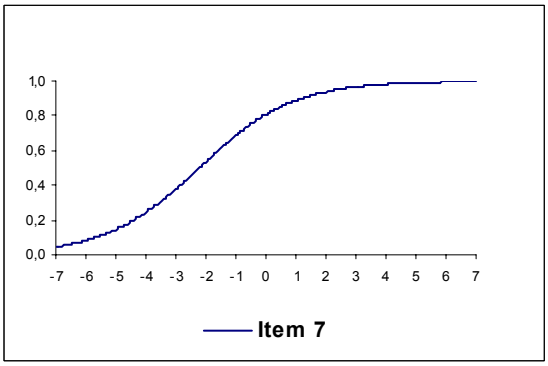
Finalmente, las curvas características de los 10 items considerados son las que se muestran en los gráficos siguientes, en los que se puede apreciar claramente cómo estos ítems difieren (en algunos casos de forma clara) en dificultad y en discriminación:

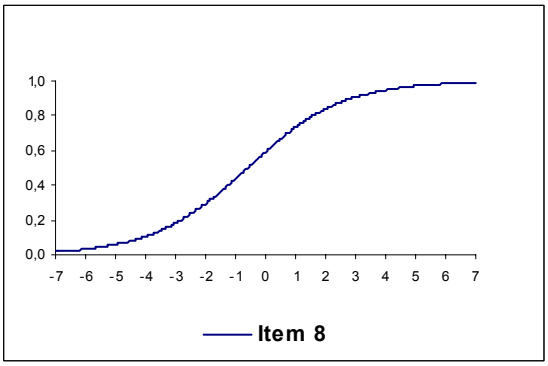

1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 1

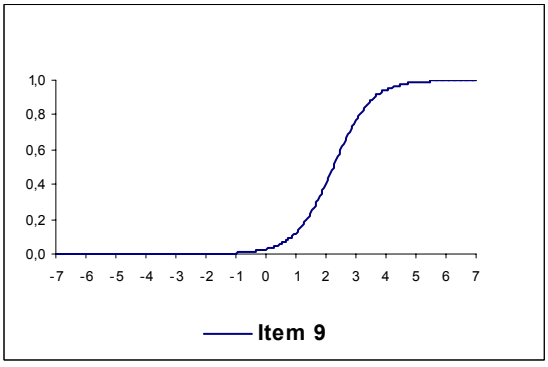

1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 2

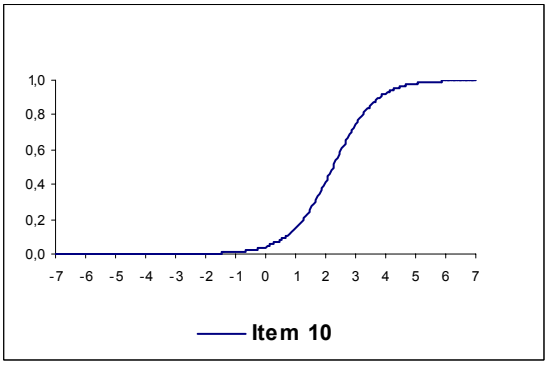

1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 3

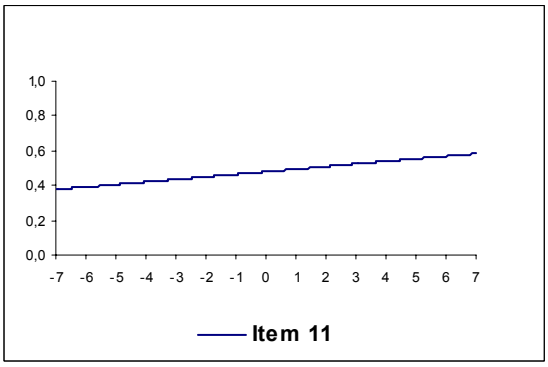

1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 5

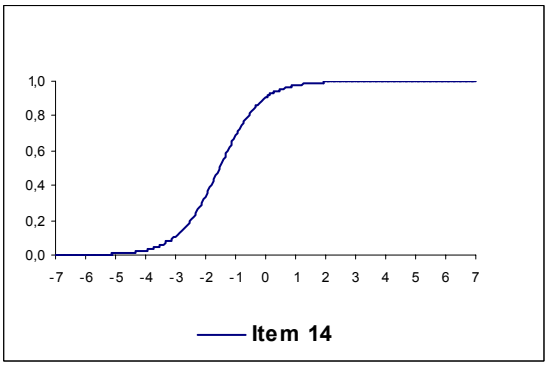

1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 7


1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 8


1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 9


1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 10


1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 11


1,0
0,8
0,6
0,4
0,2
0,0
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
Item 14

