**Modelos de diagnóstico cognitivo**

**Introducción**

El análisis cognitivo de las respuestas a los reactivos de una prueba es un tema fundamental para establecer evidencias del grado de validez de los resultados de una prueba y ha estado en boga desde hace varias décadas (Embretson, 1985; Messick, 1994), pero no es hasta apenas un par de ellas, que en el mundo se ha comenzado a aplicar en diversos escenarios educativos (Embretson, 1994; Snow y Lohman, 1993; Myslevy, 2006). Si comparamos esta metodología con los modelos tradicionales de la teoría de respuesta al reactivo (Rasch, 1960), los modelos de diagnóstico cognitivo se han desarrollado para extraer información mucho más precisa del proceso de respuesta en cada reactivo de una prueba, en comparación con el enfoque tradicional (Jurich y Bradshaw, 2014).

Por ejemplo, cuando se especifica la relación entre los atributos o habilidades cognitivas y los reactivos, los modelos de diagnóstico cognitivo describen el efecto funcional de estos reactivos sobre la inferencia del dominio de las habilidades cognitivas que subyacen al mismo (Ziecky, 2014). Al desarrollar tales modelos, se puede proporcionar información sobre el estado del dominio en múltiples dimensiones o habilidades cognitivas, que pueden ayudar y retroalimentar de manera más fina tanto a los estudiantes, pero también a los docentes ya que podrían elegir programas de capacitación apropiados, en las áreas con más oportunidades de crecimiento pedagógico (Leighton y Gierl, 2012).

Recientemente, existe una presión creciente para que las evaluaciones sean más informativas sobre los procesos cognitivos que se miden en los estudiantes. En particular, se ha generado la necesidad latente de adaptar costosas evaluaciones a gran escala, como aquellas que promueve la Organización para la Cooperación y el Desarrollo Económicos (Hopfenbeck y Maul, 2011; Pepper et al., 2018) o bien el Departamento de Educación de EE. UU. (U.S. Department of Education, 2013) para informar sobre las fortalezas y debilidades cognitivas de los estudiantes.

Estas pruebas están destinadas no solo a determinar los resultados del aprendizaje y las necesidades de los estudiantes, sino también a evaluar los programas de instrucción (efectividad escolar). En otras partes del mundo, el entusiasmo por la responsabilidad educativa y los estándares para garantizar que los estudiantes estén preparados y competitivos para entornos de trabajo basados ​​en el conocimiento, también ha fomentado el apetito por las pruebas informativas a gran escala (OCDE, 2013).

Pero queda una pregunta por responder, ¿cómo estas pruebas a gran escala informan a las partes interesadas? La información que se busca de estas pruebas se refiere esencialmente a las fortalezas y debilidades cognitivas de los estudiantes o sustentantes en su pensamiento y aprendizaje. Es decir, ¿en qué medida los puntajes de las pruebas reflejan ciertas formas de pensamiento y procesos cognitivos de orden superior asociados con el aprendizaje significativo? Por ende, en el mundo se está fortaleciendo la construcción de instrumentos y análisis de los resultados que ofrezcan información de los estudiantes o sustentantes con características tan finamente granuladas que puede retroalimentar de manera formativa al centro escolar, el aula e inclusive al propio individuo (Yamaguchi y Okada, 2018; Zhan, Jiao y Liao, 2017; Zhan, Liao y Bian, 2018).

**Antecedentes y fundamentos teóricos.**

Messick, a finales de los años ochenta (1989), anticipó la importancia de proporcionar información sobre los procesos cognitivos de los estudiantes (en oposición a los comportamientos simplemente basados ​​en el contenido) de los puntajes de las pruebas. Por lo tanto, es sensato proponer inferencias basadas en exámenes sobre los procesos cognitivos de los estudiantes solo para aumentar la probabilidad de proporcionar la instrucción más efectiva y oportuna a los estudiantes de una manera que reduzca el origen de potenciales errores. Con tal información, los docentes, directores e inclusivo los supervisores podrían modificar y adecuar ciertos conceptos erróneos de los estudiantes, el aula, la escuela o zona escolar con fines de mejora educativa.

