Volumen 22 Número 1 Junio 2025 ISSN: 1578-908X

# ACCIÓN PSICOLÓGICA

Editor jefe
**Ricardo Pellón Suárez de Puga**

Monográfico:

## Nuevos avances metodológicos en Psicología

Coordinadores:
**Belén Fernández-Castilla y**
**José Ángel Martínez-Huertas**

UNED

Open Acces online
http://revistas.uned.es/index.php/accionpsicologica

**Acción Psicológica**
REVISTA SEMESTRAL DE PSICOLOGÍA
VOLUMEN 22, NÚMERO 1, JUNIO 20245
ISSN: 2255-1271

*Acción Psicológica* es una revista semestral editada por la Facultad de Psicología de la Universidad Nacional de Educación a Distancia desde el año 2002. Publica artículos originales e inéditos de investigación, de revisión, contribuciones teóricas o metodológicas, como también estudios de casos sobre diversas áreas de la Psicología.

**Acción Psicológica**
REVISTA SEMESTRAL DE PSICOLOGÍA
VOLUMEN 22, NÚMERO 1, JUNIO 2025 ISSN: 2255-1271

*Acción Psicológica* es una revista semestral editada por la Facultad de Psicología de la Universidad Nacional de Educación a Distancia desde el año 2002. Publica artículos originales e inéditos de investigación, de revisión, contribuciones teóricas o metodológicas, como también estudios de casos sobre diversas áreas de la Psicología.

## NORMAS PARA EL ENVÍO Y PUBLICACIÓN DE TRABAJOS

***Acción Psicológica*** publica artículos originales e inéditos de investigación, de revisión, contribuciones teóricas o metodológicas, como también estudios de casos sobre diversas áreas de la Psicología.
Las normas de envío de originales se detallan en la web de la revista:
http://revistas.uned.es/index.php/accionpsicologica/about/submissions#onlineSubmissions



**Contacto:** Facultad de Psicología (UNED). C/ Juan del Rosal nº 10, 28040 Madrid, Spain. Email: accionpsicologica@psi.uned.es

**Acción Psicológica**
SEMIANNUAL JOURNAL OF PSYCHOLOGY
VOLUME 22, NUMBER 1, JUNY 2025 - ISSN: 2255-1271

*Acción Psicológica* is a semiannual journal published by the Faculty of Psychology of the Universidad Nacional de Educación a Distancia (UNED) since 2002. Publishes original research, review, theoretical or methodological contributions, as well as case studies on different areas of Psychology.

**INSTRUCTIONS FOR AUTHORS**

**Manuscript Preparation**

Prepare manuscripts according to the https://apastyle.apa.org/products/publication-manual-7th-edition
The manuscripts will be upload in formats: ".doc" or ".docx" in the journal website http://revistas.uned.es/index.php/accionpsicologica/about/submissions#onlineSubmissions Manuscripts will be accepted in English or Spanish languages. If the paper is written in English, an abstract of 100-200 words in Spanish will be required
The articles will be double-spaced in **Times New Roman, 12 point**, with all margins to 1 in. The maximum length of articles will be **6000 words** (including title, abstract, references, figures, tables and appendices). The numbering of the pages will be located in the upper right.
Other formatting instructions, as well as instructions on preparing tables, figures, references, metrics, and abstracts, appear in the *Manual*.
The articles are scholarly peer-reviewed.

*Acción Psicológica* is indexed in the following databases:

Bibliographical International: Emerging Source Citation Index (ESCI-Clarivate Analytics), Academic Search Complete, Academic Search Premier and Fuente Academica Plus (EBSCO), ProQuest Psychology Journals, ProQuest Central, ProQuest Central K-12, ProQuest Health Research Premiun Collectium, ProQuest Hospital Premiun Collection, DOAJ, FirstSearch (OCLC), PubPsych (ZPID), SciELO, Open J-Gate, Dialnet, e-Revistas, Redalyc.

Bibliographical National: Compludoc, ISOC (CSIC-CINDOC), Psicodoc, Psyke.

Web site of the journal, with information, index, abstracts and full text (in pdf format) of articles:
http://revistas.uned.es/index.php/accionpsicologica/index

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Facultad de Psicología

# ACCIÓN PSICOLÓGICA

Monográfico:

Nuevos avances metodológicos

en Psicología

Coordinadores:

**Belén Fernández-Castilla y**

**José Ángel Martínez-Huertas**

VOLUMEN 22 JUNIO 2025 NÚMEROS 1

# ÍNDICE

# CONTENTS

# INTRODUCCIÓN AL NÚMERO ESPECIAL: NUEVOS AVANCES METODOLÓGICOS EN PSICOLOGÍA

# INTRODUCTION TO THE SPECIAL ISSUE: NEW METHODOLOGICAL ADVANCES IN PSYCHOLOGY

BELÉN FERNÁNDEZ-CASTILLA[1] Y
JOSÉ ÁNGEL MARTÍNEZ-HUERTAS[1]

Cómo referenciar este artículo/How to reference this article:

Fernández-Castilla, B. y Martínez-Huertas, J. A. (2025). Introducción al número especial: nuevos avances metodológicos en Psicología [Introduction to the Special Issue: New Methodological Advances in Psychology]. *Acción Psicológica, 22*(1), 1–6. https://doi.org/10.5944/ap.22.1.44367

## Resumen

Existen muchos avances metodológicos en la literatura científica. Sin embargo, su impacto en la investigación empírica suele ser limitado. En nuestra opinión, una de las causas de esta falta de impacto es la falta de accesibilidad a muchas de estas herramientas metodológicas modernas a través de textos asequibles. En esa línea, este número especial se compone de una serie de artículos que tratan de presentar una introducción a algunos de estos avances metodológicos de manera accesible. Así, se representa una selección de líneas de investigación en metodología dentro del ámbito de la Psicología que actualmente están ganando relevancia a nivel nacional e internacional. Aunque este número especial no pretende ser una recopilación exhaustiva de todas las contribuciones metodológicas disponibles en la literatura científica, consideramos que los artículos incluidos ilustran avances significativos y tendencias emergentes en este campo. Los trabajos seleccionados pueden agruparse en tres grandes disciplinas metodológicas ampliamente estudiadas e

**Correspondence address [Dirección para correspondencia]**: José Ángel Martínez-Huertas, Facultad de Psicología, Universidad Nacional de Educación a Distancia (UNED), Madrid. C/ Juan del Rosal 10, 28040, Madrid, España.
**Email**: jamartinez@psi.uned.es

**ORCID**: Belén Fernández-Castilla (https://orcid.org/0000-0002-3451-0637) y José Ángel Martínez-Huertas (https://orcid.org/0000-0002-6700-6832).

[1] Universidad Nacional de Educación a Distancia (UNED), Madrid, España.

**Agradecimientos**: Nos gustaría agradecer a los autores el esfuerzo puesto en preparar estos artículos, a los revisores expertos por aceptar nuestra invitación altruistamente y por sus muy acertados comentarios y sugerencias, y al equipo editorial de *Acción Psicológica* por habernos dado el espacio para hacer realidad este número especial

Recibido: 5 de febrero de 2025.
Aceptado: 6 de febrero de 2025.

impartidas en Psicología: psicometría, modelado, y diseño de investigaciones. Aquí presentamos los trabajos seleccionados enmarcados en cada una de estas áreas metodológicas, y destacamos su importancia para el futuro de la investigación en Psicología.

**Palabras clave:** Metodología; Psicometría; Modelado, Diseños de investigación.

## Abstract

There are many methodological advances in scientific literature. However, their impact on empirical research is often limited. In our opinion, one of the reasons for this lack of impact is the limited accessibility of many of these modern methodological tools through easy-to-understand texts. In this regard, this special issue consists of a series of articles that introduce some of these methodological advances in an accessible way. It represents a selection of methodological research lines within the field of psychology that are currently gaining relevance at both national and international levels. Although this special issue does not intend to be an exhaustive compilation of all methodological contributions available in scientific literature, we believe that the included articles illustrate significant advances and emerging trends in this field. The selected works can be grouped into three major methodological disciplines that are widely studied and taught in psychology: psychometrics, modeling, and research design. Here, we present the selected studies within these methodological areas and highlight their importance for the future of psychological research.

**Palabras clave:** Methodology; Psychometrics; Modeling; Research designs.

## Nuevos avances psicométricos

Dentro del ámbito de la psicometría, hemos sido testigos de una evolución en la forma de abordar la medición y comprensión de los constructos psicológicos. Tradicionalmente, los modelos psicométricos se han centrado en el estudio de variables latentes (es decir, no observables) que se infieren a partir de la covarianza compartida de una serie de indicadores, asumiendo que estas subyacen y explican las relaciones observadas entre los ítems de un test. Sin embargo, en los últimos años, ha surgido un enfoque alternativo basado en los modelos de redes, que conceptualizan los constructos psicológicos no como entidades latentes, sino como sistemas dinámicos de interacciones entre los ítems (Borsboom et al., 2022). Este cambio de paradigma ha permitido explorar nuevas preguntas de investigación, como el papel de los síntomas o comportamientos específicos en el mantenimiento de los trastornos psicológicos, ofreciendo una visión más dinámica de los fenómenos psicológicos (Epskamp, 2020; Isvoranu et al., 2022). En el manuscrito de Fonseca y Muñiz, titulado "Análisis de Redes en la Medición Psicológica: Fundamentos" se presenta una introducción al análisis psicométrico de redes psicológicas, diseñado para familiarizar a los lectores con esta metodología, sus ventajas y limitaciones y los retos que enfrenta el futuro de este campo de investigación.

Por otro lado, las pruebas de elección forzosa han ganado popularidad en la psicometría por su capacidad para abordar limitaciones clave de los formatos tradicionales de respuesta como, por ejemplo, los sesgos de respuesta (e.g., sesgo de deseabilidad social o de respuesta extrema). Este enfoque, que requiere que los participantes seleccionen entre opciones igualmente atractivas o desfavorables, permite una evaluación más precisa de los constructos psicológicos, ya que reduce la subjetividad y aumenta la validez de las interpretaciones. Su creciente aplicación e integración con modelos psicométricos avanzados, como la teoría de respuesta al ítem (e.g., Thurstone IRT; Brown y Maydeu-Olivares, 2011; Multi-Unidimensional Pairwise-Preference; Stark et al., 2005), hacen que este formato de respuesta sea una herramienta metodológica fundamental para el desarrollo de instrumentos psicométricos más robustos. En el manuscrito de Abad y colaboradores, titulado "Pruebas de elección forzosa: visión actual y recomendaciones", se explican este tipo de pruebas desde la teoría de respuesta al ítem, y comentan sus principales ventajas y limitaciones. Además, los autores también ofrecen una serie de recomendaciones prácticas y pasos a seguir para el diseño e implementación de pruebas de elección forzosa,

y proporcionan a los lectores materiales prácticos para replicar los ejemplos aportados.

## El cambio de los individuos como foco de interés

Dos artículos del número especial se han centrado en el estudio del cambio de los individuos a través de distintos diseños y distintos modelos estadísticos. Por un lado, se incluye un artículo centrado en estudios de diseño de caso único, escrito por Manolov: "Diseño y análisis de datos de diseños experimentales de caso único". A diferencia de los diseños basados en grupos, que suelen enfocarse en promedios grupales, los diseños de caso único permiten una evaluación detallada de la variabilidad y los efectos de una intervención en un solo individuo, lo que resulta crucial para personalizar tratamientos y enfoques terapéuticos (Bono y Arnau, 2014; Shadish y Sullivan, 2011). Además, este diseño de investigación puede ser el único factible para estudiar la eficacia de tratamientos de condiciones poco comunes como, por ejemplo, enfermedades raras. No obstante, la interpretación de los resultados de estos estudios puede ser compleja debido a la disponibilidad limitada de datos y a su variabilidad. Precisamente por esto, el análisis estadístico se vuelve fundamental, ya que solo la aplicación de técnicas robustas en este contexto permite extraer conclusiones válidas. En este artículo, además de revisar diversos diseños de caso único y técnicas de análisis, se proporcionan criterios para orientar a los investigadores en la selección del diseño y análisis más adecuado, junto con una serie de recursos gratuitos que facilitan el análisis estadístico de este tipo de datos.

Por otro lado, el segundo artículo incluido en esta sección consiste en una revisión de los métodos estadísticos que se pueden implementar para analizar y estudiar procesos de cambio en datos longitudinales: "Una revisión de conceptos y métodos para investigar con datos longitudinales", escrito por Estrada y colaboradores. Los estudios longitudinales son fundamentales en Psicología porque permiten estudiar el cambio y la evolución de los fenómenos psicológicos a lo largo del tiempo, ofreciendo una visión más dinámica y precisa de cómo los individuos y grupos cambian, se desarrollan, o responden a intervenciones (Ferrer et al., 2018; Hoffman, 2015; Menard, 2007). En este trabajo, se introduce al lector a los aspectos clave para estudiar procesos de cambio a lo largo del tiempo, y se realiza una introducción a modelos estadísticos sofisticados y novedosos para analizar el funcionamiento y la evolución de procesos psicológicos.

## Replicabilidad y síntesis de estudios

Hace ya una década, la publicación de Open Science Collaboration (2015) reveló que solo un tercio de los estudios en Psicología lograba replicarse, marcando un hito en la preocupación de los investigadores por la replicabilidad y reproducibilidad de las investigaciones empíricas. No sólo se comenzó a evaluar la replicabilidad de los resultados de los estudios publicados en otros campos de conocimiento (e.g., Camerer et al., 2016, en Economía; y Errington, et al., 2021, en Medicina), sino que comenzaron a proliferar propuestas prácticas sobre cómo mejorar la reproducibilidad y replicabilidad de nuestros trabajos de investigación (Nosek et al., 2022), y a la vez reducir los grados de libertar del investigador para manipular, explícita o implícitamente, la metodología implementada para obtener los resultados deseados (Simmons et al., 2011). En el artículo escrito por Lecuona y colaboradores titulado "Credibilidad o barbarie: Cómo la crisis de replicación ha desatado una revolución en Psicología y otras ciencias" se hace un recorrido por los hitos más importantes de esta crisis de replicabilidad y de sus causas, y se desglosan las buenas prácticas que los investigadores pueden implementar antes, durante y después de su investigación para aumentar la probabilidad de que sus resultados sean replicables.

Además, frecuentemente se publican estudios con potencia estadística insuficiente para detectar efectos significativos en Psicología, una limitación ampliamente conocida y atribuida al empleo de muestras de tamaño reducido (Cohen, 1962). En respuesta a esta problemática, el metaanálisis, una metodología que integra cuantitativamente datos provenientes de estudios similares, se ha consolidado como una herramienta metodológica clave para paliar este problema. Al combinar los resultados de múltiples estudios, esta técnica permite aumentar el tamaño efectivo de la muestra y, en consecuencia, mejorar la potencia estadística, mejorando así la capacidad de identificar efectos

existentes en la realidad. Dentro del metaanálisis, una de las técnicas más novedosas es la combinación de metaanálisis con modelos de ecuaciones estructurales (MASEM en sus siglas en inglés: Meta-Analytic Strutural Equation Modeling; Cheung, 2015; Jak, 2015). El MASEM facilita la integración de los resultados de múltiples estudios, permitiendo la posterior estimación de un modelo de ecuaciones estructurales a partir de los datos combinados. Esta metodología ofrece la posibilidad de evaluar meta-analíticamente diversos modelos teóricos y realizar comparaciones entre modelos. En el artículo titulado "Introducción al metaanálisis multivariado con modelos de ecuaciones estructurales", López-López y colaboradores proporcionan una revisión de los enfoques analíticos disponibles dentro de la metodología MASEM, destacando las ventajas y limitaciones asociadas con cada uno de ellos. Además, el documento incluye un tutorial detallado que guía paso a paso la implementación de esta metodología.

## Hacia una modelización más formal

En este número especial, también hemos intentado ilustrar la relevancia metodológica de los modelos formales que, frente a las teorías verbales, suponen una forma operativa de estudiar fenómenos de interés en el marco de la Ciencia Psicológica. En esa línea, aunque las teorías verbales a las que estamos acostumbrados en Psicología siguen siendo muy útiles y necesarias, éstas son mucho más ambiguas que los modelos formales (e.g., Busemeyer et al., 2015; Farrell y Lewandowsky, 2010; Sun, 2023). Un modelo formal es una representación matemática, computacional o lógica que permite expresar de manera operativa los mecanismos subyacentes a un fenómeno de interés y hacer predicciones cuantificables que pueden ser evaluadas en contraposición a datos empíricos. Aunque esta perspectiva no es en absoluto novedosa, incluir explícitamente la propuesta de modelos formales en un número especial sobre avances metodológicos supone una apuesta por lo que, en nuestra opinión, será el futuro de la Ciencia Psicológica: modelos computacionales, estadísticos o matemáticos que implementen teoría psicológica para el estudio de fenómenos concretos.

En el artículo titulado “Respuestas observables y estados ocultos en redes neuronales recurrentes para razonar sobre aspectos cognitivos del lenguaje”, escrito por Jorge-Botana y colaboradores, se ilustra la utilización de modelos conexionistas, como los modelos de redes neuronales artificiales, para el estudio de características cognitivas del lenguaje. Los autores presentan cómo las redes neuronales recurrentes (Elman, 1990; Jordan, 1997) con mecanismos LSTM (Hochreiter & Schmidhuber, 1997) pueden utilizarse para estudiar cómo estos modelos aprenden a generar representaciones internas y a producir lenguaje. También se presentan distintas medidas basadas en estrategias observables y no observables que son útiles para estudiar distintos fenómenos psicológicos como la ruptura de expectativas lingüísticas. Los mecanismos implicados en este tipo de redes conexionistas son la base sobre la que se fundamentan los Grandes Modelos de Lenguaje que hoy son tan populares. Así, este artículo invita a los lectores a comprender algunos de los mecanismos que han potenciado su éxito y a pensar en la generación de hipótesis sobre los mecanismos cognitivos implicados en el funcionamiento de estos modelos.

Por otro lado, en el artículo titulado “Uso de modelos matemáticos como parte del análisis de datos en Psicología: el caso del descuento por demora”, Ramos y colaborares presentan una ilustración didáctica sobre diferentes formalizaciones de un proceso psicológico muy relacionado con la impulsividad: el descuento por demora (operativizando la relación inversa que existe entre el tiempo de demora y el valor subjetivo del reforzador). Así, se presentan modelos como el de Mazur (1987) o el Modelo Hiperbólico Generalizado (Myerson y Green, 1995), contraponiendo sus pros y sus contras. También se ilustra la evaluación de qué mecanismos ofrecen el mejor ajuste. Además, en esta ilustración se incluye un ejemplo práctico utilizando Excel, un software ampliamente utilizado y conocido, con la intención de que los investigadores aplicados se animen a implementar estos enfoques de manera sencilla en sus propios estudios y análisis.

## Conclusiones

Este número especial se presenta como un conjunto representativo de diferentes líneas de investigación sobre

avances metodológicos en Psicología y su "estado del arte". Sin embargo, los contenidos que componen este número especial no dejan de ser una selección personal de los editores invitados. Por tanto, estos artículos no son un resumen exhaustivo de toda la investigación metodológica disponible y, por desgracia, muchos temas de gran interés se han quedado fuera por limitaciones de espacio. Aun así, creemos que el interés de estos contenidos será lo suficientemente amplio como para que todos los investigadores interesados en estas metodologías puedan aprender algo nuevo. Desde este espacio, reivindicamos la necesidad de hacer accesibles muchas herramientas metodológicas modernas a través de textos asequibles (como pretende ser este número especial) para que los usuarios puedan aplicarlos en sus áreas de investigación. Esperamos que esto pueda generar un beneficio mutuo: los investigadores de áreas de contenido podrán obtener resultados más robustos utilizando métodos modernos y ello les facilitará la tarea de publicar sus investigaciones, mientras que los investigadores en metodología verán cómo se aplican los métodos que proponen y estudian para tener un impacto real en la investigación empírica. Además, esperamos que los contenidos de este número especial sean de interés para los lectores y que estos artículos tengan cierto impacto tanto en la formación de nuevos investigadores como también en investigadores más senior que quieran actualizarse en estos temas. Queda mucho por hacer.

## Referencias


Bono, R. y Arnau, J. (2014). *Diseños de caso único en ciencias sociales y de la salud* [Single-Case Designs in Social and Health Sciences]. Síntesis.

Borsboom, D. (2022). Possible Futures for Network Psychometrics. *Psychometrika, 87*(1), 253–265.

Brown, A. y Maydeu-Olivares, A. (2011). Item Response Modeling of Forced-Choice Questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. https://doi.org/10.1177%2F0013164410375112

Busemeyer, J. R., Wang, Z., Townsend, J. T. y Eidels, A. (2015). *The Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. y Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Cheung, M. W.-L. (2015). *Meta-analysis: A Structural Equation Modeling Approach*. Wiley.

Cohen, J. (1962). The Statistical Power of Abnormal-Social Psychological Research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145–153. https://doi.org/10.1037/h0045186

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Epskamp, S. (2020). Psychometric Network Models from Time-Series and Panel Data. *Psychometrika, 85*(1), 206–231.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. y Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. eLife, 10, Artículo e71601. https://doi.org/10.7554/eLife.71601

Farrell, S. y Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Current Directions in Psychological Science, 19*(5), 329–335. https://doi.org/10.1177/0963721410386677

Ferrer, E., Boker, S. M. y Grimm, K. J. (Eds.). (2018). *Longitudinal Multivariate Psychology*. Routledge.

Hochreiter, S. y Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780.

Hoffman, L. (2015). *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change.* Routledge.

Isvoranu, A. M., Epskamp, S., Waldorp, L. y Borsboom, D. (Eds.). (2022). *Network Psychometrics with R: A Guide for Behavioral and Social Scientists*. Taylor & Francis.

Jak, S. (2015). *Meta-analytic Structural Equation Modelling*. Springer. https://doi.org/10.1007/978-3-319-27174-3

Jordan, M. I. (1997). Serial order: A Parallel Distributed Processing Approach. *Advances in Psychology, 121,* 471–495. https://doi.org/10.1016/S0166-4115(97)80111-2

Mazur, J. E. (1987). An Adjusting Procedure for Studying Delayed Reinforcement. En M. L. Commons, J. E. Mazur, J. A. Nevin y H. Rachlin (Eds.), T*he Effect of Delay and of Intervening Events on Reinforcement Value* (pp. 55–73). Erlbaum.

Menard, S. (Ed.). (2007). *Handbook of Longitudinal Research: Design, Measurement, and Analysis*. Elsevier.

Myerson, J. y Green, L. (1995). Discounting of Delayed Rewards: Models of Individual Choice. *Journal of the Experimental Analysis of Behavior, 64*(3), 263–276. https://doi.org/10.1901/jeab.1995.64-263

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D. y Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology, 73*, 719–748. https://doi.org/https://doi.org/10.1146/annurev-psych-020821-114157

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science, 349,* Artículo aac4716. https://doi.org/10.1126/science.aac4716

Shadish, W. R. y Sullivan, K. J. (2011). Characteristics of Single-Case Designs Used to Assess Intervention Effects in 2008. *Behavior Research Methods, 43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Simmons, J. P., Nelson, L. D. y Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stark, S., Chernyshenko, O. S. y Drasgow, F. (2005). An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Sun, R. (2023). *The Cambridge Handbook of Computational Cognitive Sciences*. Cambridge University Press.

# DISEÑO Y ANÁLISIS DE DATOS DE DISEÑOS EXPERIMENTALES DE CASO ÚNICO

# DESIGN AND DATA ANALYSIS OF SINGLE-CASE EXPERIMENTAL DESIGNS

RUMEN MANOLOV[1]

Cómo referenciar este artículo/How to reference this article:

Manolov, R. (2025). Diseño y Análisis de Datos de Diseños Experimentales de Caso Único [Design and Data Analysis of Single-Case Experimental Designs]. *Acción Psicológica, 22*(1), 7–22. https://doi.org/10.5944/ap.22.1.42833

## Resumen

Los diseños experimentales de caso único implican el estudio intensivo de una o pocas unidades (e.g., personas) en diferentes condiciones manipuladas por los investigadores. Algunos diseños conllevan una replicación intra-sujeto (diseño ABAB, diseño de cambio de criterio y diseño de tratamientos alternantes), mientras que el diseño de línea base múltiple suele incluir replicación entre individuos. En ambos casos se dispone de varias oportunidades para demostrar el efecto de la intervención (introduciendo o retirándola) en diferentes momentos del tiempo. Asimismo, es imprescindible la replicación de los resultados en diferentes estudios para poder establecer la generalidad de las conclusiones. En cuanto al análisis de datos, actualmente se dispone de múltiples propuestas sin un consenso sobre cuáles son las opciones más apropiadas. Para favorecer la necesaria justificación de cualquier elección, se ofrece una serie de criterios organizativos que señalan en qué situaciones es más útil cada una de las propuestas comentadas. Asimismo, para acercar a los investigadores aplicados a las opciones analíticas, se comentan las páginas web gratuitas que las implementan. Finalmente, debido a que no es posible discutir con detalle todos los pormenores metodológicos, ni tampoco revisar todas las alternativas analíticas, el lector interesado es dirigido mediante múltiples referencias a las fuentes primarias.

**Palabras clave:** diseños experimentales de caso único; recomendaciones metodológicas; análisis de datos; software libre.

## Abstract

Single-case experimental designs entail the intensive study of one or few entities (e.g., individuals) in different conditions, which are manipulated by the researchers. Some designs include intra-subject replication (ABAB de-

**Correspondence address [Dirección para correspondencia]**: Rumen Manolov, Facultad de Psicología, Universitat de Barcelona, España.
**Email**: rrumenov13@ub.edu

**ORCID**: Rumen Manolov (http://orcid.org/0000-0002-9387-1926)
[1] Universitat de Barcelona.

Recibido: 21 de enero de 2025.
Aceptado: 15 de febrero de 2025.

sign, changing criterion design, and alternating treatments design), whereas the multiple-baseline design usually includes between-subjects replication. For both scenarios, there are several attempts to demonstrate the intervention effect (introducing or withdrawing the intervention) in different moments in time. Moreover, the replication of the results in different studies is necessary for establishing the generality of the conclusions. Regarding data analysis, there are currently multiple proposals, without a consensus regarding which the optimal analytical techniques are. In order to make easier the necessary justification of any data analytical technique chosen, the current text offers a series of organizing principles, which indicate in which situation each of the options is most useful. Furthermore, in order to bring applied researchers closer to the analytical options, the text refers to freely accessible websites implementing them. Finally, given that it is not possible to discuss in detail all methodological aspects, or to review all available data analytical techniques, the interested reader is directed via multiple references to the primary sources.

**Keywords:** single-case experimental designs; methodological recommendations; data analysis; software.

## Diseño y análisis de datos de Diseños Experimentales de Caso Único

Los Diseños Experimentales de Caso Único (DECU) constituyen una metodología de investigación que, en caso de cumplirse determinados requisitos, permite aportar evidencia científica sólida sobre la efectividad de una intervención (Horner et al., 2005). La característica principal de estos diseños es el estudio intensivo y longitudinal de una o más unidades (habitualmente personas, pero también pueden ser grupos considerados en su totalidad). A pesar de su denominación (también se les conoce como diseños de N =1 o diseños intra-sujeto), lo más habitual es que un estudio incluya a más de una persona (e.g., Tanious y Onghena, 2021, reportan que lo más habitual es que haya entre tres y siete personas en un DECU, con media y mediana cercanas a cuatro participantes). En los DECU se toman múltiples medidas obtenidas bajo diferentes condiciones de tratamiento (Tate y Perdices, 2019). Cada unidad se compara consigo misma, siendo uno de sus puntos fuertes la garantía sobre la validez interna de la investigación.

Las condiciones que se comparan suelen ser dos. En primer lugar, se dispone de una línea base que representa la situación habitual (problemática). Posteriormente, se introduce la intervención. También es posible comparar dos intervenciones. La línea base sirve no solo para describir la situación de partida, sino también poder predecir cómo seguiría la conducta de interés en caso de que la intervención no se introduzca o no sea efectiva. Por lo tanto, los DECU implican una comparación entre una predicción o proyección basada en la línea base y la realidad observada durante la intervención.

### *Objetivo y estructura del texto*

El objetivo del artículo es ofrecer una perspectiva general de las características metodológicas de los DECU y de las posibilidades de análisis, tanto visual como cuantitativa. En los apartados siguientes, se describen los diferentes tipos de DECU con ejemplos reales de investigaciones publicadas en diferentes ámbitos. Asimismo, se mencionan los requisitos principales, los retos y las ventajas de los DECU. Posteriormente, se presentan varias alternativas de análisis datos, ofreciendo una clasificación de éstas según diferentes criterios. Dicha clasificación que podría ser útil a la hora de escoger alguna(s) de estas opciones.

## Metodología DECU

### *Tipos principales de Diseños Experimentales de Caso Único*

Los DECU se pueden distinguir entre diseños reversibles (en los cuales la intervención se puede retirar; diseño ABAB y diseño de tratamientos alternantes) y diseños irreversibles (diseño de línea base múltiple y diseño de cambio de criterio). Otra posible clasificación es en función de si el diseño permite una comparación entre series (diseño de línea base múltiple) o no (el resto de los dise-

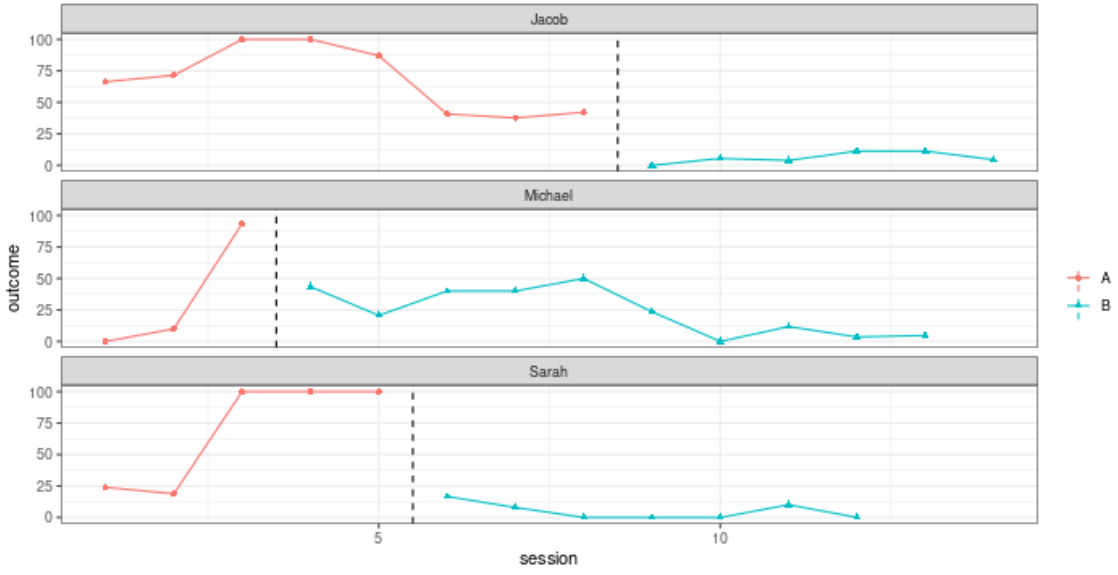
ños). A continuación, se comentan sus características principales.

El diseño más habitual es el de línea base múltiple (Tanious y Onghena, 2021), que implica la replicación de la secuencia AB (donde «A» representa la línea base y «B» la fase de intervención) en diferentes personas, conductas de la misma persona o contextos en los que se estudia la misma conducta de la misma persona. Lo fundamental es que la intervención se introduzca de forma escalonada, es decir, en diferentes momentos en el tiempo para las diferentes líneas base. En este tipo de diseños es posible que la línea base empiece en el mismo momento para todos (versión concurrente) o no (no concurrente), véase Slocum et al. (2022) para una discusión de la importancia de esta distinción en cuanto a la validez interna. La principal ventaja es que no es necesario retirar la intervención, lo que demuestra que este DECU es aplicable a situaciones en las que la intervención implica aprendizaje. Se trata de la situación probablemente más habitual en un contexto clínico, sobre todo si se trabaja desde una perspectiva cognitiva o cognitivo-conductual. Desde el punto de vista ético también parece más apropiado no retirar (ni siquiera temporalmente) una intervención que funciona. Un ejemplo de datos recogidos siguiendo un diseño de línea base múltiple puede verse en la Figura 1. Los datos provienen del estudio de Eilers y Hayes (2015) sobre el uso de ejercicios cognitivos para reducir las conductas repetitivas y restrictivas de niños con Trastorno de Espectro Autista. Es posible mejorar el diseño introduciendo aleatorización de diferentes maneras (Levin et al., 2018). Por ejemplo, se puede escoger al azar el momento en el que empieza la intervención para cada participante, entre varios posibles momentos que no se solapen entre participantes. Otra opción es determinar de antemano el momento de intervenir, pero decidir al azar qué participante comienza primero, quién segundo, etc.

El diseño ABAB o diseño de retirada o reversión (Wine et al., 2015) implica una replicación dentro del mismo participante. Se dispone de tres momentos de comparación entre fases adyacentes El diseño es aplicable

**Figura 1**

*Ejemplo de un diseño de línea base múltiple entre personas. Datos descargados de* https://osf.io/79dfs*, obtenidos originalmente por Eilers y Hayes (2015). El gráfico se ha obtenido mediante https://jepusto.shinyapps.io/scdhlm.*



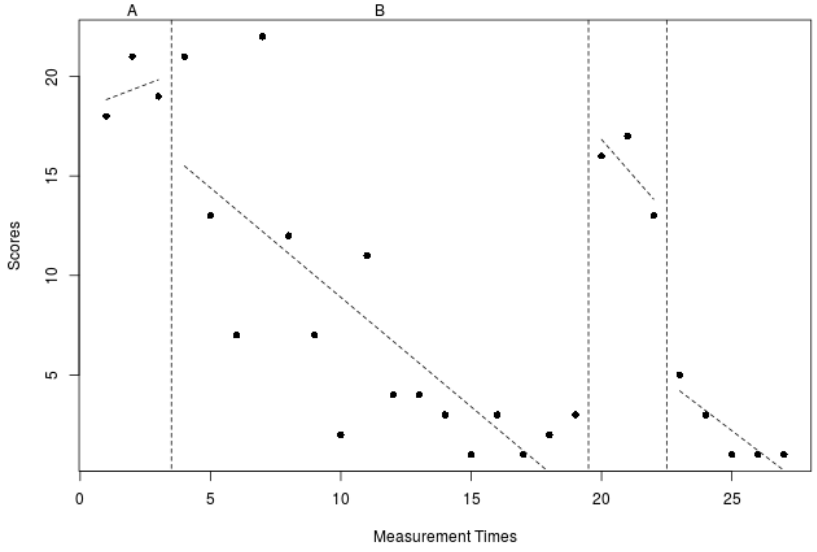

*Nota.* Los datos rojos son los correspondientes a la línea base (A), mientras que los datos azules son los correspondientes a la fase de intervención (B).

cuando la intervención no provoca un cambio permanente, sino que en ausencia de la intervención es posible volver al nivel inicial de la conducta de interés. En este sentido, la tercera fase es crítica para la inferencia del efecto del tratamiento, que sería posible en caso de que tenga lugar la reversión de la conducta de interés al nivel anterior al tratamiento. Un ejemplo se ofrece en la Figura 2: se trata de datos recogidos por Feeney e Ylvisaker (2006) de un participante que había padecido una lesión cerebral traumática, y a quien se aplicó una intervención cognitivo-conductual para tratar conducta desafiante. Es posible escoger al azar los tres momentos de cambio de fase, entre un listado de posibilidades que respeta un mínimo de longitud de cada fase (Onghena, 1992).

En el diseño de tratamientos alternantes, a diferencia de los dos diseños anteriores, la comparación principal no tiene lugar entre fases. La comparación fundamental se realiza entre las condiciones (habitualmente dos intervenciones diferentes) que están sujetas a una alternancia frecuente. Aparte de esta alternancia (que podría constituir una «fase de comparación»), puede haber -aunque no sea imprescindible- una fase inicial de línea base y una fase final en la que se aplica solo la mejor intervención (Barlow y Hayes, 1979). Dentro de la «fase de comparación», lo habitual es restringir el número de medidas consecutivas dentro de la misma condición a dos. Para que este diseño sea aplicable es necesario que la intervención no tenga efectos duraderos (e.g., intervención farmacológica). Se puede observar un ejemplo en la Figura 3: se trata de datos recogidos por Eilers y Hayes (2005), comparando dos intervenciones diferentes para reducir conductas problemáticas en niños diagnosticados con Trastorno de Espectro Autista. Como se puede apreciar, gráficamente se suelen juntar mediante una línea las medidas pertenecientes a la misma condición y se evalúa la distancia o separación entre estas líneas. Este tipo de inspección visual tiene paralelos en las cuantificaciones propias para los diseños de tratamientos alternantes (Manolov y Onghena, 2018). Asimismo, es posible implementar aleatorización de dife-

**Figura 2**

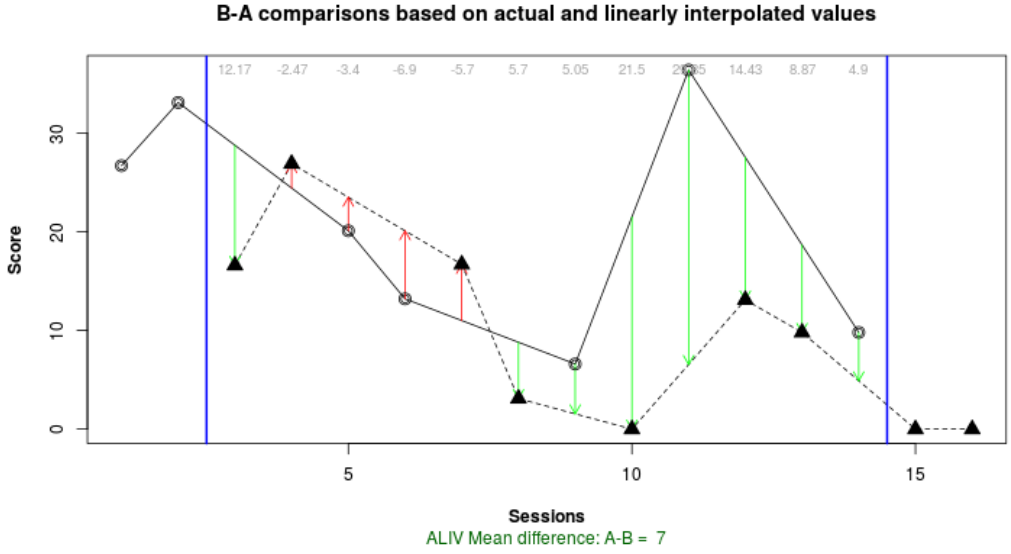
*Ejemplo de un diseño ABAB. Datos descargados de* https://osf.io/79dfs*, originalmente obtenidos por Feeney e Ylvisaker (2006). El gráfico se ha obtenido mediante https://tamalkd.shinyapps.io/scda/*




*Nota.* El primer recuadro representa la línea base inicial. El segundo recuadro es la primera introducción de la intervención. El tercer recuadro es la retirada de la intervención (i.e., la vuelta a la línea base). El cuarto recuadro es la reintroducción de la intervención.

**Figura 3**

*Ejemplo de un diseño de línea base múltiple entre personas. Datos descargados de https://osf.io/kaphj, obtenidos originalmente por Eilers y Hayes (2015). El gráfico se ha obtenido mediante https://manolov.shinyapps.io/ATDesign.*



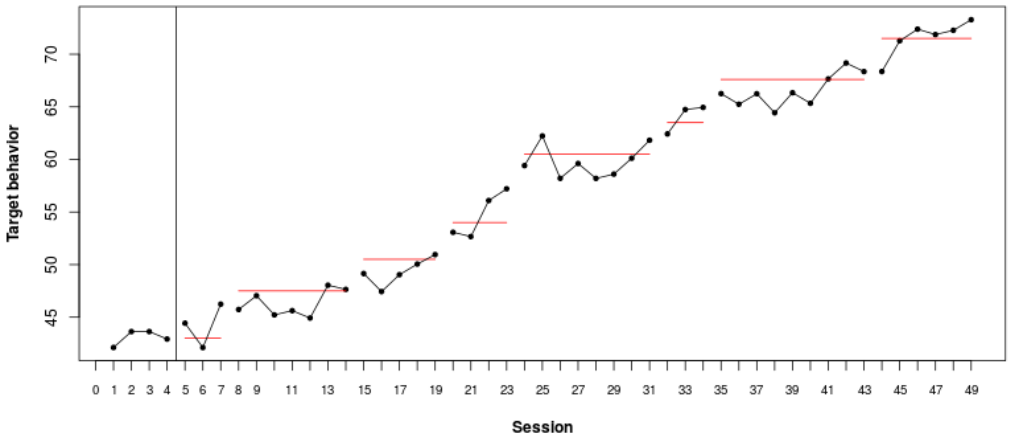

*Nota*. Las líneas verticales verdes muestran comparaciones entre condiciones que favorecen a la condición "B" (valores inferiores de la conducta indeseable). Las líneas verticales rojas muestran comparaciones que favorecen a la condición "A". A la izquierda de la primera línea azul y a la derecha de la segunda línea azul hay datos para los cuales es imposible comparar las líneas que juntan los puntos de las dos condiciones.

rentes maneras. Por ejemplo, una posible aleatorización implicaría seleccionar al azar qué condición tiene lugar en cada momento de medida, respetando la limitación de un número máximo (de dos, habitualmente) medidas consecutivas en la misma condición (Onghena y Edgington, 1994).

El diseño de cambio de criterio implica la consecución gradual de la meta final mediante el establecimiento, habitualmente de forma conjunta entre investigador y participante, de criterios cada vez más restrictivos (Hartmann y Hall, 1976). En concreto, en un contexto de modificación de conducta, con la consecución de cada criterio intermedio, el participante recibe reforzamiento. Posibles aplicaciones de este tipo de DECU incluirían ir asumiendo los diferentes pasos necesarios en una tarea compleja (e.g., que un niño aprenda a atarse los cordones) o conseguir niveles cada vez más adecuados de una conducta de interés (e.g., mejorar el cumplimiento de una intervención médica). Se ha resaltado importancia de variar la longitud de las subfases de intervención, la magnitud del cambio de criterio y también introducir retrocesos a criterios previos menos exigentes (Klein et al., 2017). Un ejemplo puede verse en la Figura 4, que corresponde a datos recogidos por Facon et al. (2008), quienes utilizaron procedimientos de aprendizaje operante para tratar un mutismo selectivo severo en un niño con discapacidad intelectual, intentando aumentar progresivamente los decibelios de sus verbalizaciones. Entre los desarrollos para los diseños de cambio de criterio hay que destacar el diseño con rango aceptable para la conducta dentro de cada subfase (McDougall, 2005) y la posibilidad de introducir aleatorización, por ejemplo, escogiendo al azar la longitud de las subfases de la intervención (Onghena et al., 2019).

**Figura 4**

*Ejemplo de un diseño de cambio de criterio. Datos descargados de* https://osf.io/kcjnu*, obtenidos originalmente por Facon et al. (2008). El gráfico se ha obtenido mediante https://manolov.shinyapps.io/ChangingCriterion.*




*Nota.* El primer recuadro representa los datos de la línea base. El segundo recuadro incluye las cuatro subfases de la fase de intervención: los criterios se marcan mediante las líneas horizontales rojas, mientras que las líneas horizontales azules son los rangos aceptables para cada subfase.

## *Recomendaciones metodológicas: rúbricas*

Se dispone de varios instrumentos para valorar el rigor metodológico o guías para sugerir cómo proceder en estudios DECU (e.g., Perdices et al., 2023; What Works Clearinghouse, 2022). Se trata de componentes de diseño que potencien la validez interna, es decir, el grado de confianza en una relación causal entre la intervención y la conducta de interés (e.g., número ocasiones para demostrar el efecto de la intervención, número de medidas, fidelidad de implementación, aleatorización). En cuanto a la validez externa o capacidad de generalizar, se requiere informar de detalles suficientes sobre el participante, la intervención y el contexto para poder valorar en qué tipo de situaciones se puede considerar que una intervención es efectiva.

## *Ventajas*

En los diseños clásicos de comparación de grupos, la aleatorización en la asignación de las personas los grupos experimentales permite asumir la equivalencia inicial de los grupos y cualquier diferencia posterior se atribuye al efecto del tratamiento. No obstante, este tipo de diseños suele implicar criterios de inclusión estrictos para que se puedan formar grupos homogéneos que suelen ser puros en cuanto a la problemática de interés (i.e., sin comorbilidades). Esto limita la posibilidad de generalizar los resultados a individuos con características menos típicas. Asimismo, inferir de un resultado general a un individuo puede resultar en una falacia ecológica, puesto que el promedio no representa necesariamente a ninguna persona en concreto. En contraste, los DECU tratan directamente con la unidad sobre la cual se quiere extraer una conclusión (i.e., el individuo) y permiten el estudio de poblaciones muy heterogéneas (e.g., personas diagnosticas con Trastorno del Espectro Autista).

Otra ventaja de los DECU es la posibilidad de estudiar el proceso de cambio, gracias a las múltiples medidas de las que se dispone. Asimismo, el estudio intensivo de la persona permite explicar posibles valores anómalos que puedan obtenerse en algún momento determinado.

Finalmente, la estructura básica de los DECU mimetiza la práctica profesional habitual: una fase de evaluación inicial seguida de una intervención. Por lo tanto, los profesionales pueden ejercer a la vez de investigadores y aprovechar su experiencia, publicándola y aportando datos para establecer la base científica de las intervenciones.

## *Retos*

La limitación fundamental relacionada con los DECU es la generalización o validez externa. La manera más segura de dar pasos hacia la generalización es la replicación (Tate y Perdices, 2019). Solo añadiendo más casos y atendiendo a sus características se puede generalizar, de forma empírica y gracias a la lógica, a otras personas de características parecidas.

Un segundo reto es conseguir que el diseño sea lo más riguroso posible para potenciar la validez interna, incorporando los componentes recomendados en los instrumentos que evalúan el rigor metodológico. Por una parte, es fundamental que haya varias demostraciones del cambio en la conducta de interés concurrente con el cambio en la condición experimental, en diferentes momentos del tiempo, para descartar la «historia» (i.e., ocurrencia de eventos externos a la intervención) como posible razón de dicho cambio. Por otra parte, la aleatorización (e.g., selección al azar del momento en que cambiar de condición experimental en un diseño ABAB o del orden en que intervenir los participantes en un diseño de línea base múltiple) se ha resaltado como un elemento fundamental para la validez interna (e.g., Jacobs, 2019).

Finalmente, un tercer reto es el análisis de datos. Se dispone de múltiples opciones analíticas sin un consenso claro respecto a cuál escoger, sobre todo en relación con la posible presencia de dependencia serial, tendencia a la mejora espontánea, o la variedad de tipos de DECU. Además, la autocorrelación (o dependencia serial entre las medidas obtenidas longitudinalmente de la misma unidad) dificulta la aplicación de pruebas inferenciales clásicas. El presente texto pretende presentar una estructura para que los investigadores aplicados puedan realizar una elección y presentar una justificación con una base sólida.

# Análisis de Datos DECU

## *Clasificación de las alternativas de análisis*

### *Formativo o Sumativo*

Una primera distinción es entre el análisis *formativo* y *sumativo* (Ledford et al., 2019). El primero forma parte de la experimentación guiada por los datos, utilizada para determinar cuándo cambiar las condiciones (i.e., decidir mientras los datos aún se están recogiendo). En cambio, el análisis sumativo sirve para documentar y comunicar el grado de efectividad de la intervención, una vez que todos los datos ya estén recogidos. El análisis formativo se lleva a cabo principalmente mediante la inspección visual de la representación gráfica de los datos (e.g., Byun et al., 2017). Las secciones siguientes se refieren a organizar opciones analíticas para análisis *sumativo*.

### *Objetivo de la evaluación de los datos*

Uno de los objetivos de la evaluación de los datos es establecer una relación funcional o causal, es decir, valorar si se puede inferir que los cambios en la conducta objeto se deben al efecto de tratamiento. Se compara visualmente el patrón de datos esperado, según el diseño, y el obtenido. Para el mismo objetivo se podría utilizar el *p*-valor de una prueba de aleatorización, que serviría para una inferencia causal tentativa (que no una inferencia poblacional basada en supuestos y modelos; Manolov y Onghena, 2018). La idea de la prueba de aleatorización es que se escoge de antemano un estadístico de prueba, y este estadístico se calcula para todas las divisiones de datos posibles (i.e., todas las aleatorizaciones o maneras de asignar los momentos de medida a diferentes condiciones). De esta manera, el *p*-valor se obtiene directamente a partir de los datos, sin necesidad de asumir una distribución (e.g., normal) para el estadístico de prueba o para los datos (Heyvaert y Onghena, 2014).

