¿Cómo utilizar la Teoría de Medida en la construcción de pruebas físicas?[1]

Antonio Hernández Mendo

Universidad de Málaga

La Teoría de Medida engloba a su vez tres grandes teorías la Teoría Clásica de los Test (TCT), la Teoría de Respuesta al Ítem (TRI) y la Teoría de la Generalizabilidad (TG). En esta conferencia el objetivo será poner de manifiesto el uso de estas tres teoría para la construcción de pruebas físicas. La Teoría de Respuesta al Ítem (TRI) es uno de los campos con mayor proyección dentro del ámbito de la medida psicológica. Lord (1980) señala que la Teoría de Respuesta al Ítem (TRI) no contradice las asunciones fundamentales de la Teoría Clásica de los Test (TCT) sino que hace asunciones adicionales que permitirán responder a las cuestiones que la TCT no lo hacía adecuadamente. Aunque la TRI se muestra como un método capaz de enfrentarse a estas deficiencias planteadas en la TCT, sigue siendo frecuente el uso de la TCT debido a su facilidad conceptual y a la sencillez del cálculo (Hamblenton y Jones, 1993). La TCT y la TRI son modelos que teóricamente se solapan y permiten una mejor compresión del funcionamiento del test más que competir entre ellos (Drasgow y Parsons, 1983). La TRI hace suposiciones de mayor potencia estadística que la TCT, en particular la independencia local y las relaciones logísticas entre las respuestas de los ítems y los rasgos subyacentes (Hambleton, Swaminathan, Rogers, 1991).

Los modelos de la TRI tienen ventajas significativas sobre los modelos de la TCT, particularmente cuando estos tienen en cuenta el sesgo del test, haciendo diferencias en cuanto al género o la raza (Hambleton, 1989; Lord, 1980).

Los conceptos claves de la TCT incluyen la dificultad del ítem (proporción de participantes con una puntuación positiva o acertada), la discriminación del ítem (la correlación del ítem con el resto del test), la fiabilidad alfa, y los cortes óptimos; todos ellos dependientes de las características de la muestra (Hambleton, Clauser, Mazor, Jones, 1993).

---

[1] Este texto pertenece al artículo de Hernández Mendo, A. (2006). Un cuestionario para la evaluación psicológica de la ejecución deportiva: estudio complementario entre TCT y TRI. *Psicología del Deporte,* 15(1), 71-93. http://psicologia.del.deporte.uma.es/archivos/11CuestiEval_Psic_EjecDeportiva.pdf

El principal problema planteado en la TCT es el relativo a la invarianza de la medida. Thurstone (1928) afirma que las mediciones de un instrumento deben ser independientes de los objetos medidos. Este inconveniente queda patente en dos problemas concretos (Bejar, 1983; Hamblenton y Swaminathan, 1985 y Muñiz, 1997): (1) la medición de las variables psicológicas no son independientes del instrumento que se utiliza para medirla; y, (2) las propiedades de los instrumentos no son independientes de los sujetos a los que se aplican. Estas cuestiones son importantes cuando se pretende establecer equivalencias entre las puntuaciones de dos test diferentes que midan una misma variable. La TCT considera los casos de un test como una muestra representativa de un universo de ítems equivalentes que permiten ser considerados indicadores similares del constructo que medimos, de ahí que se pueda utilizar como procedimiento la acumulación de puntos, lo que lleva a otra limitación de esta teoría, ya que una misma puntuación en un test puede deberse a distintos patrones de respuesta, esto impide analizar las interacciones entre los sujetos y los ítems. Además, suponer que todos los ítems son equivalentes, implica que todos los sujetos utilizan las mismas operaciones mentales, esto implica que no se tienen en cuenta las diferencias individuales ni la diferencias de dificultad de los ítems. Una limitación más de la TCT es la relativa a la fiabilidad del instrumento de medida, según esta teoría, la fiabilidad se reparte por igual a lo largo del test, esto es inverificable y no se ajusta a la realidad.

