



Hacia una visión pragmática de la medición en educación

*Towards a pragmatic vision
of measurement in education*

*Rumo a uma visão pragmática
da medição na educação*

CLAUDIA PATRICIA OVALLE RAMIREZ¹

Recibido:22/07/2024-Aprobado:01/10/2024-Publicado:30/12/2024

Para citar este artículo:

Ovalle-Ramirez, Claudia Patricia. (2024) **Hacia una visión pragmática de la medición en educación**, *Tempus Psicológico*, 8(1) - ISSN: 2619-6336
DOI: <https://doi.org/10.30554/tempuspsi.8.1.5161.2025>

¹ Universidad de Antioquia. Correo: claudia.ovalle@udea.edu.co
ORCID: orcid.org/0000-0002-3664-7290

Resumen

En el presente artículo se introduce al lector en la perspectiva pragmática de la medición que supera la visión tradicional de la medición como asignar números (Stevens, 1958). El educador ha de considerar el proceso de medición y los supuestos (teóricos y psicométricos) que tiene con respecto a dicho proceso para que la medición sea válida y accionable. Además, la medición debería pasar de centrarse en variables continuas distribuidas normalmente al uso de medidas como las ordinales y las nominales, que permiten abordar otros fenómenos de interés como los niveles de progresión y mejora. Finalmente, se plantea un ejemplo con la medición pragmática de las habilidades del siglo XXI.

Palabras clave: *Habilidades siglo XXI, Medición, Psicometría, Educación.*

Abstract

This article introduces the reader to the pragmatic perspective of measurement that goes beyond the traditional view of measurement as assigning numbers (Stevens, 1958). The educator has to consider the measurement process and the assumptions (theoretical and psychometric) that he or she has regarding that process in order for the measurement to be valid and actionable.

In addition, measurement should shift from focusing on normally distributed continuous variables to the use of measures such as ordinals and nominal that allow us to address other phenomena of interest such as levels of progression and improvement. Finally, an example is presented with the pragmatic measurement of 21st century skills.

Key words: *21st century skills, Measurement, Psychometrics, Education.*

Resumo

Neste artigo, o leitor é apresentado à perspectiva pragmática da medição que vai além da visão tradicional da medição como atribuição de números (Stevens, 1958). O educador deve considerar o processo de medição e as suposições (teóricas e psicométricas) que ele tem em relação a esse processo para que a medição seja válida e acionável. Além disso, a medição deve deixar de se concentrar em variáveis contínuas normalmente distribuídas para usar medidas como variáveis ordinais e nominais, que permitem abordar outros fenômenos de interesse, como níveis de progressão e melhoria. Finalmente, um exemplo é apresentado com a medição pragmática das habilidades do século 21.

Palavras-chave: *Habilidades do século XXI, Medição, Psicometria, Educação.*

“En las ciencias sociales, en particular, la mayoría de las evaluaciones no son medidas, sino mezclas de opinión y estimación”

(Anónimo)

Introducción

El problema de medición en Educación, no se reduce a la validez o la generalización de los resultados obtenidos, o a la aplicación de ciertas técnicas estadísticas y el cumplimiento de sus supuestos. El debate se amplía a problemas fundamentales de orden ontológico y epistemológico, que están a la base de la medición de las capacidades humanas.

Algunos autores señalan que lo psicológico no es medible y que no es en sí mismo cuantitativo. Otros autores afirman que los fenómenos psicológicos son ordinales (e.g., se puede sentir más o menos dolor) y que esto no significa que los constructos sean cuantitativos (e.g., el dolor no es numérico) (Michell, 1997). Incluso se ha postulado la inconmensurabilidad de lo psicológico (Gunther, 2009) y las dificultades de basar la ciencia de lo humano en lo cuantitativo (Tafreshi et al., 2016).

Por otra parte, hay autores que afirman que la cuantificación ha sido uno de los pilares de ciencias relacionadas con la medición humana como la Psicología (Uher, 2020) y que se requiere volver a reflexionar sobre los aspectos epistemológicos y ontológicos que subyacen a la medición: ¿qué es medir? (ó ¿cuál es su proceso, más allá de asignar valores cuantitativos o cualitativos?), ¿cuál es el estatus del objeto a ser medido?, ¿cuáles son los límites de esta acción de medir los constructos psicológicos?

A través de la historia han existido varias definiciones de medición: desde las clásicas, que indican que los objetos tienen cualidades y no son medibles de forma cuantitativa (Michell, 2003); pasando por las posturas operacionistas, que se enfocan en la medición como el producto (numérico) de una operación (medición), y que no separan la definición del concepto de su medición (Green, 2001). Así mismo, definiciones más recientes, como la de la Teoría Representacional de la Medición (Duncan & Suppes, 2002), que enfatiza que las características medibles son también nominales y ordinales (no solo continuas y normales) y que existe una correlación (no identidad) entre las propiedades empíricas y los sistemas numéricos.

El problema de la medición en campos como la Psicología o la Educación no está en la posibilidad o no de cuantificar, ya que el problema central está en la falta de claridad conceptual alrededor de los constructos y la necesidad de usar nuevas herramientas metodológicas (Franz, 2021). Esto implicaría aprehender lo

psicológico en diferentes escalas (ordinales, de intervalo, nominales, etc.) y con diversos modelos psicométricos como el de Rasch 1PL (Rasch, 1960)², 2PL, 3PL (Birnbäum, 1957)³ los modelos politómicos y los modelos conjuntos. El proceso de medición, por tanto, implicaría una visión pragmática de la medición, que permita una mejor definición del proceso (tanto a nivel teórico como metodológico) y un uso accionable de las medidas.

El presente artículo se enfoca en la medición desde una perspectiva pragmática. Se presenta el proceso de la medición, los modelos psicométricos que le subyacen y la propuesta de una nueva definición pragmática de la medición con base a los trabajos en Chile de Torres-Irribarra (2020).