Un segundo objetivo podría ser, en algunos casos, comparar los datos a un resultado final deseable, como cuando se utilizan criterios de maestría y niveles preestablecidos de rendimiento (McDougale et al., 2015). En este

sentido, es posible cuantificar el grado en qué se ha conseguido el objetivo, comparando el nivel deseado con el nivel conseguido (Ferron et al., 2020).

Un tercer objetivo es la cuantificación de la magnitud de la diferencia mediante la obtención tamaño del efecto, aunque puede presentar retos interpretativos, considerando que los criterios a seguir deberían ser específicos de cada ámbito de investigación (Vannest y Sallese, 2021). Las secciones siguientes se refieren a la cuantificación del tamaño de la diferencia entre condiciones.

Finalmente, se puede considerar una comparación más generalizada entre la situación antes y después del tratamiento a través del índice de cambio fiable (Estrada et al., 2019), en caso de utilizarse medidas con propiedades psicométricas conocidas.

#### *Escala de medida y unidad de medida*

En cuanto a la *escala de medida* de la variable dependiente, si ésta es ordinal, se pueden utilizar índices de no solapamiento (Parker et al., 2011). Cuando la escala de medida es de intervalo o razón, se pueden calcular diferencias en medias y tendencias. En estos casos, el investigador puede seleccionar las *unidades de medida* deseadas para la cuantificación resumen: porcentajes (e.g., el logaritmo de la razón de respuestas se puede convertir a un cambio porcentual; Pustejovsky, 2018), estandarizadas (e.g., la diferencia de medias estandarizada entre casos: Shadish et al., 2014; modelos multinivel tras estandarizar los datos), o no estandarizadas (e.g., modelos multinivel con los datos originales).

#### *Intra-individual o entre individuos*

Otro criterio para escoger el enfoque analítico es el nivel de análisis. Si el foco es obtener cuantificaciones separadas para cada individuo, se pueden utilizar medidas intra-individuales como la diferencia de media estandarizada (Busk y Serlin, 1992) y los índices de no solapamiento (Parker et al., 2011). Si el objetivo es obtener una única cuantificación general para varios individuos, se puede usar la diferencia de medias estandarizada entre individuos (Shadish et al., 2014) o modelos multinivel (Ferron et al., 2009, 2010).

Una clasificación parecida procede del tipo de diseño. Por ejemplo, un diseño ABAB implica una comparación intra-serie, mientras que un diseño de línea base múltiple permite tanto la comparación intra-serie, como entre-series (ver Ferron et al., 2014). En este último tipo de diseño, la comparación entre-series está ligada al inicio concurrente de la fase de línea base (Christ, 2007).

## *Técnicas analíticas y su implementación en software gratuito*

#### *Inspección Visual*

En cuanto a la inspección visual, elemento analítico que suele estar presente siempre, los desarrollos se pueden organizar en seis ámbitos. Primero, se han hecho recomendaciones sobre las características deseables de los gráficos como representaciones visuales (Dart y Radley, 2018). Segundo, se han listado aspectos de los datos a considerar (Ledford et al., 2019; Maggin et al., 2018): nivel, tendencia, variabilidad, inmediatez, solapamiento y consistencia. Concretamente, dentro de cada fase, se puede valorar el nivel, la tendencia y la variabilidad. Complementariamente, a la hora de comparar fases adyacentes, se pueden identificar cambios de nivel o cambios de tendencia, además de valorar si dichos cambios son inmediatos o demorados. Otro tipo de comparación se refiere al grado en que las diferentes fases incluyen valores parecidos (i.e., grado de solapamiento). Finalmente, al considerar varias ocasiones de demostración de efecto, se puede valorar la consistencia entre fases parecidas y la consistencia del efecto (Manolov y Tanious, 2022). Tercero, se ha propuesto usar medianas, líneas de tendencia y de variabilidad y cuantificaciones de no solapamiento que acompañen a la valoración visual de estos aspectos (Lane y Gast, 2014). Cuarto, se han propuesto protocolos que sistematicen los diferentes pasos del análisis visual, aunque sin necesariamente acudir a cuantificaciones para cada uno de los aspectos de los datos. (Wolfe et al., 2019). Quinto, se han propuesto ayudas visuales en forma de líneas de tendencia central y variabilidad superpuestas (e.g., Fisher et al., 2003). Sexto, se han propuesto gráficos para un análisis conjunto de varias comparaciones entre condiciones (Manolov, Tanious et al., 2022).

***Opciones intra-individuales***

En la presente sección las diferentes opciones analíticas intra-individuales se organizarán según el aspecto focal de los datos: nivel (media o mediana), tendencia, variabilidad, solapamiento, inmediatez y consistencia.

Los índices de *no solapamiento* (comparados en Parker, Vannest y Davis, 2011) son un grupo de cuantificaciones que se centran en la información ordinal contenida en los datos. Específicamente, comparan datos de diferentes condiciones en cuanto a cuál de ellos es superior, sin tener en cuenta la distancia (i.e., cuán superior). Algunos índices resumen los datos de la fase de línea base mediante su mejor dato (el índice con acrónimo PND; Scruggs et al., 1987) o la mediana (el índice con acrónimo PEM; Ma, 2016), mientras que otros utilizan todos los datos sin resumirlos (el índice con acrónimo NAP; Parker & Vannest, 2009). A su vez, algunos índices no tienen en cuenta una posible tendencia hacia la mejora espontánea durante la fase de línea base (PND, PEM, NAP), mientras que otros sí controlan este tipo de tendencia (los índices Tau de Parker, Vannest, Davis y Sauber 2011). En cuanto al software, la página web https://jepusto.shinyapps.io/SCD-effect-sizes proporciona explicaciones y fórmulas, además de las cuantificaciones.

Al centrarse en el *nivel*, las cuantificaciones propuestas han sido la diferencia de medias estandarizada (Busk y Serlin, 1992) y el logaritmo de la razón de respuestas (Pustejovsky, 2018). Para diferencias de medias estandarizadas intra-individuales y una cuantificación en términos de porcentaje se puede utilizar https://jepusto.shinyapps.io/SCD-effect-sizes.

En cuanto a las opciones que incorporan la *tendencia*, se trata de propuestas basadas en modelos de regresión, por ejemplo: (a) cuantificación separada del cambio de pendiente y del cambio de nivel para la primera ocasión de la fase de intervención, conocidos en inglés como modelos "piecewise" (Center et al., 1985; Moeyaert et al., 2014); y (b) cuantificación conjunta a través del promedio de las diferencias entre la tendencia proyectada de la fase de línea base y la tendencia ajustada en la fase de intervención (Swaminathan et al., 2014). En cuanto al software, el análisis de un único nivel siguiendo un modelo "piecewise" (i.e., una comparación A-B) se puede llevar a mediante http://34.251.13.245/MultiSCED/ (Declercq et al., 2020) y https://manolov.shinyapps.io/Regression/, que también incorpora el modelo de Swaminathan et al. (2014).

En cuanto a la *inmediatez* del cambio, la propuesta inicial fue comparar la media de las tres últimas medidas de la fase de línea base con la media de las tres primeras medidas de la fase de intervención, aplicable mediante https://manolov.shinyapps.io/Overlap/. Una propuesta más reciente permite identificar el momento (o los momentos) más probable(s) en que se produjo el mayor cambio y, por lo tanto, valorar si dicho momento coincide con el momento de cambio de fase (en inglés, *Bayesian Unknown Change Point Model*, Natesan y Hedges, 2017). El código de R para esta opción está disponible en https://github.com/prathiba-stat/BUCP. Otra opción es explorar diferentes latencias, teniendo en cuenta si el tipo de efecto esperado es abrupto o progresivo (Manolov y Onghena, 2022).

Finalmente, nótese que las pruebas de aleatorización (Heyvaert y Onghena, 2014) permiten seleccionar como estadístico de prueba una cuantificación centrada en el nivel (e.g., una diferencia de medias), en la tendencia (e.g., diferencia de pendientes), en la variabilidad (e.g., razón de varianzas) o en el solapamiento. Mediante dichas pruebas se asocia un *p*-valor a estas cuantificaciones para representar probabilidad de obtener una diferencia tan grande o mayor en ausencia de efecto de la intervención.

***Opciones entre individuos***

La recomendación realizada por What Works Clearinghouse (2022), pero no compartida por todos (e.g., Kratochwill et al., 2021), es utilizar una diferencia de medias estandarizada que permita obtener una única cuantificación para varios participantes. Se trata de dos procedimientos diferentes. Uno no tiene la tendencia en cuenta (utilizando estimación por el método de los momentos; Shadish et al., 2014), mientras que el otro sí modela la tendencia (utilizando estimación por máxima verosimilitud restringida; Pustejovsky et al. 2014). La estandarización se consigue teniendo en cuenta tanto variabilidad intra-individual, como entre individuos, para obtener una cuanti-

ficación comparable a la que se obtiene en diseños de comparación de grupos. La autocorrelación se modela a la hora de obtener el intervalo de confianza alrededor de la estimación puntual.

Un objetivo similar se consigue mediante los modelos multinivel de dos niveles (Ferron et al., 2009), que también permiten obtener cuantificaciones separadas para cada individuo a través de estimaciones Bayesianas empíricas (Ferron et al., 2010).

En cuanto al software, para las dos versiones de la diferencia de medidas estandarizada entre casos se puede utilizar https://jepusto.shinyapps.io/scdhlm. La potencia se puede calcular mediante https://abkpowercalculator.shinyapps.io/ABkpowercalculator/.

La web http://34.251.13.245/MultiSCED/ (Declercq et al., 2020) permite llevar a cabo un análisis de dos niveles. Otra web, https://manolov.shinyapps.io/SeveralAB/, ofrece las estimaciones Bayesianas empíricas de los efectos individuales y permite modelar la autocorrelación y varianza residual heterogénea y además de ofrecer los valores de los criterios informacionales AIC y BIC que permiten la comparación entre modelos.

Una manera diferente de combinar resultados de varios casos es valorar si un efecto puede considerar exitosamente replicado (Manolov, Tanious et al., 2022) gracias a una definición a priori del nivel deseado tras la intervención y del mínimo cambio deseable. Esta opción está implementada en https://manolov.shinyapps.io/Brinley/.

#### *Opciones entre estudios*

Debido a la importancia de la replicación para generalizar conclusiones en el contexto DECU, el metaanálisis de resultados de diferentes estudios sobre la misma problemática y con la misma intervención es necesario. Recientemente, se han distinguido dos enfoques (Declercq et al., 2022): combinar tamaños del efecto (dos etapas) y combinar datos directamente (una etapa).

En cuanto al software, http://34.251.13.245/MultiSCED/ (Declercq et al., 2020) permite implementar un modelo de tres niveles (i.e., una etapa), mientras que https://manolov.shinyapps.io/Change permite un metaanálisis de dos etapas combinando tamaños del efecto.

## Discusión

### *Recomendaciones*

#### *Planificar*

Antes de recoger y analizar datos, hay que asegurarse que el estudio puede aportar evidencia científica sólida: siguiendo las recomendaciones metodológicas (e.g., Perdices et al., 2023; What Works Clearinghouse, 2022). Asimismo, hay que valorar si con los recursos disponibles o factibles (participantes y número de momentos de medida), las técnicas analíticas funcionarían adecuadamente en términos de ausencia de sesgo, eficiencia, tasa de error Tipo I y potencia estadística.

En cuanto a la planificación, en caso de que haya aleatorización en el diseño y de que se utilice una prueba de aleatorización para inferencia causal tentativa, es necesario comprobar si el número de aleatorizaciones posibles (según el diseño) permite obtener un *p*-valor igual o inferior al alfa nominal (habitualmente 0.05). El *p*-valor no puede ser más pequeño que 1 dividido entre el número de aleatorizaciones. El cálculo del número de aleatorizaciones puede obtenerse para diferentes DECU a través de la web https://tamalkd.shinyapps.io/scda.

#### *Informar*

Se recomienda seguir las guías de publicación elaboradas por un conjunto de expertos en los DECU (Tate et al., 2016). Asimismo, una justificación es necesaria para la elección de la técnica. Esta justificación puede basarse en varios criterios: (a) el problema de investigación y la posibilidad de obtener información útil; (b) el patrón de datos esperado (e.g., la presencia de mejora espontánea durante la fase de línea base; (c) la adecuación de las propiedades estadísticas (ausencia de sesgo, mayor eficiencia, mayor potencia estadística); (d) facilidad de interpretación, más allá de meramente reportar valores. Consideramos que hay

dos justificaciones que no deberían serlo: (a) facilidad de cálculo y (b) tradición (e.g., publicaciones previas).

Aparte de valorar si un efecto es visualmente claro, si es grande o estadísticamente significativo, se ha recomendado valorar la validez social (Snodgrass et al., 2023). Se trata de una aproximación a la significación práctica: el funcionamiento del individuo después de la intervención, el mantenimiento del efecto en el tiempo, la posibilidad de que la intervención se implemente por agentes típicos y con los recursos disponibles en la práctica profesional, etc.

## *Limitaciones y aportaciones*

El objetivo de este artículo es ofrecer una amplia panorámica de las principales características de los DECU y sus diferentes tipologías. Esto se ha hecho previamente en inglés (e.g., Ledford et al., 2019; Maggin et al., 2018) y también se dispone de texto en castellano (e.g., Bono y Arnau, 2014), aunque sin incluir los últimos desarrollos a nivel de análisis de datos. Adicionalmente, también se quería ofrecer una estructura organizativa y una panorámica de las diferentes técnicas de análisis de datos disponibles. También se dispone de textos *en inglés* sobre esta temática (e.g., Maggin et al., 2019; Manolov, Moeyaert et al., 2022), pero los criterios organizativos que aquí se presentan, para guiar a la hora de escoger cómo analizar los datos, son más completos, al atender al nivel de análisis deseado, al tipo de análisis que se desea realizar, a la escala de medida de la variable de interés, y a la necesidad (o no) de considerar una tendencia en los datos.

Otra aportación del presente texto es el listado organizado de software gratuito disponible. Lo mencionado en el texto se complementa por un proyecto de *Open Science Framework* https://osf.io/t6ws6., donde se dispone de ejemplos de la manera en la que los datos han de organizarse para cada una de las webs creadas con R y Shiny.

En cuanto a las limitaciones, las restricciones de longitud del texto han impedido profundizar en los detalles técnicos de los procedimientos. Para obtener información más detallada, se invita al lector a consultar la siguiente lista de bibliografía relevante sobre DECU: https://osf.io/u9g2r.

## Referencias

Barlow, D. H. y Hayes S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12*(2), 199–210. https://doi.org/10.1901/jaba.1979.12-199

Bono, R. y Arnau, J. (2014). *Diseños experimentales de caso único en ciencias sociales y de la salud* [Single-case Experimental Designs in Social and health sciences]. Síntesis.

Busk, P. L. y Serlin, R. C. (1992). Meta-analysis for single-case research. En T. R. Kratochwill y J. R. Levin (Eds.), *Single-case Research Designs and Analysis: New Directions for Psychology and Education* (pp. 187−212). Lawrence Erlbaum.

Byun, T. M., Hitchcock, E. R. y Ferron, J. (2017). Masked visual analysis: Minimizing Type I error in visually guided single-case design for communication disorders. *Journal of Speech, Language, and Hearing Research, 60*(6), 1455−1466. https://doi.org/10.1044/2017_JSLHR-S-16-0344

Center, B. A., Skiba, R. J. y Casey, A. (1985). A Methodology for the Quantitative Synthesis of intra-Subject Design Research. *The Journal of Special Education, 19*(4), 387−400. https://doi.org/10.1177/002246698501900404

Christ, T. J. (2007). Experimental Control and Threats to Internal Validity of Concurrent and Nonconcurrent Multiple Baseline Designs. *Psychology in the Schools, 44*(5), 451–459. https://doi.org/10.1002/pits.20237

Dart, E. H. y Radley, K. C. (2018). Toward a standard assembly of linear graphs. *School Psychology Quarterly, 33*(3), 350−355. https://doi.org/10.1037/spq0000269

Declercq, L., Cools, W., Beretvas, S. N., Moeyaert, M., Ferron, J. M. y Van den Noortgate, W. (2020).

MultiSCED: A Tool for (Meta-)Analyzing Single-Case Experimental Data with Multilevel Modeling. *Behavior Research Methods, 52*(1), 177–192. https://doi.org/10.3758/s13428-019-01216-2

Declercq, L., Jamshidi, L., Fernández Castilla, B., Moeyaert, M., Beretvas, S. N., Ferron, J. M. y Van den Noortgate, W. (2022). Multilevel Meta-Analysis of Individual Participant Data of Single-Case Experimental Designs: One-stage versus Two-Stage Methods. *Multivariate Behavioral Research, 57*(2–3), 298–317. https://doi.org/10.1080/00273171.2020.1822148

Eilers, H. J. y Hayes, S. C. (2015). Exposure and Response Prevention Therapy with Cognitive Defusion Exercises to Reduce Repetitive and Restrictive Behaviors Displayed by Children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders, 19*, 18–31. https://doi.org/10.1016/j.rasd.2014.12.014

Estrada, E., Ferrer, E. y Pardo, A. (2019). Statistics for Evaluating Pre-Post Change: Relation between Change in the Distribution Center and Change in the Individual Scores. *Frontiers in Psychology, 9*, Artículo 2696. https://doi.org/10.3389/fpsyg.2018.02696

Facon, B., Sahiri, S. y Riviere, V. (2008). A Controlled Single-Case Treatment of Severe Long-Term Selective Mutism in a Child with Mental Retardation. *Behavior Therapy, 39*(4), 313–321. https://doi.org/10.1016/j.beth.2007.09.004

Feeney, T. y Ylvisaker, M. (2006). Context-Sensitive Cognitive-Behavioural Supports for Young Children with TBI: A Replication Study. *Brain Injury, 20*(6), 629–645. https://doi.org/10.1080/02699050600744194

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G. y Hibbard, S. T. (2009). Making Treatment Effect Inferences from Multiple-Baseline Data: The Utility of Multilevel Modeling Approaches. *Behavior Research Methods, 41*(2), 372–384. https://doi.org/10.3758/BRM.41.2.372

Ferron, J. M., Farmer, J. L. y Owens, C. M. (2010). Estimating Individual Treatment Effects from Multiple-Baseline Data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Ferron, J. M., Goldstein, H., Olszewski, A. y Rohrer, L. (2020). Indexing Effects in Single-Case Experimental Designs by Estimating the Percent of Goal Obtained. *Evidence-Based Communication Assessment and Intervention, 14*(1–2), 6–27. https://doi.org/10.1080/17489539.2020.1732024

Ferron, J. M., Moeyaert, M., Van den Noortgate, W. y Beretvas, S. N. (2014). Estimating Causal Effects from Multiple-Baseline Studies: Implications for Design and Analysis. *Psychological Methods, 19*(4), 493–510. https://doi.org/10.1037/a0037038

Fisher, W. W., Kelley, M. E. y Lomas, J. E. (2003). Visual Aids and Structured Criteria for Improving Visual Inspection and Interpretation of single-Case Designs. Journal of Applied Behavior Analysis, 36(3), 387–406. https://doi.org/10.1901/jaba.2003.36-387

Hartmann, D. P. y Hall, R. V. (1976). The Changing Criterion Design. *Journal of Applied Behavior Analysis, 9*(4), 527–532. https://doi.org/10.1901/jaba.1976.9-527

Heyvaert, M. y Onghena, P. (2014). Analysis of Single-Case Data: Randomisation Tests for Measures of Effect Size. *Neuropsychological Rehabilitation, 24*(3–4), 507–527. https://doi.org/10.1080/09602011.2013.818564

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. y Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children, 71*(2),

165−179. https://doi.org/10.1177/001440290507100203

Jacobs, K. W. (2019). Replicability and Randomization Test Logic in Behavior Analysis. *Journal of the Experimental Analysis of Behavior, 111*(2), 329–341. https://doi.org/10.1002/jeab.501

Kratochwill, T. R., Horner, R. H., Levin, J. R., Machalicek, W., Ferron, J. y Johnson, A. (2021). Single-case Design Standards: An Update and Proposed Upgrades. *Journal of School Psychology, 89*, 91–105. https://doi.org/10.1016/j.jsp.2021.10.006

Klein, L. A., Houlihan, D., Vincent, J. L. y Panahon, C. J. (2017). Best Practices in Utilizing the Changing Criterion Design. *Behavior Analysis in Practice, 10*(1), 52–61. https://doi.org/10.1007/s40617-014-0036-x

Lane, J. D. y Gast, D. L. (2014). Visual Analysis in Single Case Experimental Design Studies: Brief Review and Guidelines. *Neuropsychological Rehabilitation, 24*(3–4), 445–463. https://doi.org/10.1080/09602011.2013.815636

Ledford, J. R., Barton, E. E., Severini, K. E. y Zimmerman, K. N. (2019). A Primer on Single-Case Research Designs: Contemporary Use and Analysis. *American Journal on Intellectual and Developmental Disabilities, 124*(1), 35–56. https://doi.org/10.1352/1944-7558-124.1.35

Levin, J. R., Ferron, J. M. y Gafurov, B. S. (2018). Comparison of Randomization-Test Procedures for Single-Case Multiple-Baseline Designs. *Developmental Neurorehabilitation, 21*(5), 290–311. https://doi.org/10.1080/17518423.2016.1197708

Ma, H. H. (2006). An Alternative Method for Quantitative Synthesis of Single-Subject Research: Percentage of Data Points Exceeding the Median. *Behavior Modification, 30*(5), 598–617. https://doi.org/10.1177/0145445504272974

Maggin, D. M., Cook, B. G. y Cook, L. (2018). Using Single-Case Research Designs to Examine the Effects of Interventions in Special Education. *Learning Disabilities Research & Practice, 33*(4), 182–191. https://doi.org/10.1111/ldrp.12184

Maggin, D. M., Cook, B. G. y Cook, L. (2019). Making Sense of Single-Case Design Effect Sizes. *Learning Disabilities Research & Practice, 34*(3), 124–132. https://doi.org/10.1111/ldrp.12204

Manolov, R., Moeyaert, M. y Fingerhut, J. (2022). A Priori Justification for Effect Measures in Single-Case Experimental Designs. *Perspectives on Behavior Science, 45*(1), 156–189. https://doi.org/10.1007/s40614-021-00282-2

Manolov, R. y Onghena, P. (2018). Analyzing Data from Single-Case Alternating Treatments Designs. *Psychological Methods, 23*(3), 480–504. https://doi.org/10.1037/met0000133

Manolov, R. y Onghena, P. (2022). Defining and Assessing Immediacy in Single Case Experimental Designs. *Journal of the Experimental Analysis of Behavior, 118*(3), 462−492. https://doi.org/10.1002/JEAB.799

Manolov, R. y Tanious, R. (2022). Assessing Consistency in Single-Case Data Features using Modified Brinley Plots. *Behavior Modification, 46*(3), 581–627. https://doi.org/10.1177/0145445520982969

Manolov, R., Tanious, R. y Fernández-Castilla, B. (2022). A Proposal for the Assessment of Replication of Effects in Single-Case Experimental Designs. *Journal of Applied Behavior Analysis, 55*(3), 997–1024. https://doi.org/10.1002/jaba.923

McDougale, C. B., Richling, S. M., Longino, E. B. y O'Rourke, S. A. (2020). Mastery Criteria and Maintenance: A Descriptive Analysis of Applied Research Procedures. *Behavior Analysis in*

*Practice, 13*(2), 402–410. https://doi.org/10.1007/s40617-019-00365-2

McDougall, D. (2005). The Range-Bound Changing Criterion Design. *Behavioral Interventions, 20*(2), 129–137. https://doi.org/10.1002/bin.189

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N. y Van den Noortgate, W. (2014). The Influence of the Design Matrix on Treatment Effect Estimates in the Quantitative Analyses of Single-Case Experimental Designs Research. *Behavior Modification, 38*(5), 665–704. https://doi.org/10.1177/0145445514535243

Natesan, P. y Hedges, L. V. (2017). Bayesian Unknown Change-Point Models to Investigate Immediacy in Single Case Designs. P*sychological Methods, 22*(4), 743–759. https://doi.org/10.1037/met0000134

Onghena, P. (1992). Randomization Tests for Extensions and Variations of ABAB single-Case Experimental Designs: A Rejoinder. *Behavioral Assessment, 14*(2), 153–171.

Onghena, P. y Edgington, E. S. (1994). Randomization Tests for Restricted Alternating Treatments Designs. *Behaviour Research and Therapy, 32*(7), 783–786. https://doi.org/10.1016/0005-7967(94)90036-1

Onghena, P., Tanious, R., De, T. K. y Michiels, B. (2019). Randomization Tests for Changing Criterion Designs. *Behaviour Research and Therapy, 117*(6), 18–27. https://doi.org/10.1016/j.brat.2019.01.005

Parker, R. I. y Vannest, K. J. (2009). An Improved Effect Size for Single-Case Research: Nonoverlap of all Pairs. *Behavior Therapy, 40*(4), 357−367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J. y Davis, J. L. (2011). Effect Size in Single-Case Research: A Review of Nine Nonoverlap Techniques. *Behavior Modification, 35*(4), 303–322. https://doi.org/10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L. y Sauber, S. B. (2011). Combining Nonoverlap and Trend for Single-Case Research: Tau-U. *Behavior Therapy, 42*(2), 284−299. https://doi.org/10.1016/j.beth.2010.08.006

Perdices, M., Tate, R. L. y Rosenkoetter, U. (2023). An Algorithm to Evaluate Methodological Rigor and Risk of Bias in Single-Case Studies. *Behavior Modification, 47*(6), 1482–1509. https://doi.org/10.1177/0145445519863035

Pustejovsky, J. E. (2018). Using Response Ratios for Meta-Analyzing Single-Case Designs with Behavioral Outcomes. *Journal of School Psychology, 68*(6), 99−112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., Hedges, L. V. y Shadish, W. R. (2014). Design-Comparable Effect Sizes in Multiple Baseline Designs: A General Modeling Framework. *Journal of Educational and Behavioral Statistics, 39*(5), 368–393. https://doi.org/10.3102/1076998614547577

Scruggs, T. E., Mastropieri, M. A. y Casto, G. (1987). The Quantitative Synthesis of Single-Subject Research: Methodology and Validation. *Remedial and Special Education, 8*(2), 24–33. https://doi.org/10.1177/074193258700800206

Shadish, W. R., Hedges, L. V. y Pustejovsky, J. E. (2014). Analysis and Meta-Analysis of Single-Case Designs with a standardized Mean Difference Statistic: A Primer and Applications. *Journal of School Psychology, 52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Slocum, T. A., Pinkelman, S. E., Joslyn, P. R. y Nichols, B. (2022). Threats to Internal Validity in Multiple-baseline Design Variations. *Perspectives on Behavior Science, 45*(3), 619−638. https://doi.org/10.1007/s40614-022-00326-1

Snodgrass, M., Cook, B. G. y Cook, L. (2023). Considering Social Validity in Special Education Research. *Learning Disabilities Research & Practice, 38*(4), 311–319. https://doi.org/10.1111/ldrp.12326

Swaminathan, H., Rogers, H. J., Horner, R., Sugai, G. y Smolkowski, K. (2014). Regression Models for the Analysis of Single Case Designs. *Neuropsychological Rehabilitation, 24*(3–4), 554−571. https://doi.org/10.1080/09602011.2014.887586

Tanious, R. y Onghena, P. (2021). A Systematic Review of Applied Single-Case Research Published between 2016 and 2018: Study Designs, Randomization, Data Aspects, and Data Analysis. *Behavior Research Methods, 53*(4), 1371–1384. https://doi.org/10.3758/s13428-020-01502-4

Tate, R. L. y Perdices, M. (2019). *Single-case Experimental Designs for Clinical Research and Neurorehabilitation Settings: Planning, Conduct, Analysis, and Reporting*. Routledge.

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., Horner, R., Kratochwill, T., Barlow, D. H., Kazdin, A. E., Sampson, M., Shamseer, L. y Vohra, S. (2016). The Single-Case Reporting Guideline in Behavioural Interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology, 4*(1), 10–31. https://doi.org/10.1037/arc0000027

Vannest, K. J. y Sallese, M. R. (2021). Benchmarking Effect Sizes in Single-Case Experimental Designs. *Evidence-Based Communication Assessment and Intervention, 15*(3), 142–165. https://doi.org/10.1080/17489539.2021.1886412

What Works Clearinghouse. (2022). *Procedures and Standards Handbook, Version 5.0. U.S.* Department of Education, Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf

Wine, B., Freeman, T. R. y King, A. (2015). Withdrawal versus Reversal: A Necessary Distinction? *Behavioral Interventions, 30*(1), 87–93. https://doi.org/10.1002/bin.1399

Wolfe, K., Barton, E. E. y Meadan, H. (2019). Systematic Protocols for the Visual Analysis of Single-Case Research Data. *Behavior Analysis in Practice, 12*(2), 491–502. https://doi.org/10.1007/s40617-019-00336-7

# INTRODUCCIÓN AL METAANÁLISIS MULTIVARIADO CON MODELOS DE ECUACIONES ESTRUCTURALES

# INTRODUCTION TO META-ANALYTIC STRUCTURAL EQUATION MODELING

JOSÉ ANTONIO LÓPEZ-LÓPEZ[1],
RAIMUNDO AGUAYO-ESTREMERA[2], LAURA BADENES-RIBERA[3]
Y BELÉN FERNÁNDEZ-CASTILLA[4]

Cómo referenciar este artículo/How to reference this article:

López-López, J. A., Aguayo-Estremera, R., Badenes-Ribera, L. y Fernández-Castilla, B. (2025). Introducción al metaanálisis multivariado con Modelos de Ecuaciones Estructurales [Introduction to Meta-Analytic Structural Equation Modeling]. *Acción Psicológica, 22*(1), 23–40. https://doi.org/10.5944/ap.22.1.43279

## Resumen

El metaanálisis con modelos de ecuaciones estructurales (Meta-Analytic Structural Equation Modeling, MASEM) es una metodología novedosa de síntesis de la evidencia que combina las ventajas del metaanálisis tradicional con los modelos de ecuaciones estructurales, permitiendo explorar relaciones complejas entre variables a través de la integración de estudios independientes. Este tutorial ofrece una introducción accesible a la metodología MASEM, incluyendo los conceptos fundamentales, los pasos metodológicos para su implementación, y una aplicación práctica usando software estadístico libre. El

**Correspondence address [Dirección para correspondencia]**: José Antonio López López, Facultad de Psicología y Logopedia, Universidad de Murcia, España.

**Email**: josealopezlopez@um.es

**ORCID**: José Antonio López López (https://orcid.org/0000-0002-9655-3616), Raimundo Aguayo-Estremera (https://orcid.org/0000-0001-7276-9394), Laura Badenes-Ribera (https://orcid.org/0000-0002-4706-690X) y Belén Fernández-Castilla (https://orcid.org/0000-0002-3451-0637).

[1] Universidad de Murcia, España.
[2] Universidad Complutense de Madrid, España.
[3] Universitat de Valencia, España.
[4] Universidad Nacional de Educación a Distancia, España.

Recibido: 21 de enero de 2025
Aceptado: 25 de febrero de 2025

objetivo es proporcionar a los investigadores una comprensión sólida y las herramientas necesarias para aplicar MASEM en sus propios estudios. Para ello se aborda desde la preparación de datos hasta la interpretación de resultados, destacando tanto las fortalezas como las limitaciones de esta metodología.

**Palabras clave:** MASEM; Metaanálisis; Modelos de ecuaciones estructurales; Síntesis de la evidencia.

## Abstract

Meta-Analytic Structural Equation Modeling (MASEM) is a novel methodology for evidence synthesis which combines the advantages of traditional meta-analysis and structural equation modeling, allowing the exploration of complex relationships among variables through the integration of independent studies. The present tutorial offers an accessible introduction to the MASEM methodology, including the key concepts involved, methodological steps for implementation, and a practical example using free statistical software. The goal is to provide researchers with a solid understanding and with the necessary tools to apply MASEM techniques to their own studies. To that aim, the process from data preparation to results interpretation is covered, highlighting the strengths and limitations of this methodology.

**Keywords:** MASEM; Meta-analysis; Structural equation models; Evidence synthesis.

## Introducción

El incremento masivo de artículos científicos publicados cada año plantea desafíos para el seguimiento de los desarrollos y avances en áreas específicas del conocimiento. Para enfrentar esta sobrecarga de información, en la década de 1970 surgieron diversas técnicas diseñadas para sintetizar la evidencia científica acumulada en un campo de investigación determinado. Entre ellas, destaca el metaanálisis (Glass, 1976), definido como un conjunto de métodos estadísticos que permiten combinar de forma cuantitativa los resultados de estudios que investigan cuestiones similares, con el fin de generar conclusiones más robustas y precisas (Cooper et al., 2019).

La mayoría de los metaanálisis publicados hasta la fecha en Psicología se han centrado en sintetizar la evidencia disponible sobre la asociación entre dos variables. Por ejemplo, Hattie (2009) realizó una meta-revisión de más de 800 metaanálisis que estudiaban la relación bivariada entre diversas características (de los estudiantes, de los profesores, de las políticas) y el rendimiento académico. Sin embargo, muchas preguntas de investigación en Psicología son multivariantes, es decir, involucran la comprensión de cómo múltiples variables interactúan de forma compleja para influir en un determinado resultado. Por ejemplo, ¿cómo interactúan conjuntamente las características familiares, contextuales y del estudiante para determinar el rendimiento académico?

Los enfoques tradicionales de metaanálisis no permiten responder a preguntas de investigación de naturaleza multivariante. Para superar esta limitación, Becker (1992, 1995; ver también Becker y Aloe, 2019) propuso un enfoque meta-analítico basado en modelos, es decir, yendo más allá de las asociaciones bivariadas. Este enfoque permite la evaluación de teorías, y con ello ofrece una comprensión más profunda de fenómenos complejos, algo fundamental para el avance del conocimiento científico en Psicología y ciencias afines.

Una metodología clave en este contexto es el metaanálisis de modelos de ecuaciones estructurales (en inglés, *Meta-Analytic Structural Equation Modeling*, MASEM), el cual ha ido ganando relevancia a lo largo de las últimas décadas (Cheung, 2015, 2019; Cheung y Chan, 2005; Jak, 2015; Viswesvaran y Ones, 1995). Debido a su creciente importancia y a su gran potencial para la investigación psicológica, el objetivo de este trabajo es ofrecer una introducción accesible en español al MASEM, así como proporcionar herramientas para su implementación. Para lectores/as sin experiencia previa con técnicas de metaanálisis univariado, el estudio de las referencias de la siguiente

sección resultará de gran utilidad para comprender las secciones posteriores.

## MASEM: encajando las piezas

En metaanálisis, los datos individuales de los estudios no suelen estar disponibles para su integración. En consecuencia, se recurre a indicadores numéricos de la magnitud de la asociación entre las variables de interés, conocidos como tamaños del efecto (v.g., correlación de Pearson), que pueden calcularse fácilmente a partir de la información disponible en los estudios primarios (Grissom y Kim, 2012).

A la hora de combinar los tamaños del efecto en el metaanálisis, existen dos tipos principales de modelos, efectos fijos y efectos aleatorios (Hedges y Vevea, 1998). Por un lado, el modelo de efectos fijos (*Fixed-Effects Model*) asume que los tamaños del efecto varían entre estudios debido únicamente a que las muestras utilizadas son diferentes. El modelo de efectos aleatorios (*Random-Effects Model*), en cambio, asume que la variabilidad observada entre tamaños del efecto no se debe únicamente a la diferencia entre muestras (variabilidad muestral) sino también a la diferencia entre las características de los estudios (variabilidad inter-estudios). El uso de un modelo u otro dependerá de consideraciones teóricas *(¿realmente puedo considerar estos estudios como réplicas?*) y/o empíricas *(¿se observa tanta variabilidad inter-estudios como para concluir que los estudios provienen de poblaciones diferentes?*), así como del grado de generalización que se persigue (estrictamente, solo el modelo de efectos aleatorios permite generalizar las conclusiones más allá de la muestra de estudios incluidos, que se asume que representan una muestra razonablemente representativa de una población más amplia). Además, el modelo de efectos aleatorios implica una mayor complejidad, por lo que no es aconsejable para aplicaciones con muy pocos estudios. Para ampliar información sobre metaanálisis, recomendamos los manuales de Cooper et al. (2019) y Botella y Sánchez-Meca (2015).

Por otra parte, los modelos de ecuaciones estructurales (*Structural Equation Modeling,* SEM) constituyen un conjunto de técnicas estadísticas que facilitan la evaluación de las interrelaciones entre variables, ya hayan sido éstas directamente observadas o se hayan construido a partir de un conjunto de variables observadas (variables latentes). A diferencia de los métodos tradicionales como la regresión múltiple que tiende a modelar relaciones unidireccionales y analizar constructos de forma independiente, los modelos SEM permiten modelar múltiples relaciones simultáneamente, lo que los hace útiles para evaluar modelos teóricos más complejos. También constituyen una potente herramienta en la investigación psicométrica, donde una de sus aplicaciones más comunes es el análisis factorial confirmatorio, utilizado para evaluar la estructura interna de un test o cuestionario.

El punto de partida para ajustar un modelo SEM es la matriz de varianzas-covarianzas de los datos y el tamaño muestral. Una vez se ajusta el modelo deseado propuesto por el/la investigador/a, se obtiene la matriz de varianzas-covarianzas reproducida, es decir, la matriz que se desprendería del modelo estructural propuesto. El ajuste del modelo se evalúa en base a las diferencias entre la matriz de varianzas-covarianzas observada y la reproducida, utilizando para ello diversos índices de bondad de ajuste. Cuanto menor es la discrepancia entre estas dos matrices, mejor ajustarían los datos al modelo propuesto. Cuando se aplican como parte de un MASEM, los modelos SEM suelen utilizar matrices de correlaciones no transformadas (cf. Cheung y Chan, 2005) en lugar de matrices de varianzas-covarianzas, ya que las primeras permiten extraer la información en una métrica común para todos los estudios primarios. Para más información, redirigimos al lector a Kline (2023), Lei y Wu (2007), y Ruiz y col. (2010).

La combinación de los modelos SEM con las técnicas meta-analíticas ha culminado en los modelos MASEM, bajo cuyo nombre hay un conjunto de procedimientos creciente que se pueden agrupar en dos categorías generales: enfoques basados en correlaciones y basados en parámetros (ver Tabla 1). La diferencia radica en que, mientras en el primer caso se sintetizan matrices de correlaciones entre variables, en el segundo caso la síntesis se hace con los tamaños del efecto del modelo (e.g., coeficientes de correlación parcial, pesos de regresión, etc.).

**Tabla 1**

*Resumen de los principales procedimientos MASEM*

| Tipo | Procedimiento | Autores principales | Descripción general |
|---|---|---|---|
| Basado en correlaciones | Two-Stage (TSSEM) | Cheung (2015) Cheung y Chan (2005) | Se sintetizan las matrices de correlaciones de cada estudio primario y luego se ajusta el modelo SEM. |
| | One-Stage (OSMASEM) | Jak y Cheung (2020) | La síntesis de las matrices de correlaciones y el ajuste se hace en un solo paso. |
| Basado en parámetros | Two-Stage | Cheung y Cheung (2016) | Primero se ajusta el modelo SEM en cada estudio y luego se sintetizan los parámetros de interés de los modelos resultantes. |

*Nota. Este resumen no es exhaustivo, ya que existen otros procedimientos que no se abordarán en este manuscrito debido a su uso limitado.*

## Enfoques MASEM

En esta sección se presentan los principales enfoques metodológicos empleados en la actualidad para el ajuste de modelos MASEM, aplicándolos a un metaanálisis realizado por Cano-López y colaboradores (2022) sobre el modelo metacognitivo de la rumiación y la depresión (Figura 1) propuesto por Papageorgiou y Wells (2003). La base de datos completa, que incluye 15 estudios y las correlaciones entre las variables presentados en Figura 1, puede encontrarse en https://osf.io/cq74m/?view_only=87ecffb55d87455e8a50b97e4e9b5e3d

### *MASEM basado en correlaciones en dos etapas (TSSEM)*

La primera etapa de este enfoque consiste en combinar meta-analíticamente las matrices de correlaciones entre las variables integrantes del modelo examinado. Para ello, también se extraen los tamaños muestrales de los estudios primarios, con el fin de ponderar cada coeficiente de correlación de acuerdo con su precisión. En una segunda etapa, se ajusta el modelo SEM sobre la matriz combinada obtenida en el paso anterior (Cheung y Chan, 2005). A continuación, explicamos cómo realizar estos análisis utilizando R. En el material suplementario (https://osf.io/cq74m/?view_only=87ecffb55d87455e8a50b97e4e9b5e3d) también se incluye un tutorial sobre cómo realizar estos análisis utilizando una aplicación web,

**Figura 1**

*Modelo metacognitivo de la rumiación y la depresión (Papageorgiou y Wells, 2003)*

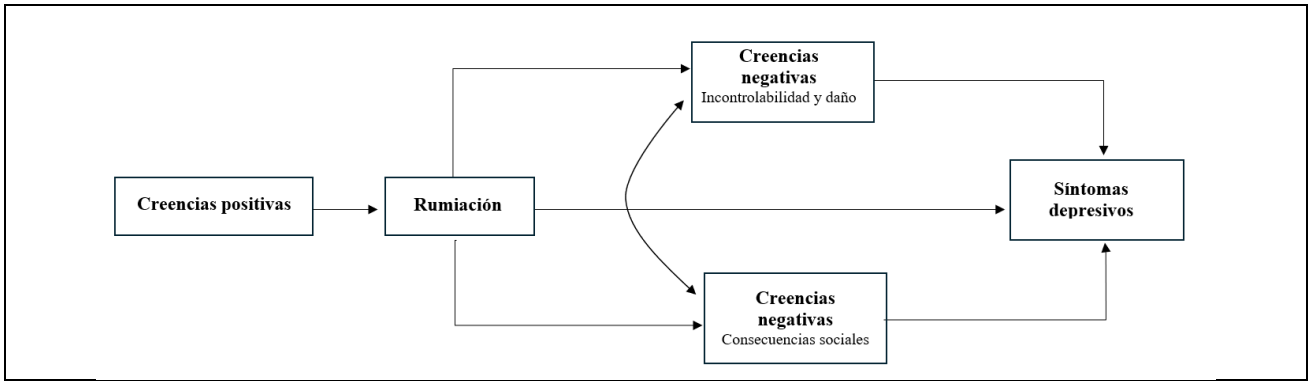


**Tabla 2**

*Matrices de correlaciones de los dos primeros estudios analizados*

| Estudio 1 | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1. Creencias positivas | 1 | | | | |
| 2. Creencias negativas (I) | - | 1 | | | |
| 3. Creencias negativas (II) | - | - | 1 | | |
| 4. Rumiación | .45 | - | - | 1 | |
| 5. Depresión | .53 | - | - | - | 1 |
| Estudio 2 | | | | | |
| | 1 | 2 | 3 | 4 | 5 |
| 1. Creencias positivas | 1 | | | | |
| 2. Creencias negativas (I) | .21 | 1 | | | |
| 3. Creencias negativas (II) | .12 | .57 | 1 | | |
| 4. Rumiación | .58 | .51 | .39 | 1 | |
| 5. Depresión | .23 | .35 | .47 | .47 | 1 |

webMASEM (Jak et al., 2021) que ha sido creada específicamente para aplicar de forma sencilla esta metodología.

En el ejemplo propuesto, comenzamos extrayendo de cada estudio las correlaciones de Pearson entre las cinco variables que conforman el modelo (Tabla 2), así como el tamaño muestral de cada estudio. Encontramos que hay estudios que no incluyen todas las correlaciones de interés, lo cual sucede habitualmente porque no se han estudiado todas las variables del modelo, en cuyo caso se asumen como *variables perdidas* y simplemente se filtran en los análisis MASEM. Un escenario diferente es el de las *correlaciones perdidas*, que se produce cuando los autores no reportan en el manuscrito alguna de estas correlaciones por no ser de su interés (este es el caso de estudio 1 de la Tabla 2, donde los autores no reportan la correlación entre rumiación y depresión a pesar de que ambas variables se habían estudiado). Este segundo escenario sí es problemático cuando se aplica un modelo de efectos fijos (cf. Jak y Cheung, 2018a), en cuyo caso habrá que considerar como perdida aquella variable de la que no se reportan todas sus correlaciones.

Una vez se han extraído las correlaciones y sus tamaños muestrales, éstas se combinan utilizando técnicas de metaanálisis multivariado para obtener una matriz de correlaciones combinada. Becker (1992) propuso inicialmente realizar esta combinación de matrices utilizando el método de estimación por *Mínimos Cuadrados Generalizados*. Sin embargo, en MASEM se realiza esta combinación utilizando modelos SEM. En caso de aplicar un modelo de efectos fijos, la matriz de correlaciones combinada se obtendría tras realizar un análisis multigrupo (Cheung y Chan, 2005). Si por el contrario se desea aplicar un modelo de efectos aleatorios, se emplean otras técnicas más complejas, cuyos detalles se pueden consultar en Cheung (2013). Como ya se ha mencionado, la decisión de ajustar un modelo u otro dependerá de consideraciones teóricas y/o prácticas.

Desde un punto de vista teórico, los estudios que se van a sintetizar en este ejemplo presentan diferencias sustanciales en cuanto a sus características y procedimientos, lo que sugiere la necesidad de aplicar un modelo de efectos aleatorios. No obstante, con fines didácticos, iniciamos el análisis utilizando un modelo de efectos fijos, con el objetivo de ilustrar cómo evaluar su ajuste y determinar, desde una perspectiva empírica, si es el modelo más adecuado para estos datos.

Para ajustar cualquiera de los modelos, se utiliza la función `tssem1` del paquete *metaSEM*. En el código de R (disponible en https://osf.io/cq74m/?view_only=87ecffb55d87455e8a50b97e4e9b5e3d), se explica paso a paso cómo preprocesar los datos para obtener los objetos que se utilizan en los

próximos comandos, que son la lista de matrices de correlaciones (`corr_lista$data`) y los tamaños muestrales (`corr_lista$n`). Este código también incluye los comandos para eliminar las variables que contienen correlaciones perdidas, paso que sólo es necesario en el caso de aplicar un modelo de efectos fijos[1]. Además, tenemos que indicar qué modelo queremos ajustar indicándolo en el comando `method`, donde `"FEM"` se referirá a un modelo de efectos fijos.

```
paso1_EF <- tssem1(Cov=corr_lista$data, n=corr_lista$n, method="FEM")
```

Después de realizar los análisis, podemos imprimir los resultados (Figura 2) con el comando `summary(paso1_EF)`.

En los resultados, primero se observan las correlaciones combinadas en forma de vector. Antes de interpretarlas, es crucial evaluar el ajuste del modelo para determinar si es adecuado aplicar un modelo de efectos fijos. Un test $\chi^2$ significativo ($\chi^2 = 212.46$, $p < .001$) indica que no podemos asumir que todas las correlaciones provienen de una misma población, lo que sugiere heterogeneidad entre los estudios. Según Hu y Bentler (1999), un RMSEA < .06 y un CFI > .95 indican un buen ajuste. En este caso, el CFI cumple (.965), pero el RMSEA es superior (.094), lo que

**Figura 2**

*Resultados al ejecutar el comando `summary (paso1_EF)`*

```
Coefficients:
      Estimate Std.Error z value             Pr(>|z|)
S[1,2] 0.413571  0.013835 29.8933 < 0.00000000000000022 ***
S[1,3] 0.235075  0.020082 11.7055 < 0.00000000000000022 ***
S[1,4] 0.141243  0.017255  8.1854  0.00000000000000222 ***
S[1,5] 0.262661  0.014847 17.6908 < 0.00000000000000022 ***
S[2,3] 0.560036  0.015706 35.6582 < 0.00000000000000022 ***
S[2,4] 0.420090  0.014667 28.6424 < 0.00000000000000022 ***
S[2,5] 0.603027  0.011066 54.4950 < 0.00000000000000022 ***
S[3,4] 0.622322  0.014298 43.5260 < 0.00000000000000022 ***
S[3,5] 0.549116  0.015895 34.5473 < 0.00000000000000022 ***
S[4,5] 0.466001  0.013482 34.5650 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Goodness-of-fit indices:
                                      Value
Sample size                        4384.0000
Chi-square of target model          212.4604
DF of target model                   59.0000
p value of target model               0.0000
Chi-square of independence model   4441.1329
DF of independence model             69.0000
RMSEA                                 0.0943
RMSEA lower 95% CI                    0.0810
RMSEA upper 95% CI                    0.1084
SRMR                                  0.0702
TLI                                   0.9590
CFI                                   0.9649
AIC                                  94.4604
BIC                                -282.2969
OpenMx status1: 0 ("0" or "1": The optimization is considered fine.
Other values may indicate problems.)
```

[1] Este código ha sido adaptado del manual de Jak (2015).

sugiere que el ajuste es cuestionable. Por lo tanto, los datos empíricos también invitan al ajuste de un modelo de efectos aleatorios.