De acuerdo con Messick (1989), la comprensión del desempeño de un estudiante en términos sustantivos, en función de los procesos mentales que los estudiantes utilizan para responder y/o resolver los reactivos de una prueba, es una característica central de la teoría de validez de constructo. En particular, considera que el enfoque sustantivo en la teoría del constructo tiene un papel definitivo en la especificación del dominio de una prueba. Sin embargo, las inferencias basadas en pruebas con fundamento en la psicología cognitiva del procesamiento de la información no son sencillas. Lo cual implica desarrollar y seguir un programa bastante riguroso de validación de constructo (para más detalles, consultar a Cronbach, 1988, Kane, 2001 y Messick, 1989). El desarrollador de la prueba debe comenzar con una teoría de constructo sumamente bien fundamentada a partir de la cual se generarán y seleccionarán los ítems, y a partir de la cual se realizarán predicciones sobre las puntuaciones obtenidas. Al considerar datos y análisis relevantes para dicho programa de validación de constructo, Messick sugirió incluir: (a) análisis críticos y lógicos para descubrir hipótesis alternativas de interpretación de puntuación, (b) análisis de correlación o covarianza para buscar evidencia convergente y discriminante de los patrones esperados de la teoría de constructo, (c) análisis del proceso para sondear la representación del constructo, (d) análisis de diferencias grupales y cambios a lo largo del tiempo, (e) receptividad de puntajes al tratamiento experimental y manipulación de condiciones, (f) generalización de la interpretación de puntajes a través de contextos y (g) amenazas a la capacidad de permanencia y generalización de las conclusiones de la investigación.

De lo anterior, se puede observar que estos pasos no solo pertenecen a los EDC, sino que también buscan identificar cómo los estudiantes de diferentes capacidades o niveles de logro representan y manipulan mentalmente la información de prueba a lo largo del tiempo, en diferentes contextos y en respuesta a intervenciones de instrucción y manipulación de distintas variables. Sin embargo, estos pasos también requieren un esfuerzo comprometido de tiempo y recursos por parte de los desarrolladores de pruebas, para comprender la psicología al presentar la prueba. Un compromiso con estos siete pasos requiere un cambio radical en cómo se ven, desarrollan e informan los resultados de las pruebas. Se requiere que reconsideremos las pruebas como un esfuerzo concreto y científico en lugar de un proceso circunstancial, donde la evidencia se recopila de manera *post-hoc* para justificar la interpretación de los puntajes de las pruebas (Borsboom, Mellenbergh, y Van Heerden, 2004).

**Modelos de Diagnóstico Cognitivo**

Los Modelos de Diagnóstico Cognitivo (CDM, por sus siglas en inglés) son un método relativamente nuevo en educación, los cuales proporcionan información de diagnóstico detallada sobre el grado de dominio de distintas habilidades cognitivas (Lee y Sawaki, 2009). Nichols (1994) utilizó el término EDC, con el propósito de capturar la integración de la psicología cognitiva en el diseño e interpretación de los resultados de evaluaciones, así como en la elaboración de diagnósticos sobre las características del estudiantado en cuanto a sus procesos cognitivos (Rupp, 2007).

La evaluación diagnóstica cognitiva se ha enfocado a la medición de estructuras específicas de conocimiento y habilidades de procesamiento de estudiantes para brindar información sobre sus fortalezas y áreas de oportunidad. Por ejemplo, para Leighton y Gierl (2007b) las ideas, teorías y métodos de la psicología cognitiva pueden contribuir al avance de la medición educativa, dado que: 1) brindan información del análisis de pruebas ya existentes para así dilucidar sus constructos subyacentes, 2) clarifican las metas de la elaboración de pruebas en términos del conocimiento y las habilidades requeridas, como verdaderos indicadores de dominio y entendimiento, y 3) mejoran las teorías de aptitudes, el logro y los aprendizajes en ámbitos diferentes. De esta manera, los aspectos cognitivos subyacentes a las pruebas en el campo educativo son de gran utilidad en la comprensión de los constructos medidos y pueden ser descritos en términos de las estrategias, los procesos y las representaciones del conocimiento que evidencian los examinados (Gierl, Leighton y Hunka, 2000; Tatsuoka, 2009).

En general los modelos de medición psicométricos proponen aproximar la ubicación de una persona respecto de una variable latente de interés; los basados en la teoría de respuesta al reactivo (IRT, por sus siglas en inglés), han contribuido en gran medida a la medición psicológica y educativa al superar algunos obstáculos técnicos, tales como el hecho de que la estimación de la habilidad de un examinado dependa de una muestra particular de los reactivos de una prueba. Sin embargo, esta forma de medir también ha evidenciado ciertas limitaciones, sobre todo en contextos educativos en los que cada vez es mayor la demanda de información sobre el procesamiento cognitivo del estudiantado.