Todas estas limitaciones y problemas impulsaron el surgimiento de nuevos modelos, algunos de ellos no eran más que extensiones del modelo lineal de Spearman asumido en la TCT y otros surgen enmarcados dentro de un nuevo marco teórico, entre los que destaca la Teoría de Respuesta al Ítem (TRI), que permitirá solventar las limitaciones de la Teoría Clásica de los Test (Bejar, 1983; Hamblenton y Van der Linden, 1982; Martínez Arias, 1995 y Muñiz, 1996). Aunque esta teoría no es reciente, su expansión se produce a partir de los años ochenta con la difusión de los ordenadores, una herramienta que será imprescindible debido a la complejidad de los cálculos matemáticos.

La TRI tiene como objetivo obtener mediciones que no varíen en función del instrumento utilizado, disponer de instrumentos de medida que no dependen de los objetos medidos, es decir, que sean invariantes respecto a los sujetos evaluados y avances técnicos como funciones de información de los ítems y del test, errores típicos

de medida diferentes para cada nivel de la variable medida y el establecimiento de bancos de ítems con parámetros estrictamente definidos.

Los modelos basados en la TRI relacionan a sujetos e ítems de modo interactivo lo que permite localizar al mismo tiempo en un continuo psicológico que representa a la variable a sujetos e ítems, el proceso de medición se puede representar como la localización de personas e ítems en un mismo contínuo (Wright y Stone, 1979; Wright y Master, 1982). Así, la posición de las personas dependerá de sus respuestas a los ítems del test, del mismo modo los ítems tendrán distintas localizaciones dependiendo de su nivel de dificultad. El concepto básico de la TRI es la Curva Característica del Ítem (CCI), que es la función matemática que relaciona la probabilidad de acertar el ítem con la competencia del sujeto, (P (θ) donde θ es el nivel de competencia). Se denomina CCI porque cada ítem se caracteriza por su curva.

La principal diferencia con la TCT la encontramos en esta curva, mientras que la TRI va a centrarse en las propiedades particulares de cada ítem, la TCT se dirige a las propiedades de la puntuación global en un test (X). La CCI es la probabilidad de acertar un ítem que solo depende de los valores de la variable medida por el ítem, de modo que los sujetos con distinta puntuación en la variable tendrán distintas probabilidades de superar un determinado ítem (Muñiz, 1997). En el eje de abscisas se representan los valores de la variable que mide el ítem y el eje de ordenadas la probabilidad de acertar el ítem (P (θ)), de modo que la (P(θ)) variará dependiendo de los valores de (θ) (nivel de competencia). A mayor nivel de competencia mayor probabilidad de acertar el ítem. Análisis pormenorizados de estas curvas pueden ayudar a (1) en la eliminación de ítems que no proporcionan información significativa sobre el rasgo de interés; (2) en la selección de ítems que proporcionen la máxima discriminación en el rasgo; y (3) se pueden usar para identificar ítems prejuiciosos o, en terminología TRI, funcionamiento diferencial del ítem (DIF). La función diferencial del ítem se produce cuando el ítem en cuestión es más discriminativo, es más difícil o es más extremo en un grupo donde es comparado con otros ítems. Considerar cuidadosamente el trazo de las curvas puede ayudar detectar prejuicios raciales, de género, etc., en un test.

Es importante diferenciar la CCI de la regresión ítem-test (que consiste en hacer corresponder las puntuaciones del test con las proporciones de aciertos en un determinado ítem). Muñiz (1997) considera que la principal diferencia de la CCI y la

regresión ítem-test estriba en que la CCI es la variable que miden los ítems (θ), no es la puntuación que obtienen los sujetos en ese test, aunque existe una relación, se podría decir que estas puntuaciones son una estimación del nivel de competencia pero no constituyen la escala θ. La CCI nunca podría ser una recta porque esto implicaría que para determinados valores del nivel de competencia (θ) existe una (P(θ)) negativa o mayor a 1, lo que es incompatible con los axiomas de la probabilidad, que establece su valor entre 0 y 1. De igual modo no podría tener una de ángulo recto porque los cambios en los seres humanos no se producen de forma brusca en un punto concreto. De modo, que generalmente estas curvas adoptan forma de S, que se puede interpretar como la gradación en el cambio del fallo al acierto, esta curva se define a partir de tres parámetros: 1) parámetro "*a"* o índice de discriminación, 2) parámetro *"b"* o índice de dificultad y 3) parámetro *"c"* o probabilidad de acertar un ítem al azar. Los distintos modelos de CCI están en función de los valores que adopten estos tres parámetros, adoptando una determinada función matemática para cada curva. A medida que se ubican más a la derecha en el eje de las abscisas significa que los ítems son más difíciles, ya que *"b"* experimenta un incremento.

