

Revisión avanzada de la Teoría Clásica de los Tests: supuestos, estimación del error, coeficientes modernos y limitaciones frente a la Teoría de Respuesta al Ítem

Psicometría II

Sesión S01 · Revisión avanzada de la Teoría Clásica de los Tests: supuestos, error de medida y coeficientes modernos frente a la TRI

01

Introducción: la Teoría Clásica de los Tests y el dilema de reportar solo el alfa

La Teoría Clásica de los Tests, conocida también como teoría del puntaje verdadero, constituye el marco psicométrico más extendido en la práctica profesional de la psicología latinoamericana desde mediados del siglo veinte. Su ecuación fundamental, expresada por Spearman en 1904 y formalizada matemáticamente por Lord y Novick en su obra canónica de 1968, plantea que el puntaje observado de una persona en una prueba equivale al puntaje verdadero más un error de medida, bajo supuestos probabilísticos sobre la distribución del error. Esta elegante simplicidad explica su dominio en la enseñanza universitaria y en los manuales técnicos de pruebas comercializadas en Colombia, México, Argentina, Chile y Perú.

Sin embargo, cuatro décadas de desarrollos teóricos posteriores, liderados por autores como Cronbach y Shavelson en 2004, Messick en 1995, McDonald en 1999, Raykov entre 1997 y 2012, Revelle y Zinbarg en 2009, y Sijtsma en 2009, han evidenciado que la aplicación acrítica del modelo clásico arrastra problemas serios. Los coeficientes de confiabilidad tradicionales se estiman bajo supuestos que rara vez se sostienen, el error de medida se trata como homocedástico cuando empíricamente es heterocedástico, y la dependencia del instrumento y de la muestra limita la generalización de resultados. Esta lectura examina cuatro ejes que estructuran la revisión crítica contemporánea del modelo clásico: los supuestos TCT y sus implicaciones prácticas, la estimación avanzada del error de medida, los coeficientes modernos de confiabilidad que superan al alfa, y las limitaciones específicas que han llevado al campo a migrar hacia la Teoría de Respuesta al Ítem.

02

Supuestos de la Teoría Clásica de los Tests y sus implicaciones

La formulación canónica de Lord y Novick en 1968 establece que el puntaje observado X se descompone en puntaje verdadero T más error E , bajo tres supuestos centrales. Primero, la esperanza matemática del error es cero dentro de cada individuo a lo largo de aplicaciones hipotéticamente repetidas; segundo, la correlación entre puntaje verdadero y error es nula; tercero, los errores entre pruebas paralelas correlacionan cero entre sí. A partir de estos supuestos se deriva la definición de confiabilidad como la razón entre la varianza del puntaje verdadero y la varianza del puntaje observado, así como la definición de pruebas paralelas, tau-equivalentes y congenéricas, propuesta por Novick y Lewis en 1967.

El alfa de Cronbach, introducido en 1951, solo constituye un estimador no sesgado de la confiabilidad bajo tau-equivalencia, es decir, cuando todos los ítems miden el mismo constructo con igual saturación factorial y difieren solo en el intercepto. En escalas

psicológicas reales adaptadas a población latinoamericana, como las versiones colombianas del Inventario Clínico Multiaxial de Millon revisadas por autores de la Universidad El Bosque y de la Universidad de los Andes, los ítems difieren sistemáticamente en su poder discriminativo, por lo cual la tau-equivalencia casi nunca se sostiene. Otra implicación menos discutida es la dependencia muestral: la confiabilidad estimada en una muestra no es una propiedad intrínseca de la prueba, sino de la interacción entre la prueba y la muestra, como recordaban Nunnally y Bernstein en 1994. Un cuarto supuesto, menos visible pero igualmente consecuente, es la unidimensionalidad implícita: Cortina en 1993 mostró con simulaciones que dos escalas con idéntico alfa de 0.80 pueden tener estructuras factoriales radicalmente distintas.

03

Estimación avanzada del error de medida

El modelo clásico presenta el error de medida como una cantidad aleatoria con esperanza cero y varianza constante, de lo cual se deriva el clásico Error Estándar de Medida, SEM, calculado como la desviación estándar de la prueba multiplicada por la raíz cuadrada de uno menos la confiabilidad. Sin embargo, la literatura avanzada ha evidenciado que el error de medida no es homocedástico, es decir, varía sistemáticamente según el nivel del rasgo latente que se evalúa. La evidencia empírica recopilada por Livingston en 1972, y ampliada por Raykov y Marcoulides en 2011, muestra que las pruebas psicológicas suelen medir con mayor precisión los niveles intermedios del rasgo y con menor precisión los niveles extremos.

Para evaluar con mayor rigor el error de medida dentro del propio marco clásico, autores como Feldt y Brennan en 1989 propusieron el Error Estándar de Medida Condicional, que estima el SEM para distintos niveles de puntaje observado. Otra estrategia es la aplicación de Teoría de la Generalizabilidad, iniciada por Cronbach, Gleser, Nanda y Rajaratnam en 1972, que descompone el error en facetas diferenciadas como evaluadores, ocasiones, ítems y momentos. Charter en 2003 defendió el uso del bootstrap no paramétrico para construir intervalos robustos del SEM cuando las distribuciones se alejan de la normalidad, condición frecuente en muestras clínicas latinoamericanas con asimetrías marcadas, y Kelley y Cheng en 2012 ampliaron esta línea ofreciendo estimadores bayesianos para el SEM.