El uso de CDM confiere algunos beneficios que otros modelos no ofrecen. En primer lugar, la mayoría de los modelos de la Teoría de Respuesta al ítem (TRI) y Modelos Rasch (MR) suponen o requieren la unidimensionalidad estadística de los datos como una condición previa para la calibración de ítems y la estimación de parámetros. Aunque los modelos Rasch multidimensionales carecen de este requisito, en la mayoría de los otros modelos la unidimensionalidad se considera un requisito para ubicar a los examinados a lo largo de un continuo hipotético, por lo que una gran ventaja de los CDM es que no requieren del supuesto de unidimensionalidad. Es entonces que adaptar modelos unidimensionales en los datos con estructuras subyacentes complejas abre la probabilidad del ajuste de los reactivos o personas de manera incorrecta, independientemente de su calidad del contenido, y con ello se elimine información útil de los examinandos o los reactivos.

En segundo lugar, y desde el punto de vista matemático, los CDM se enfocan en analizar los patrones de respuesta registrados ante los reactivos en una prueba utilizando variables categóricas latentes asociadas a los examinados, con especificaciones en torno a la relación entre las variables latentes requeridas para contestar cada reactivo (Templin y Henson, 2006). Así, el propósito principal de los CDM es la clasificación de los examinados en clases latentes dicotómicas o politómicas (usualmente referidas como atributos, habilidades, o estados de conocimiento) determinadas por vectores de indicadores binarios para cada atributo latente en cada reactivo (Chiu y Douglas, 2013; de la Torre y Douglas, 2004).

En tercer lugar, y desde el punto de vista psicológico, los CDM permiten contrastar la ejecución de los participantes en una prueba estandarizada a partir de un conjunto predefinido de atributos psicológicos asumidos como requeridos para contestar correctamente cada reactivo, de modo tal que los examinados puedan obtener retroalimentación de acuerdo a su desempeño y se puedan estimar parámetros para cada reactivo en la prueba (Rupp, Templin, y Henson, 2010).

Dependiendo del objetivo de la evaluación, los CDM se pueden utilizar para determinar si el examinado ha dominado un conjunto de habilidades (Rupp y Templin, 2008; Huo y de la Torre, 2014) o si presenta un conjunto de signos o síntomas relacionados con la presencia de un trastorno clínico (Templin y Henson, 2006). Los CDM resultan particularmente útiles cuando las respuestas observadas en una prueba o inventario se califican como dicotómicas (por ejemplo, reactivos de opción múltiple calificados como correctos o incorrectos, listas de cotejo que indican la presencia o ausencia de determinados signos y síntomas, etcétera), y cuando existe una teoría psicológica que asume a los constructos latentes como categóricos (por ejemplo, habilidades dominadas o no dominadas, o un rasgo de personalidad definido como presente o ausente). A pesar de que existen CDM que permiten la definición de los atributos latentes como politómicos, la mayoría de la investigación en este tema se ha centrado en el desarrollo de modelos para atributos latentes dicotómicos (Rupp, Templin y Henson, 2010).

Por lo tanto, de manera ideal, una teoría cognitiva debería ser utilizada dentro de la planeación de la construcción de la prueba o inventario psicológico, de modo tal que la teoría defina qué atributos latentes son requeridos para cada reactivo y describa el proceso a través del cual las habilidades están ligadas para producir una respuesta correcta (Henson, Templin y Willse, 2009). La teoría cognitiva queda especificada en la matriz Q (de la Torre, 2009), misma que es un elemento común en los distintos modelos contenidos bajo el término genérico CDM. Una matriz Q es la estructura factorial de un CDM se puede considerar como una hipótesis sobre las habilidades necesarias para hacer que cada artículo sea correcto (Li, 2011). En esta matriz se presentan tantas filas como elementos que hay en la prueba y tantas columnas como atributos que subyacen al rendimiento de la misma (Tatsuoka, 1990). La inclusión de la matriz Q en los CDM resulta un número fijo de posibles clases latentes en las que se puede clasificar a los examinados y define dicha clasificación a partir de patrones ideales de respuesta, así como de desviaciones con respecto a dichos patrones ideales (Chiu y Douglas, 2013; Templin y Henson, 2006; von Davier, 2005, 2008).

