REVISTA PSICOLOGÍA DE LA SALUD

Vol. 13. Núm. 1. 2025. Editorial
DOI: https://doi.org/10.21134/pssa.v13i1.20

¿MÁS ES MEJOR? LA PROLIFERACIÓN DE INSTRUMENTOS DE MEDIDA EN PSICOLOGÍA

Sonia Tirado Gonzalez

Editora Asociada Revista de Psicología de la Salud/Journal of Health Psychology

La psicología, como ciencia del comportamiento humano, se ha caracterizado a lo largo de su historia por la búsqueda de herramientas fiables y válidas para medir los constructos psicológicos, tarea nada sencilla dado el carácter no observable de muchos de los procesos psicológicos y el hecho de que las medidas de los síntomas no pueden asumirse como indicadores perfectos de los atributos que se están evaluando. En palabras de Edwards y Bagozzi (2000), "estos fenómenos no pueden conocerse de manera directa o con total precisión debido al error de medición y a la imperfecta lente epistemológica que proporciona un constructo". Sea como fuere, la realidad es que el uso de instrumentos de evaluación es continuo tanto en la práctica profesional como de investigación en todos los ámbitos de actuación de la psicología.

Sin embargo, los tests no solo han permitido obtener información válida y fiable de los sujetos bajo estudio, sino que han realizado contribuciones fundamentales al desarrollo del corpus teórico de nuestra disciplina. La operacionalización de los constructos teóricos en medidas empíricas, la validación de modelos mediante técnicas psicométricas que posibilitan la identificación de estructuras latentes y la retroalimentación teórica fruto del análisis de resultados novedosos o inconsistentes son, entre otras, aportaciones de los tests que han contribuido al carácter científico de la psicología. Todo ello gracias, en buena medida, al gran desarrollo metodológico y psicométrico necesario para su construcción y validación.

Precisamente este creciente desarrollo metodológico, unido a la democratización y "amigabilidad" del software de análisis de datos, ha sido una de las razones por las que, a lo largo de este recorrido, hemos presenciado un crecimiento exponencial de instrumentos de medida. Pero no es la única. Una conjunción de circunstancias inherentes a la evolución de nuestra disciplina podría estar a la base de esta proliferación, como la presión académica por la publicación de originales y la financiación de proyectos inherente a la carrera académica (Johan et al., 2024; Mischel, 2008), la creciente especialización y fragmentación de las áreas de investigación (de Aguiar-Carvalho, 2016; Henriques, 2022;), la dificultad para llegar a consensos en la definición de constructos (Borsboom et al., 2021; Fried, 2017; Shaffer, 2016), la variabilidad de directrices para el desarrollo de instrumentos (Elson et al., 2023; Flake y Fried, 2020; Hunsley y Meyer, 2003) o el predominio del pensamiento operacionalista y el enfoque de la Teoría Clásica de Tests que facilitan la creación de nuevos instrumentos sin suficiente fundamentación teórica (Borsboom, 2006; Borsboom et al., 2021; Edwards et al., 2000), entre otras cuestiones, están también en el origen de esta multiplicidad.

La psicología de la salud no es ajena a esta tendencia, tal como nos muestran diferentes revisiones sistemáticas sobre cuestionarios utilizados en este ámbito. Por citar algunos resultados recientes, encontramos en la literatura 14 cuestionarios que evalúan la adherencia a la medicación en pacientes adultos con patologías crónicas (seis de ellos en español) (Intilangelo et al., 2024), 11 instrumentos de evaluación de la calidad de Atención Primaria (únicamente en los últimos 10 años) (López y Barber, 2024), o 44 instrumentos para medir la relación médico-paciente (35 inglés y nueve en español) (Lazzaro-Salazar y Pujol-Cols, 2022), entre otros. Especialmente llamativos son los 51 instrumentos para evaluar la calidad de vida relacionada con la salud documentados por Haraldstad et al. (2019), y los nada menos que 115 cuestionarios para medir desregulación emocional en niños y adolescentes identificados por Freitag et al. (2023).

Esta proliferación de instrumentos refleja el dinamismo y la complejidad de nuestra disciplina, al tiempo que plantea importantes desafíos metodológicos, teóricos y prácticos, de manera que se ha convertido en fuente de cierta controversia, tal como refleja la literatura.