El proceso en la medición de lo humano

El proceso de medición conlleva una serie de pasos ordenados, incluyendo, la definición de los constructos a ser medidos, la creación de tareas que faciliten respuestas, la asignación de valores a las respuestas y el modelado y análisis de estas respuestas. En síntesis, el proceso de medición puede decirnos más de lo que pensamos acerca del constructo que sobre el constructo en sí mismo.

La *definición del constructo* determina la información que se requiere recoger sobre la característica psicológica y delimitará las inferencias válidas que se pueden llevar a cabo como resultado de la evaluación (Wilson & Sloan, 2000). Esta tarea no es sencilla, ya que se requieren marcos teóricos sólidos y de referentes bien elaborados.

Un ejemplo de estos marcos son las teorías de desarrollo cognitivo las cuales permiten comprender el cambio a través del tiempo (Maul et al., 2018). Estas teorías permiten establecer qué significa convertirse en experto en un dominio, facilitando que la evaluación y sus dispositivos arrojen evidencia sobre distintos niveles de dicha experticia. Nótese que en este caso el uso de una escala cuantitativa no sería suficiente para modelar la transición de novicio a experto, por lo cual el proceso de medición involucrará además órdenes y variables nominales.

2 El modelo de Rasch (1PL) permite el estudio de pruebas compuestas por ítems dicotómicos (Rasch, 1960). Supone que la probabilidad de acertar una pregunta depende solamente del poder discriminador de los ítems (que es constante para todos ellos) y de la dificultad. El parámetro de discriminación es la capacidad de cada ítem de separar a individuos que poseen niveles distintos del constructo medido. La dificultad se refiere al nivel de habilidad o actitud que debe tener una persona para tener una probabilidad de 0,5 de acertar el ítem.

3 El modelo 2PL (Birnbäum, 1957; Lord, 1952) permite ítems con distinto poder discriminador, también considera el parámetro de la dificultad. El modelo 3PL considera la dificultad, la discriminación y el pseudo-azar (probabilidad de acertar el ítem en caso de que se desconozca la respuesta).

La medición con base en variables cognitivas se ha desarrollado aún más que las evaluaciones con base a otros constructos (i.e., variables socioemocionales, competencias del siglo XXI, entre otras), y se han refinado, incluso hasta establecer la multidimensionalidad de los fenómenos evaluados y los diferentes caminos por los cuales se transita para llegar desde el nivel de novicio al de experto en una habilidad. Es decir, no son modelos simples o reduccionistas de medición, sino que conllevan una serie de presupuestos y condiciones para ser válidos y pertinentes en la medición de constructos no observables, haciendo énfasis en habilidades y conocimientos que se requieren para tener diferentes niveles de experticia.

Así, un ítem cognitivo que evalúa conceptos de física demuestra que es mejor conocedor (experto) quien comprende que la densidad se relaciona con que un objeto se hunda, comparado a quien asume que el objeto se hunde debido a su peso o por cualquier otra característica irrelevante (novicio).

Por otra parte, en la medición de los constructos (como las habilidades y el aprendizaje) debe haber correspondencia entre la tarea presentada, la respuesta emitida del sujeto y el constructo evaluado. Una forma de hacer esto es por medio de ítems de selección múltiple, de las respuestas construidas (ej., escribir un ensayo) o a través de las situaciones embebidas en un contexto (ej., juego de roles).

Una vez que se han *establecido los constructos*, se han *desarrollado los ítems* y se ha *recolectado información*, el análisis psicométrico puede constituirse en una forma de reportar resultados para la acción (e.g., intervención educativa, diagnóstico, etc.) por medio del uso de diferentes modelos de medición (Diakow et al., 2014) Por ejemplo, los modelos de medición pueden reflejar los niveles de experticia de los evaluados identificando las posibles rutas para su mejoramiento. Para esto se puede hacer uso de representaciones como los grafos de redes de nodos en modelos estadísticos de clasificación, donde cada uno de los nodos representa uno entre varios niveles de habilidad de los evaluados. Pero este es solo uno de los múltiples modelos de medición que se pueden aplicar, ya que también se encuentran modelos para ítems dicótomos y los modelos para ítems polítomos (López-Pina, 2005).

Para ítems dicótomos, un modelo que proviene de la psicometría (una rama de la psicología) es el modelo de Rasch (modelo logístico 1PL) el cual permite identificar la relación entre la capacidad (latente) del individuo y la dificultad de los ítems (preguntas). Por tanto, la probabilidad de que un individuo responda a un ítem de forma correcta se modela como función de la diferencia entre la capacidad del individuo y a la vez de la dificultad del ítem (Hontagas et al., 1998). Este modelo permite la

comparación visual entre ítems, y el análisis de su relación con la capacidad de los respondientes por medio del mapa de Wright (ver figura 1). En dicho mapa, se representa gráficamente la habilidad del individuo en la izquierda (histograma de la distribución de frecuencias de habilidad) y la dificultad del ítem a la derecha (medida con el modelo de Rasch). Por inspección visual del mapa, se puede concluir cuáles ítems deberían integrar la prueba, ya que se facilita determinar la relación entre un ítem y la capacidad del individuo en términos de una misma escala, permitiendo también la construcción de categorías de desempeño.

