Artículo
Revista digital Matemática, Educación e Internet (http://www.tec-digital.itcr.ac.cr/revistamatematica/). Vol 13, No 1. Agosto – Febrero 2013.

Aplicación del modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemática.

Karol Jiménez Alfaro.
karol.jimenez@ucr.ac.cr
Instituto de Investigaciones Psicológicas
Universidad de Costa Rica

Eiliana Montero Rojas.
eiliana.montero@ucr.ac.cr
Instituto de Investigaciones Psicológicas
Universidad de Costa Rica

Resumen. El presente trabajo pretende generar evidencias empíricas en torno a la validez de la prueba de "Diagnóstico de conocimientos y destrezas en matemática del estudiante al ingresar a la universidad", de la Escuela de Matemática de la Universidad de Costa Rica, desde la aplicación del modelo de Rasch. La muestra corresponde a 2624 examinados del 2008. Los objetivos del estudio se dirigieron primeramente a establecer evidencias de validez y confiabilidad para el instrumento. Por medio de análisis de factores exploratorio se verificó la unidimensionalidad de la escala y con el modelo de Rasch se generaron evidencias para concluir un grado aceptable de confiabilidad. Con la participación de 5 jueces expertos se establecieron niveles sustantivos de desempeño, clasificando los ítems según dificultad, y según procesos y contenidos necesarios para su resolución. Para validar las valoraciones de los jueces se contrastaron sus clasificaciones de dificultad con las estimaciones obtenidas al aplicar el modelo de Rasch, y por medio de un análisis de concordancia con la medida Kappa de Cohen se logró determinar el grupo de los 3 jueces que se acercaban más a las estimaciones de Rasch y cuyas valoraciones fueron consideradas para establecer los niveles de desempeño.

Palabras clave: Pruebas estandarizadas, Matemática, validez,confiabilidad, modelos de Rasch, juicio experto, niveles de desempeño.

Abstract. The study intended, by means of the Rasch model, to provide empirical evidences regarding the validity of the test called "Diagnostic of knowledge and skills in Mathematics of the student entering the University", developed by the School of Mathematics at the University of Costa Rica. The sample consisted of 2624 examinees in the year 2008. The research objectives first addressed the issue of establishing validity and reliability evidences for the instrument. Using exploratory factor analysis the unidimensionality of the scale was confirmed, and employing the Rasch model evidence was generated to conclude an acceptable degree of reliability. With the participation of 5 expert judges substantive levels of performance were established, classifying the items according to difficulty, and according to necessary processes and contents for their solution. To validate the judges' assessments, their difficulty classifications were contrasted with the difficulty estimations from the Rasch model, and, making use of a concordance analysis with Cohen's Kappa the group of the 3 judges that were closer to Rasch estimations was determined. These 3 judges' appraisals were considered to establish the performance levels.

KeyWords: Standardized testing, Math, validity, reliability, Rasch models, expert judgement, performance levels.

1.1 Introducción

La prueba de Diagnóstico de conocimientos y destrezas en matemática del estudiante al ingresar a la universidad, a la que en adelante se le llamará DiMa, surge como proyecto de la Vicerrectoría de Docencia de la Universidad de Costa Rica, en el año 2003 con los profesores MSc. Carlos Arce Salas y MSc. Liliana Jiménez Montero como investigadores responsables (Arce y Jiménez, 2003). El principal objetivo de esta prueba es contribuir a la solución de los problemas de bajo rendimiento académico, en los cursos de matemática de la Universidad de Costa Rica, mediante el diagnóstico de los conocimientos y destrezas en matemática con que ingresan los estudiantes a la universidad y

el seguimiento de sus resultados en los cursos universitarios de matemática.

Hasta el 2006, se aplicaba a estudiantes de primer ingreso cuyo plan de estudio incluye cursos de cálculo diferencial para las carreras de las áreas de: economía, ciencias básicas e ingenierías, ciencias de la salud y agroalimentarias. A partir del 2006 y hasta el 2008, también lo realizaron estudiantes de las carreras de computación, matemáticas e informática empresarial.

Con las mediciones realizadas con esta prueba, se pretende poder conocer hasta qué punto los resultados obtenidos, de la aplicación de la prueba en un momento determinado, proporcionan una estimación adecuada del nivel real o dominio que posee el evaluado en los contenidos que se están pretendiendo medir. Con el propósito de poder contribuir en la nivelación de los estudiantes, los responsables del proyecto ubicaban a cada estudiante en un nivel, de acuerdo con la nota obtenida en la prueba. Dependiendo del nivel, y de la sede en la que estubiera inscrita la persona, se le daba una recomendación de llevar un taller de nivelación de un mes o un curso de un semestre.

Los análisis de ítems (cálculos de los índices de dificultad y discriminación) eran basados en la Teoría Clásica de los Test (TCT). Con la aplicación del modelo de Rasch, de la Teoría de Respuesta al Ítem (TRI), además de obtener un análisis detallado de la calidad técnica y del aporte de cada ítem, independientemente del conjunto de individuos sobre el que fue aplicada la prueba, lo cual es fundamental para la conformación de un banco de ítems, da cuenta del error de medida asociado a la medición realizada, permite obtener una estimación del dominio que posee el estudiante en los conocimientos evaluados e ir más allá ya que se puede:

1. estudiar la relación entre el nivel de habilidad de los evaluados y el nivel de dificultad de cada ítem, con lo cual se podría determinar los distintos niveles de desempeño de los estudiantes y poder realizar recomendaciones más acertadas en cuanto si llevar un taller de nivelación de un mes o un curso semestral;
2. realizar evaluaciones de profesores constructores de ítems y evaluadores, con el propósito de poder ir seleccionando los más competentes y conocedores de la temática y así conformar un equipo de expertos encargados de la construcción y validación de la prueba.

El aplicar un enfoque psicométrico basado en TRI en cada proceso, tanto en el desarrollo de la prueba (construcción, juzgamiento, ensamblaje) como en su administración y calificación, permite contar con más evidencias empíricas y validaciones científicas, que deben ser consideradas en las inferencias que se derivan de los resultados, a partir del uso de la prueba, para la toma de decisiones.

1.2 Referente Conceptual

1.2.1 Los conceptos de validez y confiabilidad

Las dos propiedades fundamentales de una "buena" medición son la validez y la confiabilidad (Nunnally & Bernstein, 1995; AERA et al, 1999; Martínez et al 2006).

La confiabilidad significa precisión, consistencia, estabilidad en repeticiones. Una definición conceptual bastante ilustrativa indica que un instrumento es confiable si aplicado en las mismas condiciones a los mismos sujetos produce los mismos resultados (Nunnally & Bernstein, 1995).

La evidencia de confiabilidad es condición necesaria pero no suficiente para la validez (Babbie, 2010). En efecto, el instrumento puede medir con precisión, pero eso no implica que esté midiendo el constructo de interés. Entre los indicadores de confiabilidad que usamos con más frecuencia en psicometría se incluyen el índice de discriminación,

medido por la correlación item-total en Teoría Clásica de los Tests (TCT), el Alfa de Cronbach en TCT, la cantidad de error de medición en Teoría de Respuesta a los Items (TRI) y el modelo de Rasch, y el tamaño de la función de información en TRI y Rasch (Martínez et al, 2006; Muñiz, 2003; Prieto & Delgado, 2003).

A su vez el concepto de validez sufrió, a partir de los años 1990, una importante transformación conceptual gracias al trabajo de Samuel Messick (1989). Mientras que la definición tradicional de validez nos refería prácticamente a una tautología "un instrumento es válido si mide lo que con él se pretende medir", Messick provocó una pequeña revolución en la comunidad de la medición educativa definiendo validez como el grado de propiedad de las inferencias e interpretaciones derivadas de los puntajes de los tests, incluyendo las consecuencias sociales que se derivan de la aplicación del instrumento (Padilla et al, 2006).

El artículo seminal de Messick, publicado en la revista Educational Researcher en 1989 se tituló "Meaning and values in test validation: The science and ethics of assessment" (Significado y valores en la validación de pruebas: la ciencia y la ética de la evaluación), ha provocado la escritura de cientos de obras y textos que discuten, presentan, interpretan o critican a Messick, desde diversas ópticas.

Desde nuestra perspectiva las mayores contribuciones de Messick incluyen su definición de validez como un concepto unitario, misma que fue adoptada formalmente en los Standards for Educational and Psychological Testing, publicación conjunta de la AERA (American Educational Research Association), APA (American Psychological Associaton) y NCME (National Council on Measurement in Education), y que puede considerarse el "ISO 9000" internacional en cuanto a estándares de calidad de las pruebas educativas y psicológicas.

Así, en vez de hablar de diferentes tipos de validez, Messick indica que la idea es recolectar diferentes tipos de evidencias, de acuerdo con los propósitos y usos de los instrumentos, entre ellas evidencias de contenido, predictivas y de constructo, pero concibiendo todas esas evidencias como contribuyentes a la validez de constructo.

Otro de los más importantes aportes de Messick se refiere a su reflexión en torno a que la validez no es una propiedad intrínseca de los instrumentos, sino que se define de acuerdo al propósito de la medición, la población a la que va dirigida y el contexto específico de aplicación, así un instrumento puede exhibir un grado aceptable de validez para un propósito específico y para una población particular, pero no para otros.

Además, el proceso de validación no termina, es permanente, dado que al igual que el resto de actividades de la ciencia moderna, exige comprobaciones empíricas continuas. Igualmente nos recuerda Messick que la validez no es un rasgo dicotómico, sino una cuestión de grado, no se puede decir de manera contundente que una prueba es válida, sino más propiamente se puede afirmar que la prueba exhibe un grado aceptable de validez para ciertos usos específicos y con ciertas poblaciones.

Finalmente, Messick hace recapacitar a la comunidad de medición educativa cuando afirma que el constructor(a) del instrumEl modelo de Rasch y sus propiedades ento no solo debe poner atención a lo científico- técnico sino también a lo ético: debe preocuparse por el uso que se da a los instrumentos y por las consecuencias derivadas de la aplicación de los mismos (Messick, 1989; Padilla et al, 2006).
Desde esta perspectiva, la validez psicométrica de un instrumento es solo una parte de la sistemática y rigurosa recolección de evidencia empírica, desde diferentes dimensiones, que debe emprenderse cuando se hace la pregunta: ¿Qué tan apropiadas son las inferencias generadas a partir de los puntajes de la prueba?