Por un lado, Iliescu et al. (2024) han defendido recientemente la tesis de que la proliferación y variabilidad de las medidas psicológicas son parte natural del proceso científico y contribuyen al desarrollo teórico, a la transparencia, la replicabilidad y la validez. Cuestiones como la especificidad y adaptación de los instrumentos a diferentes contextos y poblaciones, de manera que los equipos investigadores o clínicos pueden seleccionar instrumentos que se ajusten mejor a sus necesidades específicas, considerando variables culturales, lingüísticas y contextuales; el hecho de que las relaciones entre constructos teóricos son generalizables solo en la medida en que sean confirmadas en diferentes muestras, con medidas diferentes y en diferentes contextos; o la posibilidad de "cartografiar" todas las perspectivas teóricas y capturar matices y dimensiones de los constructos psicológicos que podrían pasar desapercibidos con

instrumentos más generales, sustentan su posicionamiento. Es más, plantean que restringir esta proliferación podría obstaculizar el progreso de la psicología como ciencia.

Sin embargo, este crecimiento no está exento de riesgos y, en efecto, gran parte de la literatura lo considera una barrera para el avance de la ciencia. Según nos muestran Elson et al. (2023), la mayoría de los cuestionarios psicológicos se usan solo una o dos veces y no hay una tendencia a que los investigadores utilicen las mismas medidas a lo largo del tiempo. De hecho, aplican el concepto de "cepillo de dientes" de Mischel (2008) a los instrumentos de medida en psicología: "Ningún psicólogo que se precie quiere usar el de otra persona". Esta circunstancia pone el foco en la falta de validez de los instrumentos utilizados (Bergkvist, 2021).

En este sentido, los instrumentos no siempre alcanzan los estándares mínimos de calidad, en parte por haber sido desarrollados siguiendo lo que Flake y Fried (2020) denominan "prácticas de medición cuestionables", es decir, todas aquellas decisiones, por falta de transparencia, ignorancia, negligencia o tergiversación, que toman los investigadores, que afectan a la validez interna, externa, de conclusión estadística y de constructo. Paniagua (2023) las clasifica en tres categorías: teoría y diseño, ejecución y redacción, resultando en un total de 14 prácticas, tal como refieren en su artículo.

Esta multiplicidad de instrumentos puede producir, además, una fragmentación del conocimiento. La falta de consenso sobre la naturaleza de los constructos, la inconsistencia en su nomenclatura, el hecho de que diferentes medidas de un mismo constructo evalúen aspectos parcialmente distintos del mismo, o el desarrollo de constructos conceptualmente similares con nombres diferentes son algunas de las razones que dificultan tanto la comparación entre estudios como la acumulación sistemática de evidencia científica (de Aguiar Carvalho et al., 2016; Bergkvist, 2021; Bergkvist y Langner, 2019; Le et al., 2010; Mischel, 2008; Shaffer et al., 2016). Esta fragmentación teórica inducida por la medida dificulta la consolidación de marcos teóricos comprehensivos y sólidos de la conducta humana y obstaculiza –o ralentiza– el avance del conocimiento, cuestiones que están en la base de la crisis de replicabilidad que atraviesa nuestra disciplina (Centeno-Leyva y Dominguez, 2020; Maxwell et al., 2015; Schnep y Groeben, 2024).

En este contexto, se hace imprescindible proponer soluciones constructivas. Un aspecto clave es garantizar un procedimiento riguroso en la construcción y el análisis de los instrumentos. La selección, adaptación o elaboración de un test debe realizarse sobre la base de directrices validadas por la comunidad científica y profesional. Entre las principales destacan las de la Comisión Internacional de Tests (ITC) (Hernández et al., 2020; International Test Commission, 2012, 2014, 2017; Muñiz et al., 2013), recopiladas y traducidas por el Consejo General de la Psicología de España, o las de la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education (2018). Estas directrices proporcionan la base para evaluar la calidad de los instrumentos, tanto en su construcción y evaluación como en la interpretación de las puntuaciones obtenidas.

En el panorama internacional encontramos también la iniciativa COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments), que establece directrices para la selección de instrumentos de medida en el ámbito de la salud (www.cosmin.nl). Además de ofrecer pautas para seleccionar y evaluar estos instrumentos (Mokkink et al., 2024; Prinsen et al., 2016), COSMIN dispone de una base de datos de revisiones sistemáticas que incluye el análisis de numerosas pruebas utilizadas en investigación en salud.

Precisamente, la creación de repositorios es otra propuesta señalada en la literatura. Autores como Elson (2023) e Iliescu (2024) plantean que estos repositorios, preferiblemente de acceso abierto, deberían contener los instrumentos más utilizados, sus baremos y normas de puntuación, así como metadatos que faciliten el procesamiento automatizado. Como ventajas, señalan la agilidad en el acceso a estos instrumentos, el análisis de las decisiones tomadas en la fase de diseño y la evaluación sistemática de las prácticas de investigación, así como permitir el análisis de otras cuestiones como la coincidencia semántica, la fiabilidad o la validez convergente y discriminante en comparación con los ya registrados.