04

Coeficientes de confiabilidad modernos

El coeficiente omega de McDonald, desarrollado en 1999 y popularizado por Revelle y Zinbarg en 2009 así como por Dunn, Baguley y Brunsden en 2014, estima la confiabilidad a partir de los parámetros de un modelo factorial confirmatorio, no exige tau-equivalencia y tolera ítems congéntricos. Existen variantes como omega total, omega jerárquico y omega subescala: Revelle y Zinbarg en 2009 distinguieron omega total, que captura la varianza explicada por el factor general más los factores específicos, de omega jerárquico, que captura únicamente la

varianza del factor general y penaliza la multidimensionalidad. En la adaptación argentina de la Escala de Depresión del Centro de Estudios Epidemiológicos, CES-D, la comparación entre alfa de 0.84 y omega de 0.91 evidenció que el alfa subestimaba la consistencia interna en 0.07 puntos, diferencia clínicamente relevante.

La confiabilidad compuesta, rho C, derivada del mismo modelo factorial que el omega y defendida por Raykov en 1997 y por Bagozzi y Yi en 2012, ofrece un estimador robusto de la consistencia interna dentro del marco de ecuaciones estructurales. Bonett propuso en 2002 y refinó en 2010 intervalos de confianza para la confiabilidad basados en transformaciones logarítmicas que corrigen el sesgo de estimación en muestras pequeñas, condición habitual en investigaciones de pregrado y maestría en Latinoamérica. El GLB, Greatest Lower Bound, documentado por Ten Berge y Sočan en 2004, representa el mínimo teórico de la confiabilidad, aunque Sijtsma en 2009 advirtió que su estimación muestral es sesgada hacia arriba en muestras pequeñas. Revelle y Condon en 2019 han insistido en que ningún coeficiente es universalmente óptimo, y que la práctica avanzada exige reportar varios simultáneamente.

05

Limitaciones de la TCT frente a la Teoría de Respuesta al Ítem

A pesar de las mejoras internas al modelo clásico, la Teoría Clásica de los Tests conserva limitaciones estructurales que la Teoría de Respuesta al Ítem, TRI, supera con arquitectura matemática diferente. La primera limitación es la dependencia de la muestra: la dificultad del ítem en TCT se estima como la proporción de acierto, parámetro que cambia si la muestra es más o menos competente. En TRI, siguiendo el trabajo fundacional de Lord en 1980 y Rasch en 1960 con su modelo logístico simple, los parámetros del ítem son invariantes respecto a la muestra, propiedad conocida como invariancia paramétrica. La segunda limitación es la dependencia del instrumento respecto al puntaje del sujeto: en TCT el puntaje de una persona depende de la composición específica del test, mientras que en TRI el nivel del rasgo se estima en una escala común que permite comparar sujetos que respondieron a conjuntos distintos de ítems, propiedad crucial para la equiparación de pruebas en programas como las Pruebas Saber del ICFES en Colombia, el ENADE brasileño o el SIMCE chileno.

La tercera limitación es el tratamiento homogéneo del error: TRI entrega funciones de información por ítem y por test que permiten diseñar pruebas adaptativas computarizadas, CAT, las cuales seleccionan los ítems más informativos para cada sujeto. Pilotos de CAT con versiones adaptadas del MMPI-2 y de inventarios académicos han mostrado reducciones de hasta 60% en la cantidad de ítems administrados sin pérdida de precisión, como reportan Abad, Olea, Ponsoda y García en su texto de 2011. Una cuarta limitación es la dificultad del modelo clásico para manejar el funcionamiento diferencial del ítem, DIF, entre subgrupos, por género, por región o por nivel socioeconómico; aunque existen procedimientos clásicos como

Mantel-Haenszel, propuestos por Holland y Thayer en 1988, los métodos basados en TRI resultan más informativos y han sido aplicados intensivamente en Colombia para analizar DIF en pruebas Saber Pro entre regiones Caribe, Andina y Pacífica. Estas limitaciones explican el desplazamiento progresivo del campo hacia TRI sin abandonar TCT, pues esta sigue siendo útil como marco introductorio.

06

Cierre: síntesis y estándares profesionales para el psicólogo latinoamericano

La revisión avanzada de la Teoría Clásica de los Tests consolidada en esta sesión permite al estudiante latinoamericano operar con criterio técnico entre la herencia del modelo clásico y las alternativas contemporáneas. Los supuestos de tau-equivalencia, homocedasticidad del error y estabilidad de los parámetros entre muestras han sido matizados por cuatro décadas de literatura especializada, que culminan en recomendaciones prácticas concretas: reportar alfa acompañado de omega en sus variantes, añadir rho C cuando se trabaja con SEM, usar Bonett para muestras pequeñas, e incorporar Teoría de la Generalizabilidad cuando las facetas del error son múltiples. Los estándares internacionales de AERA, APA y NCME en su edición de 2014 hoy exigen reportar simultáneamente índices clásicos y modernos, lo cual convierte este dominio avanzado de TCT en un requisito profesional y no en un contenido histórico.

Este estándar ya opera en la región: Colombia cuenta con la Ley 1090 de 2006, que regula el ejercicio de la psicología y exige calidad técnica en los instrumentos empleados, y las revistas regionales de mayor impacto, como Revista Latinoamericana de Psicología, Universitas Psychologica y Acta Colombiana de Psicología, han endurecido en los últimos cinco años sus requisitos psicométricos, exigiendo el reporte conjunto de alfa, omega y análisis factorial confirmatorio como condición mínima para revisión por pares. La estimación de estos coeficientes modernos requiere programas especializados: R, con paquetes como psych de Revelle y lavaan de Rosseel, junto con Jamovi y JASP, han democratizado el acceso a omega y rho C en facultades de psicología latinoamericanas donde la licencia de programas comerciales resulta prohibitiva. El modelo clásico conserva valor didáctico, operativo y comunicativo, pero sus limitaciones frente a la Teoría de Respuesta al Ítem explican el desplazamiento progresivo del campo psicométrico hacia TRI, especialmente en evaluaciones de alto impacto.