1.2.2 Validación psicométrica

El proceso de recolección de evidencias empíricas para la validación de un instrumento implica normalmente la consulta a jueces expertos, pero usualmente esto no es suficiente para generar evidencia de validez sólida y suficientemente creíble, hace falta al menos una aplicación piloto del instrumento y un análisis piscométrico del instrumento y de los ítems que lo componen. Entre los métodos y modelos de análisis que utilizamos en este proceso se pueden mencionar los siguientes:

Análisis de factores exploratorio y confirmatorio

Teoría Clásica de los Tests (TCT)

Teoría de Respuesta a los Items (TRI)

Modelo de Rasch

Teoría G (Generalizabilidad)

Siendo el modelo de la Teoría Clásica de los Tests (TCT) el más antiguo y conocido, incluyendo su resultado de mayor importancia práctica, el coeficiente Alfa de Cronbach, indicador de confiabilidad en términos de consistencia interna para un instrumento (Muñiz, 2003). Este es un indicador con el que se mide la precisión de la prueba en términos de su consistencia interna y apunta hacia el grado de estabilidad de los puntajes. Estima qué proporción de la variabilidad observada en los puntajes corresponde a variancia verdadera, es decir variancia debida a diferencias en el constructo que se desea medir. Su valor máximo es 1, cuanto más se aproxime a 1 mayor es el nivel de confiabilidad. En general, los programas internacionales de pruebas educativas consideran aceptables valores de Alfa mayores a 0.8, aunque autores como Nunnally & Bernstein (1995) son más estrictos cuando se habla de pruebas de altas consecuencias en donde de toman decisiones directas sobre los examinados e indican que tales exámenes debería exhibir una confiabilidad al menos de 0.9 en la medida Alfa de Cronbach. Por otra parte, si se trata de instrumentos que van a ser utilizados solamente para procesos de investigación se puede ser más flexible en el criterio. En ese caso se consideran aceptables valores de Alfa iguales o mayores a 0.7

Por su parte, los análisis de factores exploratorio y confirmatorio son técnicas multivariadas que nos permiten explorar la dimensionalidad subyacente en los datos (Martínez et al, 2006; Nunnally & Bernstein, 1995). El análisis factorial exploratorio es una técnica de la estadística multivariada que se usa en psicometría para obtener evidencias de las dimensiones subyacentes, factores o componentes que están presentes en el instrumento. A nivel global, las cargas factoriales de los ítems (que representan el peso o nivel de importancia del ítem en cada factor) se consideran óptimas si son iguales o mayores a 0.3, en valor absoluto. Antes de realizar un análisis psicométrico con la TCT, la TRI o Rasch es importante evidenciar, utilizando el análisis factorial exploratorio, que el instrumento mide fundamentalmente solo un rasgo o constructo, pues este es un supuesto que debe cumplirse para que la aplicación de estos modelos de medición sea válida.

Finalmente, con los modelos TRI (Teoría de Respuesta a los ítems) y Rasch se obtienen estimaciones de los parámetros del ítem que son menos dependientes de la muestra de examinados y estimaciones de los niveles del constructo en los evaluados que son menos dependientes de la muestra particular de ítems aplicada. Además, en estos modelos existe una estimación específica del error de medición para cada puntaje en la prueba (a diferencia de la TCT donde se asume que el error es constante) (Martínez et al, 2006; Montero, 2001). En el caso del modelo de Rasch, las estimaciones de las habilidades de los examinados y la dificultad de los ítems están en las mismas unidades de medición, propiedad que resulta sumamente atractiva a nivel aplicado y de interpretación sustantiva, pues permite evaluar el desempeño del examinado en términos de modelos criteriales, es decir valorando en términos absolutos lo que puede o no hacer (Bond & Fox, 2001; Prieto & Delgado, 2003).

1.2.3 Teoría de Respuesta al Item (TRI)

La Teoría Clásica de los Test (TCT) nos presenta una serie de estadísticos, como el error típico de medida, los índices de dificultad y discriminación de los ítems, el coeficiente de confiabilidad de Cronbach, que representan elementos esenciales en la validación de las pruebas. Sin embargo, en concordancia con Martínez (2005) a pesar del uso tan generalizado y de la enorme utilidad práctica que se ha hecho de la TCT y de todos sus estadísticos asociados, esta teoría parte de supuestos generales débiles, de escasa plausibilidad real, que constituyen tanto su fuerza como su debilidad.

De acuerdo con Muñiz (1997) la Teoría de Respuesta al Item (TRI) nace como un nuevo enfoque en la teoría de las pruebas que permite superar algunas de las limitaciones de la Teoría Clásica de los Test.

Para Barbero (1999) la década de los 60 es cuando la TRI comienza su gran desarrollo, mucho debido a la publicación del trabajo de Rasch en 1960 titulado "*Probabilistic models for some intelligence and attainment test*", y la aparición del libro de Lord y Novick en 1968 titulado "*Statistical theories of mental test scores*".

La TRI, a diferencia de la TCT, se centra más en las propiedades individuales de los ítems que en las propiedades globales del test, de ahí su nombre. Se puede decir que uno de sus propósitos es intentar obtener la puntuación que corresponde a una persona en una dimensión o rasgo, como por ejemplo, su inteligencia, su nivel en un cierto rasgo de personalidad, su dominio en una cierta materia, etc.

Dos objetivos generales de la TRI son: 1) proporcionar mediciones de las variables psicológicas y educativas que no estén en función del instrumento utilizado, y 2)disponer de instrumentos de medida cuyas propiedades no dependan de los objetos medidos, que sean invariantes respecto de las personas evaluadas. (Muñiz, 1997, p.18).

Es importante tomar en cuenta que los modelos matemáticos planteados en la TRI especifican que la probabilidad que tiene un evaluado de contestar correctamente un ítem depende de su nivel de aptitud y de las características de los ítems. Estos modelos consideran supuestos acerca de los datos que son más restrictivos que los planteados en la TCT y "la viabilidad de estos supuestos no puede determinarse directamente, pero pueden recogerse algunas evidencias que establecen el grado de concordancia entre los supuestos del modelo y los datos" (Martínez, 2005, p.248).

Para Muñiz (1997) los dos supuestos son *la curva característica de los ítems* (CCI) y *la unidimensionalidad*, Martínez (2005) además de estos presenta como supuesto de la TRI *la independencia local*. La CCI de forma general especifica que a medida que aumente el nivel de aptitud, la probabilidad de acertar correctamente un ítem también lo hará. Su formulación es una función matemática que establece la relación que existe entre la escala de aptitud o habilidad de los sujetos evaluados (usualmente se emplea la escala estandarizada, con media 0 y desviación estándar 1) y la probabilidad de acertar correctamente un ítem. La función más utilizada como CCI es la función logística, definida por tres parámetros, específicamente:

$$\begin{aligned} P_i(\theta_s) &= c_i + (1-c_i)\frac{e^{Da_i(\theta_s-b_i)}}{1+e^{Da_i(\theta_s-b_i)}} \\ &= \frac{c_i + e^{Da_i(\theta_s-b_i)}}{1+e^{Da_i(\theta_s-b_i)}} = c_i + \frac{1-c_i}{1+e^{-Da_i(\theta_s-b_i)}} \end{aligned} \tag{1.1}$$

donde D es una constante mayor a 0, a la cual se le asigna generalmente el valor de 1.7 (para buscar semejanza con la función de distribución normal); θ_s es el valor del constructo o rasgo que se desea estimar en cada examinado, a_i es el parámetro de discriminación, b_i el de la dificultad,y c_i representa la probabilidad de acertar el ítem al azar.

Según Muñiz (1997) y Martínez (2005), dependiendo de la función matemática que define la CCI y el número de parámetros a considerar se generarán diferentes modelos, siendo los que utilizan la función logística, y que pueden ser de uno, dos o tres parámetros, los que han recibido más atención. Lo cierto es que como lo menciona Muñiz (1997), cada modelo se ajusta mejor a unas situaciones que a otras, y el uso de uno u otro dependerá del caso que se desee tratar.

Se dice que el más general y realista de ellos pero el más difícil de estimar en ocasiones, es el de tres parámetros, cuya expresión matemáticamente es la mostrada en 1.1. El modelo de dos parámetros, asume que el parámetro de azar es igual a cero, es decir, al igual que el anterior, la CCI viene dada por la función logística, pero contempla únicamente dos parámetros, el índice de dificultad y el índice de discriminación. Finalmente, se tiene el modelo de un parámetro formulado originalmente por Rasch en 1960, de ahí que sea conocido también como modelo de

Rasch. En este modelo, además de asumir que la probabilidad de acertar el ítem al azar es cero, se supone que el parámetro de discriminación es constante para todos los ítems. Su expresión matemática sería entonces

$$P_i(\theta_s) = \frac{e^{D(\theta_s - b_i)}}{1 + e^{D(\theta_s - b_i)}}. \tag{1.2}$$

En este modelo, lo usual es que se asuma para la constante D el valor de 1.

Si el modelo se ha parametrizado correctamente, $P(\theta)$ dependerá únicamente de θ_s (nivel de habilidad del sujeto examinado), es decir, "la TRI asume implícitamente en su formulación que los ítems destinados a medir la variable θ_s constituyen una sola dimensión" (Muñiz, 1997, p.25).

Sin embargo, como lo sostiene Martínez (2005), aunque se asuma que el rendimiento de un ítem es explicable por una sola aptitud o rasgo, se debe ser consciente de que no puede cumplirse completamente este supuesto dado a los múltiples factores que pueden afectar en un momento dado a las respuestas de la prueba, por ejemplo, la atención, la motivación de los examinados, el contexto, la ansiedad, etc., pero sí se puede hablar de una aptitud fundamental representada en el grupo de ítems que conforman la prueba.

Por lo tanto, es necesario poder comprobar la unidimensionalidad antes de aplicar alguno de los modelos TRI. El método más utilizado para esta comprobación es el análisis factorial. Se sabe que es muy difícil lograr encontrar una unidimensionalidad perfecta, o sea, que un solo factor explique por completo la varianza total de las puntuaciones, por tanto, "la unidimensionalidad se convierte en una cuestión de grado: cuanta más varianza explique el primer factor, "más unidimensionalidad" existirá" (Muñiz, 1997, p.26).