En el ámbito específico de la salud encontramos los repositorios PROMIS©, Neuro-QoL, AS-

CQ-Me y NIH Toolbox, impulsados por el National Institutes of Health (NIH), que incluyen instrumentos de salud física, mental y social tanto en población general como para pacientes con enfermedad crónica, algunas de ellas disponibles en español (www.healthmeasures.net/).

En el contexto español, el Consejo General de la Psicología, a través de su Comisión de Tests, ofrece una base de datos de tests evaluados mediante el Cuestionario de Evaluación de Tests, actualmente en su versión revisada (CET-R) (Hernández et al., 2016; Prieto y Muñiz, 2000). Esta herramienta permite valorar instrumentos creados según la Teoría Clásica de los Tests, que incorpora recomendaciones del modelo de revisión, descripción y evaluación de tests psicológicos y educativos de la Federación Europea de Asociaciones de Psicología (EFPA) (Evers et al., 2013). Cabe destacar que, a comienzos de este año, la EFPA ha lanzado una consulta para revisar dicho modelo, lo que subraya la importancia de mantener los estándares psicométricos actualizados y alineados con las últimas investigaciones y desarrollos.

Estas cuestiones ponen de manifiesto la necesidad de modernizar y reforzar la formación metodológica en psicología, así como de reflexionar sobre el papel de la psicometría. Tal como plantea Borsboom (2006), resulta imprescindible revisar y actualizar la formación en medición, así como fomentar la colaboración entre psicómetras y profesionales de otras áreas de la disciplina. Dicha cooperación podría aportar a la construcción de una base teórica sólida que contribuya positivamente a la construcción y el uso de herramientas de evaluación. Asimismo, también propone impulsar el uso de software que vaya más allá de la Teoría Clásica de los Tests para mejorar la modelización de las relaciones entre variables latentes y respuestas observadas. Según el autor, estos pasos ayudarían a cerrar la brecha entre la práctica cotidiana en psicología y los avances técnicos en la ciencia de la medición.

Por otro lado, dado el papel que pueden desempeñar las revistas científicas en la estandarización y control de la calidad de los instrumentos publicados, se sugiere establecer criterios más rigurosos para la publicación de nuevos tests. En este sentido, Elson et al. (2023) introducen las directrices SOBER (Standardisation Of BEhavior Research), que contemplan cuestiones como la demostración de que los instrumentos que se publican no son redundantes, que han sido validados con datos independientes y que se ha justificado cualquier modificación en los mismos, entre otras. Se plantea que los editores y revisores deberían velar por el cumplimiento de estas pautas para garantizar la validez y la transparencia de los estudios, aplicando, eso sí, los cambios en sus políticas de publicación de forma gradual. En la misma línea, Hunsley y Meyer (2003) sostienen que se debe exigir evidencia de validez incremental, es decir, demostrar lo que aporta un nuevo instrumento más allá de los ya existentes en términos de predicción y utilidad. Esto impediría la proliferación de herramientas sin un claro valor añadido. Algunos autores van más allá y plantean que solo deberían desarrollarse nuevos tests si existe una justificación sólida y no hay instrumentos previamente validados (Bergkvist y Langner, 2019; Flake y Fried, 2020), o incluso abogan directamente por el uso de instrumentos estandarizados (Bergkvist, 2021; Bergkvist et al., 2019). En cualquier caso, es preciso también implementar medidas que garanticen la reproducibilidad y favorezcan la acumulación de conocimiento a través de estudios de replicación, revisiones sistemáticas y meta-análisis de los instrumentos de medida. De este modo, se validaría la eficacia y la aplicabilidad de los ya existentes, al tiempo que se evitaría el desarrollo innecesario de otros nuevos (Maxwell et al., 2015).

En última instancia, estas propuestas descansan en la transparencia como pilar esencial de la construcción, adaptación, validación y uso de instrumentos. Tal como destacan Flake y Fried (2020), la opacidad en las decisiones metodológicas dificulta la identificación de las amenazas a la validez, la replicabilidad de los estudios y la acumulación de conocimiento. Por ello, cada vez hay más consenso a favor de prácticas de ciencia abierta tales como el pre-registro público del plan de investigación, el acceso abierto a la información y la revisión en dos fases (una previa a la recogida de datos y otra tras finalizar el estudio) (Elson et al., 2023; Fried y Flake, 2018; Hardwicke y Wagenmakers, 2023; Ludwik, et al, 2023; Nosek et al., 2019). En definitiva, la transparencia no solo asegura que las mediciones reflejen con exactitud los constructos de interés, sino que también promueve una cultura científica basada en la responsabilidad y el avance acumulativo del conocimiento.