Dada la naturaleza diagnóstica de los CDM, la matriz Q idealmente debe ser desarrollada y validada por expertos en temas relativos a los constructos medidos por la prueba (Tatsuoka, 1990). Tras la validación de la matriz Q, la lista de cotejo y el modelo de diagnóstico cognitivo elegido para el análisis ayudarán a clasificar a cada persona evaluada de acuerdo a la identificación o ausencia de alguna habilidad cognitiva. Los CDM promueven la evaluación para el proceso de enseñanza-aprendizaje en oposición a solo la evaluación de los resultados del aprendizaje (Jang, 2008). A través de la retroalimentación diagnóstica detallada, puede informar a docentes, supervisores, directivos o tomadores de decisiones el modificar la instrucción y el aprendizaje en el aula, si es necesario.

Los CDM tienen un enfoque interdisciplinario para la evaluación diagnóstica, ya que permiten investigar la relación entre los procesos psicológicos y las estrategias que subyacen al rendimiento en los reactivos de una prueba de interés y las respuestas proporcionadas a esos reactivos a través de análisis estadísticos causales, debido a que los CDM tienen su origen y pueden ser expresados como una variación del modelo general de clases latentes (Templin y Henson, 2006; von Davier, 2009). Específicamente, los CDM han sido definidos como *Modelos de variables latentes multidimensionales confirmatorias y probabilísticas con una estructura factorial simple o compleja* (Rupp y Templin, 2008).

Son modelos probabilísticos en el sentido de que cada CDM expresa el nivel de rendimiento de un encuestado dado en términos de la probabilidad de dominio de cada atributo por separado, o la probabilidad de que cada persona pertenezca a cada clase latente (Lee y Sawaki, 2009). Al igual que los modelos de Análisis Factorial Confirmatorio (AFC), los CDM también son de naturaleza confirmatoria en el sentido de que las variables latentes en CDM se definen a priori a través de una matriz Q. Los CDM también son modelos de variables latentes multidimensionales porque, a diferencia de los modelos IRT que asignan a los encuestados un único puntaje en una escala continua, estos asignan a los encuestados perfiles de habilidades multidimensionales clasificándolos como que dominan versus no dominan cada habilidad involucrada en el examen. Además, los CDM son notablemente diferentes de los modelos IRT multidimensionales en que las variables latentes en CDM son discretas o categóricas (por ejemplo, dominan vs no dominan), mientras que las estimaciones de capacidad en modelos IRT multidimensionales son continuas.

**Entrevistas cognitivas y protocolos de pensamiento en voz alta**

Dados los estándares actuales para la evaluación de la calidad psicométrica de las pruebas psicológicas y educativas, se estipula que una de las evidencias requeridas que justifican las inferencias derivadas de la aplicación de un test se refiere a las estrategias para contestar a los ítems que lo componen (American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME], 2014; Cronbach y Meehl, 1955). Este tipo de evidencias permite conocer detalladamente la forma en que los examinados resuelven los ítems de una prueba, lo cual es fundamental para establecer si esta se adecua al propósito para el que fue creada y a su marco teórico. En este sentido, esta estrategia de validación resulta necesaria para detectar posibles fuentes de varianza irrelevante al constructo que se pretende evaluar. También, posibilita averiguar si este está apropiadamente representado en los reactivos (Haladyna y Downing, 2004; Messick, 1995).

Existen diferentes metodologías para recopilar información sobre las estrategias que utilizan las personas para abordar preguntas planteadas en instrumentos de papel y lápiz (Gorin, 2006). Desde técnicas de gran precisión, como el seguimiento ocular (Allen y Horsley, 2014) hasta aquellas mediante las cuales se pueden documentar algunos contenidos mentales, que se activan durante la ejecución de una tarea (Fox, Ericsson y Best, 2011; Smith-Castro y Molina, 2011). Muchos métodos de la psicología cognitiva han mostrado ser de gran utilidad para entender los procesos mentales involucrados en la resolución de pruebas estandarizadas (Embretson y Gorin, 2001); por ejemplo, los protocolos de pensamiento en voz alta (Ericsson y Simon, 1993) constituyen una herramienta crucial para validar matrices de atributos en los CDM (Rupp, Templin y Henson, 2010; Svetina, Gorin y Tatsuoka, 2011; Wang y Gierl, 2011). Así, son muy útiles para determinar si los examinados emplean las estrategias que, según los desarrolladores de una prueba, son imprescindibles para resolver correctamente los reactivos (Cohen y Upton, 2007; Rupp, Ferne y Choi, 2006; Schroeder, 2011). Esta técnica también suele utilizarse en las fases iniciales del desarrollo de una prueba (Hopfenbeck y Maul, 2013; Karabenick et al., 2007; Willis, 2005).