Para ajustar un modelo de efectos aleatorios, utilizamos la misma función, modificando el método a `REM`. En caso de contar con un número reducido de estudios, como ocurre en este ejemplo, es recomendable emplear el comando `RE.type = "Diag"`, que supone asumir que los elementos fuera de la diagonal principal de varianzas son 0, y por lo tanto que los efectos aleatorios de las diferentes correlaciones no covarían entre sí. A pesar de que esta suposición puede no ser del todo realista, si el número de estudios es limitado, podría no ser viable estimar esas covarianzas[2] (Cheung y Cheung, 2016). El código quedaría de la siguiente manera:

```
paso1_EA <- tssem1(cons_lista$data, cons_lista$n, method="REM", RE.type = "Diag")
```

Y para obtener los resultados (Figura 3), ejecutaríamos `summary(paso1_EA)`.

**Figura 3.**

*Resultados al ejecutar el comando `summary(paso1_EA)`*

```
95% confidence intervals: z statistic approximation (robust=FALSE)
Coefficients:
               Estimate     Std.Error        lbound        ubound z value              Pr(>|z|)
Intercept1   0.2611863348  0.0212227845  0.2195904416  0.3027822281 12.3069 < 0.00000000000000022 ***
Intercept2   0.1567363798  0.0288480609  0.1001952193  0.2132775402  5.4332         0.00000005536 ***
Intercept3   0.4285081680  0.0258720939  0.3777997957  0.4792165403 16.5626 < 0.00000000000000022 ***
Intercept4   0.2609284719  0.0145639321  0.2323836895  0.2894732542 17.9161 < 0.00000000000000022 ***
Intercept5   0.6162826120  0.0269296426  0.5635014825  0.6690637416 22.8849 < 0.00000000000000022 ***
Intercept6   0.5144631575  0.0390984300  0.4378316428  0.5910946722 13.1582 < 0.00000000000000022 ***
Intercept7   0.4725051065  0.0401927328  0.3937287978  0.5512814152 11.7560 < 0.00000000000000022 ***
Intercept8   0.4091792592  0.0380579682  0.3345870121  0.4837715063 10.7515 < 0.00000000000000022 ***
Intercept9   0.4262195324  0.0360221352  0.3556174447  0.4968216201 11.8322 < 0.00000000000000022 ***
Intercept10  0.5408788265  0.0349964957  0.4722869553  0.6094706978 15.4552 < 0.00000000000000022 ***
Tau2_1_1     0.0007122902  0.0010232222 -0.0012931884  0.0027177689  0.6961               0.48635
Tau2_2_2     0.0011059318  0.0019908969 -0.0027961545  0.0050080181  0.5555               0.57856
Tau2_3_3     0.0052921247  0.0031097589 -0.0008028908  0.0113871401  1.7018               0.08880 .
Tau2_4_4     0.0000000001            NA            NA            NA      NA                    NA
Tau2_5_5     0.0000000001  0.0012646191 -0.0024786079  0.0024786081  0.0000               1.00000
Tau2_6_6     0.0075913677  0.0066750271 -0.0054914450  0.0206741803  1.1373               0.25542
Tau2_7_7     0.0099851791  0.0063484168 -0.0024574892  0.0224278475  1.5729               0.11575
Tau2_8_8     0.0005353539  0.0036079540 -0.0065361061  0.0076068138  0.1484               0.88204
Tau2_9_9     0.0051328840  0.0053905891 -0.0054324764  0.0156982445  0.9522               0.34100
Tau2_10_10   0.0088863969  0.0053646502 -0.0016281244  0.0194009181  1.6565               0.09763 .
Q statistic on the homogeneity of effect sizes: 301.8551
Degrees of freedom of the Q statistic: 72
P value of the Q statistic: 0

Heterogeneity indices (based on the estimated Tau2):
                              Estimate
Intercept1: I2 (Q statistic)    0.1797
Intercept2: I2 (Q statistic)    0.2335
Intercept3: I2 (Q statistic)    0.6690
Intercept4: I2 (Q statistic)    0.0000
Intercept5: I2 (Q statistic)    0.0000
Intercept6: I2 (Q statistic)    0.8132
Intercept7: I2 (Q statistic)    0.8375
Intercept8: I2 (Q statistic)    0.1716
Intercept9: I2 (Q statistic)    0.6848
Intercept10: I2 (Q statistic)   0.8470

Number of studies (or clusters): 15
Number of observed statistics: 82
Number of estimated parameters: 20
Degrees of freedom: 62
-2 log likelihood: -178.1773
OpenMx status1: 6 ("0" or "1": The optimization is considered fine.
Other values may indicate problems.)
```

[2] Si hay un número de estudios sustancial, se puede utilizar el comando `RE.type = "Symm"` para estimar estas covarianzas.

**Tabla 3**

*Matriz de correlación combinada, que se utilizará para ajustar un modelo SEM en la etapa 2. Entre paréntesis aparece el valor del estadístico $I^2$*

| Estudio 1 | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1. Creencias positivas | 1 | | | | |
| 2. Creencias negativas (I) | .26 *(18%)* | 1 | | | |
| 3. Creencias negativas (II) | .16 *(23%)* | .62 *(0%)* | 1 | | |
| 4. Rumiación | .43 *(67%)* | .51 *(81%)* | .41 *(17%)* | 1 | |
| 5. Depresión | .26 *(0%)* | .47 *(84%)* | .43 (68%) | .54 *(85%)* | 1 |

En los resultados, observamos que no sólo se imprime el vector que contiene las correlaciones combinadas en las líneas que comienzan con la palabra `Intercept` (en la Tabla 3 se incluyen estas correlaciones en forma de matriz), sino también un vector con las varianzas inter-estudios de estas correlaciones ($\tau^2$). Para poder evaluar la magnitud de $\tau^2$con respecto a la varianza total de cada correlación, podemos revisar los valores de los estadísticos $I^2$, que nos indicarían el porcentaje de la variabilidad total que se debe a heterogeneidad entre estudios (Higgins y Thompson, 2002). Es importante interpretar el índice $I^2$ en términos relativos (Borenstein et al., 2017).

Como puede verse en la Tabla 3, las correlaciones más altas las encontramos entre creencias negativas (I) y creencias negativas (II), seguido de rumiación y depresión. Las correlaciones combinadas en las que se observa un mayor porcentaje de variabilidad debido a heterogeneidad entre estudios son las correlaciones entre rumiación y depresión. En esta fase, también se estima la matriz de varianzas-covarianzas de las correlaciones de la Tabla 2, que puede obtenerse mediante el comando `vcov(paso1_EA)`. Esta es la matriz que se utilizará en el paso 2 para asignar pesos a las correlaciones en función de su precisión. La precisión de estas correlaciones depende del número de estudios, el tamaño muestral, y su heterogeneidad.

A continuación, podemos ejecutar el paso 2, consistente en ajustar el modelo SEM deseado sobre nuestra matriz de correlaciones combinada. Para especificarlo, empleamos el modelo de acción reticular (*Reticular Action Model*; McArdle y McDonald 1984), que implica la creación de tres matrices: En la matriz A se definen todos los coeficientes de regresión que se desean estimar, en la matriz S se establecen las varianzas (que también han de estimarse), y, por último, si el modelo incluye variables latentes, se tendrían que especificar en la matriz F. El código de R disponible en el material suplementario incorpora más instrucciones sobre cómo crear estas matrices.

Una vez creadas, podemos utilizar la función `tssem2` para obtener las estimaciones de los parámetros del modelo SEM. Al aplicar la función tssem2, aparte de indicar nuestra matriz A y S, también hay otros dos comandos a los que atender. El primero es `diag.constraints = TRUE`, el cual garantiza que, al ajustar el modelo, se tenga en cuenta que la diagonal de la matriz de correlaciones combinada está compuesta exclusivamente por 1s. El segundo comando, `intervals = "LB"`, se refiere al método utilizado para obtener los intervalos de confianza. En este caso hemos seleccionado intervalos basados en verosimilitud (likelihood-based intervals) ya que tienen un rendimiento superior a los intervalos de confianza tradicionales basados en los errores típicos (cf. Cheung, 2009), y además son el único método aplicable cuando `diag.constraints=TRUE`.

```
paso2_EA <- tssem2(cons_re, Amatrix = A, Smatrix = S, diag.constraints=TRUE, intervals = "LB")
```

Para obtener los resultados (Figura 4), ejecutamos el siguiente comando: `summary(paso2_EA)`[3]

Al igual que en el paso 1, antes de explorar las estimaciones del modelo debemos valorar su ajuste. Se puede observar que tanto el RMSEA como en CFI están dentro de los límites recomendados por Hu y Bentler (1999), por lo que el ajuste del modelo propuesto parece adecuado. Por lo tanto, podemos explorar las estimaciones del modelo (ver Figura 1). A excepción de la relación entre creencias negativas (I) y depresión, el resto de las estimaciones son estadísticamente significativas.

En MASEM, también podemos valorar la significación de los efectos indirectos entre variables, como por ejemplo la influencia indirecta de creencias positivas (PB) sobre creencias negativas I (NBI) a través de la variable rumiación (R). Para ello, dentro del comando `tssem2`, debemos especificar qué efecto indirecto queremos estimar, especificando el producto de los efectos directos que lo componen:

**Figura 4**

*Resultados al ejecutar el comando `summary(paso2_EA)`*

```
95% confidence intervals: Likelihood-based statistic
Coefficients:
            Estimate Std.Error    lbound    ubound z value Pr(>|z|)
NB12DEP     0.138858        NA -0.026484  0.298862      NA       NA
NB22DEP     0.171208        NA  0.053039  0.290443      NA       NA
R2DEP       0.420294        NA  0.332733  0.512465      NA       NA
R2NB1       0.549981        NA  0.488819  0.611863      NA       NA
R2NB2       0.418253        NA  0.344577  0.492076      NA       NA
PB2R        0.447405        NA  0.405899  0.489763      NA       NA
ErrorVarDEP 0.620617        NA  0.564849  0.672057      NA       NA
ErrorVarNB1 0.697521        NA  0.625722  0.761266      NA       NA
CorNB1NB2   0.395774        NA  0.351622  0.438576      NA       NA
ErrorVarNB2 0.825065        NA  0.757886  0.881301      NA       NA
ErrorVarR   0.799829        NA  0.760129  0.835208      NA       NA

Goodness-of-fit indices:
                                              Value
Sample size                                4384.0000
Chi-square of target model                    5.1796
DF of target model                            3.0000
p value of target model                       0.1591
Number of constraints imposed on "Smatrix"    4.0000
DF manually adjusted                          0.0000
Chi-square of independence model           1445.4738
DF of independence model                     10.0000
RMSEA                                         0.0129
RMSEA lower 95% CI                            0.0000
RMSEA upper 95% CI                            0.0311
SRMR                                          0.0195
TLI                                           0.9949
CFI                                           0.9985
AIC                                          -0.8204
BIC                                         -19.9776
OpenMx status1: 6 ("0" or "1": The optimization is considered fine.
Other values indicate problems.)
```

[3] Es importante advertir al lector de que cuando se lleva a cabo el paso 2 de MASEM con este ejemplo, se obtiene una advertencia:

```
Warning message:
In .solve(x = object$mx.fit@output$calculatedHessian, parameters = my.name) :
  Error in solving the Hessian matrix. Generalized inverse is used. The standard errors may not be
trustworthy.
```

Este mensaje advierte del cambio de método de estimación utilizado para ajustar el modelo, lo cual puede afectar a la precisión de los errores típicos. Puesto que en este caso los intervalos de confianza no están basados en los errores típicos, no supone un gran problema.

```
paso2_EA_ind <- tssem2(paso1_EA, Amatrix=A,
Smatrix=S,
   diag.constraints=TRUE, intervals="LB",
   mx.algebras=list(Ind=mxAlge-
bra(PB2R*R2NB1,name="Ind")))
```

En este caso, PB2R y R2NB1 son los nombres asignados a las asociaciones bivariadas de interés en la matriz A. Al imprimir los resultados con el comando `summary(paso2_EA_ind)`, aparece la estimación del efecto indirecto (que no es más que el producto de los dos efectos directos, 0.447 * 0.550 = 0.246) y su intervalo de confianza (IC), a través del cual podemos evaluar su significación estadística (Figura 5):

**Figura 5**

*Resultados al ejecutar el comando `summary(paso2_EA_ind)`*

```
mxAlgebras objects (and their 95% likelihood-based CIs):
             lbound  Estimate    ubound
Ind[1,1] 0.2137673 0.2460638 0.2794101
```

En este caso, el efecto indirecto de creencias positivas sobre creencias negativas (0.246) sí que sería estadísticamente diferente de 0 (IC al 95% 0.214, 0.279).

Una fase muy importante en metaanálisis es el análisis de variables moderadoras, es decir, analizar cómo las características de los estudios (e.g., edad media de la muestra) afectan a las estimaciones meta-analíticas. Desafortunadamente, el enfoque TSSEM sólo permite explorar el rol moderador de características cualitativas de los estudios (e.g., tipo de diseño de investigación empleado). Para más información sobre cómo llevar a cabo análisis de moderadores cualitativos (i.e., análisis de subgrupos) dentro de la metodología MASEM, derivamos al lector a Jak, y Cheung (2018b). Para superar esta limitación y permitir la incorporación de variables moderadoras cuantitativas, se ha propuesto el enfoque MASEM en una etapa (OSMASEM; Jak y Cheung, 2020).

## *OSMASEM: MASEM basado en correlaciones en un solo paso*

Mediante este procedimiento, el modelo SEM meta-analítico se ajusta directamente sobre los datos de los estudios primarios (es decir, omitiendo el paso 1 del enfoque anterior). El procedimiento OSMASEM siempre asume un modelo de efectos aleatorios, y una gran ventaja es que permite explorar si las características de los estudios afectan estadísticamente a todos o a algunos parámetros del modelo, ya sean estas variables cualitativas o cuantitativas (Jak y Cheung, 2020). El trabajo de Jak et al. (2021) constituye un excelente tutorial sobre cómo aplicar OSMASEM utilizando una aplicación web diseñada específicamente para ello.

Al aplicar OSMASEM, el modelo se especifica utilizando el lenguaje del paquete lavaan (Rosseel, 2012). Después, se construyen automáticamente las matrices A, S y F. Finalmente, para ajustar el modelo se aplica la función osmasem:

```
   res_osmasem<-osmasem(model.name="Modelo
metacognición",
    Mmatrix=M, Tmatrix=T, data=input, inter-
vals.type = "LB")
```

El objeto M contiene las matrices A, S y F, la matriz T especifica la estructura de las varianzas inter-estudios de las correlaciones, y el objeto input contiene las correlaciones de los estudios primarios, esta vez en forma de vectores en lugar de matrices. Para obtener los resultados, se puede ejecutar el comando `summary(res_osmasem)` y comprobar así que las estimaciones son muy similares a las que constan en la Figura 1, obtenidas con el procedimiento MASEM basado en correlaciones en dos etapas utilizando un modelo de efectos aleatorios.

En la base de datos que estamos analizando también se incluye la edad media de la muestra utilizada en cada estudio. Podríamos hipotetizar que, por ejemplo, la relación entre rumiación y depresión (que según la Figura 1 es $b = 0.420$) es más fuerte en muestras donde la media de edad es menor. La gran ventaja de OSMASEM es que nos permite contrastar este tipo de hipótesis, aunque los estudios que no reporten información sobre la variable mode-

radora tengan que ser eliminados de los análisis (por ejemplo, el estudio 14).

Derivamos a los lectores al código de R para obtener la explicación sobre cómo añadir un análisis de moderadores en la metodología OSMASEM, y recordamos que, para aquellos que no estén familiarizados con este programa, existe una aplicación web desde la que se pueden realizar estos análisis sin necesidad de utilizar código R (ver Jak et al., 2021). Mostramos a continuación los resultados que se obtienen (Figura 6), fijándonos en la última línea (`matrix A1`) que es la que contiene el efecto de la edad media sobre la relación entre rumiación y depresión ($b$ = .001, $p$ = .999). En este caso, la variable edad media no estaría moderando el efecto de rumiación sobre depresión.

## *MASEM basado en parámetros*

El MASEM basado en parámetros también consta de dos pasos, aunque diferentes a los descritos en el MASEM basado en correlaciones. En este caso, primero se ajustaría el modelo SEM deseado (e.g., Figura 1) en cada estudio, se extraerían los coeficientes de regresión y sus errores típicos, y en segundo lugar se realizaría un metaanálisis multivariado para obtener finalmente los coeficientes de regresión combinados (Cheung y Cheung, 2016). El principal problema del procedimiento es que no pueden incluirse estudios con variables o correlaciones faltantes. Esto ocurre porque los coeficientes de regresión, como tamaños del efecto parciales, dependen de las variables presentes en el modelo. Por ejemplo, la relación entre *rumiación* y *depresión* (Figura 1) varía según la inclusión o no de la variable *creencias positivas*. Si esta variable no está en el modelo, las estimaciones de ese estudio serían incompatibles con las de otro que sí incluya la variable en el modelo.

Esta limitación es importante, ya que, sobre todo en Ciencias Sociales, es muy complicado que todos los estudios incluidos en un metaanálisis analicen exactamente las mismas variables. Mismamente, si quisiéramos aplicar MASEM basado en parámetros sobre la base de datos con la que estamos trabajando, sólo podríamos utilizar cinco estudios. Aunque la aplicación de este procedimiento es muy limitada, también cuenta con algunas ventajas, como por ejemplo que el análisis de moderadores (ya sean variables cualitativas o cuantitativas) se puede hacer directamente sobre los parámetros de los modelos extraídos en cada estudio.

## MASEM psicométrico

En términos generales, se puede decir que la Psicometría se ocupa de la medición de lo psicológico. Sin embargo, a diferencia de otros procedimientos, lo característico de la Psicometría es su énfasis en las propiedades métricas exigibles a las mediciones psicológicas (Muñiz, 2018). Por ejemplo, características como la fiabilidad y la validez son requisitos que toda evaluación psicológica ha de cumplir. Dado que existen muchos estudios primarios que analizan estas propiedades, se pueden realizar estudios

**Figura 6**

*Resultados al ejecutar el comando summary(osmasem_edad)*

free parameters:

| | name | matrix | row | col | Estimate | Std.Error | A z value | Pr(>\|z\|) |
|---|---|---|---|---|---|---|---|---|
| 1 | beta1 | A0 | R | PB | 0.455257651 | 1.010275e+02 | 4.506277e-03 | 9.964045e-01 |
| 2 | beta4 | A0 | DEP | NB1 | 0.152493606 | 7.505815e+01 | 2.031673e-03 | 9.983790e-01 |
| 3 | beta5 | A0 | DEP | NB2 | 0.162110770 | 7.614170e+01 | 2.129067e-03 | 9.983013e-01 |
| 4 | beta2 | A0 | NB1 | R | 0.557737131 | 1.034079e+02 | 5.393566e-03 | 9.956966e-01 |
| 5 | beta3 | A0 | NB2 | R | 0.421398414 | 1.434558e+02 | 2.937478e-03 | 9.976562e-01 |
| 6 | beta6 | A0 | DEP | R | 0.412029122 | 9.092644e+01 | 4.531455e-03 | 9.963844e-01 |
| 7 | NB1WITHNB2 | S0 | NB2 | NB1 | 0.385431131 | 1.518898e+02 | 2.537570e-03 | 9.979753e-01 |
| 8 | beta6_1 | A1 | DEP | R | 0.001249373 | 5.927994e+02 | 2.107581e-06 | 9.999983e-01 |

meta-analíticos mediante los procedimientos estadísticos propios de esta estrategia de investigación. De todas las propiedades psicométricas, la fiabilidad ha sido la más estudiada mediante metaanálisis, denominado *Generalización de la Fiabilidad* (MA-GF). Estos estudios conllevan los objetivos propios de cualquier metaanálisis: obtener un resultado promedio (a partir de los coeficientes de fiabilidad de los estudios primarios), analizar la heterogeneidad y estudiar las fuentes de variación (Botella y Sánchez-Meca, 2015).

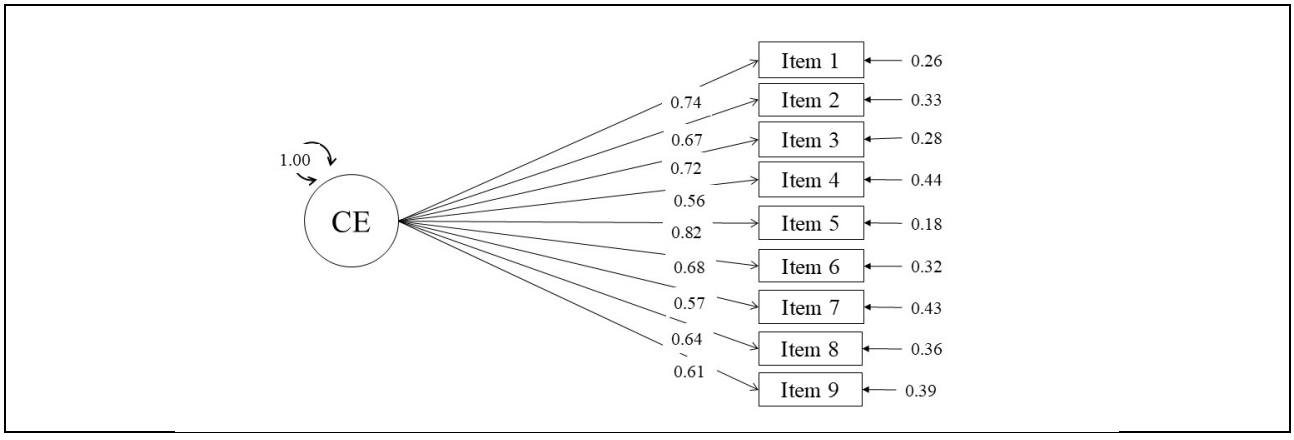
Este enfoque tradicional de metaanálisis tiene varios inconvenientes cuando se aplica al estudio de la fiabilidad (Scherer y Teo, 2020), tales como el reporte de diferentes índices de fiabilidad en los estudios primarios, la imposibilidad de tratar la dependencia estadística entre los coeficientes de fiabilidad de las puntuaciones totales y de las subescalas y la incapacidad de sintetizar el coeficiente de fiabilidad que mejor represente el modelo de medida del test. El coeficiente alfa, que es el más usado, se basa en supuestos que no se suelen cumplir (McNeish, 2018). Para superar estas limitaciones, se han ido desarrollando varios procedimientos que combinan técnicas meta-analíticas y SEM (Becker, 1992; Viswesvaran y Ones, 1995), hasta desembocar en los procedimientos actuales de MASEM aplicados al campo psicométrico (Jak, 2015; Scherer y Teo, 2020).

A continuación, se presenta el enfoque OSMASEM aplicándolo a un metaanálisis psicométrico realizado por Aguayo-Estremera y colaboradores (2024) sobre el Maslach Burnout Inventory (MBI; Maslach y Jackson, 1981). Este test evalúa el síndrome de burnout por medio de tres escalas que miden cansancio emocional, despersonalización y baja realización personal. Por razones de simplicidad solo se presenta el análisis con la primera escala, que consta de 9 ítems con un formato de respuesta tipo Likert de 7 puntos. En la Figura 7 puede verse el modelo de medida del ejemplo propuesto. La base de datos completa puede encontrarse en https://osf.io/cq74m/?view_only=87ecffb55d87455e8a50b97e4e9b5e3d.

Se empieza extrayendo, de cada estudio primario, la matriz de correlaciones inter-ítem (ver el archivo matrices.dat en el material suplementario), así como el tamaño muestral. Los estudios primarios no suelen reportar esta matriz (a menos que sean de corte psicométrico), por lo que será necesario contactar con los autores de correspondencia para conseguirlas. En nuestro ejemplo, se consiguió información de nueve estudios primarios. Una vez extraída esta información, se realizan los análisis MASEM. Como se señaló previamente, OSMASEM realiza la estimación de la matriz combinada de correlaciones

**Figura 7**

*Modelo de medida de la escala cansancio emocional (CE) del Maslach Burnout Inventory (Maslach y Jackson, 1981)*



y ajusta el modelo SEM sobre la matriz combinada en un solo paso. A diferencia del caso general, en aplicaciones psicométricas la matriz contiene las correlaciones de Pearson entre los ítems del test (no sobre las variables del modelo teórico) y el ajuste se realiza sobre el modelo de medida (en lugar de sobre el modelo teórico).

El primer paso del análisis consiste en el procesamiento de la información recabada. Las matrices de correlaciones (o, dado que son simétricas, la mitad inferior) se almacenan en un archivo .dat, que puede ser guardado como lista en R con la función `readFullMat` (o `readLowTriMat`) del paquete `metaSEM`. Los tamaños muestrales se introducen en otra lista, que luego se une a la de las matrices de correlaciones para crear el objeto que sirve para hacer los análisis posteriores. En nuestro ejemplo, las nueve matrices de correlaciones inter-ítem están en el archivo `Matrices.dat` y el objeto datos contiene en R tanto las matrices (`datos$data`) como los tamaños muestrales (`datos$n`). Un aspecto importante es comprobar que las matrices son definidas positivas mediante la función `is.pd` del mismo paquete. En caso de que haya alguna, se debe eliminar del objeto. En el ejemplo propuesto, ninguna matriz es no definida positiva, manteniéndose las nueve matrices que resultan en un total de 3493 participantes (`pattern.n(datos$data, datos$n)`). A continuación, se crea un dataframe a partir de las listas de correlaciones y de tamaños muestrales por medio de la función `Cor2data.frame`. Remitimos al lector al código que se encuentra en el repositorio para encontrar los detalles de estos análisis.

Después del procesamiento de los datos, se ajusta el modelo de medida a partir de los siguientes pasos: (a) especificar el modelo de medida en lenguaje del paquete `lavaan`; (b) crear las matrices A, S y F; (c) especificar el coeficiente de fiabilidad con la función `mxAlgebra` del paquete `OpenMX`; y (d) utilizar la función `osmasem` del paquete `metaSEM`, incluyendo el argumento `mxModel.Args` para el cálculo de la fiabilidad. En este caso, en la matriz A se especifican todas las cargas factoriales, en la matriz S se definen las varianzas (de las variables latentes) que quedan fijadas a 1 para que el modelo se pueda estimar, y, en la matriz F se indican las variables latentes (constructo) y las observadas (ítems). Para crear estas matrices se usa la función `lavaan2RAM` del paquete `metaSEM`, especificando como argumentos el nombre del objeto que contiene el modelo de medida (`model`) y las variables observadas (`obs.variables`). En el ejemplo propuesto, se calcula el coeficiente omega total (McDonald, 1999) para la estimación de fiabilidad. Teniendo todo esto en cuenta, se ejecuta el comando siguiente para el ajuste del modelo:

```
Aj_Unifactorial <- osmasem(model.name="AFC unifactorial congenérico con efectos aleatorios", Mmatrix = M0, Tmatrix = T0, mxModel.Args = list(Coef_Omega, mxCI(c("Coef_Omega"))), intervals.type = "LB", data = BBDD).
```

Y con `summary(Aj_Unifactorial, fitIndices = T)` se imprimen los resultados (Figura 8). Con `fitIndices = T` conseguimos los índices de ajuste habituales, como RMSEA, TLI y CFI. No obstante, el índice SRMR se imprime mediante la función `osmasemSRMR`. En el ejemplo[4], se observan valores adecuados para estos índices, por lo que se puede decir que el ajuste del modelo unidimensional congenérico a los datos es adecuado.

La columna `Estimate` nos muestra las cargas factoriales de los nueve ítems de la escala cansancio emocional (CE), que comprobamos que son altas y, por lo tanto, indican que los ítems son buenos indicadores del constructo. Para calcular el coeficiente de fiabilidad de la escala, ejecutamos `Aj_Unifactorial$Coef_Omega$result` y vemos que equivale a .879. Se podrían construir intervalos de confianza entorno a la estimación puntual con el procedimiento de Raykov y Marcoulides (2013).

Un aspecto a tener en cuenta del metaanálisis psicométrico con MASEM es que, a diferencia del tradicional, no solamente estudiamos la fiabilidad del test sino también la validez de su estructura interna (dimensionalidad). En

[4] Al ajustar el modelo con OSMASEM, aparece el aviso:
`The Hessian at the solution does not appear to be convex. Information matrix is not positive definite (not at a candidate optimum). Be suspicious of these results. At minimum, do not trust the standard errors.`
Por tanto, los intervalos de confianza estarían comprometidos si se basan en errores típicos.

**Figura 8**

*Resultados que se obtienen al ejecutar el comando `summary(Aj_Unifactorial, fitIndices = T)`*

```
chi-square:  χ² ( df=27 ) = 96.83264,  p = 8.474483e-10
Information Criteria:
      |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
AIC:       -1291.357              -643.3572                -642.1561
BIC:       -3009.583              -366.2239                -509.2106
CFI: 0.9396538
TLI: 0.9195383   (also known as NNFI)
RMSEA:  0.02721122  [95% CI (0.02032191, 0.03425997)]
Prob(RMSEA <= 0.05): 1
free parameters:
      name  matrix     row col    Estimate  Std.Error A      z value     Pr(>|z|)
1       L1      A0 mbi1_CE  CE   0.7380518 156.514281    0.004715555 9.962375e-01
2       L2      A0 mbi2_CE  CE   0.6660310 122.519065    0.005436141 9.956626e-01
3       L3      A0 mbi3_CE  CE   0.7167616 150.504313    0.004762399 9.962002e-01
4       L4      A0 mbi4_CE  CE   0.5552294 153.301279    0.003621818 9.971102e-01
5       L5      A0 mbi5_CE  CE   0.8194777 166.826203    0.004912164 9.960807e-01
6       L6      A0 mbi6_CE  CE   0.6817794 139.653652    0.004881930 9.961048e-01
7       L7      A0 mbi7_CE  CE   0.5668647 105.616808    0.005367183 9.957176e-01
8       L8      A0 mbi8_CE  CE   0.6370377 141.047743    0.004516468 9.963964e-01
9       L9      A0 mbi9_CE  CE   0.6114135 130.794155    0.004674624 9.962702e-01
```

efecto, la síntesis meta-analítica se realiza sobre las matrices de correlaciones inter-ítem de los estudios primarios, que lleva a una matriz combinada con la que, primero, se estima y ajusta el modelo de medida (validez de estructura interna) y, segundo, a partir de las cargas factoriales, se calcula el coeficiente de fiabilidad. Esta característica deviene en dos consecuencias importantes. Por un lado, se pueden poner a prueba diferentes modelos de medida, aunque no se hayan examinado en los estudios primarios. Asimismo, se pueden comparar modelos congenéricos y tau-equivalentes para elegir el índice de fiabilidad más apropiado según los requisitos del modelo de medida. Por otro lado, el estudio de la heterogeneidad, y posterior análisis de moderadores, no se realiza sobre los coeficientes de fiabilidad, sino sobre las cargas factoriales. Remitimos al lector al código del repositorio (disponible en https://osf.io/cq74m/?view_only=87ecffb55d87455e8a50b97e4e9b5e3d) para el estudio de la comparación de modelos y del análisis de moderadores.

## Recomendaciones finales

El MASEM es una metodología avanzada de síntesis de la evidencia cuyos dos principales ámbitos de aplicación se derivan directamente de las dos metodologías que subyacen a los modelos SEM: el análisis de vías, que a nivel de MASEM permite la combinación de estudios en los que se han examinado las relaciones entre las variables que conforman un modelo teórico; y el análisis factorial, que dentro de los modelos MASEM proporciona una herramienta flexible y elegante en el campo del metaanálisis psicométrico. Un cierto grado de familiaridad con estas metodologías será de utilidad para decidir si la mejor técnica de análisis para abordar la pregunta de investigación es MASEM, en lugar de otras alternativas para la integración de múltiples resultados de cada estudio (López-López et al., 2018).

Una práctica habitual en SEM, y por extensión en MASEM, es la evaluación de índices de ajuste para comparar distintos modelos, tal y como se ha ilustrado en los

ejemplos de este tutorial. Esto puede conllevar a veces la interpretación de los resultados del modelo de efectos fijos, especialmente en aplicaciones con pocos datos, a pesar de que el modelo de efectos aleatorios suela considerarse como una opción más realista en metaanálisis (una reflexión más amplia sobre ambos modelos puede encontrarse en Borenstein et al., 2010). Además, los supuestos clásicos de estos modelos también son de aplicación aquí, incluyendo la normalidad (multivariada) de las correlaciones, para lo cual a veces se aconseja transformarlas antes de llevar a cabo los análisis (Jak, 2015); y la independencia de las observaciones, que implica que cada matriz de correlaciones haya sido obtenida a partir de muestras de participantes distintas. Si este último supuesto no se cumple, es posible adoptar estrategias similares a las empleadas en otros ámbitos del metaanálisis (Bilici et al., 2025).

En este artículo se han presentado los enfoques del MASEM basado en correlaciones y en parámetros, cada uno de los cuales tiene sus fortalezas y limitaciones (Cheung y Cheung, 2016). Los enfoques basados en correlaciones presentan un gran potencial para su aplicación en la mayoría de las situaciones debido a su capacidad para trabajar con datos perdidos, mientras que el enfoque basado en parámetros puede resultar preferible para situaciones concretas, por ejemplo, aquellas en las que sea posible ajustar el modelo teórico a nivel de cada estudio.

Actualmente, la metodología MASEM continúa desarrollándose y mejorando en diversas direcciones, como la incorporación de variables dicotómicas (e.g., correlaciones biserial-puntuales) en los análisis (de Jonge et al., 2020) y/o el uso de datos individuales (*individual participant data*) en lugar de datos agrupados (Groot et al., 2024). Aunque su aplicación todavía se limita al ámbito de la Psicología y Ciencias de la Educación, su gran versatilidad a la hora de explorar preguntas complejas de investigación a nivel meta-analítico augura una gran expansión a otros ámbitos, como la Medicina o la Ecología, consolidándose como una herramienta clave para abordar desafíos metodológicos contemporáneos.

## Material suplementario

Las bases de datos y los códigos de R utilizados para hacer los análisis de este manuscrito, así como el tutorial sobre cómo realizar TSSEM y OSMASEM en la aplicación webMASEM, pueden encontrarse en este enlace: https://osf.io/cq74m/?view_only=87ecffb55d87455e8a50b97e4e9b5e3d.

## Referencias

Aguayo-Estremera, R., Cañadas-De la Fuente, G. R., Ariza-Castilla, T., Ortega-Campos, E., Gómez-Urquiza, J. L., Romero-Béjar, J. L. y De la Fuente-Solana, E. I. (2024). A Comparison of Univariate and meta-Analytic Structural Equation Modelling Approaches to Reliability Generalization Applied to the Maslach Burnout Inventory. *Frontiers in Psychology, 15*, Artículo 1383619. https://doi.org/10.3389/fpsyg.2024.1383619

Becker, B. J. (1992). Using Results from Replicated Studies to Estimate Linear Models. *Journal of Educational Statistics, 17,* 341–362. https://doi.org/10.3102/10769986017004341

Becker, B. J. (1995). Corrections to "Using results from Replicated Studies to Estimate Linear Models". *Journal of Educational and Behavioral Statistics, 20,* 100–102. https://doi.org/10.3102/10769986020001100

Becker, B. J. y Aloe, A. M. (2019). Model-based Meta-analysis. En H. Cooper, L. V. Hedges y J. C. Valentine, *The Handbook of Research Synthesis and Meta-analysis* (3ª ed, pp. 339-363). Russell Sage.

Bilici, Z. Ş., Van den Noortgate, W. y Jak, S. (2025). Six ways to handle dependent effect sizes in meta-analytic structural equation modeling: Is there a gold standard? *Research Synthesis Methods, 0*, 1–27. https://doi.org/10.1017/rsm.2024.10

Borenstein, M., Hedges, L. V., Higgins, J. P. T. y Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97-111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Higgins, J. P. T., Hedges, L. V. y Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8,* 5–18. https://doi.org/10.1002/jrsm.1230

Botella, J. y Meca, J. S. (2015). *Metaanálisis en ciencias sociales y de la salud* [Meta-analysis in Social and Health Sciences]. Síntesis.

Cano-López, J. B., Garcia-Sancho, E., Fernández-Castilla, B. y Salguero, J. M. (2022). Empirical Evidence of the Metacognitive Model of Rumination and Depression in Clinical and Nonclinical Samples: A Systematic Review and Meta-Analysis. *Cognitive Therapy and Research, 46*(6), 1–26. https://doi.org/10.1007/s10608-021-10260-2

Cooper, H., Hedges, L. V. y Valentine, J. C. (Eds.). (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Sage.

Cheung, M. W.-L. (2013). Multivariate Meta-Analysis as Structural Equation Models. *Structural Equation Modeling, 20,* 429–454. https://doi.org/10.1080/10705511.2013.797827

Cheung, M. W.-L. (2015). *Meta-analysis: A Structural Equation Modeling Approach*. Wiley.

Cheung, M. W.-L. (2015). metaSEM: an R Package for Meta-Analysis Using Structural Equation Modeling. *Frontiers in Psychology, 5*, Artículo 1521. https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L. (2009). Constructing Approximate Confidence Intervals for Parameters with Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 16,* 267–294. https://doi.org/10.1080/10705510902751291

Cheung, M. W.-L. (2019). Some Reflections on Combining Meta-Analysis and Structural Equation Modeling. *Research Synthesis Methods, 10*, 15–22. https://doi.org/10.1002/jrsm.1321

Cheung, M. W.-L. y Chan, W. (2005). Meta-analytic Structural Equation Modeling: A Two-Stage Approach. *Psychological Methods, 10,* 40–64. https://doi.org/10.1037/1082-989X.10.1.40

Cheung, M. W.-L. y Cheung, S. F. (2016). Random-Effects Models for Meta-Analytic Structural Equation Modeling: Review, Issues, and Illustrations. *Research Synthesis Methods, 7,* 140–155. https://doi.org/10.1002/jrsm.1166

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure Of Tests. *Psychometrika, 16,* 297–334. https://doi.org/10.1007/BF02310555

De Jonge, H., Jak, S. y Kan, K. J. (2020). Dealing with Artificially Dichotomized Variables in Meta-Analytic Structural Equation Modeling. *Zeitschrift für Psychologie, 228,* 25–35. https://doi.org/10.1027/2151-2604/a000395

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher, 5,* 3–8. https://doi.org/10.3102/0013189X005010003

Grissom, R. J. y Kim, J. J. (2012). *Effect Sizes for Research: Univariate and Multivariate Applications* (2ª Ed.). Routledge.

Groot, L. J., Kan, K. J. y Jak, S. (2024). Checking the Inventory: Illustrating Different Methods for Individual Participant Data Meta-Analytic Structural Equation Modeling. *Research Synthesis Methods, 15,* 872–895. https://doi.org/10.1002/jrsm.1735

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hedges, L. V. y Vevea, J. L. (1998). Fixed-and Random-Effects Models in Meta-Analysis. *Psychological Methods, 3,* 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T. y Thompson, S. G. (2002). Quantifying Heterogeneity in a Meta-Analysis. *Statistics in Medicine, 21,* 1539–1558. https://doi.org/10.1002/sim.1186

Hu, L. T. y Bentler, P. M. (1999). Cutoff Criteria for fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jak, S. (2015). *Meta-analytic Structural Equation Modelling*. Springer https://doi.org/10.1007/978-3-319-27174-3

Jak, S. y Cheung, M. W.-L. (2018a). Accounting for Missing Correlation Coefficients in Fixed-Effects MASEM. *Multivariate Behavioral Research, 53*, 1–14. https://doi.org/10.1080/00273171.2017.1375886

Jak, S. y Cheung, M. W.-L. (2018b). Testing Moderator Hypotheses in Meta-Analytic Structural Equation Modeling using Subgroup Analysis. *Behavior Research Methods, 50,* 1359–1373. https://doi.org/10.3758/s13428-018-1046-3

Jak, S. y Cheung, M. W.-L. (2020). Meta-analytic Structural Equation Modeling with Moderating Effects on SEM Parameters. *Psychological Methods, 25,* 430–455. https://doi.org/10.1037/met0000245

Jak, S., Li, H., Kolbe, L., de Jonge, H. y Cheung, M. W.-L. (2021). Meta-analytic Structural Equation Modeling Made Easy: A Tutorial and Web Application for One-Stage MASEM. *Research Synthesis Methods, 12,* 590–606. https://doi.org/10.1002/jrsm.1498

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling* (5ª ed.). Guilford.

Lei, P. W. y Wu, Q. (2007). Introduction to Structural Equation Modeling: Issues and Practical Considerations. *Educational Measurement: Issues and Practice, 26*, 33–43. https://doi.org/10.1111/j.1745-3992.2007.00099.x

López-López, J. A., Page, M. J., Lipsey, M. W. y Higgins, J. P. T. (2018). Dealing with Effect Size Multiplicity in Systematic Reviews and Meta-Analyses. *Research Synthesis Methods, 9,* 336–351. https://doi.org/10.1002/jrsm.1310

Maslach, C. y Jackson, S. E. (1981). The Measurement of Experienced Burnout. *Journal of Organizational Behavior, 2,* 99–113. https://doi.org/10.1002/job.4030020205

McArdle, J. J. y McDonald, R. P. (1984). Some Algebraic Properties of the Reticular Action Model for Moment Structures. *British Journal of Mathematical and Statistical Psychology, 37*, 234–251. https://doi.org/10.1111/j.2044-8317.1984.tb00802.x

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. L. Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23, 412–433. https://doi.org/10.1037/met0000144

Muñiz, J. (2018). *Introducción a la Psicometría. Teoría Clásica y TRI.* [Introduction to Psychometrics. Classical Theory and IRT]. Pirámide.

Papageorgiou, C. y Wells, A. (2003). An Empirical Test of a Clinical Metacognitive Model of Rumination and Depression. *Cognitive Therapy and Research, 27,* 261–273. https://doi.org/10.1023/A:1023962332399

Raykov, T. y Marcoulides, G. A. (2013). Meta-analysis of Scale Reliability Using Latent Variable

Modeling. *Structural Equation Modeling, 20,* Artículo 338353. https://doi.org/10.1080/10705511.2013.769396

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Ruiz, M. A., Pardo, A. y San Martín, R. (2010). Modelos de ecuaciones estructurales [Structural Equation Models]. *Papeles del Psicólogo, 31*, 34–45. https://www.redalyc.org/pdf/778/77812441004.pdf

Scherer, R. y Teo, T. (2020). A Tutorial on the Meta-Analytic Structural Equation Modeling of Reliability Coefficients. *Psychological Methods, 25*, 747–775. https://doi.org/10.1037/met0000261

Viswesvaran, C. y Ones, D. S. (1995). Theory Testing: Combining Psychometric Meta-Analysis and Structural Equations Modeling. *Personnel Psychology, 48,* 865–885. https://doi.org/10.1111/j.1744-6570.1995.tb01784.x.

# RESPUESTAS OBSERVABLES Y ESTADOS OCULTOS EN LAS REDES NEURONALES ARTIFICIALES: ¿CÓMO USARLOS PARA RAZONAR SOBRE ASPECTOS COGNITIVOS DEL LENGUAJE?

# OBSERVABLE RESPONSES AND HIDDEN STATES IN RECURRENT NEURAL NETWORKS TO REASON ABOUT COGNITIVE ASPECTS OF LANGUAGE

GUILLERMO JORGE-BOTANA[1],
JOSE ÁNGEL MARTÍNEZ-HUERTAS[2] Y
ALEJANDRO MARTÍNEZ-MINGO[2]

Cómo referenciar este artículo/How to reference this article:

Jorge-Botana, G., Martínez-Huertas, J. A y Martínez-Mingo, A. (2025). Respuestas observables y estados ocultos en las redes neuronales artificiales: ¿Cómo usarlos para razonar sobre aspectos cognitivos del lenguaje? [Observable responses and hidden states in recurrent neural networks to reason about cognitive aspects of language]. *Acción Psicológica, 22*(1), 41–56. https://doi.org/10.5944/ap.22.1.43347

## Resumen

Para estudiar los procesos psicológicos involucrados en el lenguaje, la Ciencia Cognitiva indaga sobre las representaciones internas que se manejan a la hora de comprender o producir lenguaje. También postula las operaciones que modifican esas mismas representaciones dadas unas constricciones contextuales. Así, contexto y representación interactúan para crear significados. Con arreglo a esto, se ofrecen diferentes hipótesis de como el sistema cognitivo produce lenguaje. Al igual que existen metodologías experimentales para su estudio, distintas arquitecturas de redes neuronales artificiales permiten dotar a dichas hipótesis de un aparataje formal. En estos modelos, las representaciones y las operaciones

**Correspondence address [Dirección para correspondencia]**: Guillermo Jorge-Botana, Facultad de Psicología, Univesidad Complutense de Madrid, España.
**Email**: guijorge@ucm.es

**ORCID**: Guillermo Jorge-Botana (https://orcid.org/0000-0001-5879-6783), José Ángel Martínez-Huertas (https://orcid.org/0000-0002-6700-6832) y Alejandro Martínez-Mingo (https://orcid.org/0000-0002-8375-0952).

**Agradecimientos**: esta publicación es parte del Proyecto de I+D+i PID2022-136905OB-C22 financiado por el Ministerio de Ciencia e Innovación MCIN/ AEI/ 10.13039/501100011033/ FEDER, UE.

[1] Universidad Complutense de Madrid, España.
[2] Universidad Nacional de Educación a Distancia, España.

Recibido: 12 de noviembre de 2024.
Aceptado: 22 de enero de 2025.

participantes quedan exhaustivamente caracterizadas. Las redes neuronales recurrentes (RNNs) con mecanismos LSTM y los *Transformers* destacan como arquitecturas especialmente útiles para modelar la secuencialidad contextual presente en el lenguaje. Este número especial nos brinda la ocasión para explicar el uso de sus expresiones externas (sus salidas) como de sus representaciones internas (estados ocultos) para entender en términos cognitivos el efecto que tienen los cambios de expectativas en distintas marcas temporales de las frases. Para hacerlo, se ilustra la formalización mediante una RNN Secuencia-Secuencia con codificador y decodificador y se homologan sus mediciones a los experimentos de potenciales evento-relacionados (ERPs) en un tema nuclear en el lenguaje: la composicionalidad sistemática.

**Palabras clave:** Redes Neuronales Artificiales; Redes Recurrentes; LSTM; Estados Ocultos; Sorpresividad; Lenguaje; Potenciales Evento-Relacionados.

## Abstract

In order to study the psychological processes involved in language, Cognitive Science investigates the internal representations involved in understanding or producing language. It also postulates the operations that modify those representations given contextual constraints. Thus, context and representation interact to create meanings. Accordingly, there are different hypotheses about how the cognitive system produces language. Just as there are experimental methodologies for their study, different architectures of artificial neural networks make it possible to provide these hypotheses with a formal apparatus. In these models, the representations and operations involved are exhaustively characterized. Recurrent neural networks (RNNs) with LSTM mechanisms and Transformers stand out as particularly useful architectures for modeling the contextual sequentiality of language. This special issue gives us the opportunity to explain how to use their external expressions (outputs) as well as their internal representations (hidden states) to understand, in cognitive terms, the effect that changes of expectations have on different temporal markings of sentences. To do so, we illustrate such formalization using a Sequence-Sequence RNN with encoder and decoder and relate its measures with event-related potentials (ERPs) experiments on a nuclear issue in language: systematic compositionality.

**Keywords:** Artificial Neural Networks; Recurrent Networks; LSTM; Hidden States; Surprisal; Language; Event-Related Potentials.