5.2. Análisis de países.

Tras estimar el modelo 2PL con los coeficientes de ponderación antes indicados, es posible establecer una ordenación de los 60 países estudiados en función de su mayor o menor grado de capacidad latente que, en este caso, es la calidad de vida de la infancia. Para ello, basta estimar el parámetro de capacidad (θ_j) de cada uno de ellos. Estas estimaciones se muestran en la Tabla 5.4.

Un análisis de estas estimaciones (que, recordemos, representan una medida de la calidad de vida de la infancia en cada país) permite clasificar a los 60 países analizados en 4 grandes grupos:

Países con una **excelente** calidad de vida de la infancia ($\hat{\theta}_j = 6{,}51$): son aquellos países que obtienen la máxima puntuación posible (10,7), lo que significa que se sitúan en la categoría 1 de los 10 items empleados finalmente para medir la variable latente "calidad de vida". Este grupo está formado por un total de 23 países, que son los siguientes: Alemania, Argentina, Australia, Austria, Bélgica, Canadá, Dinamarca, España, Estados Unidos, Finlandia, Francia, Grecia, Irlanda, Islandia, Italia, Japón, Luxemburgo, Noruega, Portugal, Reino Unido, Suecia, Suiza y Uruguay.

Países con una **aceptable** calidad de vida de la infancia ($1{,}47 \leq \hat{\theta}_j \leq 4{,}44$): se trata de países que responden correctamente a más de la mitad de los ítems, alcanzando puntuaciones marginales comprendidas entre 9,7 y 7,3 puntos. En este segundo grupo se encuentran Arabia Saudita, Bulgaria, Chile, China, Croacia, Hungría, Indonesia, Jamaica, Kuwait, México, Panamá, Perú, Rumania, Turquía y Venezuela.

Países con una **preocupante** calidad de vida de la infancia ($-1{,}98 \leq \hat{\theta}_j \leq 0{,}50$): en este tercer grupo se encuentran aquellos países que responden correctamente a menos de la mitad de los ítems empleados en el análisis, lo que supone puntuaciones marginales bajas, situadas entre 6,3 y 3,2 puntos. Los 11 países de este grupo son Costa Rica, Ecuador, Egipto, Filipinas, Guatemala, Honduras, Marruecos, Nicaragua, Paraguay, República Dominicana y Siria.

Países con una **pésima** calidad de vida de la infancia ($\hat{\theta}_j \leq -2{,}02$): en este último grupo se encuentran aquellos países con las puntuaciones marginales más bajas (inferiores a 3,2 puntos) como consecuencia del escaso número de indicadores que superan. Hay, incluso, 4 países (Camerún, Mozambique, Nepal y Rwanda) que no superan ninguno de los 10 items analizados, lo que se traduce en una puntuación marginal nula. Además de estos 4 países, en los que la población infantil vive en unas condiciones muy desfavorables (y en ocasiones extremas), completan este grupo Bolivia, Gambia, India, Kenya, Malí, Pakistán y Zambia.

Tabla 5.4.

Capacidades latentes estimadas (valor latente de calidad de vida de la infancia) de los países analizados