Por otro lado, los modelos de la TRI asumen que las respuestas de los evaluados (en un mismo nivel de aptitud) a un ítem son independientes a las respuestas de los otros ítems. Por tanto, bajo la TRI, no es permitido el uso de ítems encadenados en los que la respuesta de uno dependa de la de otro. Muñiz (1997) indica que "la independencia local puede expresarse diciendo que la probabilidad de que un sujeto acierte n ítems es igual al producto de las probabilidades de acertar cada uno de ellos" (p.27). También se puede hablar de independencia local de los sujetos, en el sentido de que el rendimiento de un sujeto en un test no depende del rendimiento de los otros sujetos en el mismo test.

1.2.4 El modelo de Rasch y sus propiedades únicas

De acuerdo con Prieto y Delgado (2003), el modelo de Rasch descansa sobre los siguientes supuestos:

1. El atributo que se desea medir se puede representar en una única dimensión, en la que conjuntamente se situarían a personas e ítems.

2. El nivel de la persona en el atributo (habilidad) y la dificultad del ítem determinan la probabilidad de obtener la respuesta correcta. Rasch utilizó el modelo logístico (se obtiene al despejar de la ecuación 1.2, $(\theta_s - b_i)$ y asumiendo $D = 1$)

$$\ln\left(\frac{P_{is}}{1 - P_{is}}\right) = \theta_s - b_i \tag{1.3}$$

   donde P_{is} representa la probabilidad de que la persona s responda correctamente el ítem i, θ_s es el nivel de habilidad de la persona s en el atributo que se desea medir, y b_i es el nivel de dificultad del ítem i.

Nótese que en la ecuación 1.3, $P_{is} = \frac{1}{2}$ es equivalente a $\theta_s = b_i$, es decir, la probabilidad de acertar la pregunta es $\frac{1}{2}$ cuando la habilidad del individuo iguala la dificultad del ítem. De la misma forma, $P_{is} > \frac{1}{2}$ es equivalente a $\theta_s > b_i$, es decir, la probabilidad de acertar el ítem es superior a $\frac{1}{2}$ si la habilidad del individuo está por encima de

la dificultad del ítem, y $P_{is} < \frac{1}{2}$ es equivalente a $\theta_s < b_i$, así, cuando la habilidad es menor que la dificultad del ítem, la persona tiene mayor probabilidad de fallar la respuesta que de acertarla.

Es importante observar que, al despejar P_{is} de la ecuación 1.3 se obtiene

$$P_{is} = \frac{e^{\theta_s - b_i}}{1 + e^{\theta_s - b_i}} = \frac{1}{1 + e^{b_i - \theta_s}} \tag{1.4}$$

la cual es la formulación más utilizada del modelo.

1.2.5 Sobre la escala utilizada en Rasch

En realidad, si en un modelo se asume que la probabilidad de acierto P_{is} es una función de la diferencia entre el nivel del examinado en la habilidad y la dificultad del ítem $(\theta_s - b_i)$, entonces se está midiendo al examinado y al ítem en una misma escala. Para Prieto y Delgado (2003) la escala más utilizada es la llamada "*logit*", definida por

$$\ln\left(\frac{P_{is}}{1 - P_{is}}\right) = \theta_s - b_i. \tag{1.5}$$

De acuerdo con esta expresión, se podrían obtener valores entre $]-\infty, \infty[$, pero en la práctica, según Prieto y Delgado (2003), en la gran mayoría de los casos, los valores obtenidos se encuentran en el rango $[-5, +5]$.

Dado que

$$(\theta_s + n) - (b_i + n) = \theta_s - b_i$$

la localización del punto 0 de la escala se puede elegir arbitrariamente, pero en la práctica lo común en el modelo de Rasch, de acuerdo con Prieto y Delgado (2003), es localizar el 0 en la dificultad media de los ítems que integran el test. En este caso, $\theta_s > 0$ significa que la persona s tiene probabilidad superior a $\frac{1}{2}$ de éxito en los ítems de dificultad media.

1.2.6 Sobre la escala utilizada en Rasch

En realidad, si en un modelo se asume que la probabilidad de acierto P_{is} es una función de la diferencia entre el nivel del examinado en la habilidad y la dificultad del ítem $(\theta_s - b_i)$, entonces se está midiendo al examinado y al ítem en una misma escala. Para Prieto y Delgado (2003) la escala más utilizada es la llamada "*logit*", definida por

$$\ln\left(\frac{P_{is}}{1 - P_{is}}\right) = \theta_s - b_i. \tag{1.6}$$

De acuerdo con esta expresión, se podrían obtener valores entre $]-\infty, \infty[$, pero en la práctica, según Prieto y Delgado (2003), en la gran mayoría de los casos, los valores obtenidos se encuentran en el rango $[-5, +5]$.

Dado que

$$(\theta_s + n) - (b_i + n) = \theta_s - b_i$$

la localización del punto 0 de la escala se puede elegir arbitrariamente, pero en la práctica lo común en el modelo de Rasch, de acuerdo con Prieto y Delgado (2003), es localizar el 0 en la dificultad media de los ítems que integran el test. En este caso, $\theta_s > 0$ significa que la persona s tiene probabilidad superior a $\frac{1}{2}$ de éxito en los ítems de dificultad media.

1.2.7 Ventajas del Modelo de Rasch

Siguiendo a Prieto y Delgado (2003) algunas de las ventajas más relevantes de aplicar Rasch son:

1. La *medición conjunta*: al poderse expresar los parámetros de las personas y de los ítems en las mismas unidades, se pueden representar en el mismo continuo, lo que permite analizar las interacciones entre los individuos y los ítems.

2. La *objetividad específica:* la diferencia entre las habilidades de dos personas no depende de los ítems específicos con la que sea estimada. De la misma manera, la diferencia entre las dificultades de dos ítems no depende de las personas específicas que se utilicen para su estimación. Por ejemplo, supongamos que dos personas de distinto nivel de habilidad contestan el mismo ítem, entonces se tendría: $\ln\left(\frac{P_{i1}}{1-P_{i1}}\right) = \theta_1 - b_i$ para el sujeto 1, y $\ln\left(\frac{P_{i2}}{1-P_{i2}}\right) = \theta_2 - b_i$ para el sujeto 2, entonces, la diferencia entre las habilidades de ambas personas sería

$$\ln\left(\frac{P_{i1}}{1-P_{i1}}\right) - \ln\left(\frac{P_{i2}}{1-P_{i2}}\right) = (\theta_1 - b_i) - (\theta_2 - b_i) = \theta_1 - \theta_2.$$

   De manera similar, si una misma persona de una habilidad θ_s contesta dos ítems de diferente dificultad, se tendría $\ln\left(\frac{P_{1s}}{1-P_{1s}}\right) = \theta_s - b_1$ para uno de los ítems, y $\ln\left(\frac{P_{2s}}{1-P_{2s}}\right) = \theta_s - b_2$ para el otro. La diferencia en dificultad entre estos ítems será

$$\ln\left(\frac{P_{1s}}{1-P_{1s}}\right) - \ln\left(\frac{P_{2s}}{1-P_{2s}}\right) = (\theta_s - b_1) - (\theta_s - b_2) = b_1 - b_2.$$

3. La *propiedad de intervalo:* a diferencias iguales entre un individuo y un ítem le corresponden probabilidades idénticas de una respuesta correcta, es decir, diferencias iguales en el constructo están asociadas a diferencias iguales en los puntajes.

4. El modelo de Rasch permite: a) cuantificar la cantidad de información (y la cantidad de error) con la que se mide en cada punto de la dimensión; b) seleccionar aquellos ítems que permiten incrementar la información en regiones del constructo previamente especificadas, es decir, por ejemplo para una prueba de admisión a una universidad se desea seleccionar individuos en un nivel alto del construto, por lo que se pueden utilizar los ítems con mayor información en ese nivel.

1.3 Metodología

Esta investigación se podría enmarca dentro de los estudios exploratorios en el sentido que se pretende examinar el DiMa en búsqueda de evidencias teóricas y empíricas de validez para los usos e interpretaciones que se pueden generar de los resultados de su aplicación. Además, se puede ubicar dentro de los estudios descriptivos, pues se realiza un análisis de calidad técnica detallado de los ítems que componen la prueba, aplicando TRI. También, se puede decir que contempla elementos de estudios correlacionales, pues por ejemplo se estudia cuán relacionados están los ítems entre sí, la correlación entre los ítems y la prueba, la relación entre los niveles en la habilidad de los evaluados y el nivel de dificultad de los ítems.

Se utilizó la base de datos del 2008 con el total de la población que aplicó para el DiMa en ese año (2624 casos). Para el análisis de los ítems se estudió el fórmulario 1. Un detalle importante por aclarar es que para cada año existían 4 versiones de la prueba, se diferenciaban unas de otras por el orden que los investigadores le daban a

las opciones de respuesta, pero una vez pasada la aplicación, recodificaban según la fórmula 1, y se analizaba la población completa. No fue posible poder volver a separar la población de acuerdo al número de formulario que había resuelto, por lo que los análisis se realizaron de acuerdo con el formato que tenían en la fórmula 1 de cada año.

Para la obtención de estas bases de datos se conversó con el director de la Escuela de Matemática de la UCR en el 2009, el máster Carlos Arce Salas, quién además en ese momento era el investigador principal del proyecto de investigación de esta prueba.

El aporte de los investigadores MSc. Carlos Arce y MSc. Liliana Montero, aunado con una revisión teórica que se realizó sobre tipos de pruebas, permitieron la realización de la siguiente fase, una descripción detallada del DiMa con el propósito de poder conocer más el constructo que se pretende medir con ella y realizar una mejor valoración en los análisis.

Luego, se continuó con la validación interna, lo primero fue realizar un análisis factorial exploratorio, utilizando el método de Análisis de Componentes Principales (ACP) para obtener una aproximación de los constructos subyacentes de cada uno de las pruebas, además de ver si se cumplía en grado razonable el supuesto de unidimensionalidad. Para este análisis se utilizó el programa estadístico SPSS para Windows 15.0.

Con la aplicación del modelo de Rasch, además de depurar el análisis de la calidad técnica de los ítems, y analizar los ajustes de los datos al modelo, también se pretendía poder determinar los distintos niveles de desempeño de estas poblaciones en el constructo, esto con el propósito de poder contar con más evidencias científicas para poder dar una mejor recomendación a los estudiantes de acuerdo con su desempeño en la prueba.