## Respuestas observables y estados ocultos en Redes Neuronales Recurrentes para razonar sobre aspectos cognitivos del lenguaje

Desde que se superaron las constricciones metodológicas de los modelos conductistas, la Psicología ha avanzado hacia el estudio de las representaciones internas que se generan en los razonamientos humanos (Neisser, 1967). En este contexto, la Ciencia Cognitiva propone que estas representaciones internas son clave para comprender los procesos mentales (Anderson, 2005; Pitt, 2022; Sterelny, 1990), y que su carácter emergente es el resultado de la interacción entre aspectos primitivos del entorno y la interpretación contextualizada del individuo. Estas representaciones internas se llaman emergentes porque son los símbolos que las personas manejamos que se generan a partir de aspectos primitivos de la realidad. Podríamos apelar al término emergente en tanto que han sido contextualizadas (i.e., sesgadas) a la situación interna o externa del individuo. No obstante, es ampliamente reconocido que, en general, se dedica un mayor esfuerzo a la recopilación de datos empíricos, que al desarrollo de modelos formales que describan las representaciones mentales y sus operaciones. Esto hace que las teorías psicológicas estén eminentemente sustentadas en lenguaje natural, manera legítima de describir fenómenos, pero mucho más ambigua que los modelos formales (e.g., Busemeyer et al., 2015; Farrell y Lewandowsky, 2010; Sun, 2023). Así, la propuesta de los modelos formales permite superar esta ambigüedad describiendo las representaciones mentales en términos formales y proponer qué operaciones las construyen y manejan.

Nos encontramos en el mismo escenario con los aspectos cognitivos del lenguaje. Rompiendo el marco del análisis del lenguaje como cadenas de conductas verbales del conductismo (Skinner, 1957), la Ciencia Cognitiva ha tratado de inferir qué tipo de representaciones se utilizan y qué tipo de operaciones son desplegadas sobre ellas en el momento de comprender o producir lenguaje (e.g., Anderson, 2005). Con carácter general, se trata de proponer el formato de las representaciones y qué tipo de información portan, además de estudiar su pervivencia y preeminencia en distintos formatos (e.g., modal y amodal). Así tenemos el debate entre el uso obligado o no de representaciones modales (De Vega et al., 2012), del uso de indicios emocionales y sus consecuencias (Lindquist, 2021), de la separación o no del sistema sintáctico y semántico (Kaan, 1999), etc. En este tipo de investigaciones, se aíslan propiedades que están presentes en ciertos estímulos, mientras que en otros quedan ausentes. Por ejemplo, se pueden utilizar palabras cuyo referente tiene contenido emocional frente a palabras neutras, palabras con referencia a aspectos manipulativos o sensoriomotores frente a las abstractas, frases cuya sintaxis es legítima pero con semántica confusa frente a la coincidencia de semántica plausible y sintaxis ilegítima, distintos tipos de dependencia sintáctica en términos de cercanía o lejanía, etc. (ver el manual de Belinchón et al., 2009 para entender la sutilidad de tales manipulaciones). Existen distintos paradigmas experimentales en Ciencia Cognitiva para estudiar los aspectos cognitivos del lenguaje. A modo de ejemplo, el paradigma conductual suele estudiar tiempos de reacción o registros de movimientos oculares. Otro paradigma clásico que ha madurado las últimas décadas es el de los Potenciales Relacionados con Eventos (Event Related Potencials, ERPs). Con el mismo control experimental, este paradigma trata de identificar momentos en la línea temporal que desvelen sensibilidades, entendidas como la capacidad de algunas localizaciones corticales de responder de manera diferente a esas manipulaciones. De ahí que se busque resolución temporal más que espacial, aunque las técnicas actuales puedan aunar ambos, como es el caso de la Magnetoencefalografía. Un ejemplo de ERP es el N400, actividad diferencial localizada normalmente en la zona centro-parietal que se asocia al cambio de expectativas al leer frases (Federmeier y Kutas, 1999; Kutas y Hillyard, 1980). Más tarde hablaremos de él.

En resumen, lo importante de estos paradigmas radica en su capacidad para captar los procesos cognitivos subyacentes a partir de las sensibilidades detectadas y de las representaciones mentales implicadas, así como de la información que estas contienen. En este artículo, presentamos y discutimos el aparataje que nos permite estudiar esas representaciones y sus propiedades a partir del modelado formal con ciertas arquitecturas de redes neuronales artificiales (RNAs). El objetivo de este texto no es la exhaustividad, sino fomentar el interés y la reflexión en torno a este enfoque de modelización formal.

## Las Redes Neuronales Artificiales (RNAs) en Ciencia Cognitiva

Una posibilidad para estudiar las representaciones internas es su simulación con RNAs. Una RNA puede representarse mediante una serie de nodos (neuronas artificiales) y conexiones (con pesos asociados) que reciben señales de otros nodos, las procesan y envían una respuesta a través de una función de activación concreta. Los grupos de nodos suelen agruparse en capas y la señal viaja desde la primera capa (capa de entrada) hasta la última capa (capa de salida). Esta última capa es la expresión externa de la resolución de una tarea. Será en la capa de salida donde se emitirá la predicción que hace la red a partir de una entrada concreta en función de sus pesos, funciones de activación y estructura. Este paradigma tiene una gran tradición desde los primeros estudios del conexionismo (e.g., McClelland y Rumelhart, 1989; McClelland et al., 1987; Rumelhart et al., 1986). Sin ánimo de ser exhaustivos, podemos describir su aprovechamiento en el ámbito cognitivo en varios ejes de manipulación:

***Las características de la propia red.*** Esto incluye el tipo de arquitectura de la red (el tipo de red y su funcionamiento estructural), su topología (el número de capas y nodos, además de posibles ensamblajes entre redes) y su parametrización (funciones, coeficientes de aprendizaje, optimizadores, ratios de dilución, etc.). La manipulación de estas características puede sustentar hipótesis sobre los mecanismos implicados en el procesamiento del lenguaje.

***Las muestras con las que aprende la red***. Es decir, la información con la que la red es entrenada, lo que incluiría el corpus textual (textos, frases, pares de frases, etc.) y sus características. También el tamaño es una cuestión clave, ya que puede generar un tipo de aprendizaje más estadístico o más emergente.

***Las entradas puestas bajo escrutinio una vez entrenada la red***. Consiste en aislar propiedades presentes y ausentes en ciertas entradas y comprobar el comportamiento de la red y el tipo de errores que comete. Sería una suerte de Psicología comparada persona-máquina.

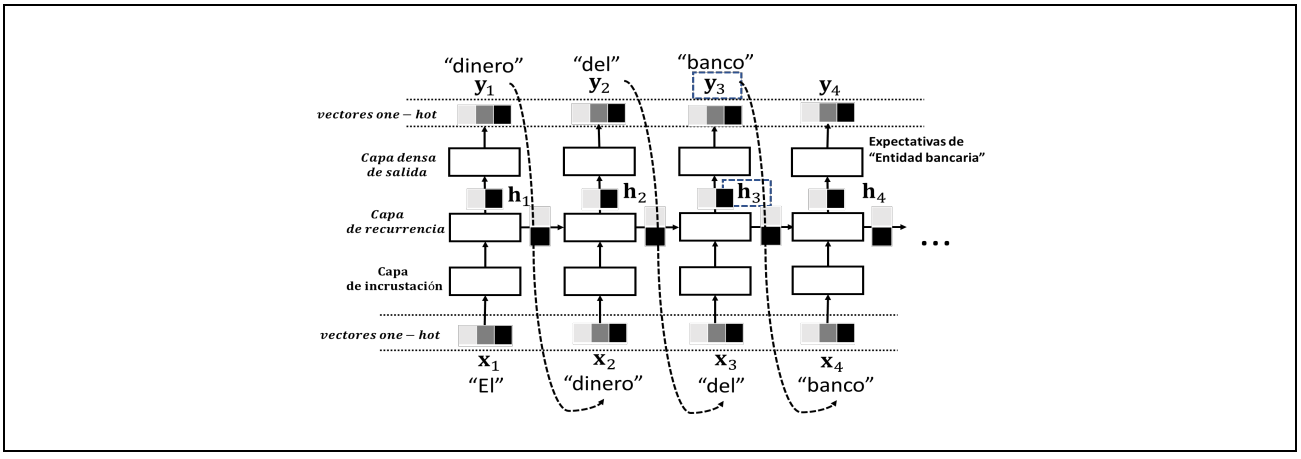
Son varias las topologías que se han propuesto para el estudio del lenguaje y aquí queremos destacar las redes neuronales recurrentes (RNN; e.g., Elman, 1990; Jordan, 1997) con mecanismos LSTM (RNN-LSTM; Hochreiter y Schmidhuber, 1997), así como los Transformers (Vaswani et al., 2017). En este artículo, ilustraremos distintas estrategias, observables y ocultas, de cómo se puede estudiar el lenguaje desde un punto de vista psicológico con una RNN.

La Figura 1 presenta un esquema del funcionamiento de una RNN. Cada palabra de una unidad de texto (normalmente, frases) será la entrada en cada marca de tiempo $t$ (siendo $t$ el orden de la palabra en la frase) como $x_t$. La tarea más comúnmente utilizada para entrenar este tipo de red es la predicción de la siguiente palabra en la secuencia (tarea autoregresiva). Introducida una palabra, la red tiene que predecir la siguiente palabra en la frase. Operativamente, la entrada $x_t$ es un vector *one-hot* (una representación binaria que codifica las palabras del vocabulario del modelo), el cual es transformado en la capa de incrustación en un vector que captura propiedades lingüísticas relevantes (como aspectos semánticos y sintácticos). En otras palabras, esta capa convierte las entradas binarias en representaciones vectoriales densas dentro de un espacio vectorial. Posteriormente, esta representación densa se procesa en la capa de recurrencia, generando un estado oculto ($h_t$) en su salida. Así, la capa de recurrencia utiliza conexiones recurrentes que permiten que la señal generada en un instante sea enviada a la misma capa en el tiempo siguiente, funcionando, así como un mecanismo de contexto temporal. Como se ve en la Figura 1, esta salida está en función de la entrada actual ($x_t$) y el estado oculto anterior ($h_{t-1}$), que codifica la información de la frase ya procesada. De esta forma, $h_t$ es tanto la salida de la capa de recurrencia, como la entrada de la misma capa en el momento inmediatamente posterior. Cada estado oculto $h_t$ se utiliza finalmente como entrada para la capa de salida, que produce un vector $y_t$ a partir de una función softmax (Jurafsky y Martin, 2023). Este vector representa la predicción del modelo en el punto temporal correspondiente para la siguiente palabra en la secuencia. La capa de salida está compuesta por tantos nodos como palabras tenga el vocabulario del modelo, y su señal puede ser expresada en forma de vector de probabilidad, indicando cada componente la probabilidad de una palabra concreta. La palabra con más probabilidad será seleccionada como la continuación más probable de la secuencia. Volviendo a la Figura 1, si estamos en el momento t = 3, la entrada es «del». La capa de recurrencia recibe información de la entrada «del» y del estado oculto anterior $h_2$. Ese estado oculto anterior lleva a su vez la información de la situación de la frase en ese punto ya que existe un contexto marcado por lo ya dicho (en este caso, «El dinero»). Por tanto, $h_3$ está en función de $h_2$ (información contextualizada por «El dinero») y la entrada actual «del» gracias a la capa de recurrencia. Es previsible que la secuencia sea continuada por un sustantivo masculino en un contexto de dinero. En este ejemplo, la continuación sería «banco» y la salida $y_3$ tendrá su mayor valor en el componente que corresponde a esta palabra. Es decir, el nodo de salida correspondiente a «banco» daría la señal (probabilidad) más grande.

Quizá lo más interesante del estado oculto $h_t$ es su naturaleza recurrente, que le permite actuar como una memoria dinámica que se actualiza en cada marca temporal ($t$). Este estado no solo acumula información sobre la frase procesada hasta el momento, sino que también integra el contexto de las palabras anteriores para formar una representación contextualizada de la secuencia. Desde una perspectiva cognitiva, $h_t$ puede interpretarse como una representación de las expectativas del modelo, es decir, una proyección de lo posible: qué palabras son más probables de aparecer a continuación según el patrón lingüístico observado.

**Figura 1**

*Representación desenrollada en el tiempo de una RNN clásica donde se muestran tanto los estados ocultos $h_t$ generados en cada momento como las salidas $y_t$ de la red en forma de probabilidades asignadas a palabras. Por ejemplo, $y_3$ es la salida y $h_3$ es el estado oculto construido a partir de «El dinero del». $h_3$ participará en construir $h_4$ para predecir $y_4$.*




## Modelando la ruptura de expectativas en la frase

Cuando ya tenemos una RNN entrenada como la de la Figura 1, podemos emplear varias estrategias para comprobar las consecuencias de procesar frases manipuladas de distintas formas. Estas consecuencias se evalúan en las llamadas muestras de evaluación. La primera estrategia podemos llamarla «observable» y consiste en comparar las salidas de la red con patrones de comportamiento sujetos a hipótesis previas. Si la red predice palabras, como la del ejemplo de la Figura 1, la comparación será frente a palabras previamente establecidas y se analizará si el modelo es capaz de predecir las palabras adecuadas. Aquí entraría, por ejemplo, la constatación de que una red (manipulada convenientemente) produce lenguaje o comete errores similares a los de las personas.

Otra de las formas de análisis dentro de esta estrategia es una evaluación sobre la previsibilidad del lenguaje que produce una persona tomando como referencia el modelo. En este caso se calcula la «sorpresividad» del lenguaje[1] (Oh y Schuler, 2023) como medida de su predictibilidad. Por ejemplo, la producción verbal de una persona con afasia no fluyente divergirá del modelo normativo representado en una red (Cong et al., 2024). Dentro de una frase, cada palabra podrá ser sorpresiva si no es esperada normativamente. Concatenar tres determinantes seguidos no es normativo, y la divergencia con el modelo normativo representado en la red es muestra de ello. Es sorpresivo que después de «El dinero del» la persona haya producido un determinante como «este», ya que su probabilidad estimada será baja y su sorpresividad notable. La forma de calcular la sorpresividad es a partir de la salida de la red ($y_t$) representada en la Figura 1. La otra estrategia podríamos llamarla «oculta», y consistiría en aprovechar los estados ocultos que la red va generando para diseñar índices que consignen la situación interna (no observable) de la frase. Estos estados pueden homologarse con las representaciones mentales construidas dadas unas constricciones o, por lo menos, poner tal homologación bajo exploración. Como hemos visto, un estado oculto $h_t$ no es más que la salida de los nodos de una capa de recurrencia (también

[1] En Psicología, la sorpresividad (o sorpresa) del estímulo es la magnitud en la que un estímulo es inesperado o incongruente con respecto a las expectativas previas del individuo. Es un concepto clásico de modelos como el de Rescorla-Wagner (Rescorla y Wagner, 1972).

llamada oculta). Es decir, es una representación interna en forma de valores numéricos. Si hay una sucesión de entradas en una línea temporal (una sucesión de palabras de una frase), se generarán tantos estados ocultos como palabras. Cada estado representa el estado de la situación de la red en el momento de introducir una nueva palabra. Como se intuye, esto es crucial en las redes neuronales que sirven para procesar lenguaje, ya que esos estados internos muestran las expectativas de la frase en cada momento.

Cabe destacar que el lenguaje tiene una clara secuenciación temporal. Las frases siguen una línea temporal que, aunque no sea monótona (ya que pueden anticiparse palabras posteriores), sí marca su orden de procesamiento tanto en comprensión como en producción (e.g., Rayner, 2012 contiene múltiples ejemplos del procesamiento secuencial del lenguaje con metodología de movimientos oculares). Topologías como las RNN-LSTM o los *Transformers* son útiles para modelar esa línea temporal implícita en las frases. Aunque estas topologías sean sensiblemente distintas, ambas van generando estados ocultos, los cuales participan de una u otra forma en generar la salida de la red. En el caso de las RNNs, los estados ocultos están en función de la palabra (entrada, $x_t$) de la marca de tiempo $t$ de la frase y del estado oculto generado en el momento anterior ($h_{t-1}$). Así es cómo las RNNs hacen que la representación de la situación de la frase sea sensible a dos circunstancias: (a) la palabra procesada en el momento actual, y (b) un contexto que recoge la parte de la frase procesada anteriormente (implícito en el estado oculto anterior). En última instancia, la salida de la red $y_t$, que indicará las palabras más probables a suceder la secuencia, se instanciará como resultado de introducir estado oculto actual ($h_t$) como entrada en la capa de salida. Todo este proceso se llama de recurrencia: siempre hay un estado oculto anterior ($h_{t-1}$) que, junto con la entrada actual $x_t$, genera un nuevo estado oculto $h_t$, y éste una salida $y_t$. En el caso de los *Transformers*, esta recurrencia se omite y se hace uso de los llamados mecanismos de autoatención (Vaswani et al., 2017). No obstante, al igual que en las RNN, también se genera un estado oculto actual.

El hecho de que podamos tener acceso a los estados ocultos de la red es muy ventajoso, ya que podemos analizarlos para estudiar la representación que se tiene de la frase en cada momento. Es una representación interna de la situación de la frase y contiene información implícita tanto de las posibles relaciones gramaticales como de la semántica expresada (o incluso de indicios sensoriomotores o emocionales, si se ha construido el modelo con ellos). Pongamos como ejemplo la frase «Fui a un banco cercano a retirar dinero». Se intuye fácilmente que, si hemos leído solo «Fui a un banco cercano a…», las expectativas generadas son aún ambiguas. Esto significa que la situación de la frase es incierta, ya que la situación puede referirse a sentarse o sacar dinero. En cualquier caso, lo esperable es continuar la frase con un verbo. El estado oculto generado por el modelo en ese momento nos podrá informar de tal fenómeno y los cambios de ese estado en lo sucesivo nos pueden dar una métrica de certidumbre. Imaginemos que manejamos la frase: «Fui a un banco cercano a rastrojos». En este nuevo caso podemos cotejar el cambio del estado oculto entre el punto en que se ha leído «Fui a un banco cercano a…" ($h_{t-1}$) y el punto en que se completa la frase "Fui a un banco cercano a rastrojos» ($h_t$). El cambio entre $h_t$ y $h_{t-1}$ puede ser notable, y de ello extraerse una métrica de cambio o de ruptura de expectativas.

Tanto el cálculo de la sorpresividad (estrategia basada en la salida $y_t$; Oh y Schuler, 2023) como del cambio en el estado oculto ($h_t$) pueden considerarse como indicadores de cambio de expectativas tanto semánticas como gramaticales. Si la sorpresividad de una palabra producida por una persona es notable, se rompen las expectativas, al igual que se rompen si la representación de la situación de la frase inferida por el estado oculto cambia de un momento a otro. Este fenómeno es precisamente lo que buscan capturar ciertos potenciales relacionados con eventos (ERPs). Así, se establece una posible conexión entre algunos ERPs y los índices derivados de estas estrategias, como se ha demostrado en investigaciones previas (e.g., Rabovsky y McClelland, 2020; Rabovsky et al., 2018).

No obstante, aun pudiendo representar ambos índices una ruptura de expectativas, se han observado diferencias entre la sorpresividad y el cambio de estado oculto. En primer lugar, se ha observado que el cambio de estado oculto está más correlacionado con el ERP N400 (Rabovsky y McClelland, 2020; Rabovsky et al., 2018), planteándose como hipótesis principal que ambos están más relacionados con la semántica. Es un hallazgo interesante para razonar sobre el impacto de la plausibilidad temática de las

frases y su confrontación con el más fino procesamiento sintáctico. Hipotéticamente, tanto el N400 como el cambio de estado oculto serían mucho más sensibles a la frase «Fui a un banco cercano a rastrojos» por su semántica inesperada, que a «Fui un dinero retirar en banco a cercano» por encontrarse los distintos elementos semánticamente relacionados. De igual manera, frases empleadas en pruebas de dislexia fonológica como «El perro es perseguido por el gato», podrían generar un N400 menor y poco cambio en los estados ocultos de la red. En ambos ejemplos se puede crear una ilusión semántica donde se interpreta la frase por plausibilidad (Rabovsky y McClelland, 2020). Sin embargo, se ha observado que la sorpresividad correlaciona en mayor manera con el P600 (Rabovsky y McClelland, 2020). La sorpresividad, al igual que el P600, es sensible a las expectativas sintácticas, aunque también en cierta medida a las semánticas (ver el trabajo de Slaats, y Martin, 2023 para una reflexión sobre este amalgamiento de fuentes de variabilidad de la sorpresividad). De esta manera, es sensible a la agramaticalidad y es en cierto modo una alarma de que algo no cuadra en la frase.

En este texto ilustramos conceptualmente la forma de conseguir los estados ocultos que se generan en los distintos momentos de las frases, y la forma de medir la sorpresividad de las palabras de las frases. Así, mostraremos la utilidad tanto de los estados ocultos y sus cambios como de la sorpresividad, en un fenómeno que inunda todos los debates de Psicología del lenguaje: la composicionalidad sistemática (e.g., Liñán, 2009; Szabó, 2001, 2020). Esta es una propiedad de los sistemas cognitivos que explica cómo somos capaces de derivar el significado de expresiones complejas a partir del significado de sus partes y las reglas que las combinan. Este concepto nos dará la oportunidad de ilustrar cómo se utilizaría un modelo de RNNs sustentado en un diseño experimental con ERPs.

## Cómo calcular índices en las estrategias observables y oculta

Vamos a analizar de manera pormenorizada una RNN con una topología meramente autogenerativa[2], asumiendo que el modelo ya ha sido entrenado. Esta topología se muestra en la Figura 1 (para más detalles se puede consultar Jorge-Botana, 2024). Hemos aludido antes a dos estrategias posibles para estudiar las expectativas lingüísticas del modelo: una observable y otra oculta.

La primera, cuyo índice es la sorpresividad (Oh y Schuler, 2023), consiste en operar en la capa de salida y sus predicciones, es decir, en la parte observable ($y_t$). Esto significa que se utilizan los vectores de salida generados por las distintas entradas de una secuencia. Si tenemos un conjunto de palabras dispuestas como secuencia $\{x_1, x_2, x_3, \ldots, x_t\}$, podemos tomar el vector $y_t$ de salida como la predicción contextualizada de una entrada en la marca de tiempo $t$, cuyo vector es $x_t$. El vector $y_t$ expresará la probabilidad que tiene cada palabra del vocabulario de ser salida en ese momento de la frase. De esa manera, $y_t$ será un vector que contiene las probabilidades de cada una de las palabras del vocabulario de suceder a la secuencia que ya se ha procesado. Si se procesa una frase como "El dinero del banco piedra", se puede estimar en el vector $y_4$ que la probabilidad de «piedra» es muy baja, acaso por la temática o porque un sustantivo no es previsible detrás de otro. Esto es posible porque $y_4$ tiene tantos componentes como palabras tenga el vocabulario, siendo un listado de probabilidades para todas las palabras que conoce el modelo. El primer componente del vector $y_4$ representa la probabilidad de que la palabra indizada como primera en el vocabulario suceda a la frase «El dinero del banco». El segundo componente del vector $y_4$ representa la probabilidad de que la palabra indizada como segunda en el vocabulario suceda a la frase. Y así con todos los componentes del vector. Así, si «piedra» ocupase la posición 5 en el vocabulario, el componente 5 de ese vector $y_4$ sería la probabilidad de que ocurra piedra detrás de «El dinero del banco»:

[2] Se usa únicamente un decodificador. En las topologías llamadas codificador-decodificador, el codificador proporciona el contexto para que el decodificador empiece a autogenerar lenguaje de manera probabilística. Si el contexto es fuerte, la autogeneración estará instigada por él. De ahí que lo que emita el decodificador no sea arbitrario sino apegado a un tema (una pregunta, una imagen, una frase, etc.). No obstante, en alguna topología solo se requiere de decodificador (por ejemplo, cuando se le da un pie para que el decodificador simplemente lo autocomplete).

$$y_5 = P(\text{w}_{\text{piedra}} | \text{w}_{\text{el}}\ \text{w}_{\text{dinero}}\ \text{w}_{\text{del}}\ \text{w}_{\text{banco}}) \quad [1]$$

Aplicando el logaritmo en negativo a lo obtenido en $y_5$, obtenemos la medida de sorpresividad. Así pues, el cálculo de la sorpresividad (S) de una palabra en una posición $t$ concreta ($w_t$) dadas una serie de n palabras previas quedaría definido como:

$$S(w_t) = -\log P(\text{w}_t | w_{t-1}\ w_{t-2}, \ldots, w_{t-n}) \quad [2]$$

En nuestro ejemplo, la sorpresividad de la palabra "piedra" sería:

$$S(w_{piedra}) = -\log\log(y_5) = -\log\log \text{P}(w_{piedra} | w_{el}\ \ w_{dinero}\ w_{del}\ w_{banco}) \quad [3]$$

Se puede calcular la sorpresividad de cualquier palabra que forme parte del vocabulario en cada momento $t$. Salta a la vista que esta estrategia es llamada observable por ser la materialización observable en la capa de salida del estado oculto $h_4$ generado en ese mimo momento. Así, $y_4$ sería la predicción a partir de la entrada $x_4$, pero contextualizada con el estado oculto $h_3$ (ver Figura 1). Tomando ese vector de salida $y_4$, podemos obtener las probabilidades asociadas a que cada palabra del vocabulario suceda después y con esto calcular su sorpresividad (en el Apéndice ofrecemos algunos enlaces a códigos que calculan la sorpresividad empleando modelos preentrenados). Esta estrategia asume que los nodos de la capa de salida tienen como tarea predecir palabras con la función *softmax* (Bridle, 1989; cuyo origen puede rastrearse hasta Boltzmann, 1868) como función de activación, dado que se trabaja con vectores one-hot para indexar las palabras del vocabulario.

No obstante, una cosa es dar una probabilidad a cada palabra y otra contar con la representación oculta de la situación de la frase en ese momento. Con la predicción ($y_t$) podemos obtener una distribución en la que algunas palabras tendrán más probabilidad de continuar la secuencia. Con la representación oculta (estado oculto $h_t$) tenemos una forma de representar la situación de la frase en el marco de un contexto. Esto permite, por ejemplo, representar dinámicamente la situación de la frase antes y después de introducir «banco» en el contexto previo de «No me queda mucho dinero en el…» o en el contexto de «Como estoy cansado me siento en el…».

Se suele formalizar la situación de la frase contextualizada en un momento concreto tomando el valor del estado oculto $h_t$ que se genera en él. De esta forma, se toma $h_t$ como la representación interna de la situación de la frase en el tiempo $t$. Si quisiéramos consignar el cambio en tal situación oculta antes y después de introducir la palabra «banco» (como en la Figura 1), podríamos calcular una simple distancia vectorial entre los estados ocultos antes y después de tal hecho:

$$Cambio\ («banco») = h_3 - h_4 \quad [4]$$

Este índice, llamado cambio de estado oculto (Rabovsky y McClelland, 2020), puede calcularse en todos los momentos de la frase como la diferencia entre el estado oculto antes y después de la palabra: $h_t - h_{t-1}$. Puede decirse que $h_4$ representa una expectativa a partir de $x_4$ y su contexto anterior. Así, de una entrada $x_4$, con un formato incierto, se consigue en $h_4$ una representación con expectativas de entidad bancaria. Piantadosi (2023) define esos estados internos como aspectos latentes de la sintaxis y la semántica que gobiernan la interpretación del texto (en el Apéndice ofrecemos algunos enlaces a códigos que ayudan a conseguir los estados ocultos en los diferentes momentos de la frase y de las diferentes capas ocultas).

En términos de plausibilidad cognitiva, es sugerente reflexionar sobre el concepto de expectativa en referencia a los estados ocultos. Cuando en una secuencia se produce una nueva entrada $x_t$ y ésta interactúa con el estado oculto anterior $h_{t-1}$, el nuevo estado oculto $h_t$ puede entenderse como las expectativas sobre qué palabras podrían acompañar a la entrada $x_t$ en la siguiente marca temporal dado ese contexto. Este estado es pues una constelación de posibilidades y, así, cada momento conlleva una constelación de palabras diferentes. Esas expectativas son una potencia que se materializa en forma de palabras en la capa de salida mediante la distribución de probabilidad $y_t$.

## Composicionalidad sistemática como núcleo del debate

El conocimiento humano no parece estar exclusivamente basado en la experiencia (o, al menos, en un mero cómputo probabilístico sobre ella). Parece haber habilidades *emergentes* que generalizan el uso de reglas sobre estructuras nunca vistas. Esto se relaciona con el conocido fenómeno de la «pobreza del estímulo» (e.g., Pearl, 2022). Este fenómeno presenta la paradoja de que, aunque las personas están expuestas a un conjunto relativamente pequeño de oraciones, sus expresiones suelen ser correctas formalmente y pueden llegar a inferir el significado de frases con palabras legítimamente combinadas, aunque nunca vistas en concurrencia. Esto significa que las personas pueden comprender y producir combinaciones lingüísticas que no han visto. Se postula que tal capacidad viene dada por la denominada composicionalidad sistemática del lenguaje (e.g., Liñán, 2009; Szabó, 2001, 2020).

Formalmente, en el concepto de composicionalidad sistemática se mantiene que la entrada no proporciona evidencia sobre todas las oraciones posibles, y que tampoco contiene reglas explícitas sobre las posibles combinatorias ni sus significados (Lasnik y Lidz, 2016). Sin embargo, las personas se desenvuelven relativamente bien en ambos casos, aunque no ven todo el lenguaje (y sus variedades) a lo largo de su vida, ni se les dice explícitamente cómo combinarlo.

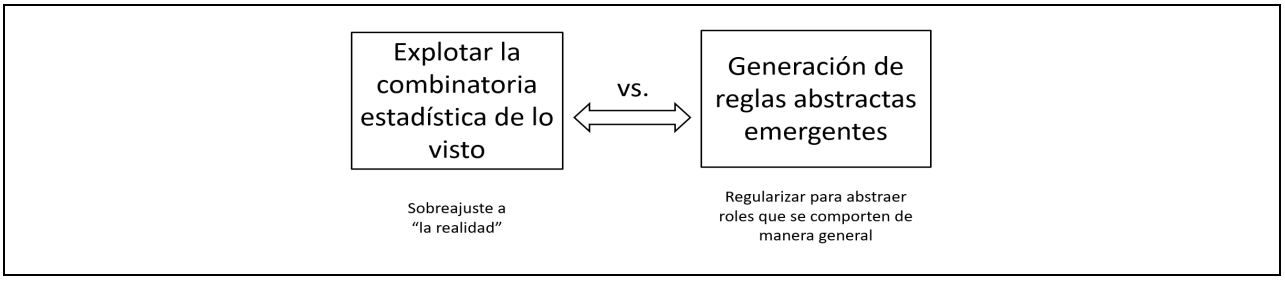
Tradicionalmente, se ha sugerido que la composicionalidad sistemática está relacionada con la sintaxis. Visto de esa forma, existirían dos sistemas, uno con representaciones semánticas de las palabras y otro que aplicaría reglas universales. Este último sistema está separado del significado de las palabras individuales (trabajo seminal de Chomsky, 1957). Es de resaltar también un componente modular introducido por Fodor y Pylyshyn (1988) a esta concepción composicional: las unidades del sistema semántica son módulos y las palabras participarían con la misma carga semántica independientemente de su rol sintáctico. En las frases «Julieta quiere a Romeo» y «Romeo quiere a Julieta», Julieta y Romeo participarían con la misma carga semántica en ambas frases. No obstante, parece complicado asumir estos supuestos (Rabovsky y McClelland, 2020; Rabovsky et al., 2018). Asumir ese componente modular implicaría que las representaciones de las palabras se recuperan antes de asignarles roles sintácticos, sin mediar ninguna contextualización más que la que posteriormente imponga la sintaxis. Sin embargo, la Julieta de la primera frase es una chica amante, y la de la segunda es amada (incluso pudiera ser que participen distintas Julietas en ambas frases), cada una con su constelación de posibilidades. Consecuentemente, algunos autores rebajan esa constricción de modularidad y hablan más bien de pseudocomposicionalidad (Rabovsky y McClelland, 2020). Esta nueva composicionalidad no asumiría que las palabras contribuyen de manera independiente a su rol sintáctico.

El fenómeno de la composicionalidad puede evaluarse a través de índices como el cambio en los estados ocultos y la sorpresividad mencionados antes. Consideremos secuencias de palabras a las que hemos sido expuestos como, por ejemplo, «Julieta ama a Romeo». Hemos visto a chicas llamadas Julieta amar a chicos llamados Romeo y, por tanto, forma parte de lo plausible. Sin embargo, podemos abstraernos de la realidad e interpretar frases como "Julieta ama el morado" o más implausibles como «Julieta ama la garrapata». La idea es jugar con la realidad (i.e., los estímulos que participan en el estudio) para confrontar su efecto en las personas y en los modelos. Tómese como ejemplo la frase que Chomsky (1957) introdujo como reto: «Las ideas verdes incoloras duermen furiosamente» (*Colorless green ideas sleep furiously*). Aunque nunca hemos sido expuestos a esa escena, podemos producir la frase y evocar su significado, incluso en contra de su plausibilidad. De manera similar, también entenderemos la frase «Julieta ama la garrapata». Lo importante es que estas frases pueden ser producidas y su significado se puede inferir, aunque la plausibilidad del lenguaje favorecerá las expectativas en algunas ocasiones y las penalizará en otras. Si un sistema no fuera capaz de abstraerse de la experiencia previa a la que ha sido expuesto y entender frases que no haya visto antes (e.g., «Julieta ama la garrapata»), estaría sobreajustado a la realidad.

Algunos antropólogos ya propusieron la existencia de algunos sistemas lingüísticos primitivos o protolenguas en el que no existía dicha generalización. Estos fueron llamados lenguajes libres de sintaxis, donde la capacidad sintác-

**Figura 2**

*La composicionalidad sistemática del lenguaje tiene como resultado la generación de reglas abstractas emergentes más que simplemente explotar la combinatoria estadística de la información procesada. Podemos pues entender esto como resultado de un proceso de regularización durante el proceso de aprendizaje*




tica se amalgama con la semántica de las palabras (Bickerton, 1995; Jackendoff, 1987, 1999). Aludimos a este caso extremo para hacer gráfico un caso en el que la composicionalidad sistemática es difícil o imposible de desplegar. En el otro extremo, podemos tener un sistema capaz de abstraer reglas aisladas de su experiencia con la realidad explotando los indicios sintácticos para dilucidar cuando un sustantivo es activo o es pasivo. En términos evolutivos, algunos autores afirman que el proceso de gramaticalización (es decir, la atribución de un papel gramatical o sintáctico a una palabra) es análogo al de la metaforización (Heine y Kuteva, 2007). Cuando dos palabras comparten algunas propiedades funcionales, el sistema cognitivo evoluciona y se apercibe de que ambas comparten un papel común y, por tanto, son intercambiables en determinados roles genéricos. Así, el sistema aprende que dos palabras comparten un papel común a través de esas propiedades funcionales compartidas a lo largo de su experiencia. Esto nos lleva también a pensar en términos piagetianos, ya que las palabras podrían ser asimiladas a una categoría abstracta como un rol sintáctico. Podríamos extender estos razonamientos también a otros ejemplos como el caso de los verbos inventados o las pseudopalabras donde, aunque no conozcamos las palabras, somos capaces de interpretarlas como sustantivos o verbos dependiendo de cómo se comportan sintácticamente.

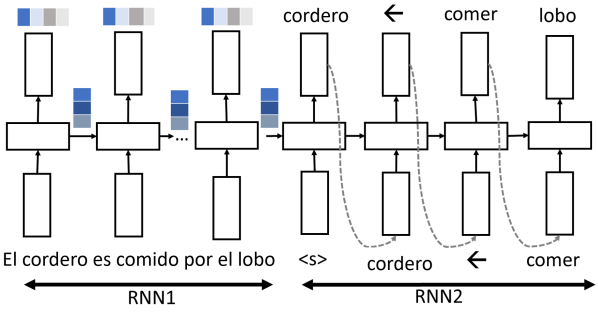
La semántica sería, según lo argumentado, una cosmología de expectativas del uso de las palabras, y estas expectativas pueden ser semánticas o incluso sintácticas en cuanto a continuidad de la frase. Este tipo de expectativas estarán constreñidas por las posibilidades de usos reales del lenguaje. Sin embargo, la composicionalidad es una manera de regularizar el conocimiento del sistema para que no se sobreajuste a la realidad (Figura 2). Tanto el término de sobreajuste como el de regularización tienen el sentido que se les da en el ámbito del aprendizaje automático. Regularizar es impedir que el modelo se sobreajuste a características irrelevantes de la muestra de entrenamiento y deje holgura para usar de manera implícita indicios mucho más genéricos como, por ejemplo, los roles sintácticos. Puede verse de manera clara esta lógica en trabajos que imponen a la función de coste usada para la actualización de los pesos una regularización basada en maximizar la independencia de los constituyentes de las frases (Nandi et al., 2024). Esto último puede considerarse una forma de instigar que los estados ocultos creados por las sucesivas palabras de una frase no se amalgamen en base a la combinatoria vista en la realidad, y el coste de la desambiguación haga que no abstraiga propiedades genéricas de la estructura composicional (Choi et al., 2017).

## Un experimento tentativo sobre los efectos de la plausibilidad

En este apartado vamos a plantear un experimento mental donde se pretenden homologar los índices de ruptura de expectativas lingüísticas (plausibilidad) de los modelos de RNNs con las sensibilidades extraídas de algunos ERPs. El objetivo es mostrar uno de los posibles paradigmas que podrían emplearse para evaluar el comporta-

**Figura 3**

*Entrenamiento del codificador-decodificador en la reconstrucción de una escena tanto en sus participantes como en sus roles pasivo-activo. La RNN1 (el codificador) proporciona a la RNN2 (el decodificador) la codificación de la frase en su último estado oculto, y la RNN2 autogenera la escena. Los pesos tanto de la RNN1 como de la RNN2 se modifican en función del éxito de la autogeneración de la escena. La RNN2 produce <s> como inicio de la escena y asigna el rol pasivo-activo a través de la dirección de una flecha (en este caso, ← indica que el verbo comer se produce de manera pasiva y que el cordero es comido por el lobo)*



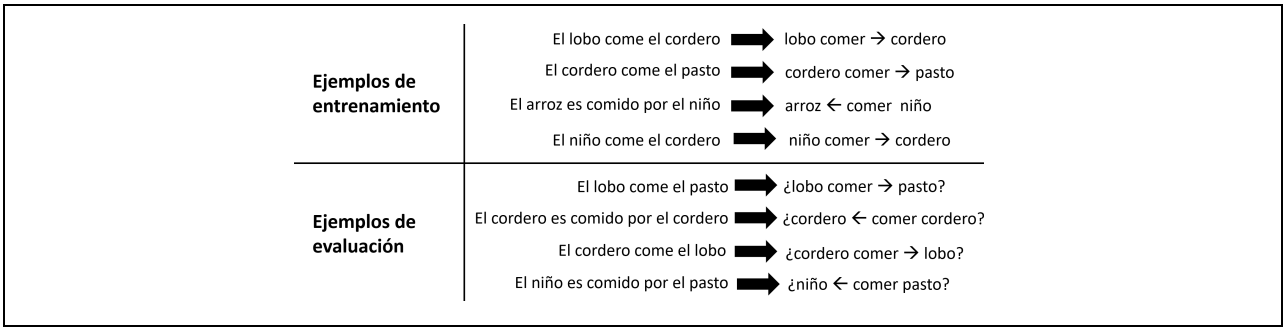

miento composicional y su correlación con ERPs (Rabovsky y McClelland, 2020 o Rabovsky et al., 2018 presentan algunos diseños completos, aunque con un modelo idiosincrático llamado *Sentence Gestalt* que no se corresponde con los modelos más empleados en el campo de Inteligencia Artificial Generativa). Este experimento mental también nos permitirá mostrar una forma tentativa de comprobar si el cambio de estado oculto en una RNN está relacionado con el cambio abrupto de la situación de la frase en términos eminentemente temáticos más que sintácticos. Se espera, según lo dicho previamente, que el cambio de estado oculto esté más alineado con el potencial N400 que con el P600.

Para ello, vamos a plantear una topología de RNN Secuencia-Secuencia con codificador y decodificador (consultar Jorge-Botana, 2024 para más detalles) como la que se presenta en la Figura 3. Durante la fase de entrenamiento, se codifica una frase en lenguaje natural en el codificador y se autogenera la reconstrucción de su escena en el decodificador, teniendo en cuenta tanto el aprendizaje de las palabras (sustantivos y verbos) como la asignación de roles pacientes y agentes. El codificador simula la transformación de la información de una frase en una representación interna, tal y como haría una persona que lee una frase e imagina una escena. Otros autores han sugerido aproximaciones similares con la intención de dotar a los modelos del lenguaje de conocimiento del mundo mediante diversas vías (véase Carta et al., 2023; Hernández et al., 2023; Ivanova et al., 2024).

La especificación de las características de la red es el primer eje de manipulación sobre el que podemos trabajar. El segundo eje de manipulación son las muestras con las que aprende la red. En este caso, se utilizaría un conjunto de pares frase-escena. Lo interesante de este corpus es que se trata de una muestra muy controlada en la que se limita el tamaño evitándose así el efecto de escala de los Grandes Modelos de Lenguaje. Pero lo más importante es que, en este conjunto, habrá sustantivos que actúan como agentes y pacientes, otros sólo como agentes, y el resto sólo como pacientes, tal y como podríamos encontrar en el mundo real (ver los ejemplos de entrenamiento en la Figura 4). Una frase plausible será, por ejemplo, «El lobo come el cordero», pues se ha presentado durante el entrenamiento y el modelo ha aprendido a generar esa escena. Así, el modelo estará sobreajustado a estas frases, pudiendo controlar la plausibilidad de las frases que, una vez entrenado, se evaluarán en el modelo. Las entradas puestas bajo escrutinio una vez entrenada la red son el tercer eje de manipulación. Dentro del conjunto de evaluación, existirán también frases no vistas cuya escena nunca se ha representado

**Figura 4**

*Entrenamiento y evaluación del modelo con pares frases-escena. El entrenamiento se hace con frases plausibles con escenas sistemáticamente correctas. La evaluación se hace con frases plausibles y no plausibles para analizar la reconstrucción de las escenas por parte del modelo*

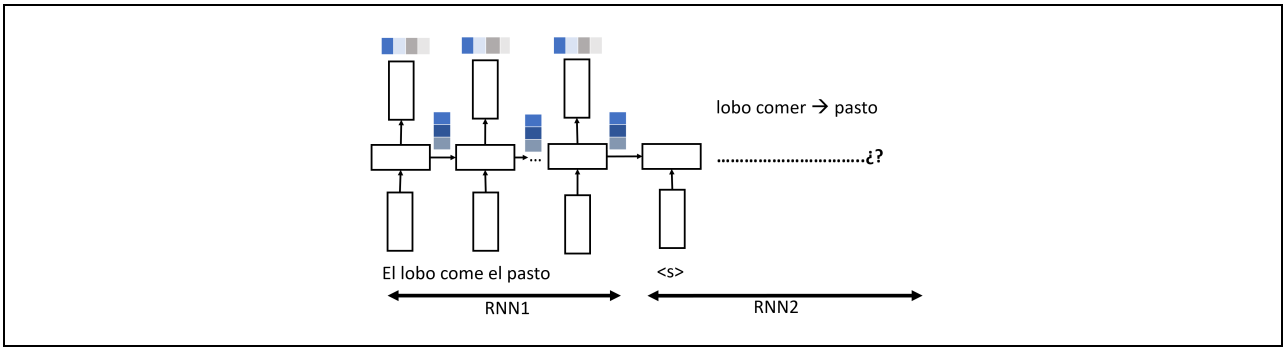



nuestro modelo (ver los ejemplos de evaluación con frases implausibles en la Figura 4) como, por ejemplo, «El cordero come el lobo» o «El niño es comido por el pasto». Por tanto, habrá distintos niveles de plausibilidad en los que solo algunos casos se podrán resolver por composicionalidad. También se podría manipular la legitimidad sintáctica de las frases de prueba cambiando el orden de algunas palabras e o intercambiando los roles pasivos y activos de los sustantivos para estudiar los roles sintácticos.

La clave de todo esto, además de ver si el modelo es capaz de reconstruir escenas de frases no vistas con distinto grado de plausibilidad (e.g., «El lobo come el pasto» o «El niño es comido por el pasto»), es medir el cambio de estado oculto en el momento de introducir en el codificador las palabras que rompen las expectativas tanto por plausibilidad como por legitimidad sintáctica (Figura 5). Con este paradigma podemos comprobar cómo la medida de cambio del estado oculto explica las representaciones del lenguaje de una manera controlada a través de esta topología de RNN. Resultados previos sugieren que el N400

**Figura 5**

*Evaluación del codificador-decodificador a través de frases plausibles e implausibles. Los estados ocultos y sus diferencias permiten consignar el cambio como ruptura de las expectativas lingüísticas ante una frase nunca vista por el modelo. Los cambios en los estados ocultos pueden ser sensibles a la ruptura de expectativas y, consecuentemente, se pueden calcular índices que cuantifiquen dicha ruptura de expectativas como:* $h_{(El\ lobo\ come\ el\ pasto)} - h_{(El\ lobo\ come\ el)}$. *En cada marca de tiempo, al introducir la siguiente palabra en el codificador, se genera un nuevo estado oculto que podría generar un cambio en las expectativas de la frase*



está más alineado con medidas como el cambio de estado oculto y que ambos son eminentemente temáticos y situacionales (Rabovsky y McClelland, 2020; Rabovsky et al., 2018). De esta manera, estos modelos nos permiten hacer predicciones sobre cómo funciona la composicionalidad y qué mecanismos hay implicados, pudiendo simular respuestas en las RNNs que son parecidas a los ERPs que se consignan en el cerebro. Además, los razonamientos sobre la plausibilidad de los modelos de redes neuronales pueden ayudar a mejorarlos y, con ayuda de las estrategias basadas en respuestas ocultas y observables, desplegar en ellos mecanismos basados en las mismas sensibilidades que se detectan en los experimentos de ERPs. Terra ignota.

## Conclusión

Las nuevas arquitecturas de red neuronal (RNN-LSTM y *Transformers*) y sus distintas topologías ponen en nuestras manos herramientas muy poderosas para formalizar las teorías cognitivas del lenguaje que describimos en lenguaje natural. Además, su aparataje permite calcular índices sobre sus salidas y analizar los estados ocultos que se van generando en cada momento de la línea temporal de una frase. Esto permite confrontar modelos y experimentos para corregir tanto las teorías como los modelos, y aplicar esas correcciones a las arquitecturas que hoy día están por debajo de los Grandes Modelos del Lenguaje.

Este texto se focaliza en la arquitectura RNN-LSTM, puesto que sus mecanismos son muy interesantes en términos de plausibilidad cognitiva. El hecho de capturar las dependencias temporales del lenguaje con mecanismos de memoria de trabajo con diferente pervivencia de la información las hace muy interesantes a nivel psicológico (la memoria de trabajo a corto plazo, a largo plazo o los propios mecanismos de olvido y aportación). Puede decirse que las RNN-LSTM son muy intuitivas para entender un posible modelo situacional de las frases. Sin embargo, actualmente, los *Transformers* (Vaswani et al., 2017) han sustituido a las RNNs-LSTM en muchas aplicaciones porque presentan ciertas ventajas: procesamiento paralelo (procesan toda la entrada de una secuencia a la vez gracias a sus mecanismos de autoatención), mayor eficiencia (mayor escalabilidad y eficiencia al no requerir recurrencia), y menor restricción secuencial (ya que pueden acceder al contexto sin tener que recorrer las secuencias en orden). No obstante, aunque de forma diferente, los *Transformers* también generan estados ocultos en cada marca de tiempo a partir de la integración de la información de las palabras de la frase. Es por ello por lo que se pueden generalizar directamente los razonamientos que hemos mostrado en el texto a esta arquitectura. Además, la capa de salida no difiere de la de las RNN-LSTM y, por tanto, los cálculos de cambio en los estados ocultos y el índice de sorpresividad serán comunes.

## Referencias

Anderson, J. R. (2005). *Cognitive Psychology and its Implications*. Macmillan.

Belinchón, M., Igoa, J. M. y Rivière, Á. (2009). *Psicología del Lenguaje. Investigación y Teoría* [Psychology of Language: Research and Theory]. Trotta.

Bickerton, D. (1995). *Language and Human Behavior.* University of Washington Press.