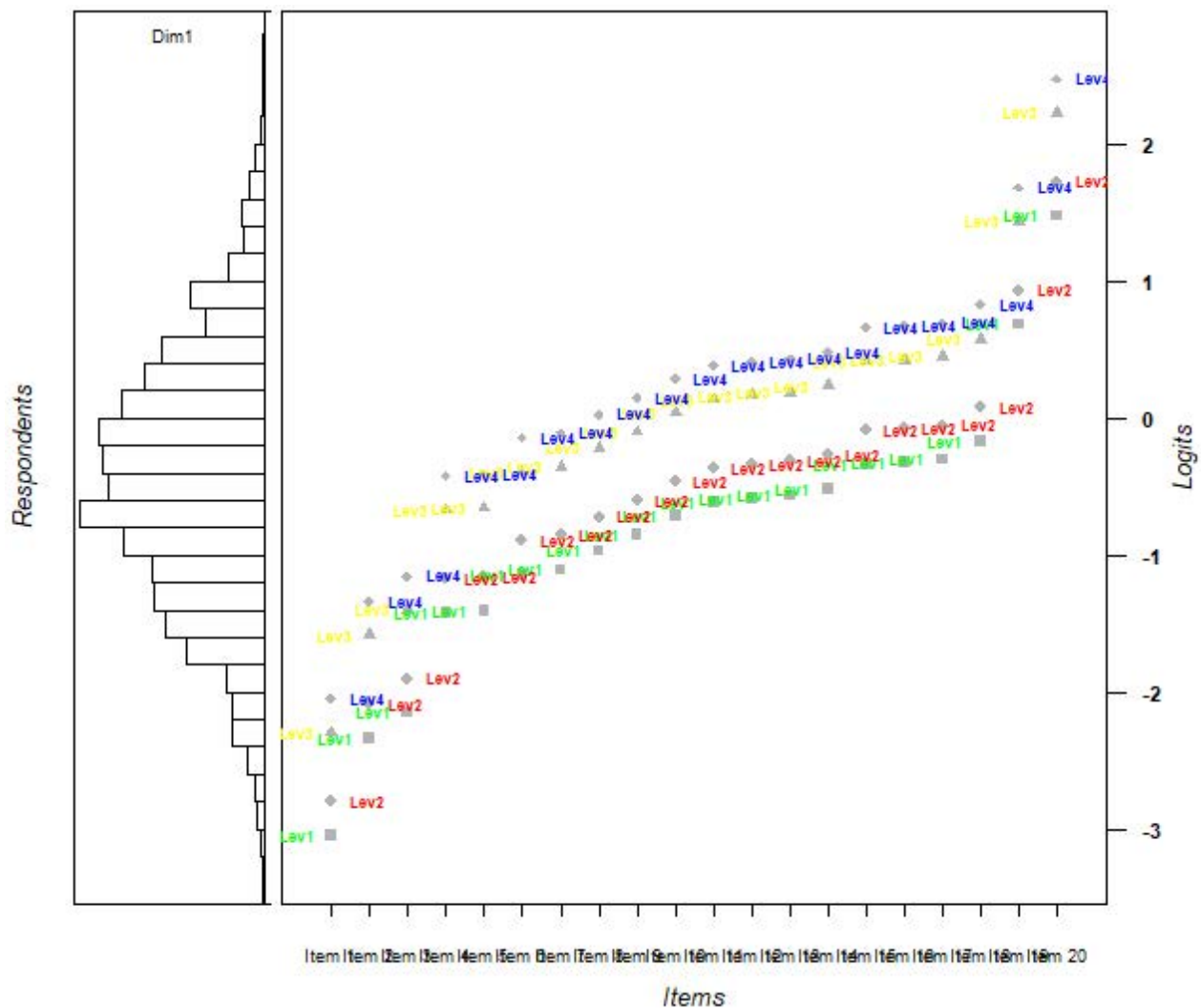


Fig. 1. Mapa de Wright, que relaciona los ítems (y su dificultad) y a los individuos (y su habilidad). Los ítems aparecen graficados como puntos a la derecha (de acuerdo con su nivel de dificultad) y las habilidades de los individuos se representan a la izquierda como una distribución de frecuencia del rasgo θ (habilidad). El mapa de Wright permite establecer categorías de sofisticación conceptual dependiendo de la dificultad y la habilidad de los individuos (a mayor dificultad y habilidad, mayor sofisticación conceptual).

Para el caso de los ítems polítomos, existen varios modelos: el modelo de respuesta graduada de Samejima (1969), el modelo de crédito parcial de Masters (1982) y el modelo de respuesta nominal de Bock (1972).

El modelo respuesta graduada (MRG) fue creado para analizar ítems politómicos ordinales. Con MRG se modela la probabilidad de responder a una alternativa de respuesta entre varias. Se pueden representar ítems que tienen hasta siete categorías (k) de respuesta, y que pueden reflejar una habilidad creciente del respondiente. Los parámetros que incluye son un parámetro de discriminación y k-1 parámetros de dificultad en orden creciente. La representación gráfica de este modelo tiene dos partes: las k-1 curvas características (que relacionan la probabilidad de responder correctamente a un ítem con el nivel de la aptitud), y para cada nivel de habilidad la probabilidad de contestar a cada una de las categorías individualmente (Samejima, 1969).

Los Modelos de Crédito Parcial o PCM (Masters, 1982), fueron creados para ítems ordinales y facilitan las comparaciones entre las categorías de respuesta de cada uno de los ítems. Así, se modela la probabilidad de que una respuesta al nivel “K” sea o no más probable que una en el nivel “k-1”. Los resultados de un modelo PCM muestran gráficamente en un mapa de Wright las posiciones relativas con respecto a la distribución de la característica (ej. habilidad) en un individuo, facilitando detectar cuántos individuos se encuentran en determinado nivel de la característica medida. Esto permite determinar niveles diferenciados de rendimiento para poder clasificar a los individuos en una escala ordinal interpretable de acuerdo con el nivel de habilidad.

Finalmente, con los Modelos de Respuesta Nominal, se pueden analizar ítems hasta con seis categorías de respuesta no ordinales, y se pueden asignar valores numéricos congruentes con el modelo para los parámetros de discriminación e interceptos (uno por cada categoría). Es decir, el modelo relaciona la probabilidad de que el sujeto escoja cada una de las “n” alternativas de respuesta a un ítem en función de su nivel de habilidad y de los parámetros de dificultad y discriminación de la alternativa escogida. Gráficamente, este modelo se representa con las curvas características de las categorías y se proporciona una tabla con los valores que corresponden a los puntos de corte que permiten establecer niveles diferenciados de desempeño.