También se deseaba ejemplificar uno de los usos del modelo de Rasch para la evaluación de jueces expertos, con el objetivo de poder identificar expertos que puedan realizar mejores valoraciones en el análisis de los ítems. Así que antes de su aplicación, se realizó un trabajo con 5 jueces expertos conocedores de las temáticas evaluadas en el DiMa. Dicho trabajo consistió en solicitarles a los jueces resolver en forma individual la prueba DiMa 2008 y clasificar cada ítem de acuerdo con: su dificultad, su contenido y procesos presentes en su solución. Los contenidos y procesos presentados a los jueces fueron una combinación entre los temas y destrezas definidas por los creadores del DiMa y la categorización de procesos mentales propuetos en la prueba de Habilidades Cuantitativas del Proyecto de Pruebas Específicas de la Universidad de Costa Rica, coordinada en ese momento por la Licda. Jeannette Villalobos.La información aportada de las valoraciones de los jueces fue tabulada y procesada para ser utilizada en el análisis posterior.

El siguiente paso fue analizar los ítems pero ahora desde el panorama de la Teoría de Respesta al Item, específicamente aplicando el modelo de Rasch. Para el análisis se utilizó el paquete computacional Winsteps versión 3.64.2, pues es exclusivo para llevar a cabo análisis de Rasch y además de contar con los elementos también generados con otros paquetes como el BILOG, se cuenta con uno más que resultaba muy valioso en el análisis que se pretendía realizar con los ítems y los aportes de los jueces expertos, el mapa de distribución conjunta de examinados e ítems.

Con el propósito de poder contar con evidencia estadística sobre los niveles de acuerdo entre los jueces según la dificultad de los ítems, los contenidos y los procesos, se aplicó el índice Kappa, el cuál fue calculado utilizando el SPSS para Windows 15.0. Para la clasificación de los ítems según nivel de dificultad se terminó considerando solo tres niveles: fáciles (agrupando los que se habían clasificado como muy fáciles y fáciles), mediano y difíciles (agrupando los clasificados como difíciles y muy difíciles). Primero se comparó la valoración de cada juez con los resultados obtenidos del análisis con Rasch; luego después de analizar los niveles de concordancia entre jueces, éstos fueron agrupados en 3 grupos, y se compararon los resultados de las valoraciones, según cada grupo de jueces, con los obtenidos en Rasch. Esto permitió la elaboración de tablas, en las que se reunió y organizó la información más relevante de cada ítem de la prueba DiMa 2008.

1.4 Resultados de los análisis

Se procedió con el análisis factorial exploratorio, con el objetivo de buscar evidencia de validez asociada a la estructura factorial y para ver si se cumple, en grado razonable, el supuesto de unidimensionalidad.

Se aplicó el estadístico de Kolmogorov-Smirnov, para comprobar si los datos se distribuían normalmente, y el resultado fue que la hipótesis de normalidad en este caso se rechaza, pues el nivel de significancia fue menor a 0.05, por lo que todas las variables (ítems en nuestro caso) no proceden de poblaciones con distribuciones normales.

Dado lo anterior, se decidió realizar un Análisis de Componentes Principales (ACP) aplicando método de extracción componentes principales, pues para este método no se demanda el cumplimiento del supuesto de normalidad.

Para el DiMa 2008 la muestra analizada fue de 2624 examinados. El valor del determinante de la matriz de correlaciones es $1.71(10^{-6})$ por lo que se puede confirmar, de acuerdo con Cea D'Ancona (2002), la existencia de intercorrelaciones elevadas entre las variables, ello permite que se pueda realizar el análisis factorial.

El índice de medida de adecuación muestral Kaiser-Meyer-Olkin (*KMO*) obtenido fue de 0.976, de acuerdo con Cea D'Ancona (2002) entre más próximo a 1 sea, indica que las correlaciones entre pares de variables pueden explicarse por otras variables, por lo que en este caso sería una evidencia más de que se puede realizar el análisis de factores.

En la figura 1.1 se presenta el gráfico de sedimentación del DiMa 2008. De acuerdo con este criterio, el número de factores está delimitado por el punto en el que se presenta un cambio importante en la trayectoria de caída de la pendiente, Catell, citado por Cea D'Ancona (2002), sugiere que se consideren todos aquellos factores situados antes de este punto, en este caso, este criterio sugiere la existencia de un componente predominante.

En tabla 1.1 se presenta un extracto de la tabla de varianza total explicada. Se puede apreciar que aproximadamente un 22.18% de la varianza total es explicada por el primer componente, ya el aporte del segundo componente principal es muy bajo (2.7%).

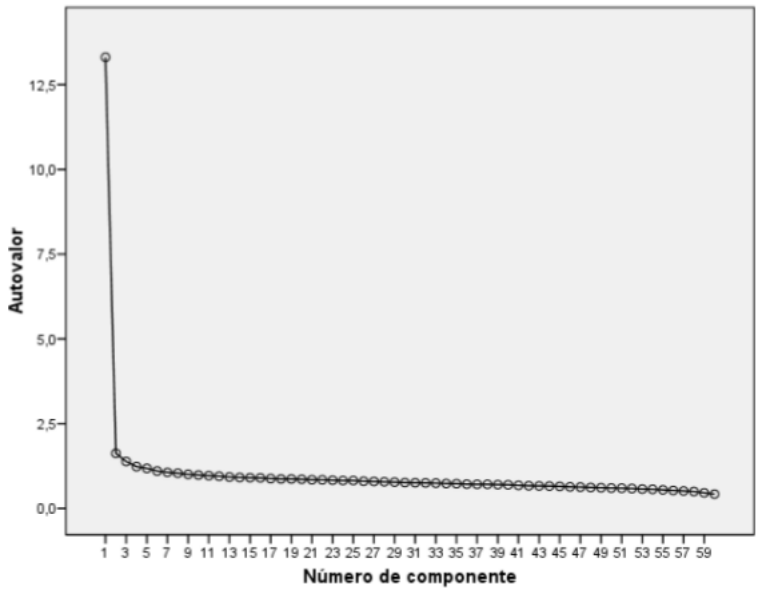

Figura 1.1 Gráfico de sedimentación DiMa 2008

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	13.308	22.180	22.180	13.308	22.180	22.180
2	1.624	2.707	24.887	1.624	2.707	24.887
3	1.384	2.306	27.193	1.384	2.306	27.193
4	1.232	2.053	29.246	1.232	2.053	29.246
5	1.182	1.970	31.216	1.182	1.970	31.216
6	1.102	1.837	33.053	1.102	1.837	33.053
7	1.062	1.771	34.823	1.062	1.771	34.823
8	1.040	1.733	36.556	1.040	1.733	36.556
9	1.005	1.675	38.232	1.005	1.675	38.232
10	0.983	1.639	39.871			
11	0.967	1.612	41.483			
12	0.950	1.583	43.066			
13	0.927	1.545	44.611			

Método de extracción: Análisis de Componentes principales.

Tabla 1.1 Varianza explicada DiMa 2008

El programa sugiere la extracción de 9 componentes, basados en el criterio del valor del autovalor superior a 1, pero debemos recordar que en nuestro caso, el objetivo de este análisis factorial no es realizar interpretaciones sustantivas de lo que significan los componentes, sólo se aplica el análisis factorial exploratorio en busca de evidencias de unidimensionalidad, y además, es muy difícil encontrar unidimensionalidad perfecta, por lo cual, como lo menciona Muñiz (1997) "la unidimensionalidad se convierte en una cuestión de grado: cuanta más varianza explique el primer factor, "más unidimensionalidad" existirá" (p.26). Lo anterior permite afirmar que en el caso del DiMa 2008 se tienen evidencias de que la prueba tendiera a ser unidimensinal, por lo que se puede continuar con los análisis de confiabilidad y también, poder aplicar el modelo de Rasch en esta muestra de datos.

1.5 Aplicación del modelo de Rasch

El estudio se realizó inicialmente a partir de las respuestas obtenidas por los 2624 individuos que realizaron la prueba, la cual está conformada por 60 ítems. En la tabla 1.2 se presentan las estadísticas de confiabilidad tanto para personas como para los ítems, obtenidas aplicando Rasch, de acuerdo con este modelo, la medida de confiabilidad de los examinados indica qué tan consistentes son los resultados, es decir, si al mismo grupo de examinados se les aplicara otro conjunto de ítems del mismo universo al que pertenece el conjunto que se está analizando, se obtendrían los mismos resultados. Para el DiMa 2008, la confiabilidad de los examinados fue de 0.93, el cual es un valor satisfactorio para este tipo de pruebas de diagnóstico.

En cuanto a la confiabilidad de los ítems, ésta lo que indica es qué tan consistentes son las estimaciones del parámetro de dificultad si el mismo conjunto de ítems se aplicara a otro conjunto de examinados con las mismas características del grupo analizado. Para este caso, el valor de la confiabilidad de los ítems es de 1, lo cual indica que las estimaciones de Rasch son muy consistentes, esto era de esperar dado que es una muestra grande de examinados.

Número Casos	Número ítems	Índice confiabilidad personas	Índice confiabilidad ítems
2624	60	0.93	1.00

	Puntajes Obtenidos	Medida estimada (personas)	Error estándar de la estimación
Media	27.00	-0.22	0.32
Desviación estándar	13.30	1.21	0.08

Tabla 1.2 Estadísticos descriptivos y de confiabilidad DiMa 2008

No se debe olvidar que uno de los objetivos con este análisis es poder identificar aquellos examinados que contestaron correctamente a ítems dentro de su nivel de habilidad y además, ítems que fueron contestados correctamente por individuos que se encuentran dentro del nivel de habilidad para hacerlo, es decir, identificar tanto los ítems como los examinados que se ajusten al modelo.