Otro supuesto que la TRI asume implícitamente en su formulación es la unidimensionalidad. Si el modelo es correcto, la probabilidad de acertar un ítem únicamente dependerá del nivel de competencia del sujeto, es decir, al tener los ítems y los sujetos valor en una única dimensión la respuesta a los ítem esta determinada fundamentalmente por el nivel de los sujetos en la variable. Para la comprobación de la unidimensionalidad de los ítems se someten a análisis factorial y se descartan los ítems que conforman factores periféricos, lo mismo se hace en análisis posteriores, hasta lograr un análisis en el que un factor explique la mayor parte de la varianza de los ítems (lo ideal seria encontrar un factor que la explique toda, algo que raramente ocurre). Es importante señalar que el supuesto de unidimensionalidad implica matemáticamente que existe independencia local entre los ítems, es decir, para un sujeto con un determinado nivel en el rasgo (unidimensionalidad) la respuesta a un ítem no está asociada con las respuestas a los demás ítems. Así, puede expresarse la independencia local como la probabilidad de que un sujeto acierte un ítem, siendo ésta igual al producto de las probabilidades de acertar cada uno de ellos.

Según el número de parámetros que se tengan en cuenta se considerará un modelo u otro dentro de la TRI. Actualmente, los modelos más utilizados en la TRI son

el modelo logístico de un parámetro, el logístico de dos parámetros y el logístico de tres parámetros. En el presente articulo se aborda únicamente el modelo logístico de un parámetro. En concreto se trabajará con el modelo de Rasch (1960), este modelo es el mas popular dentro de los modelos de la TRI, debido principalmente a su sencillez.

En 1960 el matemático George Rasch propuso un modelo que permite solventar las deficiencias de la TCT, de modo que se construyeran pruebas más adecuadas y eficientes. Este modelo, conocido como el Modelo de Rasch, se fundamenta en: (1) el atributo que se desea medir puede representarse en una única dimensión donde se sitúan conjuntamente ítems y personas; (2) el cociente entre la probabilidad de la respuesta correcta y la probabilidad de la respuesta incorrecta a un ítem es la función de la diferencia en el atributo en el nivel de la persona y el nivel el ítem. Así, cuando una persona responde a un ítem en su nivel de competencia, tendrá la misma probabilidad de dar una respuesta correcta que incorrecta, por lo que la dificultad del ítem será equivalente al nivel de competencia del sujeto. Del mismo modo, cuando la probabilidad de dar una respuesta correcta es mayor que la de dar una incorrecta la competencia del sujeto será mayor que la requerida por el ítem. (3) El modelo de Rasch es un modelo sencillo y de fácil aplicación, que al representar en una única dimensión a sujetos e ítems, nos permite hallar la dificultad de los ítems y la probabilidad de que estos sean contestados con éxito. La localización del punto 0 de la escala es arbitrario, Rasch suele situar la dificultad media de los ítems en el punto 0, de modo que interpretar los parámetros de los sujetos (nivel de competencia) es bastante sencillo ya que si estos valores son mayores a 0 indican en una alta probabilidad de responder a los ítems de dificultad media.

Los parámetros en el modelo de Rasch se estiman con un procedimiento de máxima verosimilitud, consistente en determinar los parámetros que hacen más probable las respuestas observadas. En la estimación condicional se calcula la probabilidad de las respuestas observadas a los ítems para cada puntuación conjunta de los parámetros de los sujetos (nivel de competencia), asignándole a cada persona el valor del parámetro más probable para su patrón de respuesta. Este valor de estimador de máxima verosimilitud puede ser calculado mediante el uso del programa *Acer ConQuest* (Wu, Adams & Wilson, 1998).