Después de decidir sobre el modelo de medición y una vez se cuente con resultados de medición y aplicación del modelo, dichos resultados deben garantizar una serie de características. Según Wilson y otros (2012), las mediciones deben ser “accionables”, es decir, deben ser fáciles de entender y deben orientar las acciones (por ejemplo, educativas, diagnósticas, etc.) dirigidas al desarrollo de las

competencias evaluadas. Igualmente, las evaluaciones deben estar disponibles para usuarios de diferentes niveles (ej. estudiantes, docentes, apoderados, etc.). Finalmente, deben ser evaluaciones que se ajusten y relacionen entre niveles (ej. la relación de la evaluación en aula con la evaluación de gran escala o pruebas estandarizadas).

Ejemplos de la medición de constructos

Las evaluaciones deben arrojar resultados que sirvan para la acción. El “triángulo de la evaluación” se refiere a las relaciones entre teorías psicológicas (i.e., cognitivas), observaciones sobre los evaluados, y la interpretación de las evidencias recolectadas (NRC,2001)⁴. Estas relaciones son importantes ya que mientras las teorías se refieren a escalas de desempeño (i.e., buen desempeño, desempeño regular, desempeño bajo), los modelos de medición generalmente se refieren a variables continuas latentes (i.e., modelo de Rasch, 2PL, 3PL). En este sentido, los modelos de medición en psicología puede que no traduzcan de forma aplicada lo postulado por la teoría, pero existen metodologías para aproximarse a una práctica de medición apropiada y útil.

Un caso en particular lo presentan Diakow y otros (2014) quienes diseñan una evaluación ligada a un currículo de comprensión de lectura. Originalmente, los niveles de desempeño (no comprometido, comprometido, discriminador, coordinador, innovador) de la evaluación se relacionan con los niveles de puntuación (escala 0 a 4 puntos). Sin embargo, se emplea un modelo psicométrico de medición de crédito parcial (PCM), que establece el parámetro θ , el cual indica la posición de la persona en el rasgo latente (habilidad de lectura), pero que a la vez considera la diferencia entre este rasgo latente y la dificultad del ítem. De este modo la relación entre las categorías de la teoría (los niveles de desempeño) y el modelo de medición (PCM) no se sostiene.

Entonces, como solución, se establecen nuevos puntos de corte que permitan redefinir las categorías. Esto se logra por medio de una re-parametrización que incluye la diferencia entre el rasgo latente y la dificultad y el efecto principal de pertenecer el ítem a una de las categorías propuestas por la teoría. Así, por medio de la medición, se logran unas categorías más acordes con las postuladas por la teoría. Otra alternativa, es la construcción de casos prototípicos y casos límites, que permitan caracterizar categorías parecidas a las postuladas en la teoría (Diakow et al., 2014).

4 Una de estas correspondencias en el triángulo, se refiere a la forma como los modelos psicométricos se relacionan con la teoría sustantiva para poder producir mediciones psicológicas.

Existen variados modelos psicométricos que pueden servir al propósito de hacer más “accionable” la medición, por medio de los constructos latentes y con el uso de mediciones polítomas, dicotómicas, y otras que elija el evaluador para modelar un aspecto psicológico en una muestra. En líneas generales, el proceso involucra dividir una variable continua (ej., la habilidad) en grupos de desempeño y luego se interpretarán esos grupos como niveles de habilidad cualitativamente diferenciados.

Actualmente, la psicometría ha planteado usar re-parametrizaciones que permitan mostrar el efecto principal del nivel al cual pertenece la respuesta del individuo. Este procedimiento se lleva a cabo en modelos novedosos, que aplican el modelo de Rasch al análisis de clases latentes, como el modelo LLC (“Located Latent Classes”), y L-LLC o “Level Located Latent Classes” (Lindsay et al., 1991). Estas técnicas modelan grupos de habilidad como se hace en las variables latentes, pero usan el cálculo de centroides y el cálculo de puntuaciones de corte a partir de re-parametrizaciones específicas para hacer categorías más acordes a la teoría.

La medición como una forma de transmitir incertezas

Proponemos aquí que la medición es un proceso caracterizado por su estructura, no solo por la especificación de la relación funcional que conecta sus entradas con sus salidas: lo que se requiere es una explicación de cómo el proceso hace lo que hace, no solo de lo que hace. (Maul et al., 2018, p. 612)

La medición puede caracterizarse como un proceso del modelo de caja negra donde se transforman inputs en outputs. Sin embargo, la medición por esencia produce información en forma de valores que se relacionan con las propiedades de los objetos. Esta valoración o emisión de juicios conlleva un proceso y unas condiciones que hacen de la medición una clase especial de evaluación.

Una de estas condiciones, esencial para que la medición sea tratada como un tipo especial de evaluación, consiste en que somos capaces de aportar información sobre el grado en que podemos confiar en la evaluación. Por ejemplo, a través de reportar el grado de incerteza que presenta la medida. Esta incerteza puede comunicarse en términos de objetividad (la condición de que la información transmitida se refiere a la propiedad medida y nada más) y de la intersubjetividad (la información transmitida es interpretable de la misma forma por diferentes personas en distintos contextos). Un caso puntual, es la calibración de una medida, proceso debido al

cual existen centros de metrología cuya finalidad es garantizar las certezas (por ejemplo, calibrar un kilo como 100 gramos exactos).

Pero en educación, ¿cómo podemos transmitir la incerteza que hace de la medición un tipo especial de evaluación?, ¿cómo apelar a la objetividad y la intersubjetividad propias de la medición de los constructos? De la misma manera que los métodos de medición sobre otras entidades, la medición en educación se caracteriza porque se orienta por la teoría. La medición no consiste en procesos ateóricos que produzcan datos puros, sino que se trata de compromisos con los enunciados teóricos establecidos previamente (Valsiner, 2012).