Para el modelo de Rasch, se cuenta con el índice Infit MNSQ el cual es calculado con las medias cuadráticas sin estandarizar, Wright y Linacre (1994), citados en Bond y Fox (2001), proponen como valores aceptables de Infit MNSQ, para tipos de pruebas de escogencia única, los ubicados en el rango 0.8 y 1.2; siguiendo este criterio, el único ítem del DiMa 2008 que no se ajustaría al modelo sería el p58 pues posee un valor 1.35, los otros ítems si poseen valores infit dentro del rango establecido. El ítem que resultó más difícil es el p58 con una dificultad de 1.91, en escala logit, seguido del p57 con una dificultad de 1.67, y el ítem más fácil es el p1 con una dificultad de -2.22. Se puede decir que en general se obtuvo una buena precisión de la medida realizada pues los errores asociados son cercanos a 0. De los 2624 examinados, 31 tenían un valor de Infit MNSQ superior a 1.2 y 7 poseían un valor inferior de 0.8, es decir, aproximadamente 1.45%, no se ajustaban al modelo.

Una vez detectados los ítems y personas que no cumplían con las espectativas del modelo, fueron eliminados de la base original y se volvió a correr el análisis en el Winsteps. Las nuevas estadísticas descriptivas y de confiabilidad son las que se presentan en la tabla 1.3. La confiabilidad de los examinados fue finalmente de 0.92, el cual es un valor satisfactorio para este tipo de pruebas y en cuanto a la confiabilidad de los ítems, el valor obtenido es de 0.99, lo que indica que las estimaciones del parámetro de dificultad en Rasch son muy consistentes. También se puede notar que en la estimación de la habilidad, existe una variabilidad alta y bastante simétrica, pues los puntajes oscilan entre los valores -5.51 y 5.48 en la escala logit; el promedio de ítems contestados correctamente es de 27, lo que equivale a una ubicación en la escala de la habilidad de -0.17 logits.

En las estimaciones de las dificultades de los ítems, del error estándar asociado a la medición hecha y los estadísticos de ajuste, pero ahora considerando solo los datos de los ítems y personas que si se ajustaron al modelo. Se tiene que el ítem más difícil es el p57 con una dificultad ahora de 1.74, en escala logit, mientras que el ítem más fácil es el p1 con una dificultad de -2.25. La precisión de la medición sigue siendo buena pues se continúa obteniendo valores bajos de errores estándar.

En la figura 1.2 se muestra el mapa de distribución conjunta de los individuos y los ítems en una escala logit. A la izquierda del gráfico se distribuyen los examinados según su nivel de habilidad (de arriba hacia abajo de mayor a menor puntaje en la habilidad) y a la derecha se distribuyen los ítems según su dificultad (de arriba hacia abajo de

mayor dificultad a menor dificultad). Se puede observar que el promedio del nivel de habilidad de los examinados (letra M al lado izquierdo) está muy cercano a la dificultad promedio de los ítems (fijada en 0, letra M al lado derecho); de hecho, en la tabla 1.3 se puede observar que el valor de la media de las estimaciones para la habilidad es de -0.17 logits, lo anterior indica que el conjunto de ítems resultó muy levemente difícil para esta población.

Número Casos	Número ítems	Índice confiabilidad personas	Índice confiabilidad ítems
2586	59	0.92	0.99

	Puntajes Obtenidos	Medida estimada (personas)	Error estándar de la estimación
Media	27.00	-0.17	0.33
Desviación estándar	13.40	1.30	0.14
Puntajes acierto Máx.	59.00	5.48	1.83
Puntajes acierto Mín.	0.00	-5.51	0.27

Tabla 1.3 Estadísticos con datos que ajustaron

Por otro lado, en la figura 1.2 se puede observar al lado izquierdo que existe un grupo de examinados que tiene la probabilidad de contestar correctamente todos los ítems, se debe recordar que el ítem más difícil es el p57 (dificultad de 1.74 logits, ubicado a más de una desviación estándar por encima del promedio) y analizando los resultados de las salidas obtenidas con el Winsteps, se obtuvo que de 2586 examinados 215 (un 8% aproximadamente) poseen un valor en la habilidad superior a 1.74 logits.

También, se puede apreciar que existe un grupo muy pequeño de examinados con una habilidad muy baja que tienen una alta probabilidad de fallar todos los ítems, el ítem más fácil es el p1 con dificultad -2.25, ubicado a más de dos desviaciones estándar por debajo del promedio, y existe 47 examinados, aproximadamente un 1,8% de la población examinada, que poseen un valor en la habilidad inferior a -2.25 logits, lo cual indicaría que requieren una nivelación en todos los temas evaluados en esta prueba. Pero lo que resalta más es que existe un grupo de interés de 645 examinados (un 25% aproximadamente) con habilidades entre -1.10 y -2.24 logits para los cuales no existen ítems con dificultades en ese mismo rango, que permitan hacer un diagnóstico apropiado de estos estudiantes.

Para los examinados que están en el promedio de habilidad $(\theta = -0.17)$, se puede indicar que del total de 59 ítems, tienen una baja probabilidad de contestar correctamente 36 ítems (equivalente al 61% de los ítems con dificultades superiores a -0.17), y una alta probabilidad de contestar correctamente 22 ítems (equivalente al 37,3% con dificultades inferiores a -0.17).

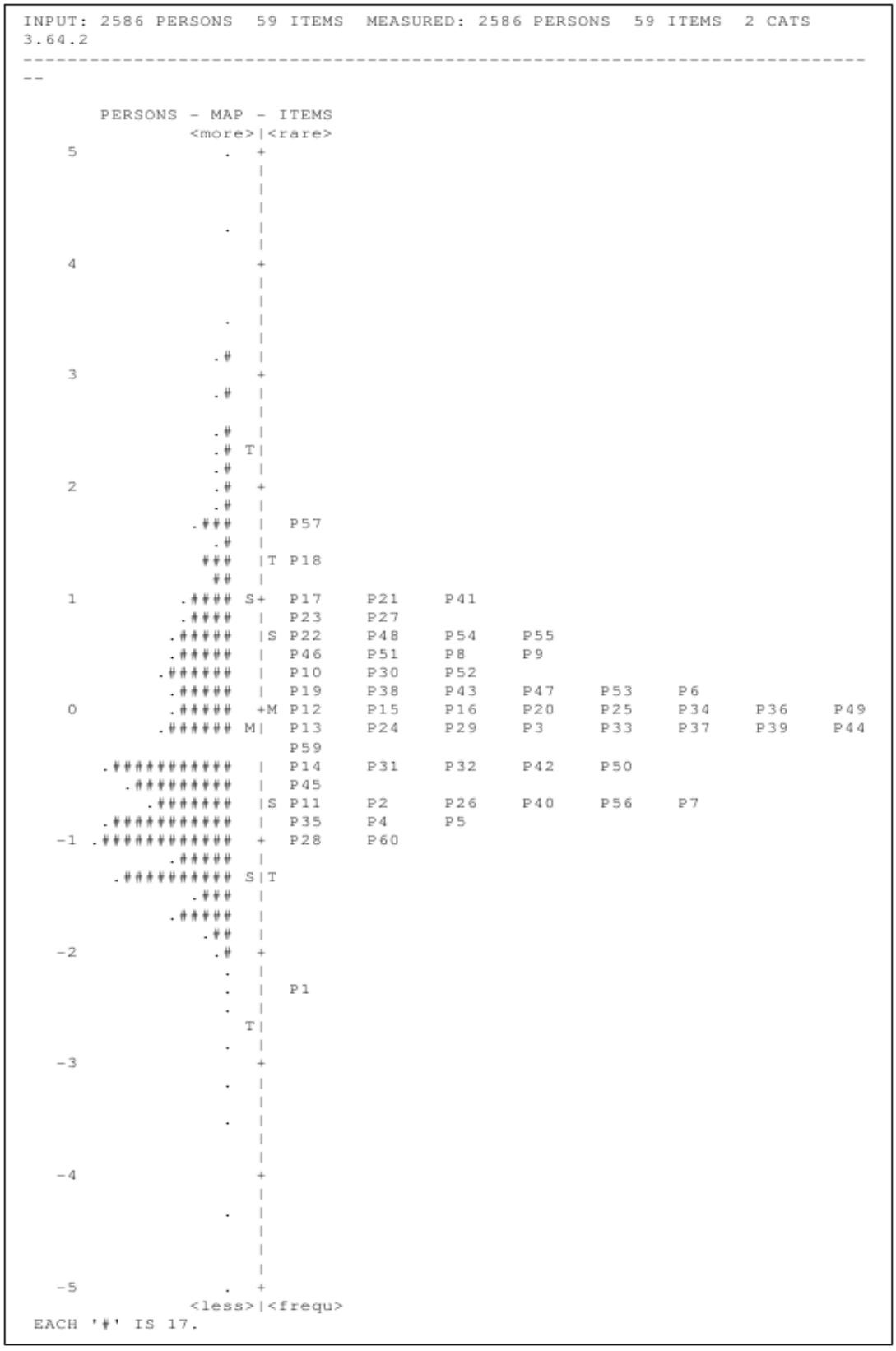

Figura 1.2 Distribución conjuntade examinados e ítems

La mayoría de los ítems se encuentran concentrados en niveles intermedios de la habilidad, casi no hay ítems que brinden información en niveles altos o bajos, es de esperar que si se trata de una prueba de diagnóstico, ésta cuente con ítems de buena calidad técnica en todos los niveles de la habilidad, en especial, se debe recordar que uno de los objetivos de esta prueba es poder ubicar a los estudiantes de acuerdo con su desempeño y dependiendo de éste, se les da la recomendación de realizar una nivelación en conocimientos básicos de matemática, la cual puede ser llevar un taller de un mes o un curso de un semestre, por lo tanto, se hace necesario poder contar con suficientes ítems que brinden información en niveles bajos de la habilidad que se está midiendo.

1.6 Establecimiento de niveles de desempeño sustantivo

Para dar una mejor recomendación a los estudiantes de si llevar un taller de un mes o un curso de un semestre es necesario contar con más evidencias científicas que brinden información de acuerdo con el desempeño en la prueba y considerando los procedimientos involucrados en la resolución de los ítems.

Por otro lado, hasta el momento la construcción de ítems para la prueba se ha orientado más que nada en la intuición y experiencia de los expertos en los temas evaluados, pero no existe una guía que describa las características que deben considerarse en la construcción del tipo de ítems que se requieren, ni existe la forma de valorar si quienes tienen a cargo esta labor son conocedores del constructo y de la población meta. Para ejemplificar cómo realizar una evaluación de jueces expertos aplicando el modelo de Rasch, se trabajó sobre las estimaciones de dificultad y descripción de contenidos y procesos de los ítems según 5 jueces. Se le solicitó a cada juez que clasificara cada ítem de acuerdo con: a) su dificultad, en muy fácil, fácil, mediano, difícil y muy difícil; la estimación la debían realizar pensando en un estudiante promedio de 11Âº año de un colegio público de nuestro país; b) su contenido y procesos presentes en su solución.