En síntesis, la proliferación de instrumentos de medida en psicología refleja tanto el dinamismo como la complejidad de nuestra disciplina, al tiempo que plantea retos que requieren res-

puestas colectivas y rigurosas que pasan, no por frenar la producción científica y la creatividad de los equipos investigadores, sino por alinear los esfuerzos de todos los profesionales implicados para alcanzar un equilibrio entre la necesaria adaptación de las herramientas a contextos y poblaciones específicas y la importancia de mantener un corpus teórico coherente y acumulativo. Para ello, es imprescindible acogerse a guías y directrices compartidas por la comunidad científica, fomentar la transparencia metodológica y apostar por una formación avanzada en medición. Del mismo modo, la implementación de prácticas de ciencia abierta y la creación de repositorios de acceso público pueden facilitar la validación, replicabilidad y comparabilidad de los tests, reforzando así la confianza en los resultados. En definitiva, el avance de la psicología como ciencia dependerá, en buena medida, de la solidez teórica de los instrumentos, la calidad y transparencia del proceso de construcción y el compromiso ético con la mejora de la práctica profesional.

Referencias

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* **(M. Lieve, Trans.). Washington, DC: American Educational Research Association. (Original work published 2014)**

Bergkvist, L. (2021). Perspectives: Measure proliferation in advertising research: Are standard measures the solution? *International Journal of Advertising: The Review of Marketing Communications*, *40(2)*, **311–323.** https://doi.org/10.1080/02650487.2020.1753442.

Bergkvist, L., & Langner, T. (2019). Construct heterogeneity and proliferation in advertising research. *International Journal of Advertising: The Review of Marketing Communications*, *38(8)*, 1286–1302. https://doi.org/10.1080/02650487.2019.1622345

Borsboom D. (2006). The attack of the psychometricians. *Psychometrika*, *71(3)*, 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16(4)*, 756-766. https://doi.org/10.1177/1745691620969647

Centeno-Leyva, S., & Dominguez Lara, S. (2020). Replicability in psychological research: a reflection. *Interacciones*, *6(3)*, e172. https://doi.org/10.24016/2020.v6n3.172

de Aguiar Rodrigues, A. C., & de Carvalho-Freitas, M. N. (2016). Theoretical fragmentation: Origins and repercussions in Work and Organizational Psychology. *Revista Psicologia Organizações e Trabalho*, *16(4)*, 310–315. https://doi.org/10.17652/rpot/2016.4.12630

Edwards, J. R., y Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological methods*, 5(2), 155–174. https://doi.org/10.1037/1082-989x.5.2.155

Elson, M., Hussey, I., Alsalti, T., & Arslan, R.C. (2023). Psychological measures aren't toothbrushes. *Communications Psychology*, *1*, 25. https://doi.org/10.1038/s44271-023-00026-9

Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA test review model for the description and evaluation of psychological and educational tests* (EFPA Board of Assessment Document 110c, Version 4.2.6). European Federation of Psychologists' Associations. https://www.efpa.eu

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3(4)*, 456-465.

Freitag, G. F., Grassie, H. L., Jeong, A., Mallidi, A., Comer, J. S., Ehrenreich-May, J., & Brotman, M. A. (2023). Systematic Review: Questionnaire-Based Measurement of Emotion Dysregulation in Children and Adolescents. *Journal of the American Aca-*

demy of Child and Adolescent Psychiatry, *62(7)*, 728–763. https://doi.org/10.1016/j.jaac.2022.07.866

Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*, *11(2)*, 130–134. https://doi.org/10.1080/17437199.2017.1306718

Fried, E. I. & Flake, J.K. (2018, March). Measurement Matters. *Observer*. Consultado en https://www.psychologicalscience.org/observer/measurement-matters.