En palabras de Maul y otros, “la medición produce información públicamente confiable porque sus resultados contienen información explícita sobre su propia calidad, se aplica también a los procesos que se consideran mediciones en ciencias sociales” (2018, p. 616). No habría entonces contradicciones entre el hecho de obtener medidas objetivas sobre fenómenos educativos. En este sentido, la medición de las entidades psicológicas debería reportar sobre el proceso de medición haciendo explícitas las incertezas, como las de tipo definicional (i.e., especificar adecuadamente cuáles son las propiedades que medimos del constructo y cuáles se dejan de medir para caracterizarlo a partir de la teoría disponible).

La medición en Psicología no es solo la asignación directa de valores o cantidades numéricas, sino que se trata de un proceso (el de medición) y de sus supuestos. Por tanto, la medición no se limitará a la característica cuantitativa de sus entradas o resultados, sino que es un proceso y una actividad dirigida a lograr conocimiento confiable, preciso y accionable (Mari et al., 2017).

En este proceso los supuestos (teóricos y psicométricos) pueden jugar un papel relevante. Por ejemplo, tenemos un grupo de estudiantes que toman una prueba diseñada para medir su razonamiento cuantitativo y se emplean modelos psicométricos como el de Rasch (1PL) para modelar dicha habilidad como una variable continua. Sin embargo, no se hace explícito por qué esta propiedad (razonamiento cuantitativo) se define como una variable continua parametrizada con un modelo 1PL, pues no se presenta la justificación específica sobre su adecuación al contexto de medición, lo cual es recomendado desde una perspectiva aplicada de la medición como proceso (ver más adelante el aparte *Hacia una visión pragmática de la medición en Psicología*).

Modelamiento de variables latentes

Los modelos de variables latentes usados en campos como la Educación, la Psicología y las ciencias sociales son también ejemplo del uso de la teoría en la medición o de una visión pragmática de la medición⁵. Las variables latentes surgen de la observación de que la correlación entre variables no solo se expresa a sí misma, sino a una propiedad emergente, dependiente de las asociaciones de las diferentes variables y que permite la medición de la causa subyacente y oculta que explica la variación entre dichas variables (Spearman, 1904). Así, se logra la medición de constructos como la inteligencia, por medio de reducir la complejidad de un conjunto de variables observadas por medio de postular una variable diferente, que no es observada y que es subyacente a las variables que se miden directamente.

Los modelos de variables latentes representan propiedades no observables, propiedades directamente observables y/o sus relaciones incluyendo las de tipo causal. El diagrama de un modelo de variables latentes expresa como un constructo provoca ciertas conductas (medidas con ítems especificados a partir de la teoría) y a su vez cómo este constructo es producido por aquello que son sus propiedades o características constituyentes (la varianza compartida). Estos modelos de medición son inferenciales más que descriptivos y permiten abordar conceptos y rasgos de tipo latente como la inteligencia, la personalidad y demás constructos.

Los modelos de variables latentes involucran al error de medición como elemento importante. Dichos modelos son una forma de hacer medición haciendo uso de la estadística, a través de constructos (inobservables) que describen los atributos (observables) que el investigador/evaluador puede medir. Aunque se conserva la idea clásica de medición como asignar un valor a un atributo, los modelos de variables latentes van un paso más allá, por medio de establecer medidas para constructos que son continuos (rasgos latentes), los de tipo categórico (clases latentes) y los de tipo ordinal (clases latentes ordenadas). Así, se supera la mirada a los constructos solamente modelados a partir de variables continuas distribuidas normalmente, como ha ocurrido en casos como el del constructo que llamamos “inteligencia” (medida a través del Coeficiente Intelectual).

Una de las características de la medición definida como modelamiento con variables latentes, consiste en que se permite caracterizar el fenómeno con diferentes escalas. Así, la definición de medición se expande más allá de la simple asignación

5 El modelamiento de variables latentes es solo una de las muchas definiciones de medición en psicología. Entre otras definiciones, se incluyen la asignación de números a un fenómeno “en su sentido más amplio, la medición es el negocio de fijar números en las cosas” o la relación entre estructuras cualitativas mapeadas en estructuras cuantitativas.

de números a determinadas características, lo cual ha sido la definición más común de medición.

Un ejemplo desde el campo de la psicología es: una escala de medición de variable latente “ansiedad” puede medir el rasgo como una variable latente categórica o continua. La escala continua medirá la ansiedad como una puntuación individual, mientras que la escala cualitativa, ordenará a los individuos en categorías como “alta ansiedad” y baja “ansiedad”. Ambas escalas pueden ser válidas y representativas del mismo fenómeno a medir, ya que ambas permiten la cuantificación, siendo en el caso de las variables cualitativas la cuantificación por medio de las cantidades ordinales las que definen grados de atributos.

Hacia una visión pragmática de la medición

Torres -Irribarra (2020) define una perspectiva pragmática de medición como una nueva definición alternativa a las posiciones clásicas y representacionales de esta actividad:

A pesar de sus diferencias, estas teorías (teorías clásicas y representacionales de la medición) asumen que existe una verdadera estructura, ya sea del atributo o de las relaciones empíricas cualitativas, que estamos tratando de recuperar o replicar. En consecuencia, una medición es buena en la medida en que se asemeja a esta verdadera estructura y se aproxima a los valores verdaderos. El pragmatismo nos invita a buscar una concepción alternativa de la medición que evite esta metáfora copiadora, y con ella, un conjunto de problemas teóricos y filosóficos relacionados con la definición y demarcación de la medición. (Torres-Irribarra, 2020, p.43).

En términos generales, la medición es definida desde esta visión pragmática como una actividad: “La medición es (i) una actividad de clasificación, ordenación o cuantificación de un conjunto de elementos (ii) de acuerdo con un modelo (iii) de un atributo relevante (iv) al servicio de un objetivo mayor” (Torres- Irribarra, 2020, p.30). Con esta definición se permiten operaciones cómo caracterizar los atributos y no solo asignarles cantidades, enfatizar en el modelo (la teoría que subyace a la medición) y esperar que la medición sea accionable (que permita el uso práctico de sus resultados para la mejora, cualificación o re entrenamiento y no solo que refleje pasivamente el mundo).