Boltzmann, L. (2012). Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten [Studies on the Equilibrium of Living Force Between Moving Material Points]. En F. Hasenöhrl (Ed.), *Wissenschaftliche Abhandlungen* (pp. 49–96). Cambridge University Press.

Bridle, J. (1989). *Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters*. In Proceedings of the 2nd International Conference on Neural Information Processing Systems (NIPS-89) (pp. 211–217). Morgan Kaufmann.

Busemeyer, J. R., Wang, Z., Townsend, J. T. y Eidels, A. (2015). *The Oxford Handbook of Computational and Mathematical Psychology.* Oxford University Press.

Carta, T., Romac, C., Wolf, T., Lamprier, S., Sigaud, O. y Oudeyer, P. Y. (2023, July). Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning. En *International Conference on Machine Learning* (pp. 3676–3713). PMLR.

Choi, H., Cho, K. y Bengio, Y. (2017). Context-Dependent Word Representation for Neural Machine Translation. *Computer Speech & Language, 45,* 149–160.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Cong, Y., LaCroix, A. N. y Lee, J. (2024). Clinical Efficacy of Pre-Trained Large Language Models through the Lens of Aphasia. *Scientific Reports, 14*(1), Artículo 15573. https://doi.org/10.1038/s41598-024-66576-y

De Vega, M., Glenberg, A. y Graesser, A. (2012). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press.

Elman, J.L. (1990). Finding Structure in Time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Farrell, S. y Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Current Directions in Psychological Science, 19*(5), 329–335. https://doi.org/10.1177/0963721410386677

Federmeier, K. D. y Kutas, M. (1999). Right Words and Left Words: Electrophysiological Evidence for Hemispheric Differences in Meaning Processing. *Cognitive Brain Research, 8*(3), 373–392. https://doi.org/10.1016/S0926-6410(99)00036-1

Fodor J. A. y Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 28*(1-2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Heine, B. y Kuteva, T. (2007). *The genesis of Grammar: A Reconstruction.* Oxford University Press.

Hernandez, E., Sen Sharma, A., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y. y Bau, D. (2023). Linearity of Relation Decoding in Transformer Language Models. *ArXiv,* https://doi.org/10.48550/arXiv.2308.09124

Hochreiter, S. y Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780.

Hubel, D. H. y Wiesel, T.N. (1968). Receptive Fields and Functional Architecture of Monkey Striate Cortex. *Journal of Physiology, 195,* 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Ivanova, A., Sathe, A., Lipkin, B., Kumar, U., Radkani, S., Clark, T., Kauf, C., Hu, J., Pramod, R., Grand, G., Paulun, V., Ryskina, M., Akyurek, E., Wilcox, E., Rashid, N., Choshen, L., Levy, R., Fedorenko, E., Tenenbaum, J. y Andreas, J. (2024). Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv* https://doi.org/10.48550/arXiv.2405.09605

Jackendoff, R. (1999). Possible Stages in the Evolution of the Language Capacity. *Trends in Cognitive Sciences, 3*(7), 272–279. https://doi.org/10.1016/S1364-6613(99)01333-9

Jackendoff, R. (1987). *Consciousness and the Computacional Mind*. MIT Press.

Jordan, M.I. (1997). Serial Order: A Parallel Distributed Processing Approach. *Advances in Psychology, 121,* 471–495. https://doi.org/10.1016/S0166-4115(97)80111-2

Jorge-Botana, G. (2024). *Redes neuronales recurrentes y Transformers para modelos cognitivos del lenguaje* [Recurrent Neural Networks and Transformers for Cognitive Language Models]. Ediciones Complutense.

Jurafsky, D. y Martin, J. H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson.

Kaan, E. (1999). Syntax and Semantics? *Trends in Cognitive Sciences, 3*(9), Artículo 322. https://doi.org/10.1016/S1364-6613(99)01376-5

Kutas, M. y Federmeier, K.D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology, 62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M. y Hillyard, S. A. (1980). Reading Senseless Sentences: Brain potentials Reflect Semantic Incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Lasnik, H. y Lidz, J. (2016). The Argument from the Poverty of the Stimulus. En I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar* (pp.221–248). Oxford Academic.

Lindquist, K. A. (2021). Language and Emotion: Introduction to the Special Issue. *Affective Science, 2*(2), 91–98. https://doi.org/10.1007/s42761-021-00049-7

Liñán, J. L. (2009). Sistematicidad, productividad y composicionalidad: Una aproximación pragmatista [Sistematicity, Productivity, and Compositionality: A Pragmatistic Approach]. *Revista de Filosofía, 34*(1), 51–75.

McClelland, J. L. y Rumelhart, D.E. (1989). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*. MIT press.

McClelland, J. L., Rumelhart, D. E. y PDP Research Group. (1987). *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models* (Vol. 2). MIT press.

Nandi, A., Manning, C. D. y Murty, S. (2024). Sneaking Syntax into Transformer Language Models with Tree Regularization. arXiv preprint arXiv. https://doi.org/10.48550/arXiv.2411.18885

Neisser, U. (1967). *Cognitive Psychology*. Prentice-Hall.

Oh, B. D. y Schuler, W. (2023). Why Does Surprisal from Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics, 11*, 336–350.

Pearl, L. (2022). Poverty of the Stimulus Without Tears. *Language Learning and Development, 18*(4), 415–454. https://doi.org/10.1080/15475441.2021.1981908

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. En E. Gibson y M. Poliak (Eds), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett* (pp.353–414). Language Science Press.

Pitt, D. (2022). Mental Representation. En E. N. Zalta y U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition). https://plato.stanford.edu/archives/fall2022/entries/mental-representation/

Rabovsky, M., Hansen, S. S. y McClelland, J. L. (2018). Modelling the N400 Brain Potential as Change in a Probabilistic Representation of Meaning. *Nature Human Behaviour, 2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

Rabovsky, M. y McClelland, J. L. (2020). Quasi-Compositional Mapping from form to Meaning: A Neural Network-Based Approach to Capturing Neural Responses during human Language Comprehension. *Philosophical Transactions of the Royal Society B, 375*(1791), Artículo 20190313. https://doi.org/10.1098/rstb.2019.0313

Rayner, K. (Ed.). (2012). *Eye Movements in Reading: Perceptual and Language Processes*. Academic Press.

Rescorla, R. A. y Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. En A. H. Black y W. F. Prokasy (Eds.), *Classical Conditioning II* (pp. 64–99). Appleton-Century-Crofts.

Rumelhart, D. E., McClelland, J. L. y PDP Research Group. (1986). *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition*. Foundations. MIT press.

Skinner, B. F. (1957). *Verbal Behavior*. Prentice-Hall.

Slaats, S. y Martin, A. E. (2023). *What's Surprising about Surprisal*. https://osf.io/7pvau/download/

Sterelny, K. (1990). *The Representational Theory of Mind.* Basil Blackwell.

Sun, R. (2023). *The Cambridge Handbook of Computational Cognitive Sciences*. Cambridge University Press.

Szabó, Z. G. (2001). *Problems in Compositionality*. Garland.

Szabó, Z. G. (2020). Compositionality. En E. N. Zalta y U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab. http://seop.illc.uva.nl/entries/compositionality/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. (2017). Attention is all you need. En U. von Luxburg, I. Guyon y S. Bengio (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Curran.

## Apéndice

Códigos que calculan la sorpresividad empleando modelos preentrenados:
https://github.com/tmalsburg/llm_surprisal
https://github.com/aalok-sathe/surprisal
https://github.com/simonepri/lm-scorer
https://github.com/TomSgrizzi/surprisal-with-psychformers
https://github.com/samer-noureddine/GPT-2-for-Psycholinguistic-Applications/blob/master/get_probabilities.py

Códigos para extraer estados ocultos:
https://stackoverflow.com/questions/48302810/whats-the-difference-between-hidden-and-output-in-pytorch-lstm
https://www.geeksforgeeks.org/difference-between-hidden-and-output-in-pytorch-lstm/

# PRUEBAS DE ELECCIÓN FORZOSA: VISIÓN ACTUAL Y RECOMENDACIONES

# FORCED-CHOICE TESTS: CURRENT PERSPECTIVE AND RECOMMENDATIONS

FRANCISCO J. ABAD[1], RODRIGO S. KREITCHMANN[2], DIEGO GRAÑA[1], PABLO NÁJERA[3] Y MIGUEL A. SORREL[1]

Cómo referenciar este artículo/How to reference this article:

Abad, F. J., Kreitchmann, R. S., Graña, D., Nájera, P. y Sorrel, M. A. (2024). Pruebas de elección forzosa: visión actual y recomendaciones [Forced-Choice Tests: Current Perspective and Recommendations]. *Acción Psicológica, 22*(1), 57–72. https://doi.org/10.5944/ap.22.1.43413

## Resumen

Este artículo tiene como objetivo ofrecer una visión actual de las pruebas de elección forzosa y proporcionar recomendaciones para su diseño y construcción. Aunque estas pruebas ayudan a superar limitaciones como los sesgos de deseabilidad social y de respuesta extrema, comunes en los formatos de respuesta graduada, presentan desafíos técnicos relacionados con la ipsatividad de las puntuaciones. Este artículo presenta modelos psicométricos basados en la Teoría de Respuesta al Ítem (TRI), como el modelo Thurstoniano de TRI para preferencias (TIRT) y el modelo de preferencia por pares multi-unidimensional (MUPP), que mejoran la estimación de los rasgos y permiten un ensamblaje óptimo de ítems en bloques. Se identifican factores de diseño del cuestionario, como la polaridad de los ítems ensamblados, que pueden afectar la calidad de las puntuaciones obtenidas. Además, se exploran los beneficios de la TRI en el desarrollo de tests adaptativos informatizados *on-the-fly*, donde los ítems se emparejan durante la prueba en

**Correspondence address [Dirección para correspondencia]**: Rodrigo S. Kreitchmann. Facultad de Psicología. Universidad Nacional de Educación a Distancia, Madrid, España.
**Email**: rschames@psi.uned.es

**ORCID**: Francisco J. Abad (https://orcid.org/0000-0001-6728-2709), Rodrigo S. Kreitchmann (https://orcid.org/0000-0001-5199-9828), Diego Graña (https://orcid.org/0009-0005-2198-5341), Pablo Nájera (https://orcid.org/0000-0001-7435-2744) y Miguel A. Sorrel (https://orcid.org/0000-0002-5234-5217).

[1] Universidad Autónoma de Madrid, España.
[2] Universidad Nacional de Educación a Distancia, España.
[3] Universidad Pontificia Comillas, España.

**Agradecimientos**: Este trabajo ha sido financiado por MICIU/AEI/10.13039/501100011033 y FEDER, UE (proyecto PID2022-137258NB-I00), por la Cátedra de Modelos y Aplicaciones Psicométricos (Instituto de Ingeniería del Conocimiento y Universidad Autónoma de Madrid) y por la ayuda PREP2022-001047, financiada por MICIU/AEI/10.13039/501100011033 y el FSE+.

Recibido: 23 de enero de 2025.
Aceptado: 23 de febrero de 2025.

función de las respuestas previas del evaluado, optimizando la precisión de las puntuaciones. Finalmente, se ofrece una guía paso a paso para la construcción de pruebas de elección forzosa, ilustrada con un ejemplo empírico y código en R de acceso abierto.

**Palabras clave:** Personalidad; Elección forzosa; Test adaptativos informatizados; *On-the-Fly*.

## Abstract

This article aims to provide a current overview of forced-choice tests and offer recommendations for their design and construction. Although these tests help overcome limitations such as social desirability and extreme response bias, common in graded response formats, they present technical challenges related to the ipsativity of the scores. This paper discusses psychometric models based on Item Response Theory (IRT), such as the Thurstonian IRT (TIRT) and the multi-unidimensional pairwise preference (MUPP) models, which improve trait estimation and enable optimal item assembly into blocks. It identifies questionnaire design factors, such as the polarity of assembled items, that can affect the quality of the obtained scores. Additionally, the benefits of IRT in developing computerized adaptive tests on-the-fly are explored, where items are paired during the test based on the examinee's previous responses, optimizing score precision. Finally, a step-by-step guide for constructing forced-choice tests is provided, illustrated with an empirical example and open-access R code.

**Keywords:** Personality; Forced-choice; Computerized Adaptive Testing; On-the-fly.

## Pruebas de elección forzosa: visión actual y recomendaciones

Aunque el origen y el debate sobre las pruebas de elección forzosa se remonta a mucho tiempo atrás (Zavala, 1965), su uso ha ido creciendo en popularidad en el campo de la evaluación psicométrica de la personalidad, especialmente en contextos laborales y educativos (Heggestad et al., 2006). Esto se debe a los problemas tradicionales de las pruebas de autoinforme que emplean un formato de respuesta de tipo Likert, tales como su menor robustez al falseamiento y la presencia de sesgos de respuesta como la deseabilidad social, aquiescencia o respuesta extrema (Cao y Drasgow, 2019; Kreitchmann et al., 2019). En términos generales, las pruebas de elección forzosa son herramientas efectivas para minimizar sesgos de respuesta y extraer inferencias precisas sobre atributos no cognitivos, como actitudes, valores, intereses vocacionales, motivaciones, competencias o estilos de aprendizaje, a partir de las preferencias relativas expresadas por los respondientes (véase Hontangas et al., 2015, para ejemplos específicos). No obstante, la construcción de una prueba de elección forzosa implica algunos desafíos técnicos que requieren el uso de modelos psicométricos avanzados en todo el proceso de desarrollo del test, desde el diseño y ensamblaje de los ítems (estímulos) en bloques, hasta la estimación del nivel de rasgo. En el presente trabajo, se recogen algunas de las herramientas y recomendaciones para el diseño óptimo de una prueba de elección forzosa, de forma éstas puedan ofrecer puntuaciones fiables y válidas.

Este trabajo complementa y amplía investigaciones previas sobre el modelado de pruebas de elección forzosa (e.g., Abad et al., 2022). Mientras que dichos estudios se han centrado en la conceptualización general y los desafíos técnicos asociados a este tipo de pruebas, el presente manuscrito adopta un enfoque más aplicado. En particular, se presenta un tutorial detallado que guía al lector en la construcción y calibración de pruebas de elección forzosa, con un énfasis especial en los test fijos (frente a los test adaptativos).

### *Las pruebas de elección forzosa y el problema de la ipsatividad*

El formato de elección forzosa se caracteriza por la presentación de bloques de uno o más enunciados, entre los que el evaluado debe indicar cuál le representa mejor, o establecer un ordenamiento, total o parcial, de estos (para una revisión completa, véase Brown y Maydeu-Olivares, 2018a; Hontangas et al., 2015, 2016). En el caso

más simple, el formato binario (PICK-PAIR), se pide a la persona que seleccione el ítem que mejor le describa de entre dos enunciados (e.g., "Me gusta innovar en lo que hago" y "Me considero una persona feliz"). Otros formatos comunes consisten en pedir a la persona que escoja el ítem que mejor le represente entre más de dos enunciados (PICK), elegir el ítem que más y el que menos le describe (MOLE, de "MOst and LEast"), u ordenar las opciones según el grado en que su descripción le representa (RANK). Respecto a la puntuación tradicional bajo la teoría clásica de los tests, en los formatos PICK y PICK-PAIR se otorga +1 a la dimensión si el enunciado escogido posee polaridad positiva (mide directamente la dimensión) o -1 si tiene polaridad negativa (mide la dimensión inversa). En el formato RANK se conceden valores que varían entre 1 y K, siendo K el número de enunciados. Por otro lado, en el sistema MOLE, se asignan las puntuaciones -1, 0 o 1, dependiendo de la selección particular y la polaridad de los ítems escogidos.

Aunque los estudios de metaanálisis han mostrado la mayor robustez de este formato al falseamiento (Cao y Drasgow, 2019; Martínez y Salgado, 2021), el problema principal de estos cuestionarios es que las puntuaciones obtenidas tendrán propiedades *ipsativas* en mayor o menor grado (Hicks, 1970). Esto quiere decir que cada puntuación de un evaluado depende en parte de sus otras puntuaciones, pues manifiesta una predominancia relativa y no absoluta de los rasgos. Por ejemplo, una persona altamente organizada y algo sociable y otra menos organizada y poco sociable podrían coincidir en sus respuestas y puntuaciones ya que ambas se perciben como más organizadas que sociables. De forma similar, una tercera persona poco sociable y nada organizada podría obtener una puntuación en sociabilidad superior a la primera persona ya que las puntuaciones reflejan preferencias relativas por los rasgos. En el caso extremo de puntuaciones totalmente ipsativas, la suma de las puntuaciones de cada individuo será un valor constante e igual para todos los evaluados, lo que imposibilita realizar comparaciones entre sujetos, como deducir que la primera persona es más organizada que la segunda y que la tercera.

La Figura 1 muestra el problema de la ipsatividad en un modelo de puntuación tradicional para el formato PICK-PAIR, con bloques de dos ítems. Se presentan las respuestas de dos evaluados en bloques que miden tres rasgos de personalidad, donde todos los ítems tienen polaridad positiva y cada bloque evalúa rasgos distintos. Suponiendo conocer los niveles de rasgo verdaderos, el primer evaluado tiene puntuaciones típicas de 1,0 en estabilidad emocional (EE), 0,5 en extroversión (EX) y 0,0 en apertura a experiencias (AP). Por lo tanto, prefiere los ítems de EE en los bloques 1 y 2, y en el bloque 3, que mide EX y AP, elige el ítem de EX. Así, su puntuación tradicional es 2 en EE, 1 en EX y 0 en AP. El segundo evaluado tiene

**Figura 1**

*Ejemplo de elección forzosa con bloques de dos ítems para dos evaluados ficticios*

**Evaluado 1**

**Nivel de Rasgo Verdadero:**

| Estabilidad Emocional (EE) | Extroversión (EX) | Apertura (AP) |
|---|---|---|
| 1,0 | 0,5 | 0,0 |

**Respuestas:**

| | |
|---|---|
| Rara vez me irrito. (EE) | Hago amigos con facilidad. (EX) ○ |
| Me gusta la innovación. (AP) ○ | Me siento cómodo conmigo mismo. (EE) |
| Sé cómo cautivar a la gente. (EX) | Disfruto escuchando ideas nuevas. (AP) ○ |

**Puntuación Ipsativa:**

| Estabilidad Emocional (EE) | Extroversión (EX) | Apertura (AP) |
|---|---|---|
| 2 | 1 | 0 |

**Evaluado 2**

**Nivel de Rasgo Verdadero:**

| Estabilidad Emocional (EE) | Extroversión (EX) | Apertura (AP) |
|---|---|---|
| -1,5 | -1,0 | -0,5 |

**Respuestas:**

| | |
|---|---|
| Rara vez me irrito. (EE) ○ | Hago amigos con facilidad. (EX) |
| Me gusta la innovación. (AP) | Me siento cómodo conmigo mismo. (EE) ○ |
| Sé cómo cautivar a la gente. (EX) ○ | Disfruto escuchando ideas nuevas. (AP) |

**Puntuación Ipsativa:**

| Estabilidad Emocional (EE) | Extroversión (EX) | Apertura (AP) |
|---|---|---|
| 0 | 1 | 2 |

puntuaciones típicas más bajas: -1,5 en EE, -1,0 en EX y -0,5 en AP. Su patrón refleja preferencia por los ítems de AP en los bloques 2 y 3, y por EX en el bloque 1, obteniendo una puntuación tradicional de 2 en AP, 1 en EX y 0 en EE. Este ejemplo evidencia que las puntuaciones ipsativas, aunque permiten medir la predominancia de rasgos en cada evaluado, no permiten comparaciones válidas entre ellos. Por ejemplo, aunque el primer evaluado tiene una puntuación típica verdadera más alta en AP, su puntuación ipsativa es inferior porque sus otros rasgos son más predominantes. En una muestra completa, las puntuaciones ipsativas en una dimensión mostrarán una covarianza negativa con las demás. Esto ocurre porque elegir más ítems de un rasgo implica seleccionar menos de los otros, generando interdependencia entre puntuaciones.

La ipsatividad plantea una serie de problemas: (a) distorsión de la dimensionalidad y estructura factorial del test (e.g., correlaciones entre dimensiones negativamente sesgadas); (b) sesgo en la validez predictiva (e.g., las correlaciones de las escalas con un criterio externo estarán sesgadas hacia cero); c) sesgos en los coeficientes de fiabilidad. Estos resultados se producen por las covarianzas negativas entre las puntuaciones de rasgo que surgen al hacer que los evaluados seleccionen una opción frente a otra de distinto rasgo. No obstante, estos problemas no son insalvables, ya que en los últimos años se han producido avances significativos tanto en el uso de modelos psicométricos que optimizan la puntuación de las pruebas como, de igual importancia, en el diseño de estas. Son numerosos los estudios que muestran que el formato de elección forzosa puede mantener propiedades psicométricas comparables a las del formato de respuesta graduada (e.g., Zhang et al., 2020) sugiriendo que, aunque estas medidas presentan limitaciones, pueden producir puntuaciones fiables y válidas. En las siguientes secciones se resume el conocimiento actual sobre ambos aspectos, haciendo, finalmente recomendaciones sobre los pasos a seguir en la construcción de pruebas de elección forzosa.

## *Modelos psicométricos para pruebas de elección forzosa*

De forma resumida, la ipsatividad ocurre cuando la elección de ítems de diferentes dimensiones da lugar a incrementos equivalentes en las puntuaciones de dichas dimensiones. Por ejemplo, si al seleccionar un ítem de extroversión incrementa la puntuación en dicho rasgo en la misma medida que elegir un ítem de responsabilidad incrementa la suya, y ocurre lo mismo para todos los ítems del cuestionario. El modelado psicométrico de respuestas de elección forzosa permite estimar parámetros de discriminación, capturando matices en la relación entre rasgos e ítems. Esto reduce la ipsatividad, ya que la relación entre la elección de un ítem y la puntuación en su dimensión no tiene por qué ser uniforme. Entre los modelos TRI más utilizados para este propósito destacan el TIRT (*Thurstone IRT*; Brown y Maydeu-Olivares, 2011) y el MUPP (*Multi-Unidimensional Pairwise-Preference*; Stark et al., 2005).

***TIRT***. El TIRT se basa en la ley de juicio comparativo de Thurstone, siendo un modelo para ítems que siguen un modelo de dominancia (i.e., la probabilidad de estar de acuerdo con un ítem sigue una relación monótona, creciente o decreciente, con el nivel de rasgo). Por ejemplo, la probabilidad de estar de acuerdo con "Soy una persona ordenada" tenderá a aumentar a medida que aumenta el nivel en el rasgo de responsabilidad. Desde este modelo, se descompone la respuesta al bloque de n ítems en términos de $n(n\text{-}1)/2$ respuestas en comparaciones binarias, modeladas a través de un modelo de TRI de ojiva normal. Para un bloque de tres ítems [*i, j, k*] cuya respuesta en formato RANK de un evaluado es $i = 1$ (mayor preferencia), $j = 2$, y $k = 3$ (menor preferencia), es posible inferir una preferencia por el ítem *i* en las comparaciones [*i, j*] e [*i, k*], y por el ítem *j* en la comparación [*j, k*]. Bajo el modelo TIRT se modelan conjuntamente estas pseudo-respuestas a las comparaciones entre pares. En la comparación [*i, j*], la probabilidad de elegir el elemento j se modelaría como:

$$P(j[i,j]|\boldsymbol{\theta}) = \Phi_N\left(\frac{(\boldsymbol{\lambda}_j' - \boldsymbol{\lambda}_i')\boldsymbol{\theta} - \gamma}{\sqrt{\psi_j^2 + \psi_i^2}}\right), \qquad [1]$$

donde $\Phi_N$ expresa la función de distribución normal acumulada, θ es un vector con los niveles de rasgo, $\lambda'_j - \lambda'_i$ es un vector que expresa la diferencia de pesos entre esos dos ítems del bloque, $\psi_j^2$ y $\psi_i^2$ reflejan las varianzas de especificidad, y γ el parámetro de umbral para la comparación [*i, j*], que se relacionaría con la dificultad para elegir el ítem j. La misma ecuación se generaliza para modelar las comparaciones [*i, k*] y [*j, k*].

En bloques de dos ítems unidimensionales, el modelo se simplifica a:

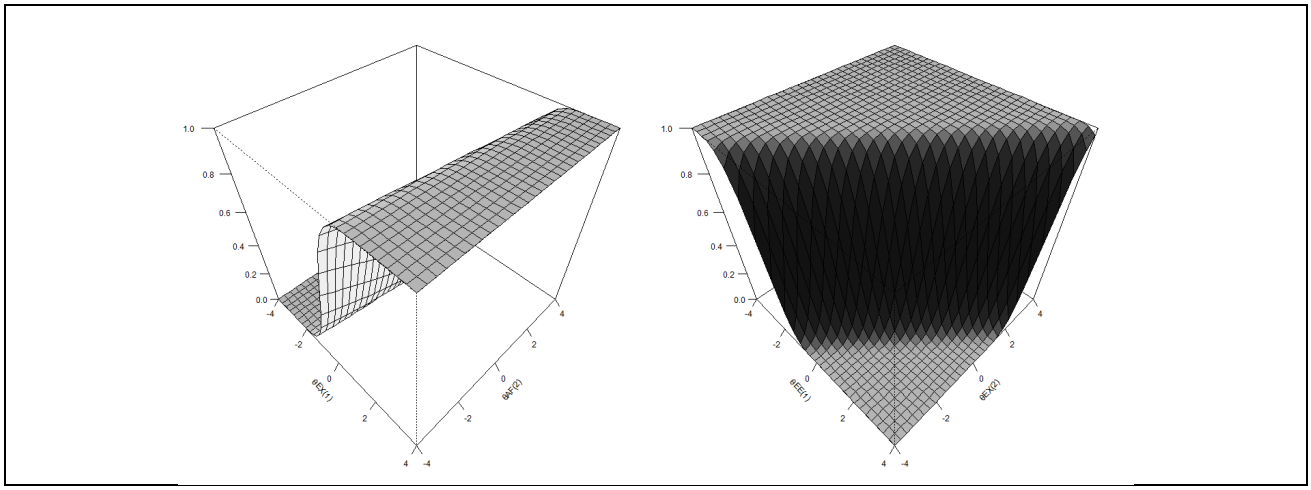
$$P(2[1,2]|\boldsymbol{\theta}) = \Phi_N\left(\frac{\lambda_{d2}\theta_{d2} - \lambda_{d1}\theta_{d1} - \gamma}{\sqrt{\psi_1^2 + \psi_2^2}}\right),$$

donde la probabilidad de elegir el ítem 2 del bloque se relaciona positiva (si el ítem 2 es directo) o negativamente (si el ítem 2 es inverso) con el nivel de rasgo medido por ese ítem ($\theta_{d2}$), pero a la vez se relaciona positiva (si el ítem 1 es inverso) o negativamente (si el ítem 1 es directo) con el nivel de rasgo medido por el ítem 1 ($\theta_{d1}$). En datos en los que los ítems no se repiten en diferentes bloques, con bloques binarios, los parámetros $\psi_1^2$ y $\psi_2^2$ no están identificados, por lo que la suma de ambos se fija a un valor. La Figura 2 muestra la representación de la superficie de respuesta de dos bloques binarios, que siguen el modelo TIRT.

En el primer bloque, se representa la probabilidad de elegir "*Presiono a las personas*" (ítem 2) frente a "*Evito las multitudes*" (ítem 1). La probabilidad está relacionada con la dimensión del primer ítem, $\theta_{EX}$ (extraversión), pero no guarda relación con la dimensión del segundo ítem, $\theta_{AF}$ (afabilidad). Como el peso del primer ítem en $\theta_{EX}$ es negativo, la relación a nivel de bloque resulta positiva: a mayor extraversión, menor es la probabilidad de elegir el ítem 1 y, por lo tanto, mayor probabilidad de elegir el ítem 2. En el segundo bloque, se representa la probabilidad de elegir "*Hago amigos fácilmente*" (ítem 2) frente a "*Mantengo la calma*" (ítem 1). En este caso, la probabilidad de elegir el segundo ítem depende de ambas dimensiones latentes, $\theta_{EX}$ y $\theta_{EE}$ (estabilidad emocional). El bloque tiene pesos similares en las dos dimensiones y, dado que ambos ítems son directos (con pesos positivos), la probabilidad de elegir el ítem 2 (indicador de extroversión) aumenta con la dimensión $\theta_{EX}$ y disminuye con $\theta_{EE}$. Además, se observa que es más fácil elegir el ítem 2 en el segundo bloque que en el primero (el umbral para elegir el ítem 2 del bloque, *γ*, es

**Figura 2**

*Representación de la superficie de respuesta en el TIRT, que representa la probabilidad de elegir el ítem 2 para un bloque con parámetros $\lambda_{EX(1)}$= -0.83, $\lambda_{AF(2)}$ = -0.09 y γ = .26 (izquierda) y para otro con parámetros $\lambda_{EE(1)}$ = 0.71, $\lambda_{EX(2)}$ = 0.61 y γ = -1.07 (derecha)*



más bajo). Esto sugiere que es más fácil afirmar la facilidad para hacer amigos (frente a mantener la calma) que afirmar que se presiona a las personas (frente a evitar las multitudes).

***MUPP***. El modelo MUPP (Stark et al., 2005) establece que la probabilidad de elegir *j*, de entre *i*, *j* o *k*, sigue el axioma de Luce:

$$P(j[i,j,k]|\boldsymbol{\theta}) = \frac{P(j)\prod_{r\neq j}^{\square} Q(r)}{\sum_{s}^{\square} P(s)\prod_{r\neq s}^{\square} Q(r)},$$

donde $P(j)$ indica la probabilidad de valorar que el enunciado del bloque le representa al evaluado y $Q(r)$ la probabilidad de valorar que el bloque *r* no le representa. La probabilidad $P(j)$ puede modelarse por diferentes modelos bajo la TRI, como son los de dominancia o de punto-ideal. En los modelos de dominancia, como el modelo logístico de dos parámetros (2PL), la probabilidad de respuesta es monotónica creciente o decreciente con el nivel de rasgo. Por otro lado, en los modelos de punto ideal la probabilidad de acuerdo con un ítem es unimodal (i. e., máxima en un nivel de rasgo no extremo). Por ejemplo, la probabilidad de estar de acuerdo con "*Soy una persona medianamente ordenada*", puede ser máxima en personas con un nivel de rasgo medio de responsabilidad, pero más bajas para niveles de rasgo más extremos (muy baja responsabilidad y nada de orden, o muy alta responsabilidad y mucho orden). Una reflexión sobre las ventajas relativas de cada modelo puede encontrarse en Brown y Maydeu-Olivares (2010). Para bloques de dos ítems, el modelo de dominancia da lugar al modelo MUPP-2PL (Morillo et al., 2016), donde el axioma de Luce puede simplificarse a:

$$P(2[1,2]|\boldsymbol{\theta}) = \psi_{logistico}\big((\mathbf{a'}_2 - \mathbf{a'}_1)\boldsymbol{\theta} + c\big), \qquad [2]$$

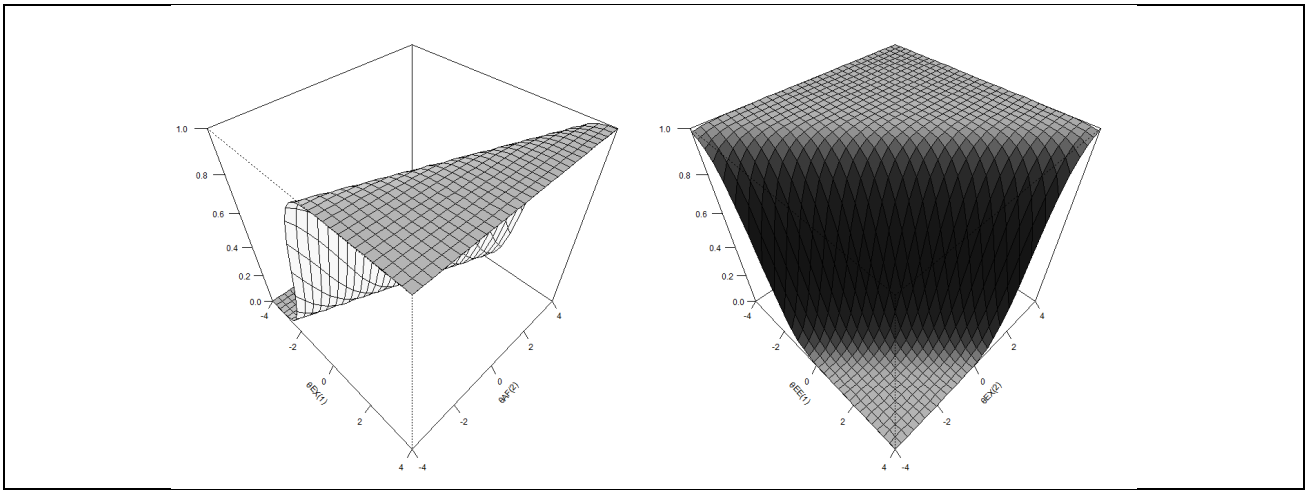
donde $\psi_{logistic}$ refleja la función logística y $\mathbf{a'}_2 - \mathbf{a'}_1$ es un vector de diferencias entre los parámetros *a* de los ítems del bloque. En el caso de que los ítems sean unidimensionales, el modelo se reduce a:

$$P(2[1,2]|\boldsymbol{\theta}) = \psi_{logistic}(\mathrm{a}_{d2}\theta_{d2} - \mathrm{a}_{d1}\theta_{d1} + c)$$

Por tanto, se establece que el *logit* de la probabilidad de elegir el ítem 2 del bloque se relaciona linealmente con los niveles de rasgo medidos por los ítems del bloque. El parámetro c determina la probabilidad de elección del ítem 2 cuando los niveles de rasgo son cero (la probabilidad sería expit(*c*); por ejemplo, si $c = 0$, la probabilidad es 0.5). La Figura 3 muestra la representación de la superficie de respuesta de dos bloques binarios según el MUPP.

**Figura 3**

*Representación de la superficie de respuesta en el MUPP-2PL que representa la probabilidad de elegir el ítem 2 para un bloque con parámetros $a_{EX(1)} = -2.81$, $a_{AF(2)} = -1.11$ y $c = -0.86$ (izquierda) y para otro con parámetros $a_{EE(1)} = 2.21$, $a_{EX(2)} = 2.25$ y $c = 4.30$ (derecha).*



En este caso, se representan las superficies de respuestas para los mismos bloques mostrados en la Figura 2, pero según la calibración del MUPP-2PL. En el primer bloque, la probabilidad de elegir el segundo ítem depende principalmente de la primera dimensión ($\theta_{EX(1)}$) y, en menor medida, de la segunda ($\theta_{AF(2)}$). Dado que ambos ítems son inversos (parámetros a negativos), la probabilidad de elegir el ítem 2 aumenta con $\theta_{EX(1)}$ y disminuye con $\theta_{AF(2)}$. En el segundo bloque, la probabilidad de elegir el segundo ítem depende en igual medida de ambas dimensiones ($\theta_{EE(1)}$ y $\theta_{EX(2)}$). Como los ítems son directos (parámetros a positivos), la probabilidad de elegir el ítem 2 aumentar aumenta con $\theta_{EX(2)}$ y disminuye con $\theta_{EE(1)}$. Además, se observa que, para niveles de rasgo iguales a cero, la probabilidad de elegir el segundo ítem es mayor en el segundo bloque que en el primero, lo que se explica por un mayor parámetro c en este bloque.

Como se observa al comparar las Figuras 2 y 3, ambos modelos, MUPP-2PL y TIRT, generan superficies de respuesta muy similares para comparaciones binarias y, en esencia, son prácticamente equivalentes (i.e., difieren únicamente en la función de enlace: *probit* o *logit* y en el procedimiento de estimación). Por lo tanto, en el caso de comparaciones binarias, la elección entre uno u otro modelo puede depender principalmente de las preferencias del investigador. Por ejemplo, una ventaja del TIRT es que sus parámetros están en una métrica fácilmente comprensible, ya que se asemejan a los pesos factoriales con los que los investigadores suelen estar familiarizados.

En escenarios más complejos, como comparaciones de más de dos elementos, las predicciones de ambos modelos pueden diferir. En el caso del TIRT, dichas comparaciones se traducen en comparaciones binarias, lo que introduce desafíos adicionales, como tratar la dependencia entre comparaciones (i.e., la preferencia por el ítem 1 frente al ítem 2 no es independiente de la preferencia por el ítem 1 frente al ítem 3) y la necesidad de imponer restricciones de igualdad en los parámetros de los ítems dentro de un mismo bloque (i.e., el peso del ítem 1 en un bloque debe ser igual en todas las comparaciones binarias que involucren dicho ítem). Por otro lado, el MUPP ofrece flexibilidad en la elección del modelo que siguen los ítems del bloque (e.g., punto ideal vs. dominancia). Sin embargo, esta flexibilidad viene acompañada de mayores desafíos en la estimación de los parámetros de los ítems (Zheng et al., 2024).

El uso de un modelo de TRI permite optimizar la estimación de los niveles de rasgo y además permite obtener medidas del error típico de estimación para cada nivel de rasgo (i. e., teniendo en cuenta que la precisión puede ser distinta en función de los bloques aplicados), así como medidas de fiabilidad marginal. Además, los modelos presentados anteriormente permiten el cálculo de los valores esperados de los parámetros de bloques a partir de los parámetros de ítems Likert. Esto ofrece la posibilidad de anticipar la capacidad informativa de los diferentes bloques posibles, con la finalidad de optimizar las pruebas de elección forzosa para una máxima precisión de cada evaluado (Kreitchmann et al., 2019, 2023).

## *Ensamblaje de bloques de elección forzosa y formato de respuesta*

Aunque los modelos psicométricos de elección forzosa permitan capturar los matices en las relaciones entre rasgos e ítems, es el diseño del cuestionario lo que garantizará que estas relaciones sean efectivamente variadas, de forma a reducir la ipsatividad en las puntuaciones estimadas. Asimismo, el diseño de los bloques afectará la robustez al falseamiento. El desarrollo de una prueba de elección forzosa optimizada requiere una serie de pasos: (a) crear un extenso banco de ítems calibrado mediante un modelo de TRI a partir de una muestra lo suficientemente amplia; (b) recopilar valoraciones de deseabilidad social de los ítems a través de un grupo de expertos; (c) utilizar un procedimiento de emparejamiento óptimo que respete las restricciones impuestas por las diferencias en deseabilidad social y, al mismo tiempo, maximice la fiabilidad; y (d) evaluar las propiedades psicométricas del test final en una muestra empírica. Con respecto al modo en que se conforman los bloques, hay ciertos aspectos a considerar:

***Restricciones de deseabilidad social***. Los ítems de los bloques deben emparejarse según su nivel de deseabilidad social. El primer paso consiste en obtener valoraciones de expertos sobre el grado de deseabilidad social de los enunciados (e.g., 1: altamente indeseable; 5: altamente desea-

ble) en un contexto específico (i.e., en función de las características de un puesto de trabajo que puedan ser más deseables). Estas medidas pueden complementarse con correlaciones entre las valoraciones promedio de deseabilidad social de los expertos y la correlación entre las respuestas reales de los ítems con escalas tradicionales de deseabilidad social. En segundo lugar, es aconsejable utilizar medidas de similaridad de la deseabilidad social de los ítems del bloque, que tengan en cuenta no sólo la equiparación del puntaje medio en deseabilidad social de los ítems, sino también el consenso entre los jueces. Pavlov et al. (2021) y Pavlov (2024) proporcionan más detalles sobre el proceso de igualación en deseabilidad social, proponiendo el uso del índice linealmente ponderado de Brennan-Prediger (BPi; e.g., Brennan y Prediger, 1981; Gwet, 2014) como medida del acuerdo entre ítems en deseabilidad social. Este índice es similar al coeficiente kappa ponderado, pero la probabilidad esperada de acuerdo por azar se calcula asumiendo una distribución uniforme. Pavlov et al. (2021) recomiendan emparejar bloques por deseabilidad social usando un BPi significativamente mayor a 0.70 (con un intervalo de confianza del 95%, ya que el valor máximo, 1, indica un acuerdo perfecto, 0 indica ausencia de acuerdo y los valores negativos indican un desacuerdo sistemático).

***Inclusión o no de bloques heteropolares (o mixtos)***. En la investigación previa, se distingue entre bloques homopolares positivos, homopolares negativos y heteropolares (mixtos). Los bloques homopolares están compuestos por ítems que miden diferentes dimensiones en la misma dirección. Por ejemplo, un bloque homopolar positivo podría estar formado por los ítems: "*Disfruto socializando con grupos grandes de personas*" (EX+) y "*Soy capaz de relajarme bajo presión*" (EE+), ambos midiendo su rasgo, extraversión y estabilidad emocional, respectivamente, de manera directa. Un bloque homopolar negativo estaría formado por ítems inversos ("*Evito participar en conversaciones grupales*" (EX-) y "*Me resulta difícil mantener la calma en situaciones estresantes*" (EE-). Este tipo de bloques suele facilitar el emparejamiento de ítems, ya que generalmente presentan una deseabilidad social similar. Los bloques heteropolares contendrían ítems que miden sus respectivas dimensiones en direcciones opuestas. Un ejemplo de bloque heteropolar, midiendo extraversión y responsabilidad (RE), sería: "*Prefiero evitar conversaciones con personas que no conozco*" (EX-) y "*Me considero alguien con mucha determinación*" (RE+). De acuerdo con Brown y Maydeu-Olivares (2011), la inclusión de bloques heteropolares es necesaria para identificar con mayor precisión la posición absoluta de una persona en el continuo de rasgos (en la Tabla 1, vimos un ejemplo de uso de bloques exclusivamente homopolares). Sin embargo, lograr una deseabilidad social equilibrada en ítems opuestos es un reto, ya que los participantes pueden identificar el ítem más deseable y sesgar sus respuestas, comprometiendo así la validez de la prueba. Llegados a este punto, parecería que nos encontramos en una encrucijada: tener que elegir entre una prueba más falseable (con bloques heteropolares) o una con puntuaciones más ipsativas (formada exclusivamente de bloques homopolares). Li et al. (2024) sugieren que es posible alcanzar un equilibrio entre la resistencia al falseamiento y las propiedades psicométricas, ya que el uso de un 20% de bloques heteropolares es suficiente para optimizar la precisión en la medición, manteniendo a la vez la resistencia al falseamiento mediante una alta coincidencia en la deseabilidad social entre los ítems del mismo bloque. Por otro lado, algunos autores han mostrado que un emparejamiento óptimo de bloques homopolares puede ser suficiente para reducir el problema de la ipsatividad, tanto mediante estudios de simulación (Kreitchmann et al., 2022; 2023), como con datos empíricos (Graña et al., 2024). En este sentido, resulta clave que se escojan ítems que miden las dimensiones en distintas magnitudes, que en la práctica puede simplificarse a emparejar ítems con mayor diferencia de pesos intra-bloque (i.e., emparejar un ítem que pesa alto en una dimensión, con otro que pesa bajo en la otra dimensión; y viceversa).

***Algoritmos de emparejamiento***. El número de bloques conformables a partir de un conjunto amplio de ítems es muy elevado. Por ejemplo, ensamblar 60 ítems en 30 bloques de 2 deriva, aproximadamente, en $2{,}92 \times 10^{40}$ cuestionarios posibles (Kreitchmann et al., 2021). Por lo tanto, se hace necesario un algoritmo para emparejarlos. Más allá de la inclusión o no de bloques heteropolares, es especialmente relevante el uso de algoritmos que optimicen el emparejamiento de bloques. Li et al. (2024) presentan un tutorial que explica cómo construir estos cuestionarios y evaluar su calidad a través de simulaciones, utilizando el paquete de R *autoFC* (Li et al., 2022).

Por su parte, Kreitchmann et al. (2021) adaptaron un algoritmo genético NHBSA (algoritmo de muestreo basado en histogramas de nodos; Tsutsui, 2006) al desafío de ensamblar ítems en bloques, proporcionando una implementación accesible a través de Shiny, con el fin de facilitar el diseño de pruebas de elección forzosa (https://psychometricmodelling.shinyapps.io/FCoptimization/). Estos procesos de ensamblaje también pueden utilizarse para generar bancos de bloques óptimos, que dispongan de un número amplio de bloques elegibles para cada nivel de rasgo, y que puedan servir de base para test adaptativos informatizados (Kreitchmann et al., 2023).

***Formato de respuesta y estructura de los bloques***. Otra decisión por tomar es el tamaño de los bloques; esto es, se pueden constituir pares, tripletas o tétradas. El uso de tripletas puede reducir la ipsatividad, pero algunos autores señalan que incrementa la carga cognitiva de las personas evaluadas al requerir más comparaciones por bloque (Sass et al., 2020); además el uso de tripletas puede dar lugar a modelos más complejos, incluso si se utiliza el TIRT, ya que este ignora las correlaciones entre pares en la estimación del nivel de rasgo.

Por otro lado, el formato de elección binaria puede dar lugar a una disminución de la fiabilidad de las puntuaciones, lo que puede esperarse como consecuencia inherente a la naturaleza dicotómica de las respuestas. Como alternativa, Brown y Maydeu-Olivares (2018b) proponen el uso de un formato de respuesta graduada, en el que los participantes expresan sus preferencias utilizando varias categorías. Este formato ya ha mostrado algunos resultados prometedores (Zhang et al., 2024), ya que combina las ventajas de los bloques de elección forzosa (mejor control de la deseabilidad social) y de las escalas tipo Likert (mayor número de categorías para diferenciar mejor las respuestas). Zhang et al. (2024) ofrecen un análisis exhaustivo de este formato, demostrando su utilidad en mejorar la precisión y la validez de las pruebas en comparación con los formatos tradicionales de elección forzosa y de Likert.

***Otros factores***. Otros factores que influyen en la ipsatividad incluyen las correlaciones entre las dimensiones evaluadas y la cantidad de dimensiones que se miden. A medida que disminuye el número de dimensiones o aumentan las correlaciones positivas entre ellas, la ipsatividad tiende a incrementarse. Esto ocurre porque una menor diferenciación entre las dimensiones reduce la capacidad de capturar de manera independiente cada rasgo. Por ejemplo, en sus estudios de simulación, Bürkner et al. (2019) concluyen que, con cinco o menos dimensiones y utilizando bloques homopolares, es difícil obtener mediciones precisas. Sin embargo, cuando se miden 30 dimensiones, los resultados mejoran significativamente, logrando una buena recuperación de los rasgos evaluados. Cabe destacar que estos autores no analizaron casos intermedios, es decir, aquellos con entre 6 y 29 dimensiones, lo que deja abierta la posibilidad de estudios adicionales para explorar el comportamiento en ese rango.

## *Calibración de los modelos*

La calibración de estos modelos puede ser más o menos compleja, en función del diseño y modelos elegidos. Si se utiliza el TIRT, con un diseño de tripletas, deben elegirse algunas restricciones para la identificación del modelo. De acuerdo, con Jansen y Schulze (2024) estas restricciones pueden ser problemáticas a la hora de recuperar los parámetros. En cuanto al software, Brown y Maydeu-Olivares (2012) proporcionan una macro en Excel (http://annabrown.name/software) que, una vez introducida la información básica sobre el diseño de bloques, genera una sintaxis de Mplus (Muthén y Muthén, 2018). Igualmente, puede utilizarse el paquete *lavaan* (Rosseel, 2012), puesto que el TIRT puede entenderse como un modelo de ecuaciones estructurales para variables categóricas (estimable, por ejemplo, con el método Mínimos Cuadrados Ponderados Robustos [WLSMV], sobre las correlaciones policóricas).

Algunos autores han encontrado problemas en estas aproximaciones, principalmente las bajas tasas de convergencia de los modelos, el coste computacional y una cierta dependencia de los valores fijados, lo que se considera problemático (ver Bürkner et al., 2019). Otra aproximación posible es el uso del algoritmo de Monte Carlo basados en cadenas de Markov (MCMC), aunque con el problema asociado del alto coste computacional. Nie et al. (2024) proporcionan una tabla de resumen de distintas aproximaciones para la calibración. Otra posibilidad con-

siste en utilizar un procedimiento en dos fases. En primer lugar, se precalibran las respuestas a los ítems Likert; estas estimaciones se integran posteriormente en la estructura del modelo, dándolos como fijos. Esa aproximación asume invarianza de los parámetros, lo que puede no ser correcto. No obstante, algunos estudios han mostrado niveles de invarianza elevados (ver Morillo et al., 2019).