País	Puntuación	Valor latente $\hat{\theta}_j$	País	Puntuación	Valor latente $\hat{\theta}_j$
Alemania	10,7	6,51	Islandia	10,7	6,51
Arabia Saudita	7,4	1,52	Italia	10,7	6,51
Argentina	10,7	6,51	Jamaica	8,6	2,78
Australia	10,7	6,51	Japón	10,7	6,51
Austria	10,7	6,51	Kenya	1,0	- 3,77
Bélgica	10,7	6,51	Kuwait	9,6	4,23
Bolivia	3,2	- 2,02	Luxemburgo	10,7	6,51
Bulgaria	9,7	4,44	Malí	1,0	- 3,77
Camerún	0,0	- 5,12	Marruecos	4,2	- 1,30
Canadá	10,7	6,51	México	8,5	2,66
Chile	9,7	4,44	Mozambique	0,0	- 5,12
China	8,6	2,78	Nepal	0,0	- 5,12
Costa Rica	6,3	0,45	Nicaragua	4,3	- 1,23
Croacia	9,7	4,44	Noruega	10,7	6,51
Dinamarca	10,7	6,51	Pakistán	2,1	- 2,75
Ecuador	6,3	0,50	Panamá	8,6	2,78
Egipto	6,3	0,45	Paraguay	5,3	- 0,46
España	10,7	6,51	Perú	7,3	1,47
Estados Unidos	10,7	6,51	Portugal	10,7	6,51
Filipinas	6,3	0,50	Reino Unido	10,7	6,51
Finlandia	10,7	6,51	R. Dominicana	5,3	- 0,42
Francia	10,7	6,51	Rumania	8,6	2,78
Gambia	1,1	- 3,65	Rwanda	0,0	- 5,12
Grecia	10,7	6,51	Siria	6,3	0,45
Guatemala	3,2	- 1,98	Suecia	10,7	6,51
Honduras	4,2	- 1,30	Suiza	10,7	6,51
Hungría	9,7	4,44	Turquía	9,7	4,44
India	1,0	- 3,77	Uruguay	10,7	6,51
Indonesia	8,6	2,78	Venezuela	7,3	1,47
Irlanda	10,7	6,51	Zambia	2,0	- 2,83

Como complemento a este análisis por países, se han estimado las probabilidades condicionadas de respuesta correcta a cada uno los ítems, $P_i(\theta_j)$, de un país del primer grupo (España), de dos países del segundo grupo (Hungría y Perú), de otros dos países del tercer grupo (Ecuador y Marruecos) y de un país del cuarto y último grupo (Mozambique). Estas probabilidades se muestran en la Tabla 5.5.

Tabla 5.5.

Probabilidad de respuesta correcta a cada uno de los ítems de 6 países

Ítems	España	Hungría	Perú	Ecuador	Marruecos	Mozambique
Item 1	1,0000	0,9999	0,9953	0,9829	0,8348	0,0283
Item 2	0,6844	0,3327	0,0571	0,0295	0,0084	0,0006
Item 3	1,0000	1,0000	0,9973	0,9887	0,8567	0,0198
Item 5	0,9997	0,9980	0,9678	0,9235	0,6898	0,0578
Item 7	0,9970	0,9881	0,9187	0,8551	0,6386	0,1203
Item 8	0,9918	0,9675	0,7980	0,6713	0,3752	0,0428
Item 9	0,9989	0,9702	0,2244	0,0583	0,0035	0,0000
Item 10	0,9972	0,9536	0,2542	0,0820	0,0074	0,0000
Item 11	0,6596	0,5969	0,5016	0,4701	0,4125	0,2994
Item 14	1,0000	0,9998	0,9867	0,9474	0,5653	0,0049

Un análisis de estas probabilidades pone de manifiesto las acusadas diferencias entre países en cuanto a calidad de vida de la infancia. Estas diferencias son especialmente palpables en los indicadores 9 y 10 (de los más difíciles y de los que más discriminan), en los que las probabilidades de responder correctamente pasan de valores superiores al 95 % en países como España y Hungría, a valores en torno al 25 % en Perú y a valores prácticamente nulos en Ecuador, Marruecos y Mozambique. Además, el elevado grado de dificultad del indicador 2 hace que la mayoría de los países analizados tengan probabilidades inferiores al 50 % de responder al mismo de forma correcta. De hecho, de los 6 países cuyas probabilidades se muestran en la anterior tabla, sólo España supera dicha probabilidad (68,44 %), reduciéndose ésta a un 33,27 % para Hungría, y a valores próximos al 5 % o claramente inferiores a este valor para los restantes 4 países. Finalmente, se puede apreciar también que las diferencias de probabilidad de algunos indicadores son mucho menos acusadas que las de otros. Un ejemplo evidente de ello son los ítems 7 y 8, pero sobre todo el item 11. Esto es así porque, según se ha demostrado con anterioridad, estos tres son los indicadores que menor poder de discriminación poseen.