Esta definición de medición es de espectro más amplio que la clásica definición de medir como “asignar números a un objeto” dada por Stevens (1958), ya que

permite operaciones de pensamiento más complejas como clasificar, ordenar y cuantificar, expandiendo la clase de fenómenos y la forma en que pueden ser medidos.

Por ejemplo, la “capacidad para razonar cuantitativamente”, puede medirse en términos de diferencias interindividuales modeladas como diferencias cuantitativas; también en términos de niveles ordenados de proficiencia o en términos de las estrategias que usan los respondientes a la prueba. Estos tres modelos pueden tener implicaciones en las hipótesis que hacemos sobre la capacidad de razonar cuantitativamente.

Según Torres-Irribarra , los psicólogos evaluadores deberían ser capaces de modelar y medir fenómenos entre estos diversos modelos:

Tenemos una tradición de medición psicológica y educativa dominada por el enfoque en el modelado de atributos cuantitativos (Birnbaum 1968; Rasch 1960/1980), una tradición separada de práctica que se centra en las medidas ordinales (Cliff & Keats 2003), y otra tradición metodológica y su comunidad profesional centrada en la clasificación (Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968). (2020, p.47)

El uso de un modelo (tanto teórico como estadístico), es importante, casi central en la definición de medición. Los constructos psicológicos solo pueden ser medidos con exactitud en cuanto estén bien definidos por la teoría. En educación, es común encontrar que a veces los constructos son “mejor medidos que definidos”, y de allí la importancia del modelo para establecer las inferencias y las consecuencias prácticas de la medición.

Los modelos deben sobrepasar la idea de un reflejo de la realidad y permitir cierta flexibilidad y alternativas para que los investigadores midan los atributos psicológicos de diferentes maneras. Un ejemplo, es el constructo de “Preparación para la Universidad” (“College Readiness”). Desde un modelo binario, se puede establecer quién está preparado para la Universidad y quién no. Desde un modelo de múltiples clases, podremos establecer no sólo quien está listo para la Universidad, sino cuales son los atributos (perfiles) que determina esta capacidad (ej., factores sociales, emocionales, cognitivos, etc.). En un modelo ordinal, se determinaría quién está más listo o menos listo para la universidad y finalmente, en un modelo continuo, se puede pensar en un puntaje de corte a partir del cual los estudiantes con cierta cantidad medida del constructo (por medio de una prueba) están listos para iniciar el pregrado.

El modelo elegido dependerá del contexto en que se plantee: por ejemplo, si necesitamos seleccionar candidatos para tomar becas universitarias, es probable que la opción continua sea más adecuada que la opción de medir la preparación para la universidad con un modelo de múltiples clases. Es importante notar que cualquiera de los modelos y sus usos es válido (Box, 1979).

Un ejemplo de la visión pragmática de la medición: el caso de las habilidades del siglo XXI

Siguiendo una visión pragmática de la evaluación se presenta a continuación el caso de la medición de las habilidades del siglo XXI (Wilson et al., 2012).

Las habilidades del siglo XXI son competencias que permiten a los individuos desarrollarse en los actuales contextos por medio de la creatividad, el autoaprendizaje, la colaboración, entre otras capacidades blandas. Sin embargo, por tratarse de constructos no observables, es necesario considerar no solo su medición, sino el proceso a través del cual se recolecta información para inferir la existencia de estas capacidades.

En otras palabras, para decir que hay una “medición” de este tipo de características, se requiere apelar a un proceso cuyos pilares son: determinar los objetivos de medición, establecer las observaciones que proveen la evidencia evaluativa, ajustar un sistema de puntuación o el espacio de resultados, y decidir sobre los modelos de medición a ser aplicados. Por tanto, el proceso de evaluación de las habilidades del siglo XXI será el mismo que el que se lleva a cabo para otros tipos de habilidades:

- (a) definir los constructos a ser medidos
- (b) crear las tareas que estimularán las respuestas del evaluado
- (c) asignar valores (códigos o puntuaciones) de las respuestas a las tareas
- (d) reunir los resultados
- (e) hacer el modelado y análisis de esas respuestas.

Sin embargo, el proceso de evaluar habilidades del siglo XXI será distinto pues se basa en modelos de evidencia y de progreso. Así, la evidencia colectada no da cuenta de las deficiencias sino del nivel de progreso, de modo que los resultados de la evaluación sirven para informar la instrucción en el aula y los currículos (mapeo de progresiones de competencia). Esta noción de competencia se basa en que los estudiantes pueden ser más hábiles en un dominio siguiendo diferentes trayectorias

de complejidad creciente con apoyo de los contextos estructurados de aprendizaje (Griffin & Care, 2015).

No es suficiente con que un estudiante sea capaz de memorizar un hecho científico, sino que debe ser capaz de enunciarlo correctamente, presentar una respuesta sofisticada, y la respuesta debe tener validez y exactitud. Esto es lo que significan las progresiones y los caminos que un estudiante tiene en su trayectoria de aprendizaje. De igual manera, la evaluación de competencias del siglo XXI no puede solo reducirse a tener ítems de recordación que nos permitan visualizar aspectos como la novedad, la validez (las conclusiones se derivan de los supuestos) y exactitud (precisión del argumento) en las afirmaciones del estudiante evaluado.

Una práctica recurrente de los educadores es interpretar la medición sin contexto. Es decir, se considera que solo pueden ser medidas respuestas concretas a ítems específicos (“ejecución específica, para contextos específicos, medidos por ítems específicos”). El caso de las Habilidades del siglo XXI es diferente, ya que se requiere recolectar evidencia de que la competencia es generalizable a múltiples contextos. Por ejemplo, la “Comunicación” es una competencia del siglo XXI que puede medirse en el contexto de la comprensión de lectura, pero también en el contexto del uso de representaciones simbólicas de la matemática. El contexto no puede ser interpretado como “ruido” o como un elemento que produce error, sino que debe invitar a comprender la multidimensionalidad del constructo.