### *Estimación del nivel de rasgo y de la fiabilidad de la prueba*

Para estimar los niveles de rasgos latentes se pueden emplear los métodos de máxima verosimilitud (MLE), a posteriori máximo (MAP) y a posteriori esperado (EAP). En pruebas con un alto número de dimensiones, es recomendable optar por MAP o MLE, ya que el costo computacional de EAP aumenta significativamente. Esto se debe a que el número de puntos de cuadratura necesarios crece exponencialmente con el número de dimensiones (e.g., con 25 puntos de cuadratura por dimensión [D], el total de puntos de cuadratura asciende a $25^D$).

### *Invarianza de los parámetros*

La comprobación de la invarianza en los parámetros de los ítems al ensamblarse en bloques es esencial para lograr un ensamblaje exitoso y hacer viable la creación de pruebas de elección forzosa *on-the-fly* en contextos de selección de personal. En este caso, hablamos de un proceso de ensamblaje en tiempo real, permitiendo seleccionar dinámicamente el bloque más adecuado para cada evaluado según sus respuestas anteriores. Este proceso de ensamblaje en tiempo real optimizaría tanto la precisión como la eficiencia del cuestionario.

La invarianza es importante porque asegura que las propiedades psicométricas de los bloques sean consistentes y no dependan de las combinaciones específicas de ítems en las que se presentan. Este aspecto puede analizarse mediante estrategias tradicionales de funcionamiento diferencial de los ítems, utilizando modelos de TRI. Por ejemplo, Morillo et al. (2019) emplean test de razón de verosimilitudes, comparando un modelo restringido (donde todos los parámetros son idénticos entre las versiones Likert y de elección forzosa) con modelos en los que un parámetro particular varía entre versiones. Lin y Brown (2017), por su parte, estiman los parámetros de los ítems en bloques de tres y los comparan con los obtenidos en bloques de cuatro, tras realizar una equiparación de parámetros para asegurar su comparabilidad. Estos estudios encontraron cierto grado de invarianza en bloques binarios (Morillo et al., 2019), aunque también evidencian que los parámetros pueden variar en función del contexto de los bloques (Lin y Brown, 2017). La invarianza tiende a ser más estable para los parámetros de discriminación que para los de umbral (Lin y Brown, 2017). Estos resultados sugieren que el problema puede mitigarse mediante un diseño cuidadoso de los bloques, evitando combinaciones de ítems que puedan inducir comparaciones no deseadas. Es esperable que el problema del contexto sea mayor para bloques no binarios. En cualquier caso, una solución sería implementar correcciones a través de modelos TRI que consideren las variaciones de parámetros derivadas del contexto.

## Pasos para la construcción de pruebas de elección forzosa

A partir de los aspectos relevantes definidos anteriormente, podemos plantear una serie de pasos recomendados a seguir a la hora de construir una prueba de elección forzosa.

***Construcción y calibración del banco de ítems***. Con la finalidad de poder posteriormente diseñar la prueba de elección forzosa en base a sus propiedades psicométricas, se recomienda partir de un banco de ítems administrados individualmente en formato de respuesta graduada (grado de acuerdo) sobre una muestra en la que se minimicen sesgos como la deseabilidad social. A través de valoraciones de expertos, se obtiene información acerca de la deseabilidad social de los ítems para el contexto en el que se busca utilizar la prueba, que servirá para equiparar la deseabilidad de los ítems que formarán parte de un mismo bloque.

***Definición de especificaciones de la prueba.*** Respetando las limitaciones del banco de ítems, se determina el número de bloques deseado, la representatividad de las di-

mensiones en ellos, y las restricciones de deseabilidad social. La elección del modelo depende del tipo de ítems (dominancia o punto ideal) y del formato de respuesta (pares o tríadas). Por ejemplo, el modelo TIRT admite bloques con más de dos ítems, mientras que el MUPP-2PL se limita a pares, por otro lado, el MUPP permite la utilización de ítems de punto ideal, mientras el TIRT se limita a ítems de dominancia.

***Ensamblaje de la prueba***. Para el ensamblaje óptimo de un test, cabe utilizar la información acerca de las propiedades psicométricas de los ítems en la calibración del banco para predecir las propiedades psicométricas (discriminación, umbral, información) de los bloques construidos. Así, es posible generar una prueba que maximice la fiabilidad de las puntuaciones, a la vez que cumple con restricciones de deseabilidad social y de contenido de los bloques definidas en la fase de especificación. En esta fase, cabe considerar también la posibilidad de la aplicación adaptativa de la prueba, de forma que sea posible administrar a cada evaluado aquellos bloques que permitan estimar su puntuación con la mayor precisión. El ensamblaje de los bloques en una administración adaptativa puede hacerse en directo (*on-the-fly*), considerando todas las posibles combinaciones de ítems que cumplan con criterios de equiparación en deseabilidad social.

***Administración y calibración de la prueba***. Los datos recogidos de las respuestas de elección forzosa deben utilizarse para calibrar los parámetros en este formato, buscando identificar posibles dependencias contextuales de los ítems. Por ejemplo, la discriminación o el parámetro de umbral de un ítem puede variar según con qué ítem éste se empareje.

***Cálculo de puntuaciones y evaluación de la calidad métrica.*** Con los datos de la prueba, se estiman las puntuaciones de los sujetos, evaluando su fiabilidad, así como el ajuste del modelo y la invarianza de los parámetros. En pruebas adaptativas, la invarianza es crucial para garantizar que las puntuaciones estimadas se basen en parámetros consistentes.

## Ilustración empírica

Con la finalidad de ilustrar los principales pasos en la construcción de una prueba de elección forzosa, se incluyen datos de ejemplos y códigos de R a través de https://osf.io/a5tsg/?view_only=2b1a4c6747b8437592c0050c5e5dad35. Asimismo, se incluye un tutorial comentado sobre cada apartado del código, con resultados e interpretaciones.

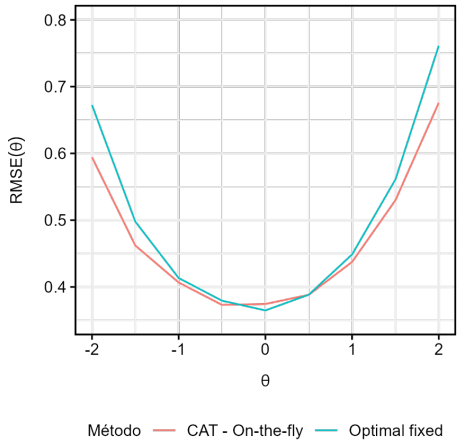
En el ejemplo, se parte de una selección de los datos del IPIP-NEO recogidos por Johnson (2014). El conjunto original de datos incluye 300 ítems del IPIP-NEO y 307313 participantes (https://osf.io/tbmh5). En nuestro caso, se seleccionaron los datos de una muestra aleatoria de 1000 participantes sin valores perdidos, recogidos en Estados Unidos y de entre 19 y 24 años. Se seleccionaron los 20 ítems que más claramente pesaban en su dimensión teórica (de acuerdo con el análisis factorial en una muestra aleatoria distinta, de 30000 participantes). Los datos de deseabilidad social de los ítems del IPIP-NEO se han tomado del trabajo de Hughes et al. (2021; https://osf.io/8gfxs/).

Así, se ha logrado un banco de 100 ítems midiendo los Cinco Grandes (EE: estabilidad emocional, EX: extraversión; AP: apertura a la experiencia; AF: afabilidad; RE: responsabilidad), con pesos factoriales con media de 0.66 y desviación típica de 0.11, ajuste próximo a lo aceptable (RMSEA = 0.062; CFI = 0.891), y alta fiabilidad de las puntuaciones en todas las dimensiones, entre 0.90 para *apertura* y 0.95 para *extroversión*.

A partir del banco, se ha construido una prueba de 40 bloques binarios utilizando el modelo MUPP y el algoritmo de ensamblaje disponible en https://psychometric-modelling.shinyapps.io/blockAssemblySD/. A través de un estudio simulación se encuentra una fiabilidad empírica esperada media-alta con el cuestionario generado (EE: 0.81; EX: 0.83; AP: 0.74; AF: 0.71; RE: 0.78). Se ilustra también la utilización de un *test* adaptativo *on-the-fly*, con el que se obtienen fiabilidades mayores que para un test de elección forzosa fijo (EE: 0.84; EX: 0.86; AP: 0.75; AF: 0.74; y RE: 0.81), y especialmente mejores estimaciones para niveles de rasgo más alejados del promedio (Figura 4).

**Figura 4**

*Promedio del RMSE (raíz del error cuadrático medio) para cada nivel de rasgo (valor promedio a través de los rasgos) y para cada tipo de test (fijo óptimo o adaptativo on-the-fly)*



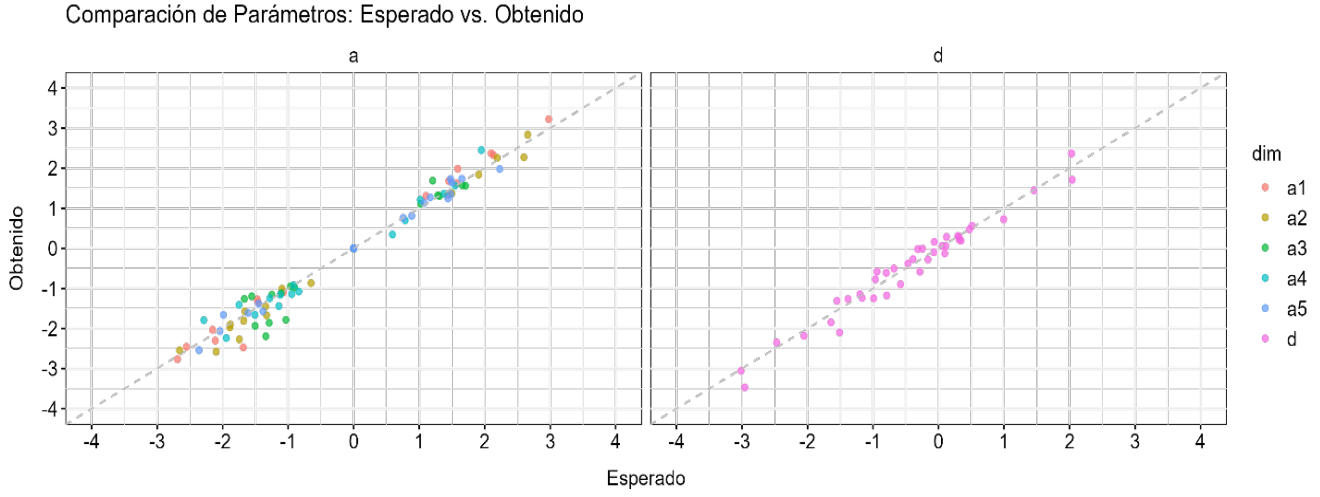

Por último, en la ilustración, con datos simulados, se observa el cumplimiento de la invarianza de los parámetros, ya que los parámetros de discriminación (a) y umbral (d) estimados con los datos de respuestas de elección forzosa se aproximan a su valor esperado a partir de la generalización de los parámetros de los ítems bajo el modelo MUPP-2PL (Figura 5).

## Discusión

Los modelos recientes de TRI para elección forzosa, como el TIRT y el MUPP, han mejorado la precisión en la estimación de los rasgos evaluados. Además, estos modelos permiten el ensamblaje óptimo de pruebas mediante

**Figura 5**

*Comparación de parámetros esperados y obtenidos en los bloques*



algoritmos de emparejamiento de ítems, en los que se controla la similitud en deseabilidad social de los ítems emparejados, al tiempo que se maximiza la fiabilidad de la prueba completa. Nuestra perspectiva es que, si bien el uso de bloques homopolares y heteropolares puede reducir considerablemente la ipsatividad de las puntuaciones, la implementación de algoritmos de optimización podría alcanzar el mismo objetivo empleando únicamente bloques homopolares, más robustos a la deseabilidad social. Además, el uso de la TRI ofrece una ventaja adicional: facilita la construcción de test adaptativos informatizados y la implementación de pruebas adaptativas *on-the-fly*, que ajustan la prueba en tiempo real y de forma dinámica para cada evaluado, maximizando la precisión de las puntuaciones. Sin embargo, el beneficio de estos sistemas avanzados depende en gran medida de la disponibilidad de un banco de ítems amplio (e.g., con variabilidad en los parámetros de los ítems) y de consideraciones sobre el costo computacional, que puede incrementarse considerablemente en bancos de gran tamaño. Paralelamente, el emparejamiento adecuado de ítems sigue siendo un aspecto crítico para su efectividad. La construcción de estas pruebas es más exigente, tanto por las consideraciones técnicas mencionadas como por el hecho de que, en formatos de bloques de opción binaria, la información por unidad tiende a ser menor, lo cual suele requerir pruebas de mayor longitud. Finalmente, el éxito del ensamblaje y la construcción de pruebas adaptativas *on-the-fly* dependerá también del supuesto de invarianza de parámetros en distintos contextos de aplicación, un aspecto que aún requiere mayor exploración en futuras investigaciones.

En este artículo, además de discutir los diferentes aspectos relevantes en la evaluación con pruebas de elección forzosa, se presentan recomendaciones de los pasos a seguir que han mostrado ser eficaces para la construcción de pruebas de elección forzosa de personalidad (Graña et al., 2024). Además, a través de la accesibilidad de código abierto en R, se busca facilitar a que profesionales aplicados puedan seguir con facilidad estas recomendaciones para así construir pruebas robustas a los sesgos de respuestas, con máxima fiabilidad y validez).

## Referencias


Abad, F. J., Kreitchmann, R. S., Sorrel, M. A., Nájera, P., García-Garzón, E., Garrido, L. E. y Jiménez, M. (2022). Construyendo test adaptativos de elección forzosa "On the Fly" para la medición de la personalidad [Building Adaptive Forced Choice Tests "On the Fly" for Personality Measurement]. *Papeles del Psicólogo, 43*(1), 29–35. https://doi.org/10.23923/pap.psicol.2982

Brennan, R. L. y Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement, 41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Brown, A. y Maydeu-Olivares, A. (2010). Issues that Should not Be Overlooked in the Dominance versus Ideal Point Controversy. *Industrial and Organizational Psychology, 3*(4), 489–493. https://doi.org/10.1111/j.1754-9434.2010.01277.x

Brown, A. y Maydeu-Olivares, A. (2011). Item Response Modeling of Forced-Choice Questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. https://doi.org/10.1177%2F0013164410375112

Brown, A. y Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT Model to Forced-Choice Data Using Mplus. *Behavior Research Methods, 44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A. y Maydeu-Olivares, A. (2018a). Modelling forced-choice response formats. En P. Irwing, T. Booth y D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 523–569). Wiley. https://doi.org/10.1002/9781118489772.ch18

Brown, A. y Maydeu-Olivares, A. (2018b). Ordinal Factor Analysis of Graded-Preference Questionnaire Data. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Bürkner, P.-C., Schulte, N. y Holling, H. (2019). On the Statistical and Practical Limitations of Thurstonian IRT Models. *Educational and Psychological Measurement, 79*(5), 827–854. https://doi.org/10.1177/0013164419832063

Cao, M. y Drasgow, F. (2019). Does Forcing Reduce Faking? A Meta-Analytic Review of Forced-Choice Personality Measures in High-Stakes Situations. *Journal of Applied Psychology, 104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Graña, D. F., Kreitchmann, R. S., Abad, F. J. y Sorrel, M. A. (2024). Equally vs. Unequally Keyed Blocks in Forced-Choice Questionnaires: Implications on Validity and Reliability. *Journal of Personality Assessment,* 1–14. https://doi.org/10.1080/00223891.2024.2420869

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Heggestad, E. D., Morrison, M., Reeve, C. L. y McCloy, R. A. (2006). Forced-choice Assessments of Personality for Selection: Evaluating Issues of Normative Assessment and Faking Resistance. *Journal of Applied Psychology, 91*(1), 9–24. https://doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some Properties of Ipsative, Normative, and Forced-Choice Normative Measures. *Psychological Bulletin, 74*(3), 167–184. https://doi.org/10.1037/h0029780

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D. y Abad, F. J. (2015). Comparing Traditional and IRT Scoring of Forced-Choice Tests. *Applied Psychological Measurement, 39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Hontangas, P. M., Leenen, I. y de la Torre, J. (2016). Traditional Scores versus IRT Estimates on Forced-Choice Tests Based on a Dominance Model. *Psicothema, 28*(1), 76–82. https://doi.org/10.7334/psicothema2015.204

Hughes, A. W., Dunlop, P. D., Holtrop, D. y Wee, S. (2021). Spotting the "ideal" Personality Response: Effects of Item Matching in Forced Choice Measures for Personnel Selection. *Journal of Personnel Psychology, 20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

Jansen, M. T. y Schulze, R. (2023). Linear Factor Analytic Thurstonian Forced-Choice Models: Current Status and Issues. *Educational and Psychological Measurement*, *84*(4), 660–690. https://doi.org/10.1177/00131644231205011

Johnson, J. A. (2014). Measuring Thirty Facets of the Five Factor Model with a 120-item public Domain Inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Kreitchmann, R. S., Abad, F. J. y Sorrel, M. A. (2022). A Genetic Algorithm for Optimal Assembly of Pairwise Forced-Choice Questionnaires. *Behavior Research Methods, 54,* 1476–1492. https://doi.org/10.3758/s13428-021-01677-4

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D. y Morillo, D. (2019). Controlling for Response Biases in Self-Report Scales: Forced-choice vs. Psychometric Modeling of Likert Items. *Frontiers in Psychology, 10,* Artículo 2309. https://doi.org/10.3389/fpsyg.2019.02309

Kreitchmann, R. S., Sorrel, M. A. y Abad, F. J. (2023). On Bank Assembly and Block Selection in Multidimensional Forced-Choice Adaptive Assessments. *Educational and Psychological Measurement, 83*(2), 294–321. https://doi.org/10.1177/00131644221087986

Li, M., Sun, T. y Zhang, B. (2022). autoFC: An R Package for Automatic Item Pairing in Forced-Choice Test Construction. *Applied Psychological Measurement, 46*(1), 70–72. https://doi.org/10.1177/01466216211051726

Li, M., Zhang, B., Li, L., Sun, T. y Brown, A. (2024). Mix-keying or Desirability-Matching in the Construction of Forced-Choice Measures? An Empirical Investigation and Practical Recommendations. *Organizational Research Methods, 0*(0). Advance Online Publication. https://doi.org/10.1177/10944281241229784

Lin, Y. y Brown, A. (2017). Influence of Context on Item Parameters in Forced-Choice Personality Assessments. *Educational and Psychological Measurement, 77*(3), 389–414. https://doi.org/10.1177%2F0013164416646162

Martínez, A. y Salgado, J. F. (2021). A Meta-Analysis of the Faking Resistance of Forced-Choice Personality Inventories. *Frontiers in Psychology, 12,* Artículo 732241. https://doi.org/10.3389/fpsyg.2021.732241

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P. y Ponsoda, V. (2019). The Journey from Likert to Forced-Choice Questionnaires: Evidence of the Invariance of Item Parameters. *Journal of Work and Organizational Psychology, 35*(2), 75–83. https://doi.org/10.5093/jwop2019a11

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J. y Ponsoda, V. (2016). A Dominance Variant under the Multi-Unidimensional Pairwise-Preference Framework: Model Formulation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 40*(7), 500–516. https://doi.org/10.1177/0146621616662226

Muthen, L. K. y Muthen, B. O. (2018). *Mplus User's Guide* (8ª ed.). Muthen & Muthen.

Nie, L., Xu, P. y Hu, D. (2024). Multidimensional IRT for Forced Choice Tests: A Literature Review. *Heliyon, 10*(5), Artículo e26884. https://doi.org/10.1016/j.heliyon.2024.e26884

Pavlov, G. (2024). An Investigation of Effects of Instruction Set on Item Desirability Matching. *Personality and Individual Differences, 216,* Artículo 112423. https://doi.org/10.1016/j.paid.2023.1124233

Pavlov, G., Shi, D., Maydeu-Olivares, A. y Fairchild, A. (2021). Item Desirability Matching in Forced-Choice Test Construction. *Personality and Individual Differences, 183,* Artículo 111114. https://doi.org/10.1016/j.paid.2021.111114

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sass, R., Frick, S., Reips, U.-D. y Wetzel, E. (2020). Taking the Test Taker's Perspective: Response Process and Test Motivation in Multidimensional Forced-Choice versus Rating Scale Instruments. *Assessment, 27*(3), 572–584. https://doi.org/10.1177/1073191118762049

Stark, S., Chernyshenko, O. S. y Drasgow, F. (2005). An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Tsutsui, S. (2006). Node Histogram vs. Edge histogram: A Comparison of Probabilistic Model-Building Genetic Algorithms in Permutation Domains. *2006 IEEE International Conference on Evolutionary Computation,* 1939–1946. https://doi.org/10.1109/CEC.2006.1688544

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*(2), 117–124. https://doi.org/10.1037/h0021567

Zhang, B., Luo, J. y Li, J. (2024). Moving Beyond Likert and Traditional Forced-Choice Scales: A Comprehensive Investigation of the Graded Forced-Choice Format. *Multivariate Behavioral Research, 59*(3), 434–460. https://doi.org/10.1080/00273171.2023.2235682

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S. y White, L. A. (2020). Though Forced, Still Valid: Psychometric Equivalence of Forced-Choice and Single-Statement Measures. *Organizational Research Methods, 23*(3), 569–590. https://doi.org/10.1177/1094428119836486

Zheng, C., Liu, J., Li, Y., Xu, P., Zhang, B., Wei, R., Zhang, W., Liu, B. y Huang, J. (2024). A 2PLM-RANK Multidimensional Forced-Choice Model and its Fast Estimation Algorithm. *Behavior Research Methods, 56,* 6363–6388. https://doi.org/10.3758/s13428-023-02315-x

# UNA REVISIÓN DE CONCEPTOS Y MÉTODOS PARA INVESTIGAR CON DATOS LONGITUDINALES

# A REVIEW OF CONCEPTS AND METHODS FOR RESEARCH WITH LONGITUDINAL DATA

EDUARDO ESTRADA[1], PABLO F. CÁNCER[2*] Y NURIA REAL-BRIOSO[1*]

Cómo referenciar este artículo/How to reference this article:

Estrada, E. Cáncer, P. F. y Real-Brioso, N. (2025). Una revisión de conceptos y métodos para investigar con datos longitudinales [A review of Concepts and Methods for Research with Longitudinal Data]. *Acción Psicológica, 22*(1), 73–86. https://doi.org/10.5944/ap.22.1.43402

## Resumen

En Psicología, entender cómo los fenómenos se influyen unos a otros y se desarrollan a lo largo del tiempo es clave para comprender sus causas, consecuencias y mecanismos subyacentes. En este artículo, presentamos una revisión de los aspectos fundamentales para investigar procesos de cambio longitudinal. Comenzamos explorando los tipos de preguntas que guían este tipo de investigaciones, diferenciando entre el interés por los resultados finales de un proceso (e.g., ¿se reducen los síntomas tras la terapia?), el propio desarrollo del fenómeno (¿cómo evolucionan las capacidades cognitivas durante la infancia?), las diferencias entre individuos (¿por qué algunas personas aprenden más rápido que otras?) y los procesos individuales de cambio (¿cómo fluctúa el afecto de un individuo a lo largo del tiempo?). Posteriormente,

**Correspondence address [Dirección para correspondencia]**: Eduardo Estrada, Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, España.
**Email**: eduardo.estrada.rs@gmail.com

**ORCID**: Eduardo Estrada (https://orcid.org/0000-0003-0899-4057), Pablo F. Cáncer (https://orcid.org/0000-0001-9279-8440) y Nuria Real-Brioso (https://orcid.org/0000-0002-3890-5062).

[1] Universidad Autónoma de Madrid, España.
[2] Universidad Pontificia Comillas, Madrid, España.

**Agradecimientos**: Trabajo financiado por la Agencia Estatal de Investigación española (PID2023-148585NB-I00/MCIN/AEI/ 10.13039/501100011033 /FEDER, UE). NRB financiada por Ayuda FPU (FPU22/03300), Ministerio de Universidades.
**Nota**: * Ambos autores han contribuido por igual.

Recibido: 18 de noviembre de 2014.
Aceptado: 23 de enero de 2025.

abordamos la naturaleza dinámica de los fenómenos y sus posibles patrones de cambio, como trayectorias de crecimiento o fluctuaciones en torno a un equilibrio. También discutimos los diseños de investigación adecuados para capturar estas dinámicas. Finalmente, repasamos los principales modelos estadísticos disponibles para estudiar el funcionamiento y desarrollo de estos procesos. Esperamos que esta revisión y las referencias a la literatura proporcionadas sean de utilidad para investigadores interesados en el estudio de procesos de cambio.

**Palabras clave:** investigación longitudinal; cambio; medidas repetidas.

## Abstract

In psychology, understanding how phenomena influence one another and develop over time is key for grasping their causes, consequences, and underlying mechanisms. In this article, we present a review of the fundamental aspects involved in investigating longitudinal processes of change. We begin by exploring the types of questions that guide such investigations, distinguishing between interest in the outcomes of a process (e.g., do symptoms decrease after therapy?), the development of the phenomenon itself (how do cognitive abilities evolve during childhood?), differences between individuals (why do some people learn faster than others?), and individual processes of change (how does an individual's affect fluctuate over time?). We then address the dynamic nature of phenomena and their potential patterns of change, such as growth trajectories or fluctuations around a stable equilibrium. We also discuss research designs appropriate for capturing these dynamics. Finally, we review the main statistical models available to study the functioning and development of these processes. We hope that this review and the references to the literature provided will be useful for researchers interested in studying processes of change.

**Keywords:** longitudinal research; Change; Repeated measures.

En numerosos ámbitos relacionados con la Psicología, la Educación, la Medicina, y otras Ciencias Sociales y de la Salud existe un gran interés por estudiar el cambio. De hecho, cuando realizamos cualquier acción sobre nuestro entorno, incluyendo una intervención psicológica, educativa o sanitaria, generalmente esperamos provocar algún cambio (o detener uno que ya se está produciendo). Por lo tanto, evaluar el cambio resulta fundamental para responder a preguntas de investigación de todo tipo. En este trabajo hacemos una breve revisión de métodos disponibles para evaluar el cambio, centrándonos en preguntas habituales en investigación en Psicología.

## Tipos de preguntas en investigación longitudinal

Por *pregunta de investigación longitudinal* nos referimos a aquella que implica algún aspecto relacionado con el cambio y que, por tanto, para ser contestada, requiere al menos dos mediciones de la misma variable en puntos temporales distintos, generalmente tomadas en los mismos participantes. A continuación, nos detenemos en tres aspectos clave relativos a las preguntas longitudinales.

### *Sobre el resultado vs. sobre el proceso*

Un ejemplo de pregunta *centrada en el resultado* sería ¿cuál de estos dos programas de entrenamiento genera un mayor aprendizaje? En cambio, una pregunta centrada en el proceso sería ¿cómo ha ido cambiando el conocimiento de los participantes a lo largo de las semanas que duró el programa? (Estrada et al., 2020).

Para responder a preguntas centradas en el resultado suelen ser suficientes pocas medidas repetidas: como mínimo una antes del inicio del programa (pre), y otra al final (post). En algunos contextos es habitual incluir también medidas de seguimiento, por ejemplo, a los 6 o 12 meses tras finalizar el programa. En cambio, para responder a preguntas sobre procesos sería necesario en este ejemplo contar con medidas repetidas durante el programa de en-

trenamiento (e.g., una medida semanal o al terminar cada sesión).

## Sobre el grupo vs. sobre el individuo

Las preguntas sobre el grupo implican examinar el cambio conjunto de un grupo de casos. Para contestarlas es necesario calcular valores muestrales—que se pueden usar como estimadores de los parámetros poblacionales correspondientes. Por ejemplo, la media en cada medida repetida (¿la media de síntomas de ansiedad es menor al final de la intervención que a la mitad?), o la varianza (¿se observa la misma variabilidad en las diferencias pre-post en el grupo tratado y el control?).

Las preguntas sobre el individuo implican examinar el cambio de un caso particular, que puede formar parte de una muestra o ser el único para el que existe información (en escenarios con $n = 1$). Si los únicos datos disponibles son dos medidas pre-post para un caso, es necesario contar con información adicional para valorar el cambio observado. Por ejemplo, la diferencia individual estandarizada (*Standardized Individual Difference*, SID) compara la diferencia $D = X_{post} - X_{pre}$ observada para el caso *i* con la desviación típica de las diferencias del grupo de referencia para ese caso, $SID_i = Di / Sd_{(D)}$. Por otro lado, el índice de cambio fiable (reliable change index, RCI), compara el cambio con el error típico de medida del instrumento empleado, $RCI_i = D_i / Err.Tip$. Con ello, se intenta decidir si el cambio es demasiado grande como para considerarse solo una fluctuación debida errores en la medición. Después, para ambos estadísticos se establecen puntos de corte basados en la distribución de las puntuaciones para decidir cuándo un valor de $SID_i$ o $RCI_i$ implican un cambio fiable (Ferrer y Pardo, 2019). En caso de trabajar con $n = 1$, tanto $Sd_{(D)}$ como *Err.Tip* se pueden obtener de fuentes de información externas, como por ejemplo el manual del test que se ha utilizado para medir, o estudios previos realizados con muestras provenientes de la misma población que el caso bajo estudio. Si se trabaja con una muestra y se quieren tomar decisiones sobre casos particulares que la componen, es posible calcular $Sd_{(D)}$ y *Err.Tip* a partir de las propias puntuaciones de la muestra.

Por otro lado, si existen varias medidas repetidas para un caso, es posible realizar otros análisis, como por ejemplo valorar: (a) si el cambio en la variable sigue una determinada tendencia (no cambia, es creciente, decreciente, lineal, curvilíneo, cíclico, etc.), (b) si existe un nivel estable al que el caso tiende (i.e., puntos de equilibrio), y cuál es dicho nivel para cada caso, (c) si, en general, el caso muestra mayor o menor discrepancia en torno a su nivel esperado en cada momento (i.e., más o menos variabilidad intra-individual, ver siguiente sección). La aparición de dispositivos que permiten registrar comportamientos y estados psicológicos muchas veces al día de forma poco intrusiva ha facilitado la recolección de este tipo de datos, en los que existen muchas medidas repetidas para cada persona (ver sección sobre trayectorias estables, más adelante).

## Sobre diferencias intra-indivividuales vs. inter-individuales

Al elegir un modelo estadístico para caracterizar el cambio, suele ser interesante incluir parámetros que permitan capturar uno o ambos tipos de diferencias. Matemáticamente, tanto las diferencias entre sujetos como los cambios en cada sujeto a lo largo del tiempo se estiman mediante varianzas.

Las *diferencias interindividuales* son aquellas que se observan entre distintos casos, e.g.: ¿distintos estudiantes muestran distintas velocidades de aprendizaje?, ¿existen diferencias estables en el estado de ánimo de los participantes (i.e., hay un componente de rasgo en las puntuaciones observadas)? ¿cómo de homogéneo era el nivel de comprensión lectora de los participantes al iniciar el estudio? Por el contrario, las *diferencias intraindividuales* son aquellas que se observan entre las distintas medidas repetidas de los mismos casos. Por ejemplo, si evaluamos el estado afectivo de una persona durante dos semanas, podríamos preguntarnos si su afecto negativo es relativamente estable, o por el contrario se observan grandes altibajos (o bien qué participantes son más inestables).

Existen modelos estadísticos, como por ejemplo el *Random-Intercept Cross-Lagged Panel Model* (RI-

CLPM, Mulder y Hamaker, 2021) que están diseñados específicamente para identificar qué parte de la variabilidad observada en las medidas repetidas de una muestra se debe a varianza interindividual (diferencias estables entre casos), y qué parte se debe a varianza intraindividual (fluctuaciones de los casos en torno a su media).

Esta distinción es muy importante no solo desde el punto de vista de las varianzas, sino también respecto a las *covarianzas* (o correlaciones) entre variables. Numerosos trabajos han mostrado que dos variables que estén correlacionadas cuando ambas se han medido una vez en distintos casos no necesariamente estarán correlacionadas si se miden repetidamente en un mismo caso. Esta es una implicación muy importante derivada de que los procesos psicológicos y sociales *no suelen ser ergódicos* (Hunter et al., 2024).

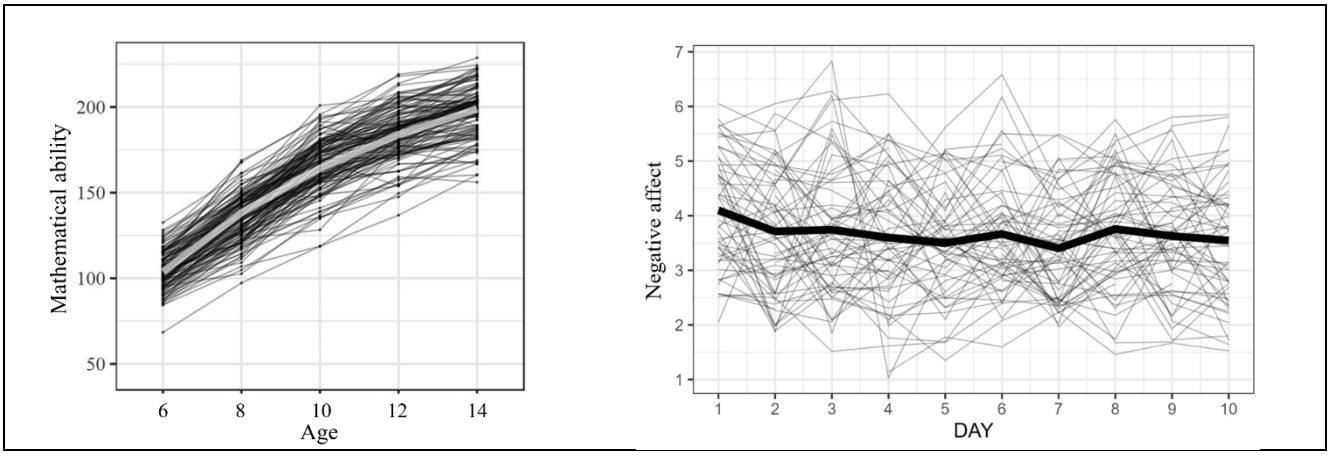
## Tipos de trayectorias longitudinales y diseños de investigación relacionados

Las características del fenómeno bajo estudio determinan qué preguntas tiene sentido plantear, cómo se debe diseñar la recogida de datos, y qué herramientas estadísticas permiten contestar las preguntas. Se pueden distinguir dos grandes tipos de fenómenos longitudinales: los que siguen *trayectorias de desarrollo*, en las que existe un crecimiento o decrecimiento del fenómeno de interés, y los que siguen *trayectorias estables* en las que el nivel esperado no cambia a lo largo del tiempo, aunque se observan fluctuaciones en torno a ese nivel. La Figura 1 muestra un ejemplo de trayectorias de cada tipo.

Ejemplos típicos de *trayectorias de desarrollo* son el incremento de la capacidad lectora durante la infancia y adolescencia, o el declive de la memoria durante la vejez. Puesto que los procesos de desarrollo psicológico suelen ocurrir durante periodos largos, es habitual que sean necesarios varios años para estudiarlos. Por ejemplo, obtener las trayectorias de capacidad lectora mostradas en el panel izquierdo de la Figura 1 requiere seguir a los mismos participantes (i.e., *la misma cohorte*) durante al menos 7 años. Este tipo de estudios se suele denominar «de cohorte». También es posible comparar el cambio entre varias cohortes distintas (e.g., provenientes de distintos países o distintos centros de investigación, o de un mismo país, pero nacidas en generaciones distintas). Debido a la dificultad que entraña mantener un estudio durante tantos años, se han propuesto algunas interesantes alternativas como los diseños longitudinales acelerados, o diseños de secuencias de cohorte (*accelerated longitudinal designs, cohort-sequential designs*; Bell, 1953; Estrada y Ferrer,

**Figura 1**

*Ejemplo de trayectorias de desarrollo (izquierda) y estables (derecha) para varios casos de una muestra. La línea gruesa indica la media grupal en cada medida repetida*



2019). En estos diseños se incluyen cohortes de distintas edades, y cada una se sigue durante una fracción del rango de edad de interés. Después, se agrega la información longitudinal y transversal.

En contraste, en las *trayectorias estables* no se esperan grandes cambios en el nivel medio a lo largo del periodo analizado. Aquí, el interés suele estar en las fluctuaciones mostradas por los casos en torno a dicho nivel esperado en cada momento. Algunos ejemplos son: el afecto negativo observado en una o más personas a lo largo de varios días (panel derecho de la Figura 1), la velocidad de procesamiento a lo largo de varios ensayos en una tarea de laboratorio, o el dolor percibido por pacientes crónicos a lo largo de varios días. En estos casos, la duración del estudio es mucho menor (e.g., siete días), y se toman medidas con una frecuencia muy superior (e.g., uno o más al día). Los avances tecnológicos permiten recoger numerosas medidas repetidas cada día mediante los dispositivos móviles de los participantes (e.g., Mestdagh et al., 2023). Estas investigaciones se realizan mediante métodos de *muestreo de experiencias* (*Experience Sampling Methods*, *ESM*; Fritz et al., 2024), que suelen incluir cuestionarios sencillos que los participantes responden varias veces al día. El término ESM se suele utilizar cuando se pregunta a los participantes sobre sus sensaciones internas en ese momento (e.g., emociones, afecto, activación, dolor, etc.).

Existen otros términos relacionados con *ESM*, aunque no exactamente equivalentes. Cuando el interés está en variables de tipo biomédico (e.g., actividad física, tasa cardíaca, conductancia de la piel, etc.), se le suele llamar *evaluación ambulatoria* (*ambulatory assessment*). En contraste, cuando el énfasis está en otros datos relativos al medio en el que se encuentra la persona (e.g., geolocalización, climatología, o eventos relevantes que puedan desencadenar una conducta de interés), se le suele denominar *evaluación ecológica momentánea* (*Ecological Momentary Assessment*). Los datos obtenidos mediante estas técnicas se suelen denominar *datos longitudinales intensivos* (*Intensive Longitudinal Data*) para hacer referencia a la gran cantidad de medidas repetidas que proporciona cada caso, ya exista información para uno o más casos.

No obstante, también existen variables con trayectorias estables cuyo estudio se prolonga a lo largo de varios años, y que no fluctúan tanto como para requerir más de una o dos medidas por año. Algunos ejemplos pueden ser los rasgos de personalidad, la autoeficacia o la satisfacción con el trabajo. En estos estudios (que se suelen llamar «de panel») también se sigue a un mismo grupo de personas durante periodos relativamente largos de tiempo. Una diferencia entre estos datos y los intensivos es que los datos de panel suelen tener muchos más participantes (e.g., 100) y muchas menos medidas repetidas (e.g., cinco).

Otro aspecto importante a tener en cuenta es que ciertos fenómenos presentan *componentes cíclicos* en sus trayectorias. Esto significa que el cambio se produce de manera periódica. Es decir, se repite a intervalos regulares de tiempo (Ernst et al., 2024). Algunos ejemplos pueden ser el nivel de activación o energía a lo largo de 24h (muchas personas se sienten más activas por la mañana que por la tarde), o la motivación de los estudiantes a lo largo del curso académico (mayor al inicio del semestre, con un progresivo descenso después).

## Una clasificación de modelos longitudinales

Probablemente la primera distinción que debe valorarse a la hora de elegir un modelo es si, para abordar la pregunta de interés, es necesario un modelo longitudinal dinámico o estático.

### *Modelos estáticos vs. dinámicos*

Un *modelo longitudinal estático* expresa el estado de una o más variables en un momento concreto, y generalmente lo pone en función del tiempo. Un ejemplo sencillo es el llamado modelo de curva de crecimiento (*growth curve model*):

$$Y_{i,t} = \beta_0 + \beta_1 \cdot t + \epsilon_{i,t}, \qquad [1]$$

donde $Y_{i,t}$ es el nivel en la variable *Y* del individuo *i* observado en el momento *t*. Por ejemplo, *Y* podría ser el nivel de *comprensión lectora* mostrado por los participantes, mientras que t podría ser la edad, *medida en meses*.

Dicho nivel se expresa como una función lineal del tiempo en la que $\beta_0$ y son $\beta_1$ la intersección y la pendiente, respectivamente, mientras que $\varepsilon_{ti,t}$ es el error de predicción para ese caso en ese momento. A menudo, se permite que la intersección y la pendiente sean variables aleatorias, por lo que se añade un subíndice i a estos dos términos para expresar diferencias entre individuos en comprensión lectora cuando $t = 0$ meses (intersección $\beta_0$) y en la tasa de cambio por cada incremento de tiempo de una unidad, que en este ejemplo sería un mes (pendiente $\beta_1$). En este modelo, la edad, *t*, sirve como predictor de la variable comprensión lectora, Y. Si se conoce la edad en meses, se puede conocer el estado del sistema (es decir, el nivel de la variable dependiente). Esto es diferente en un modelo dinámico, en el que el punto temporal *t* puede ser necesario, pero no suficiente, para determinar el estado del sistema (Voelkle et al., 2018).

Paradójicamente, aunque este modelo se denomine curva de crecimiento, su formulación habitual solo permite describir el cambio como una función lineal. Es decir, no permite caracterizar trayectorias curvilíneas a lo largo del tiempo. Una forma de hacerlo es añadir un término cuadrático a la ecuación y convertirla en un polinomio:

$$Y_{i,t} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \epsilon_{i,t}, \quad [2]$$

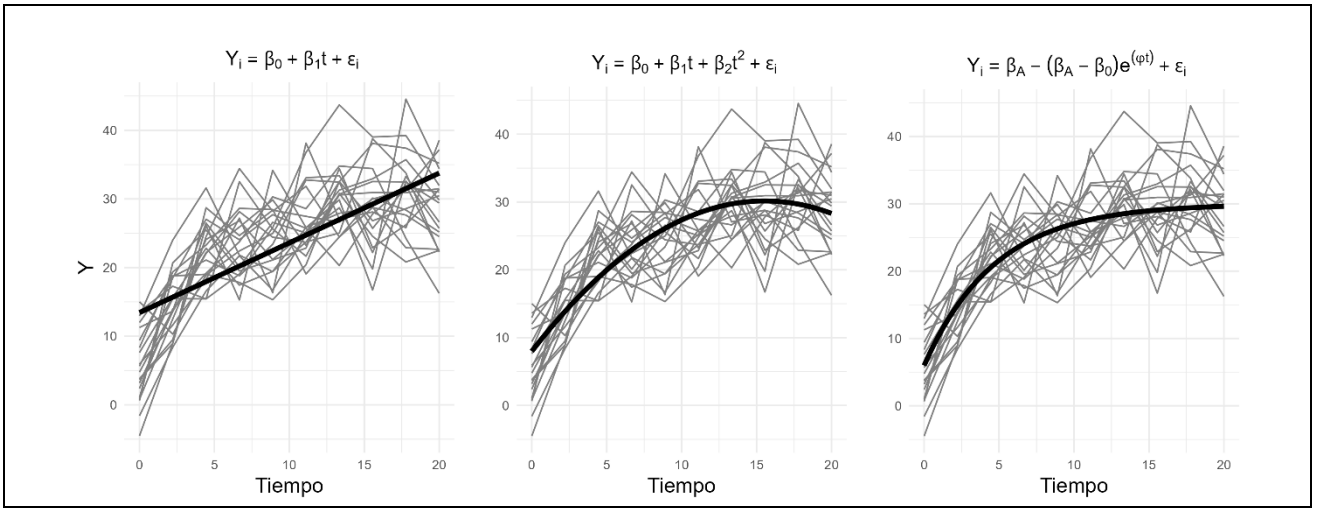
Otra posibilidad es elegir otra función para expresar la relación entre la comprensión lectora, Y, y la edad, t, como por ejemplo una función exponencial del tipo:

$$Y_{i,t} = \beta_A - (\beta_A - \beta_0) \cdot e^{\varphi \cdot t} + \epsilon_{i,t}, \quad [3]$$

En la expresión anterior, $\beta_0$ es el nivel de Y cuando $t = 0$, mientras que $\beta_A$ es la asíntota en *Y*, o el nivel hacia el cual la trayectoria tiende (o del cual se aleja) a medida que avanza el tiempo. La letra *e* representa la constante de Euler, y el coeficiente $\varphi$ representa la tasa a la que la diferencia entre $\beta_0$ y $\beta_A$ se reduce (o amplía, dependiendo del signo de $\varphi$) por cada cambio de una unidad en el tiempo *t* (para más detalles, ver Cáncer et al., 2021). Ambas ecuaciones, cuadrática y exponencial, tienen tres parámetros cada una. Sin embargo, si el objetivo es simplemente caracterizar una trayectoria curvilínea de crecimiento o decrecimiento, creemos preferible utilizar una ecuación exponencial ya que sus parámetros se pueden interpretar con mayor facilidad en términos del fenómeno bajo estudio (ver Figura 2).

**Figura 2**

*Datos de desarrollo modelados mediante una trayectoria lineal (izda.), cuadrática (centro) y exponencial (dcha.)*



En cambio, un modelo longitudinal dinámico es aquel que pone el cambio observado en las variables, para un intervalo de tiempo determinado, en función del nivel alcanzado en las propias variables. Dicho de otra forma, los cambios en el sistema son una función del pasado de dicho sistema (Voelkle et al., 2018). Un ejemplo sencillo es un modelo autorregresivo de orden 1 (AR-1):

$$Y_t = \phi \cdot Y_{t-1} + \epsilon_t, \qquad [4]$$

donde $Y_t$ es el nivel de la variable *Y* en el tiempo *t*. Por ejemplo, *Y* podría ser el nivel de ansiedad manifestado por el participante, mientras que t podía ser el día en el que se preguntó, donde $t = 0$ corresponde a la primera medida, recogida el mismo día para todos los casos. $\varphi$ es el *coeficiente autorregresivo* que pone en relación la ansiedad un día, $Y_t$, con el nivel de ansiedad del día anterior (*t-1*), y $\varepsilon t$ es el error de predicción en el momento *t*. Nótese que en la ecuación anterior se define el nivel, no el cambio, en la variable *Y* para el momento *t*. Sin embargo, dicho nivel puede expresarse como el cambio ocurrido desde *t-1* hasta *t* (este cambio se expresa como ΔYt), y la ecuación anterior puede expresarse como:

$$\Delta Y_t = (\phi - 1) \cdot Y_{t-1} + \epsilon_t.$$

Un modelo longitudinal dinámico puede combinar componentes estáticos y dinámicos para describir el cambio. Estos modelos tienen varias características interesantes. La primera es que no es necesario tener una idea clara sobre cuál es la relación funcional entre el tiempo y la variable de interés. Puesto que el modelo especifica el mecanismo del cambio, y no necesariamente la forma de este, una misma ecuación o modelo puede describir trayectorias crecientes, decrecientes, estables, con cambio acelerado o decelerado, como resultado de los distintos conjuntos de valores que tomen sus parámetros (e.g., Cáncer et al., 2021). Otra característica interesante es que, si el modelo describe el cambio en más de una variable a la vez, se pueden incluir e interpretar no solo parámetros autorregresivos, sino también efectos dinámicos *cruzados* que capturan el efecto del nivel de cada variable sobre el cambio en la otra. Un ejemplo sencillo de esto es un modelo de vector autorregresivo de orden 1 (VAR-1), como el que se muestra en la siguiente ecuación matricial:

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \cdot \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{Y,t} \\ \epsilon_{X,t} \end{bmatrix}, \qquad [5]$$

que se puede reexpresar como las siguientes dos ecuaciones lineales:

$$\begin{aligned} Y_t &= \phi_{11} \cdot Y_{t-1} + \phi_{12} \cdot X_{t-1} + \epsilon_{Y,t}, \\ X_t &= \phi_{22} \cdot X_{t-1} + \phi_{21} \cdot Y_{t-1} + \epsilon_{X,t}, \end{aligned} \qquad [6]$$

Este modelo VAR-1 constituye la extensión a dos variables, *X* e *Y*, del modelo AR-1 presentado anteriormente. Se llama VAR a cualquier modelo de este tipo que incluya 2 o más variables. Estos modelos están estrechamente emparentados con los modelos de redes psicométricas (Borsboom et al., 2021; Epskamp, 2020).