5.3. Consideraciones adicionales.

Aunque calcula los residuos y los residuos estandarizados para todos los países y todos los indicadores objeto de análisis, el software estadístico empleado en el presente trabajo (Winsteps) no computa, a partir de dichos residuos, un estadístico chi-cuadrado de Pearson de ajuste global del modelo 2PL a los datos empíricos, debido a que sus grados de libertad están indeterrminados como consecuencia de la imposición de restricciones en el proceso de estimación de los parámetros. Por otra parte, cuando se utilizan modelos TRI no se trata tanto de determinar si el ajuste del modelo a los datos es más o menos perfecto (o más o menos aceptable) como de verificar si dicho ajuste es admisible en la generalidad de los casos. Es decir, ¿ son los residuos (estandarizados) lo suficientemente grandes como para poner en duda la validez del modelo empleado ?.

Para dar respuesta a esta cuestión, se muestran en el Anexo 2 todos los residuos estandarizados para los 60 países y los 15 indicadores originalmente empleados. Se observa que los residuos estandarizados de algunos países son especialmente elevados para los ítems 4, 6, 12, 13 y 15, pero no debe olvidarse que estos 5 indicadores han sido eliminados del análisis por su discriminación negativa, lo que también podría explicar, al menos en parte, los elevados valores que alcanzan estos residuos. Sin embargo, si se analizan los residuos estandarizados para los 10 indicadores restantes, se podrá comprobar que éstos son, en general, relativamente pequeños en términos generales. De hecho, en la Tabla 5.6. se muestran aquellos residuos estandarizados mayores que 2 en valor absoluto. Si se tiene en cuenta que sólo 11 de los 600 residuos estandarizados calculados para 60 países y 10 items son estadísticamente significativos, es evidente que se puede concluir que el ajuste del modelo 2PL a los datos empíricos es, en general, aceptable, por lo que podemos dar por válidos los análisis anteriormente realizados.

Tabla 5.6.
Residuos estandarizados superiores a 2 en valor absoluto

País	Indicador	Valor observado	Valor esperado	Residuo estandarizado
Arabia Saudita	Item 2	1	0,019	7,105
Arabia Saudita	Item 7	0	0,977	- 6,466
Bolivia	Item 11	1	0,032	5,528
India	Item 5	1	0,174	2,181
Kenya	Item 7	1	0,174	2,181
Kuwait	Item 10	0	0,879	- 2,693
Malí	Item 8	1	0,040	4,894
Nicaragua	Item 5	0	0,727	- 1,633
Rep. Dominicana	Item 5	0	0,857	- 2,453
Rep. Dominicana	Item 11	1	0,140	2,483
Zambia	Item 8	1	0,097	3,056

Otra cuestión de gran importancia en la teoría de respuesta al item es la relativa a la unidimensionalidad del modelo. Para ello, se puede realizar un análisis de componentes principales de los residuos estandarizados al objeto de determinar dimensiones "secundarias" en el modelo 2PL. Es decir, partiendo de que la primera dimensión (o componente) que subyace en los datos es la propia variable latente que se está cuantificando a través del modelo TRI, el empleo del análisis de componentes principales para explicar la varianza de los residuos permite verificar si existen otras dimensiones significativas. Para el caso que nos ocupa, los resultados de este análisis se muestran en la Tabla 5.7., en la que se aprecia que la variable latente explica el 87,1 % de la varianza residual, lo que deja bien a las claras que estamos ante la dimensión "dominante", hasta el punto de que las dimensiones secundarias explican únicamente el 12,9 % restante (un 2,5 % el primer factor; un 2,1 % el segundo factor; un 1,7 % el tercero; un 1,3 % el cuarto; y un 5,3 % los restantes 11 posibles factores). A la vista de estos resultados, la unidimensionalidad del modelo 2PL estimado no ofrece dudas.

Tabla 5.7.