El modelo de Rasch (1960) se caracteriza por: (1) La *medición conjunta*, los parámetros de personas e ítems se expresan en las mismas unidades y se localizan en el

mismo continuo, de lo que se deduce que: a) no todos los ítems estiman la misma proporción del constructo (por lo que, no se mantiene el supuesto de invarianza de los ítems defendida por la TCT) y b) la interpretación de las puntuaciones no se fundamentan en las normas del grupo, sino en la identificación de los ítems que la persona tiene una alta o baja probabilidad de resolver correctamente, así, si los sujetos tienen un nivel alto de competencia, se estimaran con mayor precisión los parámetros de los ítems difíciles. (2) *Objetividad específica* (Rasch, 1977), una medida solo puede ser considerada válida y generalizable si no depende de las condiciones específicas con la que ha sido obtenida. Así, la puntuación de las personas no dependen de los ítems administrados. (3) *Propiedades del intervalo,* a diferencias constantes entre sujetos e ítems le corresponde la misma probabilidad de una respuesta correcta, la métrica intervalar tiene gran importancia por ser condición necesaria para realizar análisis paramétricos (medias, análisis de varianza, regresión, etc.) y por que garantiza la invarianza de las puntuaciones diferenciales a lo largo de un continuo. (4) *Específidad del error típico de medida* que permite cuantificar la cantidad de información con la que se mide en cada punto de la dimensión y seleccionar los ítems que permiten incrementar la información en regiones del atributo previamente especificada. Esto es una diferencia con la TCT supone que los test miden con la misma fiabilidad en todas las regiones de la variable, supuesto que desde otros modelos ha sido rechazado.

Bibliografía

Bejar, I.I. (1983). Introduction to Ítem Response Models and their assumptions. In R.K. Hamblenton (Ed.), *Applications of ítem response theory.* Vancouver: Educational Research Instiotute of British Columbia.

Drasgow, F. y Parsons, Ch. K. (1983). Application of unidimensional item response theory models to mutidimensional data. *Applied Psychological Measurement*, *7* (2), 189-199.

Hamblenton, R.K. y Jones, R.W. (1993). Comparison of classical test theory and ítem response theory and their applications to test development. *Educational Measurement: Issues and Practice,* 12 (3), 38-47.

Hamblenton, R.K. y Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Kluwer Academic Publishers.

Hamblenton, R.K. y van der Linden, W.J. (1982). Advances in item response theory and applications: an introduction. *Applied Psychological Measurement,* 6,(4), 373-378.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In Linn, Robert L (Ed), *Educational measurement* (pp.147-200). New York, NY, England: Macmillan Publishing Co, Inc American Council on Education.

Hambleton, R. K., Clauser, B.E., Mazor, K. M., Jones, R.W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9* (1), 1-18.

Hambleton, R. K., Swaminathan, H., Rogers, HJ.(1991). *Fundamentals of item response theory. Measurement methods for the social sciences series,* Vol. 2. Thousand Oaks, CA, US: Sage Publications, Inc.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: LEA.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: LEA.

Martinez Arias, R. (1995). *Psicometría: teoría de los test psicológicos y educativos.* Madrid: Síntesis.

Muñiz, J. (1996). *Teoría clásica de los test.* Madrid: Pirámide.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems.* Madrid: Pirámide.

Muñiz, J. (1997). *Teoría de respuesta a los ítems.* Madrid: Pirámide.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test.* Copenhague: The Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. En M. Glegvad (De.), *The Danish Yearbook of Philosophy* (pp. 59-94). Copenhagen: Munksgarrd.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology,* 33, 529-554.

Wright, B.D. y Masters, G.N. (1982). *Rating scale analysis.* Chicago: Mesa Press.

Wright, B.D. y Stone, M. H. (1979). *Best test desing.* Chicago: MESA.

Wu, A., Adams, R.J. y Wilson, M. R. (1998). *Acer ConQuest.* Melbourne: Acer Press.