Es importante recordar que los contenidos y procesos propuestos a los jueces fueron una combinación entre los temas y destrezas definidas por los creadores del DiMa y la categorización de procesos mentales propuetos en la prueba de Habilidades Cuantitativas del Proyecto de Pruebas Específicas de la UCR. De manera resumida, los contenidos eran: (C1) Operatoria con números reales, (C2) Algebra, (C3) Función exponencial y función logarítmica, (C4) Funciones algebraicas, (C5)Trigonometría. En cuanto a los procesos que los jueces debían indicar si estaban presentes o al menos eran los más representativos en la solución del ítem, una breve descripción fue la siguiente:

Pr1 Procesos aritméticos: cálculos sencillos con operaciones aritméticas. Se pide explícitamente que se efectúe la operación, sin relacionar la operación con otros conceptos y usando la notación y vocabulario que es usual en los materiales didácticos de mayor difusión.

Pr2 Procesos comparativos: agrupar, comparar, discriminar, relacionar. Incluye ítems que requieren operaciones del tipo anterior, pero que se presentan en conjunto con otro concepto, expuesto o no explícitamente.

Pr3 Procesos algebraicos: aplicación de leyes, sustitución, aplicación de fórmulas, despejar variables. Se presentan en forma explícita, leyes de potencia y radicales, propiedades de logaritmos o de la función exponencial, identidades trigonométricas, o algún teorema o definicición básica, para que se reconozca su validez.

Pr4 Interpretación: interpretar, traducción de lenguaje verbal al algebraico. Capacidad para plantear ecuaciones, o trasladar una ecuación a una forma equivalente. Leer y poder entender la definición de un concepto nuevo.

Pr5 Visualización espacial: extraer información de un dibujo. En algunas ocasiones, dada la descripción algebraica de una función se requiere que se visualice el gráfico de la función, para reconocer algunos elementos de las funciones. En otros casos, dado el gráfico de una función en una variable real, se requiere reconocer los

elementos de la función (el ámbito, los intervalos de monotonía, entre otros).

Pr6 Proceso deductivo: de una ley general, se infieren afirmaciones para casos particulares, es decir, de una generalidad se pasa a particularidades.

Pr7 Proceso Inductivo: obtener conclusiones generales a partir de premisas que contienen datos particulares.

Pr8 Uso de hipótesis: se requiere asumir ciertas hipótesis para llegar a una solución.

Pr9 Razonamiento en contexto negativo (uso de negación en el planteamiento del ítem).

Uno de los objetivos del trabajo era determinar distintos niveles de desempeño en el constructo basados en los resultados obtenidos con el modelo de Rasch, como en el caso del DiMa 2008, los resultados obtenidos al aplicar Rasch muestran que no hay suficientes ítems en niveles muy altos, ni en bajos de la habilidad que puedan dar información, por lo que para poder continuar con el estudio, fue necesario considerar aquellos ítems ubicados a una distancia mayor a 0.55 logits por debajo de la media en dificultad de los ítems, como ítems relativamente fáciles (Nivel 1, 12 de 59 ítems), y aquelleste proceso Pr6 os ubicados a una distancia mayor a 0.55 logits por encima de esta media como ítems relativamente difíciles (Nivel 3, 11 de 59 ítems), el resto de ítems (Nivel 2, 36 ítems) se consideraron de dificultad media. Es necesario recordar que en el mapa de distribución conjunta de examinados e ítems, se ubican los ítems de abajo hacia arriba del más fácil al más difícil, de acuerdo con las estimaciones de dificultad obtenidas en el modelo de Rasch.

La estimación de dificultad de los ítems emanada por cada juez, se trasladó al mapa obtenido en Rasch. Por ejemplo, en la figura 1.3 se muestra la representación de las valoraciones de uno de los jueces. Se dibujaron líneas para identificar los tres niveles que se establecieron según las dificultades obtenidas en Rasch (la parte superior representa el nivel 3, la que está en medio de las líneas el nivel 2, y la parte inferior el nivel 1) y se identificó con color rojo, aquellos ítems clasificados como fáciles por el juez, con color verde los clasificados como medianos y con color azul los difíciles.

En la tabla 1.4 se indica la cantidad de ítems en los que cada uno de los jueces concuerda con la estimación de dificultad obtenida con el modelo de Rasch, se puede apreciar que el juez 5 es el que coincide más con lo estimado por el modelo (51% de acierto), mientras que el juez 1 es el que más se aleja a la clasificación obtenida con el modelo de Rasch.

Jueces	Niveles de Dificultad según Rasch			Total ítems concordancia Rasch-Juez	
	Nivel 1 12 ítems	Nivel 2 36 ítems	Nivel 3 11 ítems		
J1	11	9	2	22	37%
J2	6	19	3	28	47%
J3	3	16	7	26	44%
J4	5	17	5	27	46%
J5	7	21	2	30	51%

Tabla 1.4 Concordancia en dificultad, Rasch-Jueces

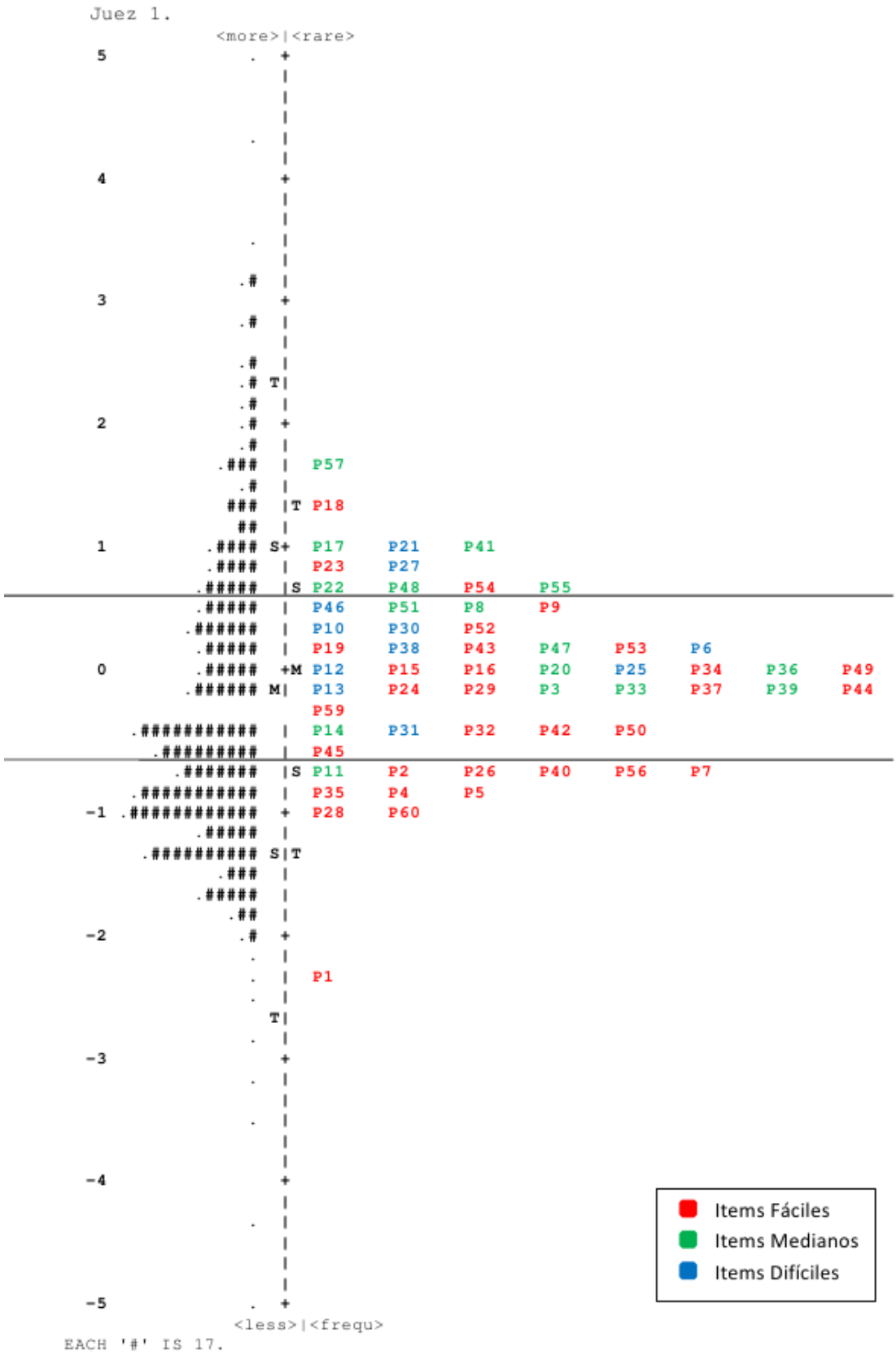

Figura 1.3 Clasificación del ítem por dificultad Juez 1 vrs Rasch

Aquí podemos valorar la importancia de uno de los usos del modelo de Rasch, pues vemos como se puede implementar en la evaluación de los jueces, se debe recordar que los resultados procedentes del modelo son los obtenidos a partir de los datos observados en la muestra analizada, mientras que los resultados de los jueces son tan solo estimaciones, valoraciones a su criterio. Es importante resaltar que la aplicación del modelo de Rasch permite de esta manera identificar aquellos jueces que parecen conocer más la población a evaluar y el constructo que se está pretendiendo medir.

Con el propósito de poder contar con evidencia estadística sobre los niveles de acuerdo entre los jueces según la dificultad de los ítems, se aplicó el índice Kappa, el cual brinda una medida de la concordancia entre dos jueces, al clasificar de forma individual un conjunto de ítems en un mismo conjunto de categorías; el valor de este índice se considera moderado si es superior a 0.41 (Landis y Koch, 1977). En el caso de este estudio los resultados obtenidos muestran que solo las parejas de jueces J2-J4 y J3-J4, se acercan a un valor moderado en este índice, 0.397 y 0.434 respectivamente, además, con el juez 1 se obtienen los valores más bajos al compararlo con los jueces 2, 3 y 4.