Las habilidades del siglo XXI son habilidades que poco o nada se reflejan en el currículo, comparado con otras habilidades como hacer operaciones numéricas o leer textos breves. Debido a la poca definición y especificidad de las habilidades del siglo XXI, se requiere usar tests con un mayor número de tareas, que permitan hacer inferencias sobre las capacidades del individuo y que disminuyan la varianza de la interacción individuo/prueba.

Por ejemplo, las *valoraciones entre pares*, pueden ser tareas que permitan evidenciar la competencia de “solución de problemas” por medio de las metacogniciones y el rol del estudiante que debe evaluar la calidad de otro compañero. Así mismo, la “creatividad”, es otra competencia cuyo constructo es difícil de medir (existen múltiples pruebas para evaluarla), pero se requerirían tareas auténticas que nos permitan hacer inferencia sobre un constructo poco o nada definido. La definición más aceptada de esta habilidad del siglo XXI es la de “solucionar problemas por medio de la novedad”, lo cual hace que sea una tarea compleja poder evaluarla.

Otras habilidades del siglo XXI, como el “trabajo en grupo” e “interactuar socialmente”, también pueden ser evaluadas por medio de un proceso que involucra la

definición del constructo y sus observaciones, el espacio de respuestas y el modelo de medición (Griffin & Care, 2020).

Por ejemplo, la *definición del constructo* de la competencia “trabajo en equipo” debería incluir aspectos como son el “liderazgo”, “el compromiso”, etc., que pueden ser medidos sobre el individuo mientras interactúa con un grupo, o con varios grupos si se quiere hacer un muestreo más amplio y generalizable de la habilidad. También puede emplearse en las *observaciones* el reporte de cada uno de los participantes en el grupo sobre cada uno de los sujetos involucrados en la solución de la tarea que se aplique. Luego, se establece *el espacio de respuestas* (el conjunto de resultados categóricos en los que se clasifica la ejecución del evaluado) que puede ser la valoración de los productos del trabajo en dos escalas: una para la ejecución individual y otra para el desempeño grupal.

A continuación, se establecerá el *modelo de medición* más conveniente. Por ejemplo, recientemente se habla de la medición empleando “testlets” (Wang & Wilson, 2005). Un testlet es un conjunto de elementos que comparten un estímulo común (por ejemplo, un pasaje de comprensión de lectura) o una figura. Otro nombre para un testlet es un “paquete de ítems” adoptado en pruebas educativas y psicológicas.

La adaptación de modelos psicométricos de respuesta al ítem (IRT) a las respuestas del testlet ignora la posible dependencia entre los elementos dentro de un testlet. El análisis de respuesta al ítem tiende a sobreestimar la precisión de las medidas obtenidas de los testlets y arroja una estimación sesgada de la dificultad y la discriminación de los ítems. Por esto, el modelo de medición que es aplicable a las competencias del siglo XXI puede tener características distintas a los modelos clásicos y de IRT (Teoría de respuesta al ítem). Entre otros modelos aplicables a las competencias del siglo XXI están: tratar el testlet como un super ítem único, puntuarlo de forma politómica y aplicar modelos de respuesta politomos como el modelo nominal, el modelo de crédito parcial y los modelos graduados de respuesta.

Finalmente, debido a que la evaluación es un proceso, o actividad con propósito, se requiere poder compartir la información obtenida por un estudiante para establecer sus fortalezas y debilidades en sus habilidades blandas o habilidades del siglo XXI. Para el propósito de hacer pública la información se pueden usar *mapas de Wrigth* (ver arriba figura 1). Los mapas de Wrigth permiten proveer evidencia visual y de alta calidad. Son representaciones empíricas que permiten comprender cómo se desenvuelve o evoluciona la ejecución de un sujeto a través del tiempo de forma incremental y sofisticada. Los mapas muestran un ordenamiento de las actividades de evaluación desde la más básica a la más difíciles.

Los mapas colocan al estudiante y a las tareas en la misma escala, permitiendo la interpretación del resultado en términos de lo que el estudiante sabe o hace y dónde tiene mayor dificultad.

Conclusiones

En el presente artículo se ha argumentado que la medición es algo más que asignar números a un objeto (Stevens, 1958). Siguiendo a Briggs (2021), se puede decir que, bajo la perspectiva pragmática, la medición se caracteriza como una actividad de clasificación, ordenación, o cuantificación de un conjunto de elementos según un modelo de un atributo relevante en servicio de un objetivo mayor.

En Colombia podríamos diversificar la evaluación y cambiar las escalas y el proceso evaluativo para que sea más transparente y progresivamente mejorado. Por ejemplo: ¿por qué se cuantifica el examen del ICFES? ¿Por qué no mejor se transforma en una prueba que le ayude a entender al estudiante en que categoría de desempeño está y en que está fallando al aprender el currículo o los estándares de competencia del MEN? ¿por qué no diversificamos la prueba del ICFES? Los niños de educación rural, reformatorios, escuelas públicas, minorías y de colegios de baja calidad no pueden competir con las puntuaciones que obtienen los niños de colegios privados. Es decir, potencialmente se les colocaría una barrera de injusticia educativa.

La perspectiva pragmática depende de la capacidad de establecer criterios para el éxito de una actividad de medición que es externa a lo instrumental. Una limitación de este enfoque puede ser la falta de entrenamiento y conocimiento de los educadores en aspectos de medición como la existencia de múltiples modelos psicométricos que permiten modelar diferentes fenómenos y que facilitan la toma de decisiones con base en evaluaciones construidas para tal fin.

Alfred Binet fue comisionado para construir una escala que permitiera escoger a los niños que recibirían educación pública en Francia. Sin tener aún un conocimiento o un modelo refinado de inteligencia, y debido a sus propósitos prácticos, Binet construyó sus escalas de inteligencia, pero advirtió de la importancia de identificar que estas pruebas no tenían valor predictivo y tampoco podrían abarcar la historia pasada del individuo. Esto no se ha respetado en la actualidad y se usan las pruebas de inteligencia para rotular y dar pocas esperanzas hacia el futuro a los niños y jóvenes.