Haraldstad, K., Wahl, A., Andenæs, R., Andersen, J. R., Andersen, M. H., Beisland, E., Borge, C. R., Engebretsen, E., Eisemann, M., Halvorsrud, L., Hanssen, T. A., Haugstvedt, A., Haugland, T., Johansen, V. A., Larsen, M. H., Løvereide, L., Løyland, B., Kvarme, L. G., Moons, P., Norekvål, T. M., ... LIVSFORSK network (2019). A systematic review of quality of life research in medicine and health sciences. *Quality of life research*, *28(10)*, 2641–2650. https://doi.org/10.1007/s11136-019-02214-9

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature human behaviour*, *7(1)*, 15–26. https://doi.org/10.1038/s41562-022-01497-2

Henriques, G. (2022). *A new synthesis for solving the problem of psychology: Addressing the Enlightenment Gap*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-18493-2

Hernández, A., Hidalgo, M.D., Hambleton, R.K., Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema*, *32(3)*, 390-398. doi: 10.7334/psicothema2019.306

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G. y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, *37*, 192-197.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological assessment*, *15(4)*, 446–455. https://doi.org/10.1037/1040-3590.15.4.446

Iliescu, D., Greiff, S., Ziegler, M. et al. (2024). Proliferation of measures contributes to advancing psychological science. *Communications Psychology*, *2 (19)*. https://doi.org/10.1038/s44271-024-00065-w

International Test Commission (2012). *International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores*. [www.intestcom.org]

International Test Commission (ITC) (2014) ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. *International Journal of Testing*, *14:3*, 195-217, DOI: 10.1080/15305058.2014.918040

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests* (Second edition). [www.InTestCom.org]

Intilangelo, A., Majic, S., Palchik, V., & Traverso, M. L. (2024). Validated medication adherence questionnaires and associated factors in chronic patients: Systematic review. Cuestionarios validados de adherencia a la medicación y factores asociados en pacientes crónicos: revisión sistemática. *Farmacia hospitalaria: órgano oficial de expresión científica de la Sociedad Española de Farmacia Hospitalaria*, *48(4)*, 185–192. https://doi.org/10.1016/j.farma.2024.01.001

Johann, D., Neufeld, J., Thomas, K., Rathmann, J., & Rauhut, H. (2024). The impact of researchers' perceived pressure on their publication strategies. *Research Evaluation*, 00: 1-16. https://doi.org/10.1093/reseval/rvae011

Lazzaro-Salazar, Mariana, & Pujol-Cols, Lucas. (2022). Instrumentos estandarizados para medir la relación médico-paciente: una revisión sistemática de la literatura internacional e iberoamericana. *Revista médica de Chile*, *150(4)*, 512-531. https://dx.doi.org/10.4067/S0034-98872022000400512

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112(2)*, 112–125. https://doi.org/10.1016/j.obhdp.2010.02.003

López y Barber (2024). Revisión sistemática de los instrumentos de evaluación de la calidad de Atención Primaria utilizados en los últimos 10 años. *Atención Primaria 56(9)*:1203046. https://doi.org/10.1016/j.aprim.2024.103046

Ludwig, T., Altenmüller, M. S., Schramm, L. F. F., & Twardawski, M. (2023). Evading Open Science: The Black Box of Student Data Collection. *Social Psychological Bulletin*, *18*, 1-31. https://doi.org/10.32872/spb.9411

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70(6)*, 487–498. https://doi.org/10.1037/a0039400

Mischel, W. (2008). The toothbrush problem. *APS Observer*, *21*. Consultado en https://www.psychologicalscience.org/observer/the-toothbrush-problem.

Mokkink, L. B., Elsman, E. B.M., & Terwee, C. B. (2024). Cosmin guideline for systematic reviews of patient-reported outcome measures version 2.0. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*. Advance online publication. https://doi.org/10.1007/s11136-024-03761-6

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: Segunda edición [International Test Commission Guidelines for test translation and adaptation: Second edition]. *Psicothema*, *25(2)*, 151–157.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in cognitive sciences*, *23(10)*, 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Paniagua, D., Sánchez-Iglesias, I., Miguel-Alvaro, A., Casas-Aragonez, N., Aparicio-Garcia, M. E., & Aguayo-Estremera, R. (2023). Prácticas Cuestionables en Estudios de Validez de Instrumentos de Medición Psicológica: Comunalidades y Unicidades de la Crisis de Replicabilidad en el Campo de la Psicometría. *Revista Iberoamericana de Diagnóstico y Evaluación Psicologica*, *66(5)*, 23–34. https://doi.org/10.21865/RIDEP66.5.02

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, *77*, 65-72.

Prinsen, C. A., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials*, *17(1)*, 449. https://doi.org/10.1186/s13063-016-1555-2

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the Problem of Construct Proliferation: A Guide to Assessing the Discriminant Validity of Conceptually Related Constructs. *Organizational Research Methods*, *19(1)*, 80-110. https://doi.org/10.1177/1094428115598239

Schnepf, J., & Groeben, N. (2024). *The replication crisis as mere indicator of two fundamental misalignments: methodological confirmation bias in hypothesis testing and anthropological oversimplification in theory-building* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.4797750