En la Ecuación 6, *Y* podría representar la *ansiedad*, mientras que *X* podría representar el cansancio. Ahora no existe un solo parámetro autorregresivo, sino cuatro. En el tiempo *t* (un determinado día), el nivel en cada variable es, en parte, función del nivel en ella misma en *t-1*. Estos efectos autorregresivos o auto efectos están capturados por los coeficientes $\varphi_{11}$ (para *Y* o ansiedad), y $\varphi_{22}$ (para *X* o cansancio). Pero, además, el nivel de cada variable también se pone en función del nivel previo de la otra variable, mediante los parámetros $\varphi_{12}$ (efecto de *X* sobre *Y*) y $\varphi_{21}$ (efecto de *Y* sobre *X*). Estos son los llamados pesos o efectos *cruzados* (*cross-loadings*). Su inclusión permite describir cómo es la relación dinámica entre las dos variables: ¿ambas se influyen mutuamente? ¿hay un peso cruzado que sea claramente mayor que el otro? ¿los dos procesos siguen dinámicas independientes?

## *Modelos en tiempo discreto vs. en tiempo continuo*

Un *modelo en tiempo discreto* (DT) es aquel que contempla el paso del tiempo en incrementos que tienen siempre la misma amplitud. Por ejemplo, los modelos dinámicos descritos en el apartado anterior están basados en diferencias de tiempo constantes entre una medición y la siguiente. Es decir, independientemente de la unidad de medida usada para el tiempo (segundos, días, semanas o años), y del tiempo en el que se haga una determinada observación ($t = 2$, $t = 15$ o $t = 100$), la ecuación pone en re-

lación el nivel en ese momento con el nivel en el momento inmediatamente anterior (*t-1* = 1, *t-1* = 14 o *t-1* = 99). Por tanto, se asume que los intervalos entre observaciones son iguales para todos los casos y todas las observaciones consecutivas de cada caso. Es importante destacar que también existen modelos dinámicos que no están especificados en tiempo discreto (ver Tabla 1).

En cambio, un *modelo en tiempo continuo* (CT) trata el tiempo como una variable continua. En consecuencia, no asume que el intervalo temporal entre dos observaciones será siempre de la misma amplitud; ni para distintos casos ni para distintas medidas repetidas de un mismo caso. En la Tabla 1, basada en la Figura 1 de Voelkle et al. (2018), mostramos una clasificación de algunos marcos de modelado estadístico habituales en función de cómo tratan el tiempo y de si son estáticos y dinámicos. En la siguiente sección explicaremos brevemente los marcos de modelado más habituales, de los mencionados en la Tabla 1.

## Marcos de modelado estadístico del cambio

### *Modelos lineales mixtos: curvas de crecimiento*

Estos modelos constituyen una extensión de la ecuación de regresión lineal «clásica». El modelo de curva de crecimiento especificado en la Ecuación 1 es un ejemplo básico de estos modelos. El nivel de la variable *Y* se expresa como una función lineal del tiempo *t*. Dos aspectos interesantes de estos modelos para analizar medidas repetidas son que permiten: (a) tener en cuenta que las medidas repetidas de un mismo caso no son independientes, y (b) capturar diferencias entre casos tanto en el nivel de la variable dependiente *Y* cuando $t = 0$ (intersección $\beta_0$) como en la tasa de cambio en *Y* por cada incremento de una unidad en el tiempo *t* (pendiente $\beta_1$). Esto permite que cada individuo i pueda tener su propia intersección ($\beta_{0i}$) y pendiente ($\beta_{1i}$). Es decir, la Ecuación 1 que relaciona capacidad lectora (*Y*) con edad (*t*) se puede extender a la siguiente expresión:

**Tabla 1**

*Una clasificación de modelos longitudinales (adaptada y ampliada de Voelkle et al, 2018)*

| | **Tiempo Discreto** | **Tiempo Continuo** |
|---|---|---|
| Modelos estáticos | – *Modelos de ecuaciones estructurales* (SEM): por ejemplo, curva de crecimiento latente, curva de crecimiento con bases latentes (Grimm et al., 2017). | – *Modelos lineales mixtos*: esto es, regresión multinivel con la variable de interés como dependiente y el tiempo como predictor: curva de crecimiento (Hoffman, 2015). |
| Modelos dinámicos | – *(V)AR*: Modelos autorregresivos, tanto univariados como multivariados o "de vector" (Ernst et al., 2024).<br>– *Modelos de redes longitudinales* (Borsboom et al., 2021; Epskamp, 2020)<br>– *Modelos de ecuaciones estructurales (SEM) con parámetros autorregresivos*: por ejemplo, modelos de cambio latente (LCS-DT), RI-CLPM (Usami et al., 2019).<br>– *Modelos de estado-espacio en DT* (Hunter, 2018)<br>– *Dynamic SEM* (McNeish y Hamaker, 2020) | – *Modelos de ecuaciones diferenciales* (Mongin et al., 2024).<br>– *Modelos de estado-espacio en CT*.<br>– *Modelos de ecuaciones estructurales en CT* (ctSEM, Driver y Voelkle, 2018). |

*Nota*. **DT** (Tiempo Discreto); **CT** (Tiempo continuo)

$$Y_{i,t} = \beta_{0i} + \beta_{1i} \cdot t + \epsilon_{i,t}, \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \sim N\left(medias = \begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, varianzas = \begin{bmatrix}\sigma_0^2 & \sigma_{0,1}\\ \sigma_{0,1} & \sigma_1^2\end{bmatrix}\right). \quad [7]$$

Ahora la intersección y pendiente tienen un subíndice *i*, que refleja el hecho de que cada caso puede tener su propio nivel inicial y pendiente. Por lo tanto, no solo estimamos un valor para ellas, sino que para cada una asumimos una distribución de valores (habitualmente normal) con sus respectivas medias ($\mu_0$ y $\mu_1$) y varianzas ($\sigma_0^2$ y $\sigma_1^2$), además de la covarianza entre ellas ($\sigma_{0,1}$). Las dos medias se denominan «efectos fijos», mientras que las dos varianzas se denominan «efectos aleatorios». Estos últimos permiten cuantificar las diferencias interindividuales en las intersecciones y pendientes: cuanto más cercanas sean a cero, más homogéneos serán los casos en sus niveles iniciales y en sus pendientes, respectivamente. Estos modelos se llaman *mixtos* precisamente porque incluyen tanto efectos fijos como efectos aleatorios (Figura 3).

## *Modelos longitudinales de ecuaciones estructurales (Structural Equation Modeling, SEM*

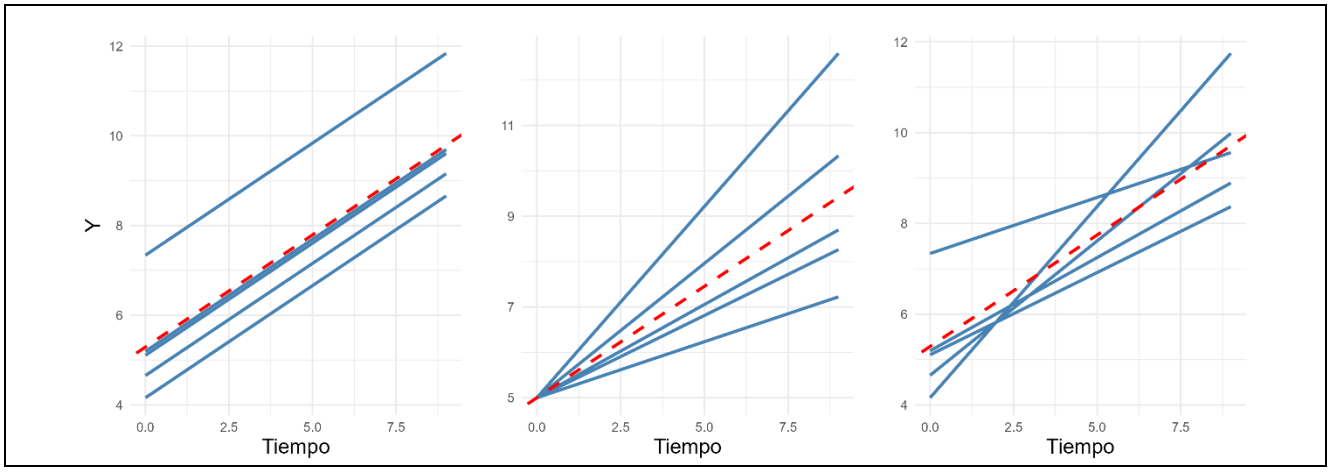
Estos modelos son enormemente versátiles y se usan tanto en investigación longitudinal como de otros tipos. Para analizar el cambio longitudinal es necesario que las medidas repetidas de los mismos casos estén recogidas en distintas columnas en la base de datos, una para cada momento (i.e., datos en «formato ancho»). Esto implica que se trata el tiempo de forma discreta.

En un modelo SEM longitudinal, es habitual incluir *variables observadas* (columnas en nuestra base de datos) y *variables latentes*. Estas últimas son variables no observadas que están vinculadas a las observadas mediante ecuaciones lineales que especifican el peso de las latentes sobre las observadas. El conjunto de parámetros que determina la relación entre las variables latente y observadas se llama «modelo de medida».

Es muy frecuente representar un SEM mediante una figura con símbolos estandarizados que se denomina *diagrama de rutas* (*path diagram*). Las variables latentes están representadas por círculos. Las variables observadas

**Figura 3**

*Curvas de crecimiento con intersecciones aleatorias (izda.), pendientes aleatorias (centro) e intersecciones y pendientes aleatorias (dcha.). La línea roja discontinua muestra la trayectoria media*

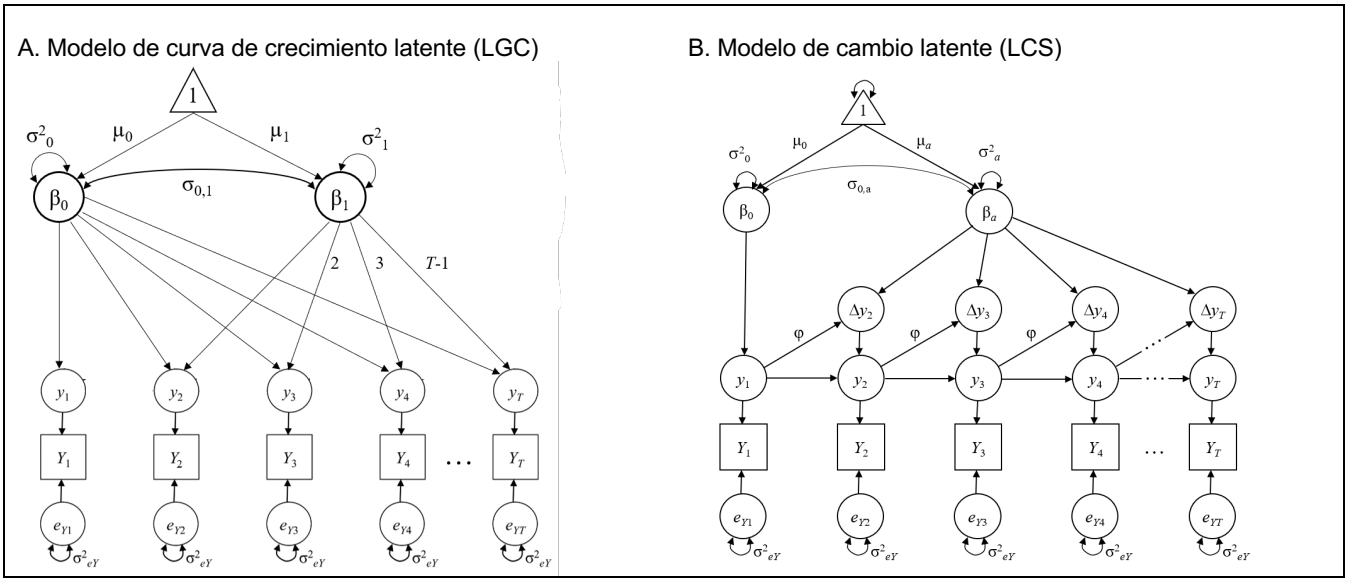


por rectángulos. Las flechas con una sola punta representan pesos de regresión (la punta señala la variable dependiente), mientras que las flechas de dos puntas representan varianzas (si van de una variable a sí misma) o covarianzas (si conectan dos variables distintas). Un triángulo representa una constante con valor 1. Se estimará una media o intersección propia (es decir, un «efecto fijo») para cualquier variable que reciba un peso desde esta constante. Los parámetros a estimar se representan mediante flechas que van acompañadas de letras, generalmente griegas. Cuando una flecha va acompañada de un número, quiere decir que el valor de ese parámetro se fija a ese número. Las flechas que no tienen letras ni números se fijan al valor 1.

El panel A de la Figura 4 representa un diagrama de rutas de un modelo SEM que es muy similar a la *curva de crecimiento* de la Ecuación 7, pero en este caso se trata de una *curva de crecimiento latente* (*Latent Growth Curve Model*, LGC). El término «latente» hace referencia a que, para cada edad *t*, se considera que las puntuaciones observadas de capacidad lectora, $Y_t$, son la suma de dos componentes. El primero son las puntuaciones verdaderas de capacidad lectora, que constituyen una variable no observada, $y_t$. El segundo son los errores de medida, $e_{Yt}$. En este modelo, se asume que esos errores tienen media cero, varianza $\sigma^2_{eY}$, y no correlacionan con ninguna otra variable. A menudo se asume que la varianza de los errores de medida es invariante para las distintas medidas repetidas. Esto implica asumir que la fiabilidad del instrumento no cambia con el tiempo, pero no implica asumir que un determinado caso tenga siempre un error de medida de la misma magnitud o signo. Esta partición, en cada medida repetida, de la varianza observada en dos partes (verdadera y error) es el «modelo de medida». Esta estructura se puede extender para, por ejemplo, permitir que la variable latente sea medida por varios indicadores observados en cada punto temporal, cada uno de ellos con un peso factorial que puede ser distinto. Conceptualmente, esto es equivalente a aplicar un modelo factorial confirmatorio en cada medida repetida. En este escenario, es necesario evaluar si la relación entre los indicadores observados y la variable latente es constante a lo largo del tiempo para que las puntuaciones latentes se puedan comparar a través de las distintas medidas repetidas. A esto se llama invarianza métrica *longitudinal o invarianza factorial longitudinal* (Widaman et al., 2010).

**Figura 4**

*Ejemplos de modelos SEM longitudinales univariados*



Sobre las variables latentes, que representan el nivel verdadero en cada medida repetida, se especifica un modelo de cambio, que puede ser tanto estático como d inámico. En el panel A de la Figura 4 se representa una curva de crecimiento (*modelo estático*) en la que el nivel latente en el momento *t* se debe a un nivel inicial o intersección latente y una pendiente, también latente. Igual que en la Ecuación 7, estas dos variables tienen sus respectivas medias (efectos fijos, $\mu_0$ y $\mu_1$) y varianzas (efectos aleatorios, $\sigma_0^2$ y $\sigma_1^2$), y pueden covariar entre ellas ($\sigma_{0,1}$). De hecho, si la varianza error se fija a cero y las ocasiones de medida son discretas y uniformemente espaciadas, ambos modelos son matemáticamente equivalentes y proporcionan los mismos resultados. Para aprender más sobre curvas de crecimiento estimadas en un marco SEM, posibles extensiones, y sus relaciones con modelos lineales mixtos, puede consultarse Grimm et al. (2017).

El panel B de la Figura 4 representa otro modelo SEM longitudinal, en este caso *dinámico*, denominado *modelo de puntuaciones de cambio latente*, o *modelo de cambio latente* (*Latent Change Score Model*, LCS). Una diferencia importante con el LGCM es que, para cada medida repetida a partir de la segunda, se crea una variable latente que recoge cualquier diferencia entre esa medida y la anterior: estos son los llamados «cambios latentes»: $\Delta y_t = y_t - y_{t-1}$. El mecanismo o estructura que caracteriza el cambio se especifica sobre los cambios latentes, y no los niveles latentes. El modelo permite varias especificaciones; la que representamos aquí es el llamado LCS «dual». Para cada punto temporal, el cambio se debe a dos influencias distintas: por un lado, el efecto del nivel alcanzado en la medida anterior, llamado *auto-efecto,* $\varphi$, y por otro lado una magnitud constante que se añade en cada medida repetida, llamado *componente aditivo*, que tiene una media ($\mu_a$) y una varianza que recoge diferencias entre casos, ($\sigma_a^2$). Este componente a veces se llama «pendiente latente», pero no es un nombre preciso ya que el efecto del componente aditivo en *t* se va propagando a los siguientes puntos temporales a través del auto-efecto. Esta acumulación da lugar a una trayectoria exponencial similar a la descrita en la Ecuación 3 (Cáncer et al., 2021). La inclusión de un auto-efecto que pone en relación el cambio con el nivel del tiempo anterior es lo que convierte al LCS es un *modelo dinámico*.

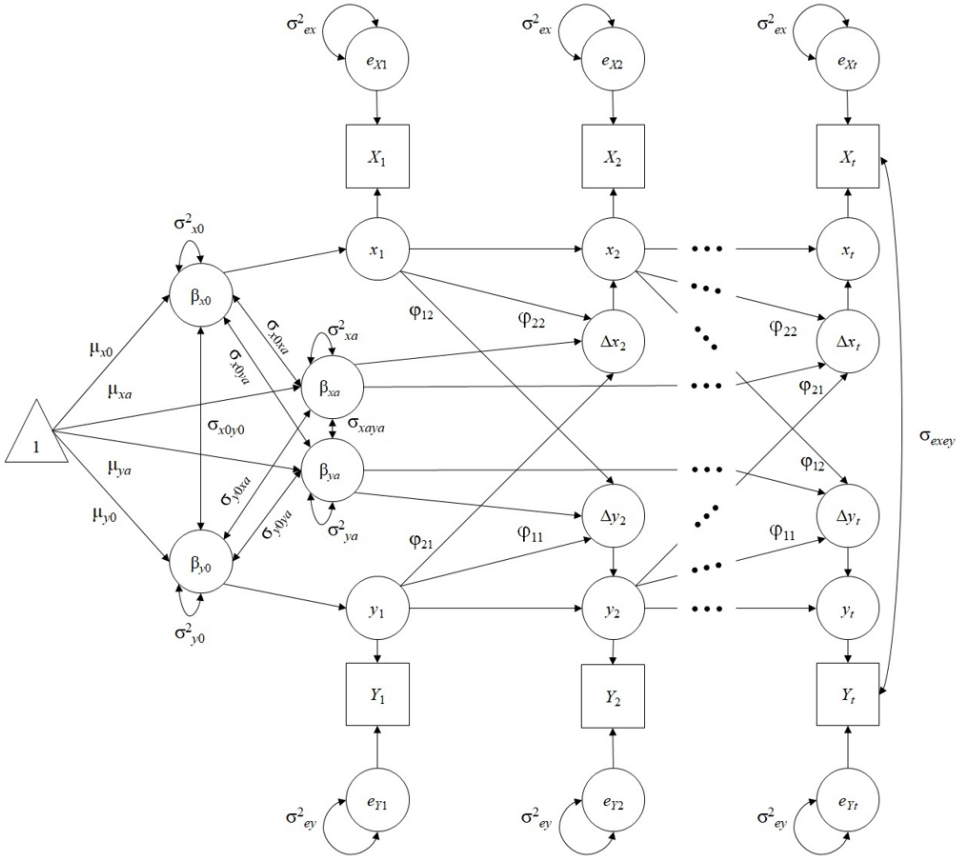
Los modelos presentados en la Figura 4 son univariados: describen el cambio en una sola variable. Es posible extenderlos para incluir dos o más variables medidas repetidamente. Sin embargo, el número de parámetros a estimar aumenta considerablemente con cada nueva variable, y quizá por ello lo más frecuente es encontrar aplicaciones bivariadas (y rara vez con tres variables). La Figura 5 presenta el diagrama de rutas de un LCS bivariado. Para profundizar sobre modelos LGC bivariados puede consultarse Estrada et al. (2020).

Utilizar un modelo BLCS permite estudiar varios aspectos del cambio de ambas variables, y su relación entre ellas: a) ¿existe correlación entre los momentos iniciales (intersecciones latentes) de ambas variables?, b) ¿entre los componentes aditivos?, c) ¿relaciones cruzadas entre intersección de una variable y componente aditivo de la otra?, d) ¿qué variable tiene un auto-efecto más fuerte?, e) ¿existen influencias cruzadas entre el nivel en una variable y el cambio posterior en la otra? Las dos últimas preguntas implican analizar la dinámica de ambas variables, y en concreto la pregunta (e) implica analizar la relación dinámica entre las variables. Para profundizar, consúltese Cáncer et al., (2021).

Tanto el LGC como el LCS se aplican a trayectorias de desarrollo porque permiten modelar cambios en el nivel a lo largo del tiempo. Existen otros modelos SEM longitudinales que están diseñados para aplicarse a trayectorias estables, tanto en datos de panel como longitudinales intensivos. Un ejemplo es el *random intercept cross-lagged panel model* (RI-CLPM), que hemos mencionado anteriormente. Puede profundizarse sobre este modelo en Mulder y Hamaker (2021). También puede consultarse la interesante revisión de Usami et al. (2019), sobre las diferencias y similitudes entre todos los modelos SEM que se pueden usar para estudiar las relaciones dinámicas entre dos variables.

**Figura 5**

*Diagrama de rutas de un modelo LCS bivariado (BLCS)*




## *Otros marcos de modelado dinámico con variables latentes*

Los SEM del apartado anterior también pueden especificarse como *modelos de estado-espacio* (*state-space models*, SSM). Estos modelos utilizan ecuaciones diferenciales (si se modela en tiempo continuo) o ecuaciones de diferencia (en tiempo discreto) para caracterizar trayectorias latentes de las cuales solo se tienen observaciones en puntos temporales concretos (vinculados a la trayectoria latente mediante un modelo de medida similar al de un SEM). Puede profundizarse sobre SSM en Hunter (2018). Tienen dos ventajas importantes respecto a un SEM: (a) Son más cómodos de utilizar si existen muchas medidas repetidas, y lo más importante, (b) permiten especificar modelos de cambio en tiempo continuo para variables latentes. Esto último no es posible con SEM tradicional, si bien la reciente aparición de *SEM en tiempo continuo* lo ha hecho posible (*ctSEM*, Driver y Voelkle, 2018).

## Conclusión

El uso de datos longitudinales permite observar el desarrollo y el cambio a lo largo del tiempo, lo que resulta fundamental para la comprensión de los procesos psicológicos. En este trabajo hemos revisado conceptos básicos para distinguir tipos de preguntas longitudinales, tipos de trayectorias, diseños de investigación habituales y marcos

de modelado estadístico que permiten contestar a dichas preguntas. Confiamos en que, cuanto mayor sea la popularidad y comprensión de estas herramientas, la Psicología alcanzará progresivamente un mayor conocimiento del comportamiento humano.

## Referencias


Bell, R. Q. (1953). Convergence: An Accelerated Longitudinal Approach. *Child Development, 24*(2), 145–152. https://doi.org/10.2307/1126345

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R. y Waldorp, L. J. (2021). Network Analysis of Multivariate Data in Psychological Science. N*ature Reviews Methods Primers, 1*(1), 1–18. https://doi.org/10.1038/s43586-021-00055-w

Cáncer, P. F., Estrada, E., Ollero, M. J. F. y Ferrer, E. (2021). Dynamical Properties and Conceptual Interpretation of Latent Change Score Models. *Frontiers in Psychology, 12,* Artículo 696419. https://doi.org/10.3389/fpsyg.2021.696419

Driver, C. C. y Voelkle, M. C. (2018). Hierarchical Bayesian Continuous Time Dynamic Modeling. *Psychological Methods, 23*(4), 774–799. http://dx.doi.org/10.1037/met0000168

Epskamp, S. (2020). Psychometric Network Models from Time-Series and Panel Data. *Psychometrika, 85*(1), 206–231. https://doi.org/10.1007/s11336-020-09697-3

Ernst, A. F., Albers, C. J. y Timmerman, M. E. (2024). A Comprehensive Model Framework for between-Individual Differences in Longitudinal Data. *Psychological Methods, 29*(4), 748–766. https://doi.org/10.1037/met0000585

Estrada, E. y Ferrer, E. (2019). Studying Developmental Processes in Accelerated Cohort-Sequential Designs with Discrete- and Continuous-Time Latent Change Score Models. *Psychological Methods, 24*(6), 708–734. https://doi.org/10.1037/met0000215

Estrada, E., Sbarra, D. A. y Ferrer, E. (2020). Models for Dyadic Data. En A. G. C. Wright y M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology* (pp. 350–368). Cambridge University Press. https://doi.org/10.1017/9781316995808.033

Ferrer, R. y Pardo, A. (2019). Clinically Meaningful Change. *Methodology, 15*(3), 97–105. https://doi.org/10.1027/1614-2241/a000168

Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E. y Bringmann, L. F. (2024). So You Want to Do ESM? 10 Essential Topics for Implementing the Experience-Sampling Method. *Advances in Methods and Practices in Psychological Science, 7*(3), 1–27. https://doi.org/10.1177/25152459241267912

Grimm, K. J., Ram, N. y Estabrook, R. (2017). *Growth Modeling: Structural Equation and Multilevel Modeling Approaches*. Guilford Press.

Hoffman, L. (2015). *Longitudinal Analysis: Modeling within-Person Fluctuation and Change*. Routledge.

Hunter, M. D. (2018). State Space Modeling in an Open Source, Modular, Structural Equation Modeling Environment. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 307–324. https://doi.org/10.1080/10705511.2017.1369354

Hunter, M. D., Fisher, Z. F. y Geier, C. F. (2024). What Ergodicity Means for you. *Developmental Cognitive Neuroscience, 68*, 1–16. https://doi.org/10.1016/j.dcn.2024.101406

McNeish, D. y Hamaker, E. L. (2020). A Primer on Two-Level Dynamic Structural Equation Models for Intensive Longitudinal Data in Mplus. *Psychological Methods, 25*(5), 610–635. https://doi.org/10.1037/met0000250

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kilani, G., Tuerlinckx, F., Kuppens, P. y Dejonckheere, E. (2023). m-Path: An Easy-to-use and Highly Tailorable Platform for Ecological Momentary Assessment and Intervention in Behavioral Research and Clinical Practice. *Frontiers in Digital Health, 5,* Artículo 1182175. https://doi.org/10.3389/fdgth.2023.1182175

Mongin, D., Uribe, A., Cullati, S. y Courvoisier, D. S. (2024). A Tutorial on Ordinary Differential Equations in Behavioral Science: What does Physics Teach us? *Psychological Methods, 29*(5), 980–1002. https://doi.org/10.1037/met0000517

Mulder, J. D. y Hamaker, E. L. (2021). Three Extensions of the Random Intercept Cross-Lagged Panel Model. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(4), 638–648. https://doi.org/10.1080/10705511.2020.1784738

Usami, S., Murayama, K. y Hamaker, E. L. (2019). A Unified Framework of Longitudinal Models to Examine Reciprocal Relations. *Psychological Methods, 24*(5), 637–657. https://doi.org/10.1037/met0000210

Voelkle, M. C., Gische, C., Driver, C. C. y Lindenberger, U. (2018). The Role of Time in the Quest for Understanding Psychological Mechanisms. *Multivariate Behavioral Research, 53*(6), 782–805. https://doi.org/10.1080/00273171.2018.1496813

Widaman, K. F., Ferrer, E. y Conger, R. D. (2010). Factorial Invariance Within Longitudinal Structural Equation Models: Measuring the Same Construct Across Time. *Child Development Perspectives, 4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

# ANÁLISIS DE REDES EN LA MEDICIÓN PSICOLÓGICA: FUNDAMENTOS

# NETWORK ANALYSIS IN PSYCHOLOGICAL MEASUREMENT: FUNDAMENTALS

EDUARDO FONSECA-PEDRERO[1] Y JOSÉ MUÑIZ[2]

Cómo referenciar este artículo/How to reference this article:

Fonseca-Pedrero, E. y Muñiz, J. (2025). Análisis de Redes en la Medición Psicológica: Fundamentos [Network Analysis in Psychological Measurement: Fundamentals]. *Acción Psicológica, 22*(1), 87–100. https://doi.org/10.5944/ap.22.1.43296

## Resumen

El análisis de redes ha ido ganando terreno en los últimos años para analizar datos psicológicos multivariantes. El propósito de este trabajo es llevar a cabo una exposición introductoria del análisis psicométrico de redes psicológicas. En primer lugar, se contextualizan los enfoques de redes. En segundo lugar, se aborda el núcleo central de la metodología de análisis psicométrico de redes: estimación de la estructura de la red, descripción de la red y análisis de la estabilidad de la red. En tercer lugar, se comentan algunas aplicaciones al campo de la Psicología y la Psicometría. En cuarto lugar, se mencionan algunas críticas al análisis de redes, lo que permite plantear algunos de las retos a los que se enfrenta este enfoque psicométrico. Se finaliza con una breve recapitulación y posibles líneas de investigación futuras.

**Palabras clave:** Análisis de redes; Modelo de red; Estimación de redes; Medición; Psicometría.

## Abstract

The use of network analysis to analyze multivariate psychological data has become popular in recent years. The purpose of this work is to introduce network analysis for the measurement of psychological variables. First, the net-

**Correspondence address [Dirección para correspondencia]**: Eduardo Fonseca Pedrero, Facultad de Letras y de la Educación, Universidad de La Rioja, España.
**Email**: eduardo.fonseca@unirioja.es

**ORCID**: Eduardo Fonsea-Pedrero (https://orcid.org/0000-0001-7453-5225) y José Muñiz (https://orcid.org/0000-0002-2652-5361).

[1] Universidad de la Rioja, España.
[2] Universidad Nebrija, España.

Recibido: 8 de noviembre de 2024.
Aceptado: 27 de enero de 2025.

work approaches are contextualised. Second, the core of the psychometric network analysis methodology is addressed: network structure estimation, network description and network stability analysis. Thirdly, some applications to the field of psychology and psychometrics are discussed. Fourthly, possible criticisms of network analysis are mentioned, to indicate some of the challenges faced by this psychometric approach. We finish with a brief recapitulation and possible lines of future research.

**Keywords:** Network Analysis; Network Model; Network Estimation; Measurement; Psychometrics.

## Introducción

El uso de análisis de redes para estudiar datos psicológicos de naturaleza multivariante ha ido ganando terreno en los últimos años entre los investigadores y los profesionales de la Psicología (e.g., Borsboom et al., 2021). El análisis de redes se presenta como un nuevo método en la identificación e inferencia de atributos psicológicos, que pretende ir más allá de los modelos clásicos de variables latentes, como el análisis factorial o las ecuaciones estructurales, entre otros. Esta nueva forma de analizar el comportamiento humano y los procesos psicológicos abre todo un abanico de posibilidades, ya que permite el uso de formas alternativas de analizar datos, maneras diferentes de modelar y analizar las relaciones entre variables, o diseñar nuevas formas de intervención. Por todo ello, no es de extrañar que los modelos de redes y el análisis de redes han suscitado un creciente interés entre los profesionales de la Psicología.

El objetivo de este trabajo es realizar una breve introducción al análisis psicométrico de redes psicológicas. En primer lugar, se conceptualiza el modelo de redes, así como sus implicaciones para el campo de la Psicología. En segundo lugar, se aborda el núcleo central de la metodología de análisis psicométrico de redes: estimación de la estructura de la red, descripción de la red y análisis de la estabilidad de la red. En tercer lugar, se comentan algunas aplicaciones al campo de la Psicología y la psicometría. En cuarto lugar, se mencionan posibles críticas al análisis de redes, lo que permitirá otear algunos de los retos a los que se enfrenta este enfoque. Se finaliza con una breve recapitulación y se reflexiona sobre posibles líneas de investigación futuras.

Para un análisis más detallado y especializado del análisis de redes se pueden consultar excelentes trabajos previos, tanto en inglés (Borsboom, 2017; Borsboom y Cramer, 2013; Isvoranu et al., 2022; McNally, 2016), como en español (Fonseca-Pedrero, 2017, 2018). También se pueden consultar tutoriales (Costantini et al., 2015, 2019; Epskamp, Borsboom et al., 2018; Epskamp y Fried, 2018; Hevey, 2018; Huth et al., 2023; Jones et al., 2018), y otros excelentes recursos en la web (http://psychosystems.org/, https://www.youtube.com/watch?v=C6-BgnWGwfA), o software estadístico libre con módulos específicos para el análisis de redes, por ejemplo, JASP (https://jasp-stats.org/, https://jasp-stats.org/2018/03/20/perform-network-analysis-jasp/) o R (e.g., qgraph; Isvoranu et al., 2022; Jones et al., 2018).

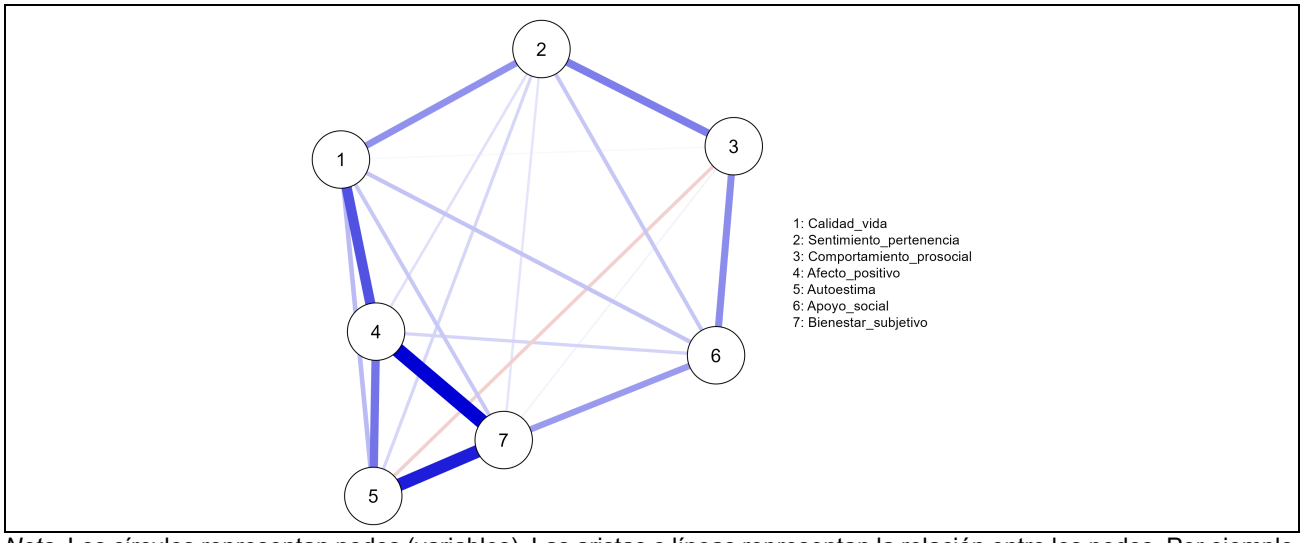
## Contextualización

Los modelos de red (network models) permiten llevar a cabo representaciones estadísticas de las relaciones entre variables, es decir, constituyen estructuras estadísticas creadas para configurar redes a partir de los datos. Una red es un modelo abstracto representado gráficamente que contiene nodos (nodes) y aristas (edges). Los nodos simbolizan los objetos o variables de estudio, mientras que las aristas representan las conexiones entre los nodos, esto es, la «línea» que los conecta (véase Figura 1). La representación gráfica de nodos y aristas se denomina grafo.

De una forma general el análisis de redes se podría concretar en las técnicas estadísticas para estimar, analizar, e interpretar la red psicológica. Representa un enfoque rela-

**Figura 1**

*Ejemplo de red psicológica estimada: fortalezas psicológicas en la adolescencia*




*Nota*. Los círculos representan nodos (variables). Las aristas o líneas representan la relación entre los nodos. Por ejemplo, el nodo nº 1 la calidad de vida subjetiva, el nodo nº 2 el sentimiento de pertenencia al centro educativo, etc. Los nodos se corresponden con los indicadores psicométricos de fortalezas psicológicas. A mayor valor del coeficiente, mayor grosor de la línea y, por lo tanto, la asociación más fuerte entre nodos. Color azul de la arista indica relación positiva entre nodos (variables). Color rojo de la arista indica relación negativa entre nodos (variables).

tivamente reciente en Psicología, si bien no es algo nuevo desde un punto de vista científico. Se ha utilizado extensamente en otras áreas como el estudio de las relaciones sociales bajo la denominación de teoría de grafos (Borgatti et al., 2009; Newman, 2010; Vega-Redondo, 2007), o en el ámbito económico (Goyal, 2023; Jackson, 2008). No obstante, no ha sido hasta hace poco que se ha rescatado este enfoque para modelar otros fenómenos psicológicos, y muy especialmente los trastornos psicológicos (Borsboom, 2017). El profesor Denny Borsboom de la Universidad de Ámsterdam y su equipo de investigación han estimulado un enfoque diferente en la conceptualización de los problemas psicopatológicos, tales como la depresión o la psicosis (Borsboom y Cramer, 2013; Schmittmann et al., 2013).

El uso de los modelos de redes surge como respuesta epistemológica a ciertas dificultades de las que adolece la psicopatología clásica, como el modelo biomédico que se postula desde los principales sistemas taxonómicos. Así, el Manual diagnóstico y estadístico de los trastornos mentales 5ª edición-Texto Revisado (DSM-5-TR, por su abreviatura en inglés; APA, 2022) considera que los síntomas y signos que refieren las personas tienen su origen en una supuesta causa latente denominada «trastorno mental». Se hipotetiza, por ejemplo, que las manifestaciones fenotípicas tales como las experiencias alucinatorias o las creencias delirantes son debidas a un trastorno subyacente que los causa, denominado, en este caso, esquizofrenia. A esta interpretación se le conoce como «modelo de causa común» (Borsboom y Cramer, 2013), y conduce a determinados problemas como el reduccionismo (e.g., los problemas psicológicos tienen una única causa), el razonamiento tautológico (e.g., repetición de un mismo argumento o hecho expresado de distintas maneras; véase el caso de que alguien que refiere alucinaciones auditivas se le diagnóstica de psicosis, y posteriormente se argumenta que tiene psicosis porque escucha voces) o la reificación del diagnóstico (creer que el nombre que le damos al trastorno, véase psicosis, explica todo sobre un fenómeno o es el fenómeno en sí).

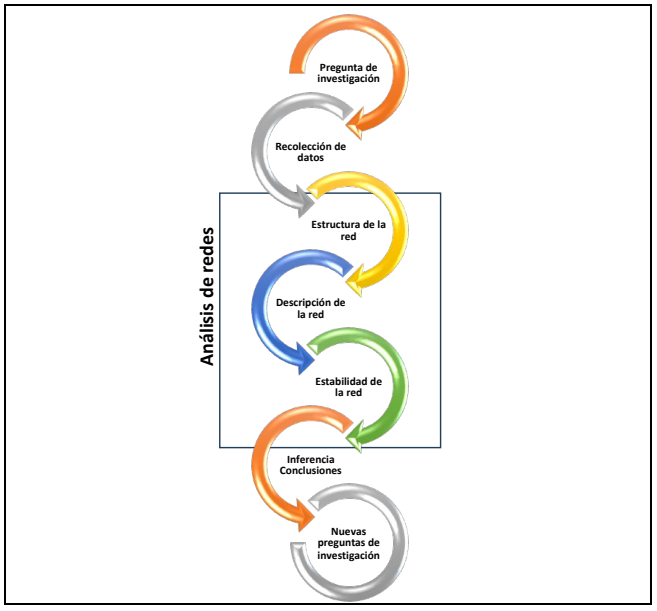
El modelo de redes, sin embargo, conceptualiza los trastornos psicológicos como constelaciones dinámicas de conductas (signos, síntomas, rasgos, conductas, etc.) que se encuentran interrelacionados (o no) de forma causal, esto es, impactan o interactúan mutuamente entre sí. Los síntomas no serían meras consecuencias pasivas de una supuesta variable latente, no observable directamente. Desde esta perspectiva el trastorno psicológico, si lo hubiere, habría que comprenderlo en las propias relaciones funcionales establecidas entre los elementos de la red, tanto horizontal (dentro del nivel psicológico), como verticalmente (entre niveles de análisis). No habría que buscar más allá. Tampoco habría que indagar sobre supuestas averías intrapsíquicas o alteraciones en los circuitos cerebrales (Fonseca-Pedrero y Al-Halabí, 2024; Al-Halabí y Fonseca-Pedrero, 2024; Pérez Álvarez, 2018). A fin de cuentas, todo está relacionado con todo, pero no del todo.

El modelo de redes, al igual que ocurre en muchos campos científicos, entiende que los fenómenos se caracterizan y entienden mejor a nivel de sistemas. Esto significa que para comprender determinados fenómenos hay que centrarse en la organización de los componentes del sistema y no tanto en el funcionamiento de sus componentes individuales. Dichos componentes se pueden representar en una red. Esa visión entronca con la idea de que la conducta es compleja por naturaleza, por lo que su comprensión requiere de modelos más sofisticados que vayan más allá de una visión lineal, estática y unicausal que permitan analizar y entender todo el abanico de comportamientos que conforman la diversidad humana.

Siguiendo a Borsboom et al. (2021), en la Figura 2 se recoge un esquema de los pasos a seguir en las investigaciones que utilizan el enfoque de redes. El núcleo de la metodología de análisis de redes psicométricas reside en tres pasos: (a) estimación de la estructura de la red, (b) descripción de la red y (c) análisis de la estabilidad de la red. Estos pasos se asientan en preguntas de investigación sustantivas y procedimientos de recopilación de datos. Los resultados encontrados se combinan con consideraciones metodológicas generales y conocimientos específicos del ámbito para apoyar la inferencia científica.

**Figura 2**

*Pasos del proceso de investigación en el análisis de redes (modificado de Borsboom et al., 2021)*




## *Estimación de la estructura de la red*

Las redes en Psicología necesitan ser estimadas (Epskamp y Fried, 2018), y dicha estimación parte de una matriz de correlaciones que pueden ser, básicamente, de tres tipos: simples, parciales y parciales regularizadas.

Las correlaciones simples (conocida también como red de asociación), es la representación gráfica derivada de la matriz de correlaciones de Pearson. Las correlaciones parciales, o red de concentración, permiten ver la correlación entre el nodo *A* y el nodo *B* controlando el efecto del resto de nodos de la red. Una red de correlación parcial, en la que muestra las correlaciones condicionales (controlando) a todas las demás variables de la red, es más fácil de interpretar, esto es, dos nodos están conectados si y solo si existe una covarianza entre esos nodos que no puede explicarse por ninguna otra variable en la red. Dicha red se llama *Pairwise Markov Random Field* (PMRF). Cuando los datos continuos tienen una distribución normal multivariante, el análisis de las correlaciones parciales se implementa mediante el *Gaussian Graphical Model*. Si los

datos son mixtos (categóricos y continuos) se utiliza el *Mixed Graphical Model*. Si son datos son binarios entonces se utiliza el *Ising Model*.

La estimación de la red se realiza mediante un algoritmo denominado *Fruchterman-Reingold*. Las correlaciones parciales regularizadas, implementan un procedimiento de regularización, que en esencia permite extraer una red estable y de fácil interpretación que necesita de menos parámetros a estimar. En este caso se puede estimar la red con el *Least Absolute Shrinkage and Selection Operator* (LASSO) o con una variación denominada *Graphical*-LASSO (G-LASSO; Epskamp, Borsboom, et al., 2018). El uso del LASSO requiere establecer un parámetro de ajuste denominado hiperparámetro ($\lambda$). La regularización no está exenta de problemas, por lo que en determinadas situaciones no funciona adecuadamente. En estos casos se deben utilizar formas alternativas para maximizar la sensibilidad y la especificidad a la hora de estimar los *edges*[1] (Epskamp y Fried, 2018). La elección del método de estimación no se debe ver como algo trivial ya puede tener un gran impacto tanto en la topología de la red como en las inferencias que posteriormente se hagan a partir de los resultados encontrados (Epskamp et al., 2017).

Utilizando el algotitmo Fruchterman-Reingold la disposición espacial de los nodos no es fácil de interpretar. Existen en la literatura otros diseños de visualización de datos multivariantes que van más allá de los algoritmos dirigidos por fuerza y que permiten interpretar la posición de los nodos de una forma más correcta. A este respecto Jones et al. (2018) proponen el escalamiento multidimensional, los componentes principales y el *eigenmodel*.

Obviamente, el procedimiento para la estimación de la red depende de la naturaleza de los datos, por ejemplo, si estos son transversales o longitudinales, o si tienen una estructura multinivel o no. Para casos donde los datos son de tipo longitudinal y/o presentan una estructura multinivel se pueden encontrar otros procedimientos como *Graphical-VAR* o *Multilevel-VAR*. También se pueden hacer estimaciones de redes mediante estadística bayesiana (McNally, 2016).

[1] Véase: http://psychosystems.org/glasso_developments

## *Descripción de la red*

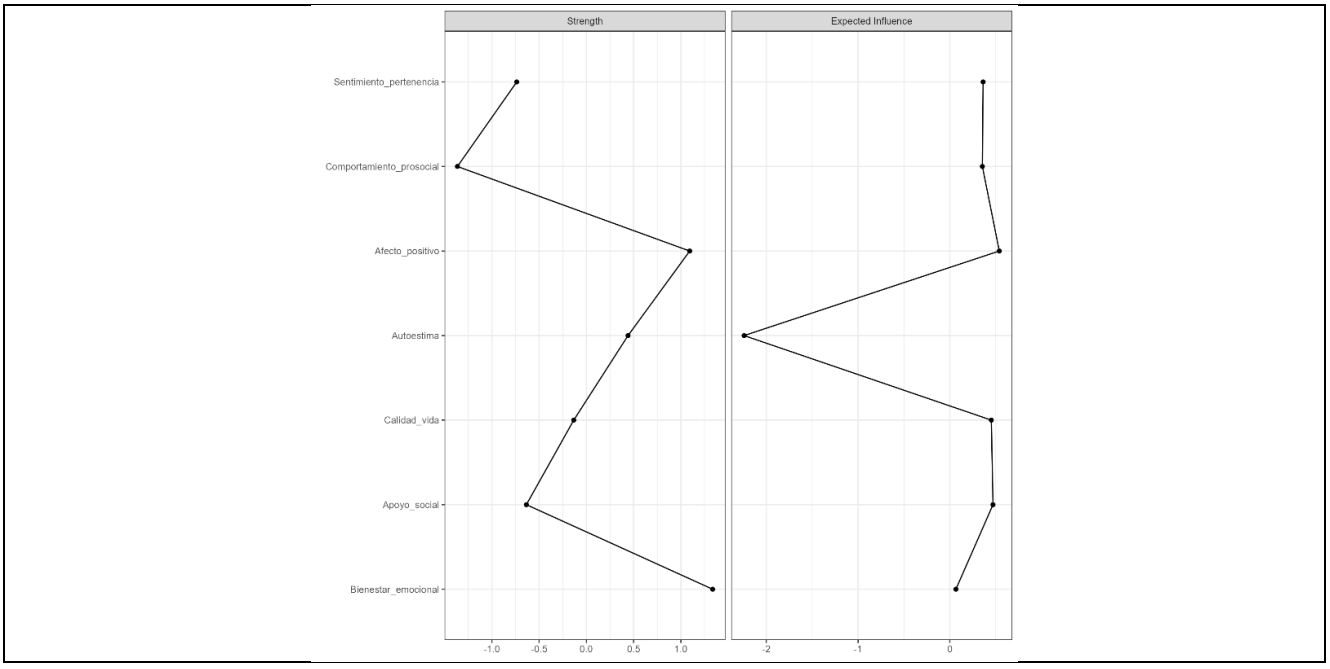
A partir de la red estimada se pueden realizar diferentes inferencias que permitan comprender tanto su estructura como examinar la importancia relativa de los nodos dentro de ella. Para analizar la estructura de la red se pueden encontrar diferentes índices: (a) distancia y longitud de la trayectoria más corta (¿puede influir el nodo *X* rápidamente en el nodo *Y*?); (b) centralidad (¿cuál es el nodo más importante en la red?); y (c) conectividad y agrupamiento (¿en qué medida los nodos están bien conectados?). Aquí únicamente se expondrán las medidas de centralidad, quienes estén interesados en profundizar en las medidas de inferencia de la red puede consultar trabajos más especializados (Costantini et al., 2015; Hevey, 2018).