De la misma manera, sin entrar en el debate sobre la conmensurabilidad o no de los constructos, se puede decir que es necesario hacer mediciones con un criterio

pragmático en educación, que funcionen como un modelo de constructo y que sepan identificar abiertamente sus supuestos teóricos y psicométricos y por tanto sus limitaciones y restricciones. Esto evitaría la rotulación y la falta de información que padecen los aprendices.

Referencias

- Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems*. USAF School of Aviation Medicine.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51. <https://doi.org/10.1007/BF02291411>
- Box, G. (1979). Robustness in the strategy of scientific model building. En Launer, R. L.; Wilkinson, G. N. (Eds.), *Robustness in Statistics*. Academic Press.
- Briggs, D. (2021). Book review: A pragmatic perspective of measurement by David Torres Iribarra. *Integrative Psychological and Behavioral Science*, 56(1), 1-10. <https://doi.org/10.1007/s12124-021-09635-7>
- Diakow, R., Torres-Iribarra, D. & Wilson, M. (2014). Some Comments on Representing Construct Levels in Psychometric Models. En: Millsap R.E., van der Ark L.A., Bolt D.M., Woods C.M. (Eds). *New Developments in Quantitative Psychology. Springer Proceedings in Mathematics & Statistics*. Springer.
- Duncan, R. & Suppes, P. (2002). Representational Measurement Theory. En H. Pashler & J. Wixted (Eds.), *Stevens' Handbook of Experimental Psychology* (pp. 1-41). John Wiley & Sons, Inc.
- Formann, A. (1995). Linear logistic latent class analysis. En G. H. Fischer & I. W. Molenaar. *Rasch models: Foundations, recent developments, and applications* (pp. 239-256). New York: Springer.
- Franz, D. (2021). Are psychological attributes quantitative? is not an empirical question: Conceptual confusions in the measurement debate. *Theory & Psychology*, 1(1), 1-20. <https://doi.org/10.1177/09593543211045340>
- Green, C. (2001). Operationism again: What did Bridgman say? What did Bridgman need?. *Theory & Psychology*, 11(1), 45-51 <https://doi.org/10.1177/0959354301111003>
- Griffin, P. & Care, E. (2015). *Assessment and teaching of XXI century skills*. Springer.
- Hontagas, P., Ponsoda, V., Olea, J. & Revuelta, J. (1998). Representación de funciones características de ítems dicotómicos y politómicos. *Psicothema*, 10(2), 475-479. <https://www.psicothema.com/contenido?num=19981002>
- Lindsay, B, Clogg, C. & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86(413), 96-107 <https://www.tandfonline.com/doi/abs/10.1080/01621459.1991.10475008>
- López-Pina, J. (2005). Ítems politómicos vs. dicotómicos: un estudio metodológico. *Anales de Psicología*, 21(2), 339-344. <https://revistas.um.es/analesps/article/view/26941>
- Mari, L., Maul, A., Torres-Iribarra, D. & Wilson, M. (2017). Quantities, Quantification, and the Necessary and Sufficient Conditions for Measurement. *Measurement*, 100, 115-121. <http://dx.doi.org/10.1016/j.measurement.2016.12.050>
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174. <https://doi.org/10.1007/BF02296272>
- Maul, A., Mari, L., Torres Iribarra, D. & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, 116, 611-620. <https://doi.org/10.1016/j.measurement.2017.08.046>

- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–383. <https://doi.org/10.1111/j.2044-8295.1997.tb02641.x>
- Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology*, 13(1), 5–31. <https://doi.org/10.1177/0959354303013001758>
- National Research Council [NRC]. (2001). *Knowing what students know: the science and design of educational assessment*. National Academy Press.
<https://nap.nationalacademies.org/catalog/10019/knowning-what-students-know-the-science-and-design-of-educational#toc>
- Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 17, 1-169 <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1968.tb00153.x>
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101. <https://pubmed.ncbi.nlm.nih.gov/3322052/>
- Stevens, S. (1958). Measurement and man. *Science*, 127(3295), 383–389.
<https://doi.org/10.1126/science.127.3295.383>
- Tafreshi, D., Slaney, K. & Neufeld, S. (2016). Quantification in psychology: Critical analysis of an unreflective practice. *Journal of Theoretical and Philosophical Psychology*, 36(4), 233-249.
<https://doi.org/10.1037/teo0000048>
- Torres-Irribarra, D. (2020). *A pragmatic perspective on measurement*. Springer Briefs in Psychology.
- Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, 19(5), 579–599. <https://doi.org/10.1177/0959354309341926>
- Trendler, G. (2013). Measurement in psychology: A case of ignoramus et ignorabimus? A rejoinder. *Theory & Psychology*, 23(5), 591–615. <https://doi.org/10.1177/0959354313490451>
- Uher, J. (2020). Measurement in metrology psychology and social sciences: data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Quality and Quantity*, 54(3), 975–1004. <https://doi.org/10.1007/s11135-020-00970-2>
- Valsiner, J. (2012). La dialéctica en el estudio del desarrollo. En M. Carretero & J. A. Castorina, (Eds.). *Desarrollo cognitivo y educación. Los inicios del conocimiento* (pp. 137-157). Buenos Aires: Paidós.
- Wang, W. C. & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement* 29(2), 126–149. <https://doi.org/10.1177/0146621604271053>
- Wilson, M., Bejar, I., Scalise, K., Templin, J., Dylan, W. & Torres-Irribarra, D. (2012). Perspectives on Methodological issues. En Griffin P. & Care E. (Eds.). *Assessment and Teaching of 21st Century Skills*. Springer.
- Wilson, M. & Sloane, K. (2000). From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education*, 13(2), 181-208. https://doi.org/10.1207/S15324818AME1302_4