Otros cálculos estadísticos relativos a los residuos estandarizados

a) **Análisis de componentes principales para los indicadores**		
Tabla de varianza de residuos estandarizados:	Valor	Porcentaje
- Total varianza residual	116,0	100 %
- Varianza explicada por la variable latente	101,0	87,1 %
- Total varianza no explicada	15,0	12,9 %
- Varianza (no explicada) explicada por:		
- Factor 1	2,8	2,5 %
- Factor 2	2,5	2,1 %
- Factor 3	2,0	1,7 %
- Factor 4	1,5	1,3 %
- Factor 5	1,3	1,1 %
- Resto de factores (factores 6 a 15)	4,9	4,2 %

b) **Mayores correlaciones estandarizadas residuales entre los indicadores**	
Indicadores	Correlación residual
4 – 5	0,54
2 – 15	0,45
1 – 3	0,42
11 – 14	0,39
2 – 10	- 0,41
4 – 14	- 0,43
5 – 14	- 0,48
9 – 11	- 0,50
2 – 7	- 0,83

Finalmente, el análisis de la correlación entre los residuos estandarizados de los indicadores es útil en tanto en cuanto permite detectar indicadores que podrían ser localmente dependientes (cuando la correlación entre los residuos de dos indicadores sea suficientemente elevada). De esta forma, se podría sospechar de la existencia de otras dimensiones latentes cuando la mayor parte de estas correlaciones sean muy elevadas. Según se puede apreciar en la Tabla 5.7., sólo las correlaciones entre los ítems 2 y 7, entre los items 4 y 5, y entre los ítems 9 y 11 son superiores a 0,5 en valor absoluto. Por consiguiente, no se puede afirmar que la dependencia local entre ítems sea tan generalizada como para sospechar de la posible presencia de otras dimensiones latentes.

6. Referencias bibliográficas.

Andersen, E.B. (1980): *Discrete statistical models with social science applications.* Ed. North-Holland, Amsterdam.

Birnbaum, A. (1968): "Some latent trait models and their use in inferring an examinee´s ability". En F.M. Lord y M.R. Novick, *Statistical theories of mental test scores* (capítulos 17 a 20). Reading, MA: Addison-Wesley.

Bock, R.D. y Aitkin, M. (1981): "Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm". *Psychometrika*, nº 46, pp. 443-459.

Bock, R.D. y Lieberman, M. (1970): "Fitting a response model for *n* dichotomously scored items". *Psychometrika*, nº 35, pp. 179-197.

Fischer, G.H. y Molenaar, I.W. (1995): *Rasch models. Foundations, recent developments and applications.* Ed. Springer-Verlag, New York.

Fondo de las Naciones Unidas para la Infancia – UNICEF (2002): *Estado Mundial de la Infancia 2001*.

Glas, C.A.W. (1988): "The derivation of some tests for the Rasch model from the multinomial distribution". *Psychometrika*, nº 53, pp. 525-546.

Hambleton, R.K., Swaminathan, H. y Rogers, H.J. (1991): *Fundamentals of item response theory*. Sage Publications, Newbury Park, California.

Hattie, J. (1985): "Assessing unidimensionality of tests and items". *Applied Psychological Measuremets*, nº 9, pp. 139-164.

Lord, F.M. (1952): *A theory of test scores.* Psychometric monograph, nº 7. Iowa City, IA: Psychometric Society.

Lord, F.M. (1980): *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. (1984): "Standard errors of measurement at different ability levels". *Journal of Educational Measurement*, nº 21, pp. 239-243.

Martin-Löf, P. (1973): *Statistika modeller* (statistical models). Anteckningar fran seminarier lasaret 1969-1970, utarbetade av Rolf Sundberg. Obetydligt andrat nytryck, Oktober 1973. Stockholm: Institutet for Forsakrings-matematik och Matematisk Statistisk vid Stockholms Universitet.

McDonald, R.P. (1981): "The dimensionality of tests and items". *British Journal of Mathematical and Statistical Psychology*, nº 34, pp. 110-117.

Novick, M.R. (1966): "The axioms and principal results of classical test theory". *Journal of Mathematical Psychology*, nº 3, pp. 1-18.

Rasch, G. (1960): *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research. Copenhagen, Dinamarca.

Spearman, C. (1904): "The proof and measurement of association between two things". *American Journal of Psychology*, nº 15, pp. 72-101.

Van der Linden, W.J. y Hambleton, R.K. (1997): *Handbook of modern item response theory*. Ed. Springer-Verlag, New-York.

Van der Wollenberg, A.L. (1982): "Two new tests statistics for the Rasch model". *Psychometrika*, nº 47, pp. 123-139.

Yen, W. M. (1981): "Using simulation results to choose a latent trait model". *Applied Psychological Measurements*, nº 5, pp. 245-262.