Para poder analizar cuales son los procesos más representativos en cada uno de los ítems, se requiere identificar en cuales de ellos existe más concordancia entre jueces y entre jueces y Rasch. Por los resultados obtenidos, se decide estudiar 3 subgrupos de jueces:

- Grupo 1, el formado por los jueces J2, J3, J4 y J5, sin considerar J1 porque cuando se estudió la concordancia de cada juez con el modelo Rasch, fue el que más se alejaba de lo estimado por el modelo.

- Grupo 2, el formado por J2, J4 y J5, porque cuando se estudiaron en forma individual, fueron los que se acercaron más a las dificultades estimadas con el modelo de Rasch.

- Grupo 3, el formado por J2, J3 y J4, pues cuando se hizo el cálculo de los índices de concordancia entre jueces (índice Kappa) según la dificultad del ítem, fue con estos jueces con los que se obtuvieron valores moderados.

El siguiente paso era identificar, para cada grupo, en cuales ítems había más concordancia y en cuales no se llegaba a un acuerdo entre jueces. Para esto, se procedió, en cada grupo, a medir la variabilidad entre jueces, se consideraron estos como sujetos y los ítems como variables, donde los datos correspondían a 1 si el juez lo clasificó como fácil, 2 si lo calificó como de mediana dificultad y 3 si lo consideró difícil, se calcularon los estadísticos descriptivos media y desviación estándar; así, si el ítem se ubicaba a menos de 0.6 desviaciones estándar de la media respectiva, se consideraba que había consenso entre las respuestas dadas por los jueces del respectivo grupo porque existía poca variabilidad. Si existía concordancia, en términos de las calificaciones hechas por los jueces en la dificultad, entonces se consideró, para el grupo 1, la clasificación de dificultad, según lo indicado por al menos 3 de los 4 jueces; para los grupos 2 y 3, la clasificación de dificultad, según lo indicado por 2 de los 3 jueces.

Una vez clasificados los ítems de acuerdo con el consenso entre jueces según su dificultad, se trasladó esta información al mapa de distribución conjunta entre examinados e ítems obtenido en Rasch, volviendo a identificar con líneas los tres niveles establecidos según Rasch, como se hizo por ejemplo en la figura 1.3, e identificando nuevamente con color rojo, aquellos clasificados como fáciles por el grupo, con color verde los medianos, con color azul los difíciles y los de color negro son los ítems en los que no se logró un consenso en dificultad en el grupo. En la figura 1.4 se puede observar la representación del grupo 3. De los tres grupos, justamente el 3, conformado por los jueces J2, J3 y J4, fue el que más se acercó a lo estimado con el modelo de Rasch (53% de acierto) según la dificultad de los ítems. Por todo lo anterior, para el estudio de los contenidos y procesos presentes en los ítems se decidió trabajar con lo propuesto por éste grupo.

Con el propósito de poder contar con evidencia estadística sobre los niveles de acuerdo entre los jueces según el contenido en que clasificaron los ítems, se aplicó el índice de Kappa para los jueces del grupo 3, los valores obtenidos oscilaron entre 0.821 y 0.860, los cuales son considerados muy buenos, es decir, existe bastante acuerdo entre estos jueces con respecto a la categoría de contenido en la que clasificaron los ítems.

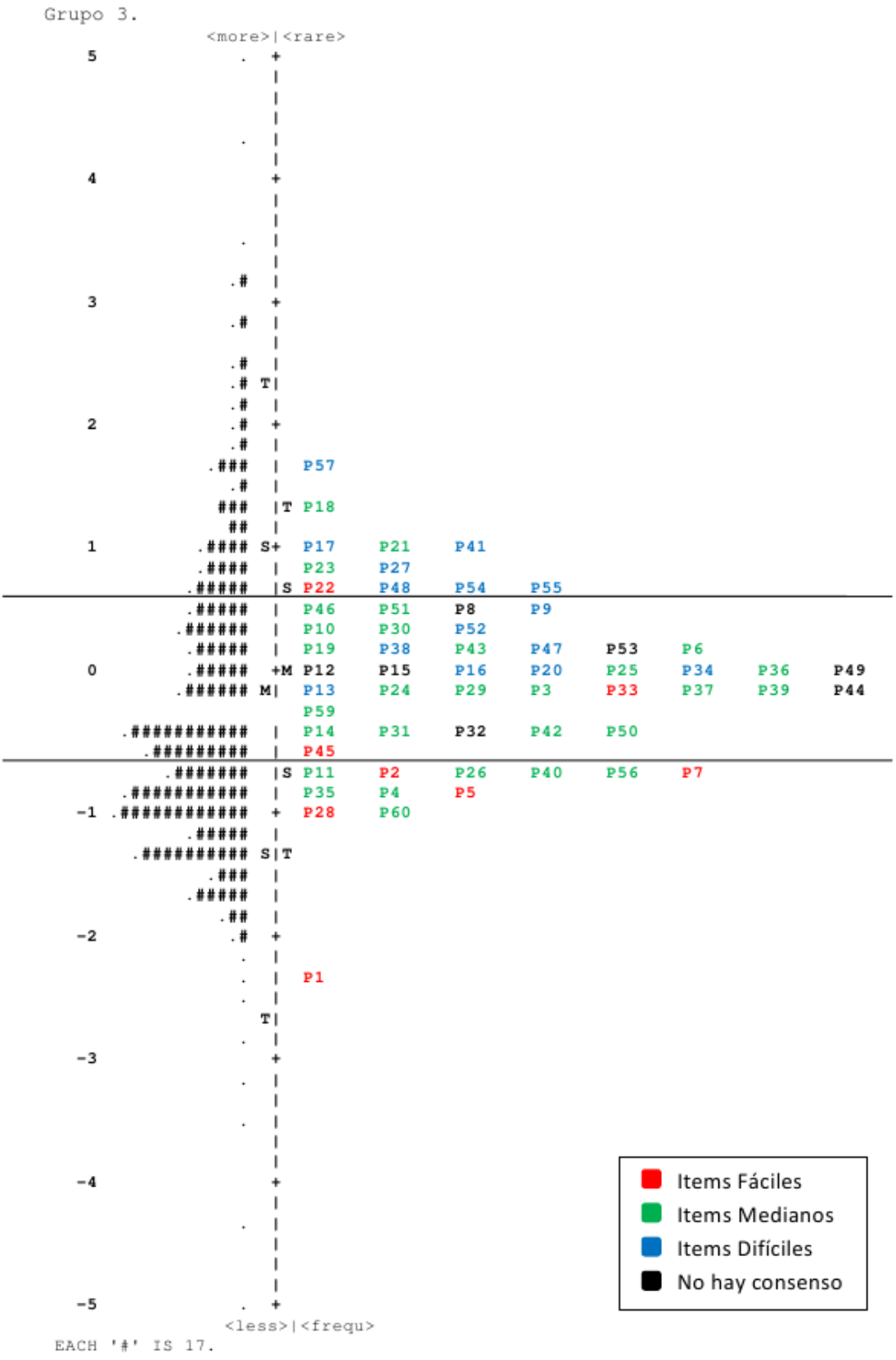

Figura 1.4 Clasificación del ítem por dificultad, Grupo 3 vrs Rasch

Las tablas 1.5, 1.6 y 1.7 son una primera aproximación a lo que se le puede llamar una tabla de especificaciones para la construcción de ítems que pueden formar parte de esta prueba, además, dan una idea de los distintos niveles de desempeño o competencia en el constructo medido, con lo cual se le podría indicar al examinado en el informe que se le da no solo en cuales contenidos debe fortalecerse sino que también a cuales procesos debe prestarle más atención. En el mapa de la distribución conjunta entre ítems y examinados, se puede observar, de acuerdo con la posición del examinado, cuales ítems representan un grado de mayor dificultad para la persona, y teniendo ubicados estos ítems en los distintos niveles de dificultad, se puede valorar los contenidos y procesos presentes en ese nivel.

En estas tablas se reunió y organizó la información más relevante de los ítems: dificultad obtenida según Rasch, contenido en el que quedarían clasificados y los procesos más representativos en la resolución de cada ítem, según la coincidencia de al menos 2 de los 3 jueces. Para este grupo de jueces, los procesos Pr7, Pr8 y Pr9 no están presentes en ninguno de los ítems analizados, por lo que no se consideraron en las tablas.

Items ubicados en el Nivel 1 según Rasch

Item	Indice de Dificultad	Conte-nido	Pr1	Pr2	Pr3	Pr4	Pr5	Pr6
p11	-0.62	C2			1			
p2	-0.68	C1	1					
p40	-0.71	C4				1	1	1
p26	-0.72	C2		1	1			
p56	-0.74	C5					1	
p7	-0.75	*	1		1			
p5	-0.79	C1	1					
p35	-0.81	C4					1	
p4	-0.82	C1	1					
p60	-0.96	C5			1	1	1	
p28	-1.09	C2		1	1			
p1	-2.22	C1	1					
		Total	5	2	5	2	4	1
		Porcentaje	**42%**	17%	**42%**	17%	33%	8%

* En este ítem no hubo consenso entre jueces.

Tabla 1.5 Tabla de especificaciones. Nivel 1.

En la tabla 1.6 se puede apreciar que para los ítems que pertenecen al nivel 1, según el criterio de estos jueces los procesos más representativos en su solución son el Pr1 (procesos aritméticos) y el Pr3 (procesos algebraicos), lo cual indicaría que si se desea confeccionar ítems que midan en niveles bajos de este constructo, se requiere que estén presentes estos procesos.