La necesidad de identificar con cierto nivel de detalle las estrategias utilizadas por los examinados para responder a los ítems de un test estandarizado es relativamente reciente en comparación con la larga trayectoria que posee este tipo de pruebas (Hunt, 2011). Esta línea de investigación se ha construido a partir de la indagación (mediante entrevistas semiestructuradas y autorreportes verbales) sobre las destrezas que utilizan las personas para resolver ítems en pruebas estandarizadas.

Por ejemplo, Farr, Pritchard y Smitten (1990) mediante entrevistas introspectivas y retrospectivas lograron determinar que, pese a las diferencias entre las estrategias empleadas para resolver una prueba estandarizada, los examinados se interesaban fundamentalmente por leer las preguntas primero, para luego enfocarse en el pasaje correspondiente.

Por su parte, mediante la técnica de entrevista, Powers y Wilson (1995) lograron descubrir que un grupo de examinados a quienes se les presentaron reactivos de comprensión lectora del Scholastic Aptitude Tests (SAT) sin los respectivos textos o pasajes obtuvieron notas altas en la escala de puntuación; inclusive arriba de lo esperable por azar, y observaron que la estrategia más utilizada era la de examinar la consistencia entre las preguntas y reconstruir el tema original de estas. A partir de lo anterior, los autores mencionan que estos reactivos no muestran una relación directa con la comprensión de lectura, porque el texto principal resulta prescindible para contestarlos correctamente. En otro estudio, Rupp, Ferne y Choi (2006) analizaron los datos extraídos de diez entrevistas cognitivas con lectores no nativos del inglés, a los cuales se les proporcionaron algunos pasajes con varias preguntas de selección única provenientes de una prueba de inglés, logrando identificar las siguientes estrategias para responder a las preguntas: a) leer superficialmente el pasaje, luego las preguntas y buscar palabras claves en el texto principal para escoger la opción correcta; b) leer todo el pasaje, buscar palabras claves, leer todas las opciones, buscar términos claves en las opciones y responder a las preguntas; c) leer primero las preguntas, luego el texto principal y contestar, y d) leer las preguntas y buscar palabras claves en el texto principal para ir respondiendo a las preguntas, según los términos que los lectores iban encontrando.

En el campo de la teoría de la validez, tanto Messick (1995) como Kane (2006) propusieron a los encuestados que piensen en voz alta mientras responden a las versiones piloto de los instrumentos de evaluación con el fin de recopilar evidencia de validez de sus procesos de respuesta. Por su parte, en las normas AERA, APA y NCME (2014) se señalan cinco líneas de evidencia para la validez de las evaluaciones: 1) contenido de la prueba; 2) procesos de respuesta; 3) estructura interna; 4) relaciones con otras variables; y, 5) las consecuencias de la prueba. Las referencias a "pruebas" reflejan el predominio de las pruebas estandarizadas en la evaluación educativa, pero las Normas pretenden referirse a las evaluaciones en general. A su vez, Castillo-Díaz y Padilla (2013) argumentaron que los estándares ofrecían muy poca orientación sobre los métodos de recolección de evidencia de validez de los procesos de respuesta y encontraron que esta línea de validez rara vez se busca, pero que con entrevistas cognitivas se podría abordar mucho mejor.

Con base en múltiples evidencias, se resalta que los autorreportes verbales en voz alta son sumamente útiles para aportar evidencias sobre el constructo medido por un test y las posibles fuentes de varianza irrelevante, que pudieran afectar la medición de este (Castillo y Padilla, 2013; Cui y Roduta, 2013; Ercikan et al., 2010; Taylor y Dionne, 2000) siempre y cuando se utilice correctamente (Cromley y Azevedo, 2006; Leighton, 2004; Leighton y Gierl, 2007a) y en un contexto relevante (Leighton y Gierl, 2007b).

Faltan las referencias, jiji