Items ubicados en el Nivel 2 según Rasch

Item	Indice de Dificultad	Conte-nido	Pr1	Pr2	Pr3	Pr4	Pr5	Pr6
p8	0.47	C2	1	1	1			
p46	0.43	C4		1	1			1
p51	0.41	C3			1	1		1
p9	0.41	C2		1	1			1
p52	0.37	C3		1	1	1		
p10	0.3	C1		1	1			1
p30	0.28	C4		1		1		
p43	0.18	C4				1	1	
p47	0.16	C3			1	1		
p6	0.12	C1		1	1			
p38	0.12	C4			1	1		
p19	0.1	C2			1			1
p53	0.09	C3		1	1	1		
p25	0.01	C2		1	1			
p20	0.01	C2		1	1			1
p12	0,00	C2		1	1	1		1
p15	-0.02	C2		1	1			
p16	-0.02	C2		1	1			1
p36	-0.03	C4			1	1		
p34	-0.06	C4			1			1
p49	-0.09	C3		1			1	
p29	-0.12	C4					1	
p39	-0.13	C2			1	1		
p24	-0.14	C2		1	1			
p44	-0.18	C4		1		1		
p59	-0.2	C5		1	1	1		1
p3	-0.23	C1	1					
p13	-0.23	C2		1	1	1		1
p33	-0.25	C4				1	1	
p37	-0.27	C4		1	1	1		
p50	-0.3	C3			1	1		1
p14	-0.32	C3		1	1			1
p31	-0.33	C4			1	1		
p32	-0.41	C4			1	1		
p42	-0.45	C2	1		1	1		
p45	-0.48	C4		1	1	1		
		Total	3	21	29	20	4	13
		Porcentaje	8%	**58%**	**81%**	**56%**	11%	**36%**

Tabla 1.6 Tabla de especificaciones. Nivel 2.

En el caso de los ítems del nivel 2 (ver tabla 1.6), los procesos más relevantes serían Pr2 (procesos comparativos), Pr3 (procesos algebraicos) y Pr4 (interpretación), es decir, pareciere que para este tipo de ítems, además de requerir un dominio en los procesos presentes en los del nivel 1, se requiere también realizar comparaciones, agrupaciones interpretaciones de definiciones de conceptos nuevos, traducir del lenguaje verbal al algebraico y viceversa, saber como plantear ecuaciones.

Items ubicados en el Nivel 3 según Rasch								
Item	Indice de Dificultad	Conte-nido	Pr1	Pr2	Pr3	Pr4	Pr5	Pr6
p57	1.67	C5		1		1		1
p18	1.29	C2			1			
p17	1.03	C2		1	1			1
p41	0.97	C2		1	1	1		
p21	0.96	C2			1			1
p27	0.78	C2	1	1	1			1
p23	0.72	C2		1	1	1		
p54	0.61	C5		1	1	1	1	
p55	0.6	C5		1	1	1	1	
p22	0.57	C2	1		1			
p48	0.55	C3			1	1		
		Total	2	7	10	6	2	4
		Porcentaje	18%	**64%**	**91%**	**55%**	18%	**36%**

Tabla 1.7 Tabla de especificaciones. Nivel 3.

Finalmente, en la tabla 1.7 se muestran la clasificación en contenidos y procesos realizada por los jueces del grupo 3, para los ítems del nivel 3 (ítems relativamente difíciles), aquí se puede apreciar de manera muy general que los procesos más representativos en este caso son Pr2 (procesos comparativos), Pr3 (procesos algebraicos) y Pr4 (interpretación), lo cual no marcaría una diferencia con los ítems del nivel 2, y esto era de esperar, pues hay que recordar que en realidad no se trata de ítems difíciles y que se escogieron los que se consideraban relativamente difíciles con el propósito de poder ilustrar uno de los usos del modelo de Rasch. Lo que sí se puede apreciar en la tabla 9 es como para aquellos ítems que tienden a ser los más difíciles en este nivel, el proceso Pr6 (proceso deductivo) tiende a estar más presente, una hipótesis podría ser entonces que, para contar con ítems que midan en niveles altos de la habilidad, se requiere la presencia de este proceso Pr6.

1.7 Conclusiones y recomendaciones

Al aplicar el modelo de Rasch, se obtiene que la medida de confiabilidad de los examinados y la correspondiente a los ítems resultaron bastante consistentes y existe una medición bastante precisa en cuanto a la dificultad de los ítems.

En los mapas de las distribuciones conjuntas de los individuos y los ítems, se puede observar que el promedio del nivel de habilidad de los examinados está muy cercano (por debajo) a la dificultad promedio de los ítems, la mayor parte de la población se ubicó por debajo de la dificultad promedio de los ítems, esto indica que la prueba resultó levemente difícil para los examinados. Pero, la gran mayoría de los ítems se encuentran concentrados en niveles intermedios de la habilidad, casi no hay ítems que brinden información en niveles altos o en niveles bajos, al tratarse de una prueba de diagnóstico es deseable disponer de ítems con niveles óptimos, en cuanto a calidad técnica, en todos los niveles de la habilidad; en especial para el caso de esta prueba, se requieren más ítems en niveles bajos de la habilidad, ya que uno de los objetivos del DiMa es poder hacer recomendaciones a los estudiantes que no

obtuvieron un buen desempeño, para lograr una nivelación.

Además, con el uso del modelo de Rasch, al poder tener a los evaluados e ítems representados en un mismo continuo, se pudo analizar las interacciones entre éstos y las interpretaciones de las puntuaciones se pudieron hacer identificando los ítems que el examinado tiene mayor o menor probabilidad de acertar. Según la TCT, si dos individuos poseen la misma calificación, ésta lo que indica es que ambos acertaron la misma cantidad de ítems, pero de acuerdo con Rasch, podemos identificar si uno de ellos acertó mayor cantidad de ítems de niveles altos en el constructo.

El utilizar el modelo de Rasch permitió hacer una evaluación de 5 jueces quienes analizaron y clasificaron los ítems, según su dificultad, contenido y procesos presentes en su solución. Con los resultados obtenidos fue posible elaborar tablas con la información más relevane de los ítems del DiMa 2008: dificultad del ítem, según Rasch; contenido, procesos más representativos presentes en la solución del ítem. Al poder identificar los jueces que más concuerdan con los resultados del modelo de Rasch, se puede hacer una mejor escogencia de evaluadores de ítems que conocen mejor el constructo, y que en el momento de hacer la valoración de ítems experimentales, serán más certeros.

Este estudio realizado con los jueces y el uso del modelo de Rasch, constituye una primera aproximación para la construcción de una tabla de especificaciones tan necesaria para el trabajo de construcción de ítems, sin embargo, uno de los problemas mayores en este estudio fue la ubicación de la mayoría de los ítems en niveles intermedios, por tanto, para poder generar los tres niveles de desempeño fue necesario realizar ajustes ad hoc, los cuales no serían necesarios si se contaran con suficientes ítems en los tres niveles estándar establecidos en Rasch (entre -3 y -1 fáciles, entre -1 y 1 medianos, entre 1 y 3 difíciles), por lo cual, más allá de los resultados obtenidos, su mayor valor está en el aporte metodológico, procedimental y de interpretación desarrollado.

Es importante continuar con este análisis, pero estudiando con más detalle los procesos involucrados en la solución de los ítems, desde la psicología cognitiva, para lograr identificar qué procesos deparan diferentes niveles de dificultad en los ítems y poder identificar las características que comparten los ítems de un nivel de dificultad similar, así, se le podría indicar a un constructor de ítems los procesos que deben estar presentes en un ítem para que resulte con el nivel de dificultad que se desea y elaborar pruebas más adaptadas a las necesidades; pues con Rasch se sabe que no se asume el supuesto de que la prueba mide con la misma confiabilidad siempre, sino que si se tienen ítems fáciles, se sabe que los parámetros de los sujetos de niveles bajos en la habilidad se estimarán con mayor precisión, o si se cuenta con examinados ubicados en niveles altos de la habilidad, con estos se podrá estimar los parámetros de los ítems difíciles con mayor precisión.
Por todo lo expuesto anteriormente, se puede afirmar que el aplicar este enfoque psicométrico, que incluye el modelo de Rasch, en el proceso de construcción, juzgamiento y calificación de los ítems y de la prueba, permite dar un soporte más científico al DiMa y por tanto, aporta más evidencias de validez, que deben ser consideradas en las inferencias que se derivan de los resultados, a partir del uso de la prueba, para la toma de desiciones.

Bibliografía

[1] AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education). (1999). *The Standards for Educational and Psychological Testing.* Washington: AERA (American Educational Research Association).

[2] Arce C. y Jiménez L. (2003). *Diagnóstico de conocimientos y destrezas en matemática del estudiante al ingresar a la Universidad*. Propuesta de Investigación. Sistema de formulación de proyectos 2003-2004 UCR.

[3] Babbie, E. (2010). *The Practice of Social Research*. Belmont, California: Wadsworth.

[4] Barbero, M. (1999). Desarrollos recientes de los modelos psicométricos de la teoría de respuesta a los ítems. *Psicothema*, *11*(1), 195-210. Recuperado de http://www.doredin.mec.es/documentos/017199930016.pdf

[5] Bond, T. & Fox, C. (2001). *Applying the Rasch model: fundamental measurement in the human Sciences.* Mahwah, New Jersey: LEA.

[6] Cea D'Ancona, M. (2002). *Análisis multivariable*. España: Editorial Síntesis, S.A.

[7] Landis J. R., Koch G. G. (1977). The measurement of observer agreement for categorical data.*Biometrics*, 33, 159-174.

[8] Martínez, M. R. (2005). *Psicometría: Teoría de los Test Psicológicos y Educativos*. Madrid: Editorial SINTESIS, S.A.

[9] Martínez, M. R., Hernández M.J. & Hernández, M.V. (2006). *Psicometría*. Madrid: Alianza Editorial.

[10] Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.

[11] Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

[12] Montero, E. (2001). La teoría de respuesta a los ítems: una moderna alternativa para el análisis psicométricos de instrumentos de medición. *Revista de Matemática: teoría y aplicaciones*. Centro de Investigaciones en matemática pura y aplicada (CIMPA) y la Escuela de Matemática de la Universidad de Costa Rica. 7(1-2), 217-228.

[13] Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los ítems*. Madrid: Ediciones Pirámide, S.A.

[14] Muñiz. J. (2003). *Teoría Clásica de los Tests*. Madrid: Ediciones Pirámide, S.A.

[15] Nunnally, J.C. & Bernstein, I.J. (1995). *Teoría psicométrica* (3ra ed). México, D.F.: Editorial McGrawHill Latinoamericana.

[16] Padilla J.P. et al (2006). La evaluación de las consecuencias del uso de los tests en la teoría de validez. *Psicothema*, 18(2), 307-312.

[17] Prieto, G. & Delgado A.R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, 15(1), 94-100.

[18] Prieto, G. y Delgado, A. (2010). Fiabilidad y Validez.*Papeles del Psicólogo*, *31*(1), 67-74. Recuperado de http://www.papelesdelpsicologo.es/pdf/1797.pdf