En concreto las medidas de centralidad se preguntan cuál es el nodo más importante en la red. Permiten analizar la importancia relativa del nodo dentro de la red en función del patrón de conexiones, esto es, en una red estimada no todos los nodos son igualmente importantes. Un nodo es central si tiene muchas conexiones. Un nodo es periférico, se encuentra en la parte externa de la red, si tiene pocas conexiones. Para saber si el nodo es central (importante o influyente) en la red se deben tener en cuenta: (a) la fuerza (*strenght centrality*); (b) la cercanía (*closenness centrality*); y (c) la intermediación (*betweeness centrality*). Los programas estadísticos permiten generar figuras y tablas para examinar los valores de estos índices de centralidad. Dicha representación gráfica arroja valores estandarizados (puntuaciones *Z*) referidos a la fuerza, cercanía y/o intermediación de los nodos, aspecto que informa sobre la importancia relativa de cada nodo en la red.

En la actualidad se está cuestionando si estas medidas de centralidad tienen sentido a la hora de estimar redes con variables psicológicas (Bringmann et al., 2019; Hallquist et al., 2021), por ello se han propuesto otras como la influencia esperada o la predictibilidad. La influencia esperada se refiere a la suma de todas las aristas de un nodo. Esta medida de inferencia mejora la centralidad de fuerza que usa la suma de los pesos absolutos (es decir, las aristas negativas se convierten en aristas positivas antes de sumarse), lo que distorsiona la interpretación si hay bordes

**Figura 3**

*Índices de centralidad en una red psicológica estimada*




negativos. La predictibilidad es una medida absoluta de interconexión que proporciona la varianza de cada nodo que se explica por todos sus nodos vecinos. En la Figura 3 se presenta una tabla de centralidad con los valores de fuerza e influencia esperada.

## *Análisis de la estabilidad de la red*

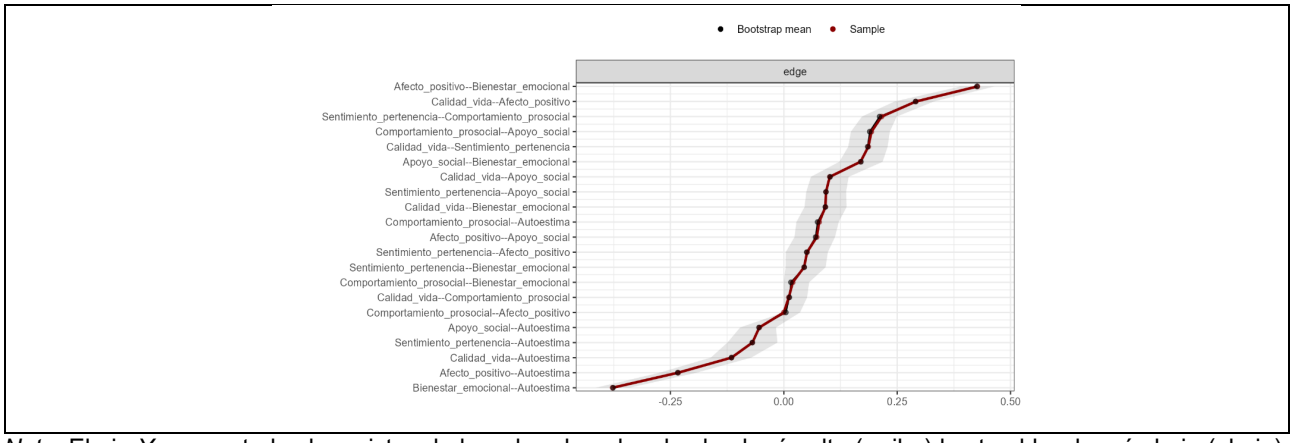
Evaluar la precisión de las conexiones de red estimadas, investigar la estabilidad de los índices de centralidad, y comprobar si las conexiones de red y las estimaciones de centralidad para distintas variables difieren entre sí, son cuestiones nucleares que conciernen a la replicabilidad (Epskamp y Fried, 2018). En los análisis de redes de variables psicológicas se debe comprobar la precisión con la que se estiman las redes y la estabilidad de las inferencias de la estructura de la red.

En el tutorial de Epskamp et al. (2018) se describen procedimientos en el paquete R *bootnet* que permiten a los investigadores medir la precisión de las estimaciones de peso de borde (Figura 4) y la estabilidad de las métricas de centralidad (Figura 5). La precisión de las ponderaciones de los bordes se estima calculando intervalos de confianza (IC) del 95% para sus estimaciones. Para cada peso de la arista, el valor real del parámetro correspondiente estará dentro del IC en el 95 % de los casos. Además, aunque no se aborda en esta introducción, también se puede evaluar si el peso de un borde difiere significativamente de otro, por ejemplo, si la asociación entre el nodo 1 y el nodo 2 es mayor que la asociación entre el nodo 5 y el nodo 6 (e.g., Epskamp et al., 2018).

Para estimar la estabilidad de las medidas de centralidad, Epskamp et al. (2018) recomendaron el *bootstrap* de subconjuntos mediante *casedropping*. Este procedimiento implica ir descartando aleatoriamente un porcentaje cada vez mayor de participantes del conjunto total de datos (primero sobre el 95 % de la muestra, luego sobre el 90 %, y así sucesivamente) y luego volver a calcular las medidas de centralidad y computar el coeficiente de correlación. Se

**Figura 4**

*Precisión de los pesos de las aristas*

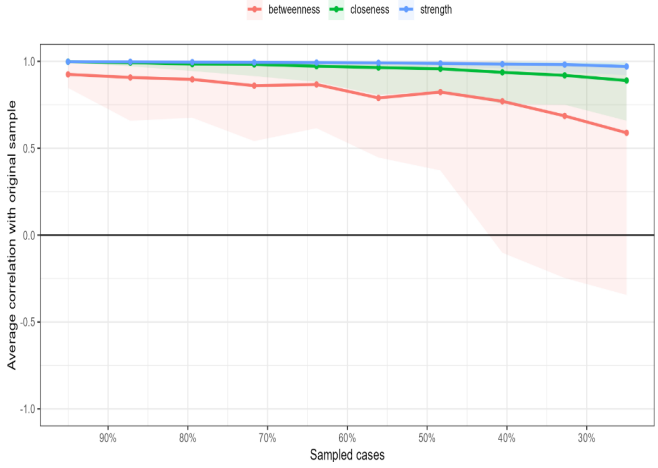



*Nota*. El eje *Y* recoge todas las aristas de la red, ordenados desde el más alto (arriba) hasta el borde más bajo (abajo). Dichas aristas se refieren a la relación entre dos nodos, esto es, «bienestar_emocional-autoestima» se refiere a la relación entre el nodo bienestar emocional y el nodo autoestima. Los puntos rojos representan el peso de las artistas en la red. El área gris representa el IC del 95% del peso de las aristas (cuanto más pequeño mayor es la precisión del peso de las aristas).

trata de examinar si los índices de centralidad permanecen estables a medida que se va perdiendo participantes. Si el valor de la correlación cambia considerablemente pondría en cuestión la estabilidad de las medidas de centralidad. Para cuantificar la estabilidad de los valores de centralidad de una red se propone el coeficiente de estabilidad de correlación. Se ha sugerido que un coeficiente de estabilidad de 0.7 o superior entre la estimación original de la muestra completa y las estimaciones de los subconjuntos muestrales podría ser un umbral útil a considerar. Se recomienda igualmente que el coeficiente no sea inferior a 0.25, y preferiblemente superior a 0.5.

**Figura 5**

*Coeficiente de estabilidad de correlación para los índices de centralidad*



## Aplicaciones al campo de la Psicología y la Psicometría

El análisis de redes ha tenido un gran desarrollo dentro del campo de la Psicología clínica y la psicopatología (Mcnally, 2021; Robinaugh et al., 2020), si bien se ha extendido a otras áreas como la resiliencia (Scheffer et al., 2018), la personalidad (Costantini et al., 2019), la inteligencia (Kan et al., 2020), la relación con enfermedades físicas (Isvoranu et al., 2022) o la educación (Álvarez-Díaz et al., 2022), por mencionar sólo algunas. El análisis de redes tiene una enorme potencialidad para ayudar a comprender y responder a algunos de los desafíos más importantes que tenemos que acometer profesional y socialmente, por ejemplo, la conducta suicida (Antypa et al., 2024; Fonseca-Pedrero et al., 2024).

También se han desarrollado nuevos métodos que pueden ser de utilidad para el investigador y el clínico, como pudieran ser las redes bayesianas (Briganti et al., 2022), el *Exploratory Graph Analysis* (Golino y Epskamp, 2017), la comparación de redes (Van Borkulo, 2018), la invarianza de medición (Hoekstra et al., 2023), los *Moderated Network Models* (Haslbeck et al., 2021) o los estudios intensivos. A continuación, nos referimos brevemente a las potencialidades asociadas al uso de redes bayesianas y a las redes temporales.

En la actualidad la mayoría de los estudios realizados en este campo utilizan análisis de redes con datos transversales (redes ponderadas y no dirigidas). Las redes bayesianas pueden ayudar a superar algunas de estas limitaciones pues son modelos que incluyen interacciones dirigidas para realizar inferencias causales sobre constructos psicológicos. En esencia, las redes bayesianas son modelos gráficos probabilísticos que representan las relaciones de independencia condicional entre variables como un grafo acíclico dirigido (DAG), en el que las aristas pueden interpretarse como efectos causales que conectan nodos de forma causal. Investigar bajo la óptica de redes bayesianas puede ser de interés de cara a obtener una aproximación causal de las variables psicológicas.

Por sus implicaciones prácticas y clínicas, mención especial merecen las redes resultantes de series temporales (también denominadas redes dinámicas ideográficas) (e.g., Bringmann, 2024). Este tipo de redes suelen inferirse a partir de datos de series temporales recogidos mediante evaluación ambulatoria. Utilizando los métodos de estimación adecuados se pueden calcular dos tipos redes: la red temporal y la red contemporánea. La red temporal, es una dirigida y ponderada, que representa cómo las variables en un punto temporal (tiempo $t$) predicen las variables en la siguiente ventana temporal (tiempo $t + 1$), incluyendo cómo una variable evaluada en el primer punto temporal predice su valor en el siguiente. La red contemporánea representa las aristas (no dirigidas y ponderadas) que conectan los nodos dentro de la misma ventana de medición. Normalmente, esta red presenta aristas que representan las correlaciones parciales entre nodos en esta ventana de medición tras ajustar todas las demás variables de la ventana, así como todas las demás variables de la ventana de medición anterior. Este tipo de redes extraídas de diseños longitudinales permiten revelar cómo se desarrollan las interacciones entre los sistemas a lo largo del tiempo, establecer posibles relaciones causales o dar respuesta a la cuestión de la ergodicidad, es decir, si las relaciones que se establecen entre las variables a nivel de grupo son las mismas que operan a nivel intraindividual (Fonseca-Pedrero, 2018; Mcnally, 2021).

La combinación de los métodos psicométricos de redes ha abierto una variedad de nuevas vías para conceptualizar y estudiar los fenómenos psicológicos. Igualmente, los modelos de redes se presentan como un enfoque complementario a otros modelos psicométricos, como la teoría de respuesta a los ítems (Epskamp et al., 2017; Kan et al., 2020; Marsman et al., 2018). A pesar de los antecedentes divergentes de estos acercamientos, trabajos previos han estudiado la equivalencia entre ellos e ilustran las oportunidades que germinan de estas posibles conexiones (Kan et al., 2020; Marsman et al., 2018). No obstante, desde un punto de vista epistemológico, algunos autores se cuestionan si la equivalencia matemática es lo mismo que la equivalencia ontológica (Mcnally, 2021).

## Algunos desafíos y controversias en el Análisis de Redes

La investigación de los análisis de redes en Psicología se encuentra en estos momentos en su infancia, por lo que es necesario seguir trabajando en la construcción de modelos sólidos y refutables, e incorporar nuevas evidencias científicas (Borsboom, 2017). Obviamente, el modelo de redes no está exento de ciertas limitaciones y algunos autores han realizado algunas reflexiones cautelares (e.g., Neal et al., 2022). Pretenden con ello no dejar pasar por alto algunas de las principales críticas asociadas a estos modelos y métodos. Aquí se comentan brevemente algunas de ellas relativas a la selección del modelo, el diseño del estudio, la fiabilidad de la estimación y la interpretación de las medidas, así como otras de calado más conceptual.

Primero, se debe distinguir aquellos campos científicos que permiten un análisis bajo esta perspectiva respecto a los que no. No partimos de cero, en la literatura ya existen numerosos métodos multivariantes adecuados para responder a determinadas preguntas de investigación, tales como los modelos de ecuaciones estructurales. Por tanto, los análisis psicométricos de redes deben utilizarse en aquellos casos en los que se ajusten mejor que otros métodos a los problemas planteados.

Segundo, hasta la fecha, la mayoría de los estudios publicados en análisis de redes utilizan diseños transversales y nomotéticos, por ello, habría que ser cauteloso, incluso abstenerse de hacer inferencias sobre causalidad cuando se utilicen este tipo de datos. No cabe duda de que una de las líneas futuras más prometedoras en el análisis de redes son los modelos de redes ideográficas de medidas repetidas (Mansueto et al., 2023). El modelaje de redes psicológicas personalizadas tiene claras implicaciones para la investigación y la práctica clínica (Epskamp et al., 2018).

Tercero, la correcta interpretación de una red psicológica no debe centrarse únicamente en su inspección visual. Un problema a evitar en las redes psicológicas es precisamente la sobre interpretación a la hora de su visualización (Jones, Mair, y McNally, 2018). Este aspecto se refiere especialmente al diseño y a la colocación de nodos en el grafo, por ejemplo, cuando los nodos de la red se agrupan en un clúster o un nodo está en el centro de la red y otros en la periferia. Nótese que la ubicación del nodo dentro de una red es solo una de las muchas formas igualmente «correctas» de colocar los nodos en la red, es decir, con la misma muestra la colocación de los nodos en la red, en una nueva estimación, podría ser diferente. Algo parecido, mutatis mutandis, a lo que ocurre, por ejemplo, con la ubicación de los factores en el Análisis Factorial. Por ello, hay que ser cauteloso a la hora de realizar una interpretación visual de los nodos y su posición en la red. Para una adecuada interpretación de la red psicológica existen diferentes procedimientos, por ejemplo, el análisis de las comunalidades[2] (Golino y Epskamp, 2017), o la predictibilidad (Haslbeck y Fried, 2017).

Cuarto, existe un amplio debate sobre la replicabilidad de las redes psicométricas estimadas. Algunos autores afirman que se están utilizando métodos que producen estimaciones poco fiables (Forbes et al., 2017; Neal et al., 2022), mientras que otros opinan justo lo contrario (Borsboom et al., 2017; de Ron et al., 2022). Dado el estado de la cuestión se aconseja actuar con cierta cautela a la hora de extraer conclusiones para la investigación o la práctica clínica antes de que los resultados se hayan replicado rigurosamente. También se deben considerar otros factores como el sesgo de Berkson. Según trabajos previos parece ser que, por el momento, las dudas sobre la replicabilidad de los resultados se han resuelto de forma tranquilizadora (de Ron et al., 2022). No cabe duda de que se debería seguir estudiando la replicabilidad y reproducibilidad de las redes psicológicas estimadas y analizar la necesidad de incorporar el error de medición.

Quinto, la mayoría de las métricas de centralidad utilizadas en el campo de la Psicología se desarrollaron originalmente en el análisis de redes sociales con sus consabidas peculiaridades (e.g., actores sociales, redes no ponderadas). Igualmente, aún no existen directrices claras para interpretar los índices de centralidad en las redes psicomé-

[2] Véase: http://psych-networks.com/r-tutorial-identify-communities-items-networks/

tricas (Hallquist et al., 2021). Se recomienda, por lo tanto, implementar alternativas más apropiadas que los índices de centralidad, así como avanzar en el desarrollo de nuevas medidas de inferencia e indicadores de ajuste.

Finalmente, desde un punto de vista conceptual, el análisis de redes con su impresionante y elegante tecnología podría ir en detrimento de análisis cualitativos narrativos y clasificaciones prototípicas más que politéticas (Fonseca-Pedrero, 2018, Pérez Álvarez, 2018). Las redes psicológicas suponen y a la vez tienden a homogeneizar los síntomas, rasgos, etc., cuando éstos podrían ser cualitativamente distintos, aspecto que requiere de un análisis fenomenológico de sus diferencias cualitativas. Por lo tanto, no se debe olvidar nunca que la metodología es la herramienta que nos permite exprimir y potenciar la información que se extrae de los datos, no se debe confundir el fin con los medios, el método debe estar al servicio de los temas sustantivos y problemas de la psicológica y no a la inversa. Como le gustaba recordar a nuestro maestro José Luis Pinillos, los datos sin conceptos son ciegos y los conceptos sin datos vacíos.

## Recapitulación

El propósito de este trabajo fue realizar una introducción al análisis psicométrico de redes. Se ha tratado de presentar, de forma sencilla, este fértil acercamiento a los investigadores y profesionales de la Psicología, esperando que les sirva de motivación para seguir profundizando en el tema. El análisis de redes ha emergido con la meta de dar respuesta a ciertos problemas de los que adolece algunas áreas de la Psicología actual como pudiera ser superación de la noción de causa latente subyacente a los trastornos de la salud mental. El modelo de redes y su aplicación psicométrica, el análisis de redes, representan un avance en el abordaje, comprensión y medición de los fenómenos psicológicos. Su correcto uso y su utilidad depende del objetivo de estudio y de los intereses del clínico o el investigador.

El modelo de redes constituye un enfoque prometedor en la forma de conceptualizar los fenómenos psicológicos, por ejemplo, entendiéndolos como sistemas dinámicos complejos. Es esencial para la Psicología incorporar diferentes ópticas y perspectivas que ayuden a repensar, en cierto modo, el comportamiento humano, en sentido amplio. No cabe duda de que la comprensión y estudio de la conducta humana es una labor compleja, donde operan una infinita cantidad de variables procedentes de múltiples niveles de análisis: biológico, psicológico y social (Fonseca-Pedrero et al., 2023; Piazza et al., 2024). En cualquier caso, el modelo de redes ayude a cambiar o no el actual abordaje epistemológico y metodológico de la Psicología, esperemos que sí, al menos se presenta como un nuevo enfoque a partir de la cual observar, medir, analizar, comprender e intervenir en los fenómenos psicológicos. Obviamente el análisis de redes no se debe ver como algo incompatible a los grandes modelos psicométricos (Muñiz, 2010, 2018), sino como un enfoque complementario con otros acercamientos teóricos y metodológicos (Epskamp et al., 2017; Marsman et al., 2018).

Muchas líneas de investigación interesantes se abrirán paso en los próximos años (Borsboom, 2022). Primero, sería interesante desplazarse hacia modelos de redes multinivel que permitan integrar múltiples niveles de análisis. A este respecto, también se tiene que explorar la forma en que las variables externas a la propia red (por ejemplo, variables biológicas o sociales) afectan a la estructura y la dinámica de la red. Segundo, como ha mencionado anteriormente, sería conveniente analizar el comportamiento humano desde una perspectiva dinámica (longitudinal), personalizada (ideográfica) y contextual (ser-en-el mundo). La conducta humana es compleja por naturaleza, por lo que su comprensión requiere de modelos más sofisticados que vayan más allá de una visión lineal, estática y unicausal que permitan analizar y entender todo el abanico de comportamientos que conforman la diversidad humana. La evaluación ambulatoria podría ayudar a este fin (Elosua et al., 2023; Fonseca-Pedrero et al., 2022). Tercero, se debe seguir trabajando en la incorporación de avances psicométricos, como pudieran ser, el desarrollo de nuevos índices de inferencia, métodos de estimación o el establecimiento de *guidelines* y estándares (Burger et al., 2023). Cuarto, sería interesante hacer programas libres más amigables que pudieran ser utilizados por los profesionales de la Psicología en su quehacer diario. Quinto, en el horizonte también se encuentra la modelización dinámica de redes computacionales que permitan refutar teorías. Todas estas propuestas facilitarían su implementación tanto en la in-

vestigación como en la práctica clínica. Futuros estudios determinarán la verdadera utilidad y calado del análisis de redes en Psicología, queda mucho por hacer, pero, como reza el clásico, todas las grandes caminatas empiezan por un pequeño paso.

## References


Al-Halabí, S. y Fonseca-Pedrero, E. (2024). Editorial for Special Issue on Understanding and Prevention of Suicidal Behavior: Humanizing Care and Integrating Social Determinants. *Psicothema, 36*(4), 309–318. https://doi.org/10.7334/psicothema2024.341

Álvarez-Díaz, M., Gallego-Acedo, C., Fernández-Alonso, R., Muñiz, J. y Fonseca-Pedrero, E. (2022). Análisis de Redes: una Alternativa a los Enfoques Clásicos de Evaluación de los Sistemas Educativos [Network analysis: An alternative to classic approaches for education systems evaluation]. *Psicología Educativa, 28*, 165–173 https://doi.org/10.5093/psed2021a16

Antypa, N., Kivelä, L. M. M., Fried, E. I. y Van Der Does, W. (2024). Psychological Medicine Examining Contemporaneous and Temporal Associations of Real-Time Suicidal Ideation using Network Analysis. *Psychological Medicine, 54*(12), 3357–3365. https://doi.org/10.1017/S003329172400151X

American Psychiatric Association. (2022). Diagnostic and Statistical Manual of Mental Disorders (5ª ed., rev.). Autor.

Borgatti, S. P., Mehra, A., Brass, D. J. y Labianca, G. (2009). *Network Analysis in the Social Sciences. Science, 323*, 892–896. https://doi.org/10.1126/science.1165821

Borsboom, D. (2017). A Network Theory of Mental Disorders. *World Psychiatry, 16,* 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D. (2022). Possible Futures for Network Psychometrics. *Psychometrika, 87*(1), 253–265. https://doi.org/10.1007/S11336-022-09851-Z

Borsboom, D. y Cramer, A. O. J. (2013). Network analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology, 9,* 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L. J. y Cramer, A. O. J. (2017). False alarm? A Comprehensive Reanalysis of "Evidence that Psychopathology Symptom Networks have Limited Replicability" by Forbes, Wright, Markon, and Krueger (2017). *Journal of Abnormal Psychology, 126*(7), 989–999. https://doi.org/10.1037/abn0000306

Briganti, G., Scutari, M. y McNally, R. J. (2022). A Tutorial on Bayesian Networks for Psychopathology Researchers. *Psychological Methods, 28*(4), 947–961. https://doi.org/10.1037/met0000479

Bringmann L. F. (2024). The Future of Dynamic Networks in Research and Clinical Practice. *World Psychiatry, 23*(2), 288–289. https://doi.org/10.1002/wps.21209

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T. W. y Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology, 128*(8), 892–903. https://doi.org/10.1037/ABN0000446

Burger, J., Isvoranu, A. M., Lunansky, G., Haslbeck, J. M. B., Epskamp, S., Hoekstra, R. H. A., Fried, E. I., Borsboom, D. y Blanken, T. F. (2023). Reporting Standards for Psychological Network Analyses in Cross-Sectional Data. *Psychological Methods,*

*28*(4), 806–824. https://doi.org/10.1037/MET0000471

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J. y Cramer, A. O. J. (2015). State of the aRt Personality Research: A Tutorial on Network Analysis of Personality Data in R. *Journal of Research in Personality, 54*, 13–29. https://doi.org/10.1016/j.jrp.2014.07.003

Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S. y Perugini, M. (2019). Stability and Variability of Personality Networks. A Tutorial on Recent Developments in Network Psychometrics. *Personality and Individual Differences, 136*, 68–78. https://doi.org/10.1016/j.paid.2017.06.011

De Ron, J., Robinaugh, D. J., Fried, E. I., Pedrelli, P., Jain, F. A., Mischoulon, D. y Epskamp, S. (2022). Quantifying and Addressing the Impact of Measurement Error in Network Models. *Behaviour Research and Therapy, 157,* Artículo 104163. https://doi.org/10.1016/J.BRAT.2022.104163

Elosua, P., Aguado, D., Fonseca-Pedrero, E., Abad, F. J. y Santamaría, P. (2023). New Trends in Digital Technology-Based Psychological and Educational Assessment. *Psicothema, 35*(1), 50–57. https://doi.org/10.7334/psicothema2022.241

Epskamp, S., Borsboom, D. y Fried, E. I. (2018). Estimating Psychological Networks and their Accuracy: A Tutorial Paper. *Behavior Research Methods, 50,* 195–212. 1–18. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S. y Fried, E. (2018). A Tutorial on Regularized Partial Correlation Networks. *Psychological Methods, 23*, 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Kruis, J. y Marsman, M. (2017). Estimating Psychopathological Networks: Be Careful what you Wish for. *PLoS ONE, 12*(6), Artículo e0179891. https://doi.org/10.1371/journal.pone.0179891

Epskamp, S., Rhemtulla, M. y Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika, 82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A. M., Riese, H. y Cramer, A. O. J. (2018). Personalized Network Modeling in Psychopathology: The Importance of Contemporaneous and Temporal Connections. *Clinical Psychological Science, 6*(3), 416–427. https://doi.org/10.1177/2167702617744325

Fonseca-Pedrero, E. (2017). Análisis de redes: ¿una nueva forma de comprender la Psicopatología? [Network Analysis: A New Way of Understanding Psychopathology?]. *Revista de Psiquiatría y Salud Mental, 10*(4), 206–215. https://doi.org/10.1016/j.rpsm.2017.06.004

Fonseca-Pedrero, E. (2018). Análisis de redes en Psicología [Network Analysis in Psychology]. *Papeles del Psicólogo, 39*, 1–12. https://doi.org/10.23923/pap.psicol2018.2852

Fonseca-Pedrero, E. y Al-Halabí, S. (2024). Sobre la conducta suicida y las conductas adictivas [On suicidal Behaviour and Addictive Behaviours]. *Adicciones, 36*(2), 121–128. https://doi.org/10.20882/adicciones.2074

Fonseca-Pedrero, E., Díez-Gómez, A., de la Barrera, U., Sebastian-Enesco, C., Ortuño-Sierra, J., Montoya-Castilla, I., Lucas-Molina, B., Inchausti, F. y Pérez-Albéniz, A. (2024). Suicidal Behaviour in Adolescents: A Network Analysis. *Spanish Journal of Psychiatry and Mental Health, 17*(1), 3–10. https://doi.org/10.1016/J.RPSM.2020.04.007

Fonseca-Pedrero, E., Pérez-Albéniz, A., Díez-Gómez, A., Al-Halabí, S., Lucas-Molina, B. y Calvo, P. (2023). Profesionales de la Psicología en Contextos Educativos: Una Necesidad Ineludible [Psychology Professionals in Educational Contexts: An

Unavoidable Necessity]. *Papeles del Psicólogo, 44*(3), 112–124. https://doi.org/10.23923/pap.psicol.3018

Fonseca-Pedrero, E., Ródenas-Perea, G., Pérez-Albéniz, A., Al-Halabí, S., Pérez, M. y Muñiz, J. (2022). La hora de la evaluación ambulatoria [The Time of Ambulatory Assessment]. *Papeles del Psicólogo, 43,* 21–28. https://doi.org/10.23923/pap.psicol.2983

Forbes, M. K., Wright, A. G. C., Markon, K. E. y Krueger, R. F. (2017). Evidence that Psychopathology Symptom Networks have Limited Replicability. *Journal of Abnormal Psychology, 126*(7), 969–988. https://doi.org/10.1037/abn0000276

Golino, H. F. y Epskamp, S. (2017). Exploratory Graph Analysis: A New Approach for Estimating the Number of Dimensions in Psychological Research. *PLoS ONE, 12*(6), Artículo e0174035. https://doi.org/10.1371/journal.pone.0174035

Goyal, S. (2023). *Networks: An Economics Approach.* MIT Press.

Hallquist, M. N., Wright, A. G. C. y Molenaar, P. C. M. (2021). Problems with Centrality Measures in Psychopathology Symptom Networks: Why Network Psychometrics Cannot Escape Psychometric Theory. *Multivariate Behavioral Research, 56*(2), 199–223. https://doi.org/10.1080/00273171.2019.1640103

Haslbeck, J. M. B., Borsboom, D. y Waldorp, L. J. (2021). Moderated Network Models. *Multivariate Behavioral Research, 56*(2), 256–287. https://doi.org/10.1080/00273171.2019.1677207

Haslbeck, J. M. B. y Fried, E. I. (2017). How Predictable are Symptoms in Psychopathological Networks? A Reanalysis of 18 Published Datasets. *Psychological medicine, 47*(16), 2767–2776. https://doi.org/10.1017/S0033291717001258

Hevey, D. (2018). Network Analysis: a Brief Overview and Tutorial. *Health Psychology and Behavioral Medicine, 6*(1), 301–328. https://doi.org/10.1080/21642850.2018.1521283

Hoekstra, R. H. A., Epskamp, S., Nierenberg, A. A., Borsboom, D. y McNally, R. J. (2024). Testing Similarity in Longitudinal Networks: The Individual Network Invariance Test. *Psychological Methods*. https://doi.org/10.1037/met0000638

Huth, K. B. S., de Ron, J., Goudriaan, A. E., Luigjes, J., Mohammadi, R., van Holst, R. J., Wagenmakers, E. J. y Marsman, M. (2023). Bayesian Analysis of Cross-Sectional Networks: A Tutorial in R and JASP. *Advances in Methods and Practices in Psychological Science, 6*(4). https://doi.org/10.1177/25152459231193334

Isvoranu, A., Epskamp, S., Waldorp, L. y Borsboom, D. (2022). *Network Psychometrics with R A Guide for Behavioral and Social Scientists*. Routledge.

Isvoranu, A. M., Abdin, E., Chong, S. A., Vaingankar, J., Borsboom, D. y Subramaniam, M. (2021). Extended Network Analysis: From Psychopathology to Chronic Illness. *BMC Psychiatry, 21*(1), Artículo 119. https://doi.org/10.1186/s12888-021-03128-y

Jackson, M. O. (2008). *Social and Economic Networks.* Princeton University Press.

Jones, P. J., Mair, P. y McNally, R. J. (2018). Visualizing Psychological Networks: A Tutorial in R. *Frontiers in Psychology, 9,* Artículo 1742. https://doi.org/10.3389/fpsyg.2018.01742

Kan, K. J., De Jonge, H., Van Der Maas, H. L. J., Levine, S. Z. y Epskamp, S. (2020). How to Compare Psychometric Factor and Network Models. *Journal of Intelligence, 8*(4), 1–10. https://doi.org/10.3390/JINTELLIGENCE8040035

Mansueto, A. C., Wiers, R. W., van Weert, J. C. M., Schouten, B. C. y Epskamp, S. (2023). Investigating the Feasibility of Idiographic Network Models. *Psychological Methods, 28*(5), 1052–1068. https://doi.org/10.1037/MET0000466

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., Maas, H. L. J. van der y Maris, G. (2018). An Introduction to Network Psychometrics: Relating Ising Network Models to Item Response Theory Models. *Multivariate Behavioral Research, 53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

McNally, R. J. (2016). Can Network Analysis Transform Psychopathology? *Behaviour Research and Therapy, 86,* 95–104. https://doi.org/10.1016/j.brat.2016.06.006

Mcnally, R. J. (2021). Network Analysis of Psychopathology: Controversies and Challenges. *Annual Review of Clinical Psychology, 17*, 31–53. https://doi.org/10.1146/ANNUREV-CLINPSY-081219-092850

Muñiz, J. (2010). Las teorías de los test: teoría clásica y teoría de respuesta a los ítems [Test Theories: Classical Theory and Item Response Theory]. *Papeles del Psicólogo, 31*(1), 57–66.

Muñiz, J. (2018). *Introducción a la Psicometría* [Introduction to Psychometrics]. Pirámide.

Neal, Z. P., Forbes, M. K., Neal, J. W., Brusco, M. J., Krueger, R., Markon, K., Steinley, D., Wasserman, S. y Wright, A. G. C. (2022). Critiques of Network Analysis of Multivariate Data in Psychological Science. *Nature Reviews Methods Primers 2,* Artículo 90. https://doi.org/10.1038/s43586-022-00177-9

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Piazza, G. G., Allegrini, A. G., Eley, T. C., Epskamp, S., Fried, E., Isvoranu, A. M., Roiser, J. P. y Pingault, J. B. (2024). Polygenic Scores and Networks of Psychopathology Symptoms. *JAMA Psychiatry, 81*(9), 902–910. https://doi.org/10.1001/JAMAPSYCHIATRY.2024.1403

Pérez-Álvarez M. (2018). La psicoterapia como ciencia humana, más que tecnológica [Psychotherapy as a Human Science, more than Technological one]. *Papeles del Psicólogo, 40*(1), 1–14 https://doi.org/10.23923/pap.psicol2019.2877

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R. y Borsboom, D. (2020). The Network Approach to Psychopathology: A Review of the Literature 2008-2018 and an Agenda for Future Research. *Psychological Medicine, 50*(3), 353–366. https://doi.org/10.1017/S0033291719003404

Scheffer, M., Elizabeth Bolhuis, J., Borsboom, D., Buchman, T. G., Gijzel, S. M. W., Goulson, D., Kammenga, J. E., Kemp, B., van de Leemput, I. A., Levin, S., Martin, C. M., Melis, R. J. F., van Nes, E. H., Michael Romero, L. y Olde Rikkert, M. G. M. (2018). Quantifying Resilience of Humans and Other Animals. *Proceedings of the National Academy of Sciences of the United States of America, 115*(47), 11883–11890. https://doi.org/10.1073/PNAS.1810630115

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., y Borsboom, D. (2013). Deconstructing the Construct: A Network Perspective on Psychological Phenomena. *New Ideas in Psychology, 31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Van Borkulo, C. D. (2018). Network Comparison Test: Permutation-Based Test of Differences in Strength of Networks. Retrieved from github.com/cvborkulo/ NetworkComparisonTest

Vega-Redondo, F. (2007). *Complex Social Networks.* Cambridge University Press.

# USO DE MODELOS MATEMÁTICOS COMO PARTE DEL ANÁLISIS DE DATOS EN PSICOLOGÍA: EL CASO DEL DESCUENTO POR DEMORA

# USE OF MATHEMATICAL MODELS AS PART OF DATA ANALYSIS IN PSYCHOLOGY: THE CASE OF DELAY DISCOUNTING

SERGIO RAMOS[1], GABRIELA E. LÓPEZ-TOLSA[1], ANTONIO MARTÍNEZ-HERRADA[1], FERNANDO MOLINES[1], MARLON PALOMINO[1] Y RICARDO PELLÓN[1]

Cómo referenciar este artículo/How to reference this article:

Ramos, S., López-Tolsa, G. E., Martínez-Herrada, A., Molines, F., Palomino, M. y Pellón, R. (2025). Uso de modelos matemáticos como parte del análisis de datos en Psicología: el caso del descuento por demora [Use of Mathematical Models as Part of Data Analysis in Psychology: The case of Delay Discounting]. *Acción Psicológica, 22*(1), 101–114. https://doi.org/10.5944/ap.22.1.43727

## Resumen

El uso de modelos matemáticos en Psicología básica ha experimentado un crecimiento significativo en los últimos años, tanto en la formulación teórica como en el análisis de datos. Sin embargo, el incremento de la complejidad de estos modelos, sumado a la falta de formación matemática en Psicología ha ampliado la brecha entre la investigación y la Psicología aplicada, dificultando el acceso a literatura experimental a profesionales ajenos a este enfoque. El presente artículo expone la aplicación de modelos matemáticos empleados en la investigación del paradigma del descuento por demora, con el propósito de analizar la elección impulsiva y sus implicaciones prácticas en contextos del mundo real, buscando familiarizar a los lectores con el uso de estos modelos matemáticos en el ámbito de la Psicología básica y aplicada. En primer lugar, se presenta una breve introducción teórica sobre los

**Correspondence address [Dirección para correspondencia]**: Gabriela E. López-Tolsa, Facultad de Psicología, Universidad Nacional de Educación a Distancia, España.
**Email**: glopez@psi.uned.es

**ORCID**: Sergio Ramos (https://orcid.org/0000-0002-2967-0783), Gabriela E. López-Tolsa (https://orcid.org/0000-0001-8997-2831), Antonio Martínez-Herrada (https://orcid.org/0000-0002-7742-5635), Fernando Molines (https://orcid.org/0000-0002-7136-4333), Marlon Palomino (https://orcid.org/0000-0002-1026-9631) y Ricardo Pellón (https://orcid.org/0000-0002-4099-7621).

[1] Universidad Nacional de Educación a Distancia, España.

Recibido: 13 de diciembre de 2024.
Aceptado: 31 de enero de 2025.

modelos matemáticos, incluyendo una breve descripción de su proceso de construcción y ajuste. Después se presenta el desarrollo y evolución de los diferentes modelos matemáticos desarrollados en el paradigma de descuento por demora. Más adelante se incluye una guía breve para ajustar estos modelos utilizando Excel. Por último, se destacan otros modelos de interés y se realiza una reflexión sobre los retos que presentan, su aplicabilidad y su dirección a futuro.

**Palabras clave:** Modelos matemáticos; Descuento por demora; Análisis cuantitativo de la conducta; Psicología.

## Abstract

The use of mathematical models in basic psychology has experienced significant growth in recent years, both in theoretical formulation and data analysis. However, the increasing complexity of these models, combined with the lack of mathematical training in psychology, has widened the gap between research and applied psychology, making it difficult for professionals outside this approach to access experimental literature. This article presents the application of mathematical models used in research within the delay discounting paradigm, with the aim of analyzing impulsive choice and its practical implications in real-world contexts. The goal is to familiarize readers with the use of these mathematical models in the field of both basic and applied psychology. First, a brief theoretical introduction to mathematical models is presented, including a short description of their construction and fitting process. Then, the development and evolution of various mathematical models designed within the delay discounting paradigm are discussed. Next, a short guide is provided on how to fit these models using Excel. Finally, other relevant models are highlighted, along with a reflection on the challenges they present, their applicability, and future directions.

**Keywords:** Mathematical Models; Delay Discounting; Quantitative Analysis of Behavior; Psychology.

## Introducción

La ciencia busca explicar la realidad (o lo más aproximado a ella) y, en la medida de lo posible, predecir fenómenos con base en hallazgos previos. Una herramienta clave para el quehacer científico es el uso de modelos matemáticos, pues permite expandir teorías de manera más eficiente que a través de la mera experimentación (Cavagnaro et al., 2013). Los modelos matemáticos permiten simplificar la realidad de forma que se puedan observar las relaciones entre variables claves, sugieren predicciones de forma simple y clara, que requerirían de mucha experimentación para llegar a ellas, y permiten explorar escenarios complejos de replicar (Cavagnaro et al., 2013). La Psicología, como ciencia, no es la excepción, por lo que el uso de modelos matemáticos brinda una herramienta para avanzar en el entendimiento de los distintos fenómenos psicológicos, con el fin último de entender el comportamiento para diseñar estrategias que mejoren el bienestar de las personas (Mazur, 2006).

Los modelos matemáticos son una descripción matemática de los procesos psicológicos que presenta un avance en términos de precisión respecto a las descripciones derivadas del modelado verbal (Cavagnaro et al., 2013; Mazur, 2006). El modelado verbal se refiere al uso de explicaciones y predicciones cualitativas basadas en el análisis estadístico centrado exclusivamente en la prueba de hipótesis y la interpretación de los valores p (Cavagnaro et al., 2013; véase también Benjamin et al., 2018; Wagenmakers et al., 2011). Un ejemplo sería: “las personas con TDAH que consumen drogas con frecuencia toman decisiones más impulsivas que las que no lo hacen”. Este enunciado plantea una hipótesis que puede ser probada, pero no especifica la magnitud de la diferencia. En contraste, el modelado matemático no solo permite hacer predicciones más precisas, sino también examinar, por ejemplo, qué factores están detrás de esas decisiones impulsivas (como la sensibilidad a la magnitud frente a la

sensibilidad a la demora); también permite explorar si las diferencias son constantes en diferentes contextos, identificando así elementos clave en las teorías que intentan explicar diversos fenómenos psicológicos (Cavagnaro et al., 2013; Mazur, 2006). Es decir, no sólo predeciría qué grupo es más impulsivo, sino en qué medida, y qué aspectos contextuales podrían aumentar o disminuir esas diferencias.

El uso de modelos matemáticos ha estado presente en Psicología desde la década de 1950, sin embargo, los recientes avances tecnológicos han favorecido que cada día se construyan y prueben más modelos (Batchelder, 2010; Cavagnaro et al., 2013). Entre las ventajas de usar modelos, Mazur (2013) identifica: construcción de teorías precisas y claras; identificación de distinciones teóricas y potenciales aplicaciones; prueba de aspectos específicos de la teoría; creación de marcos comunes en los que se describen fenómenos diferentes; y mejora en la comunicación entre la Psicología básica y aplicada ya que proporciona un marco formal y cuantitativo que facilita la traducción de términos teóricos en términos prácticos utilizados en la Psicología aplicada.

Existen diferentes tipos de modelos matemáticos, pero aquí nos centraremos solo en los modelos algebraicos, debido a que son ampliamente utilizados en Psicología básica por su versatilidad y ajuste a los datos empíricos. En este tipo de modelos se introducen los datos recogidos experimentalmente y, tras resolver una ecuación, se pueden derivar o estimar uno o más parámetros que se interpretan desde la teoría psicológica que respalda el modelo. Este tipo de modelos son una generalización de la regresión lineal, ya que predicen un resultado a partir de una combinación ponderada de las variables de entrada (los datos experimentales). No obstante, también pueden incluir términos no lineales y parámetros que describen de forma explícita factores psicológicos específicos (Cavagnaro et al., 2013). La construcción de un modelo, sin embargo, no es una tarea sencilla, ya que su formulación debe seguir principios teóricos generales bien fundamentados, con base en los cuales se describirán y predecirán hallazgos específicos (Killeen, 2023). Una vez que tenemos claros los principios teóricos, se plantean las predicciones que debe realizar nuestro modelo, y se escribe una ecuación que estará compuesta por parámetros observables y no observables. Los parámetros observables son aquellos que corresponden a los datos que podemos obtener experimentalmente, mientras que los no observables son aquellos que se derivan de los observables, y son precisamente los que nos proporcionan la información teórica. Además, es importante dar cuenta de las pequeñas diferencias que pueda haber entre sujetos y/o mediciones, es decir, la variabilidad, y a la estructura de esa variabilidad se le llamará error (Cavagnaro et al., 2013).

Una vez que se tiene el modelo, se ajustará a distintos conjuntos de datos (e.g., los datos de cada participante), lo que significa que se buscará que el resultado del modelo se parezca lo más posible a los datos – y nunca al revés –, de forma que sea posible extraer conclusiones con los parámetros no observables que derivemos de ellos. En la siguiente sección se analizará un ejemplo específico de un modelo matemático ampliamente utilizado en la actualidad, el modelo de descuento por demora, empleado para medir la impulsividad. Posteriormente, se presentará una breve guía sobre cómo ajustar dicho modelo utilizando Excel.

Por último, es crucial evaluar constantemente los modelos y comparar diferentes modelos para identificar el que mejor explica los datos. Si bien la meta final debería ser encontrar el modelo que describe perfectamente los datos, esta es una tarea prácticamente imposible, como lo resumía George E. P. Box en su conocida frase “todos los modelos están mal, pero algunos son útiles” (Cavagnaro et al., 2013, p. 449; ver, Box, 1976, para la reflexión completa), por lo que la meta real debería ser encontrar aquel que, además de acercarse lo más posible a la verdad, permita avanzar en la teorización del fenómeno, contribuyendo así a la mejora de la práctica psicológica.

## El descuento por demora como ejemplo de modelos matemáticos en Psicología

Existe una gran cantidad de modelos matemáticos que se usan actualmente en investigación básica y aplicada en Psicología (Mazur, 2006). Para ilustrar cómo algunos fenómenos psicológicos pueden estudiarse mediante el ajuste de modelos matemáticos, hemos centrado nuestro análisis en los modelos de descuento por demora, dado que

son unos de los más sencillos, validados y utilizados en la disciplina, y que tienen relación con la importante influencia de factores psicológicos en la toma de decisiones y la conducta impulsiva (Franck et al., 2015; Vanderveldt et al., 2016).

La impulsividad, definida como la preferencia por recompensas pequeñas e inmediatas en lugar de grandes pero demoradas, es un tema relevante en Psicología, debido a que es una tendencia común en trastornos como el TDAH y las adicciones (Evenden, 1999; Rung y Madden, 2018). Una de las principales variables que controla esta preferencia es el tiempo de demora para recibir el reforzador. Existe una relación inversa entre el tiempo de demora y el valor subjetivo del reforzador, de forma que a medida que el tiempo de entrega aumenta, el valor subjetivo del reforzador disminuye. Este fenómeno se conoce como descuento por demora. Una mayor sensibilidad a la demora – menor tolerancia al tiempo de espera – está asociada con un mayor grado de descuento, lo que caracteriza al individuo como más impulsivo. Este fenómeno se estudia tanto en seres humanos como en animales no humanos mediante tareas en las que el sujeto debe elegir entre dos opciones: una que ofrece un reforzador pequeño e inmediato y otra que proporciona un reforzador grande, pero demorado. El valor de la demora se ajusta de acuerdo con diferentes criterios para estudiar cambios en la preferencia (Renda et al., 2021; Sosa y dos Santos, 2019; Vanderveldt et al., 2016).

Con el objetivo de cuantificar la impulsividad y describir la relación matemática entre el valor del reforzador y la demora, se han desarrollado diversos modelos basados en fundamentos teóricos que han evolucionado con la evidencia empírica. Estos modelos se dividen principalmente en dos categorías: exponenciales e hiperbólicos, cada uno con diferentes supuestos y predicciones.

Los modelos exponenciales fueron los primeros en ser adoptados por economistas para estudiar las elecciones de los individuos. Estos modelos asumen que el valor del reforzador disminuye de manera constante y proporcional con cada unidad de tiempo, lo que implica una relación lineal entre el valor y la demora (Green y Myerson, 2004, 2010). La fórmula más básica es la derivada del Modelo de Utilidad Descontada (Samuelson, 1937):

$$V = Ae^{-kD},$$

donde, $V$ representa el valor subjetivo del reforzador demorado, medido como el porcentaje de veces que se elige la opción demorada; $A$ es la cantidad de la recompensa, medida a través del porcentaje máximo de veces que es elegida en ausencia de demora; $D$ es el tiempo de demora, que típicamente se mide en segundos, minutos, días o meses dependiendo del contexto experimental; y $k$ es un parámetro estimado, sin unidad, que representa la tasa de descuento. Según este modelo, los individuos mantendrán preferencias consistentes en el tiempo, independientemente de la cantidad de la recompensa. Sin embargo, este enfoque ha sido criticado porque no considera fenómenos observados en las experiencias diarias de las personas, como los cambios de preferencia, donde los individuos modifican sus elecciones dependiendo de la demora o el contexto (Green y Myerson, 1996; Vanderveldt et al., 2016). Las personas a menudo planean levantarse temprano para hacer ejercicio, pero al sonar la alarma apagan el despertador y optan por seguir durmiendo. Del mismo modo, pueden decidir seguir una dieta equilibrada para mejorar su salud, pero al llegar a la caja del supermercado terminan comprando el dulce estratégicamente colocado para tentar a la clientela. Según las predicciones de este modelo, si la persona decide la noche anterior que se levantará temprano para hacer ejercicio, esa preferencia debería permanecer invariable al momento de sonar la alarma. De forma similar, la elección de alimentos saludables debería mantenerse firme incluso cuando el dulce se presenta como una recompensa inmediata. Además, el modelo supone que la sensibilidad al tiempo es constante, lo que no se alinea con los datos empíricos (Green y Myerson, 1996).

En contraste, los modelos hiperbólicos, desarrollados posteriormente por profesionales de la Psicología, introducen un enfoque más dinámico. Estos modelos asumen que el valor del reforzador no disminuye de manera constante, sino que varía según la demora. La fórmula más básica de este enfoque fue propuesta por Mazur (1987):

$$V = \frac{A}{(1 + kD)},$$

donde, *V, A, D* y *k* representan las mismas variables que en la ecuación anterior. A diferencia del modelo exponencial, este modelo predice que la disminución del valor subjetivo será más pronunciada en demoras cortas y se atenuará a medida que la demora aumente, estableciendo una relación no lineal entre el valor y el tiempo. Además, permite predecir preferencias inconsistentes en el tiempo, un fenómeno ampliamente observado en los datos experimentales (McKerchar et al., 2009). En este caso, el parámetro *k* es crucial, ya que permite cuantificar el grado de impulsividad: valores más altos de *k* indican mayor preferencia por recompensas inmediatas, reflejando una mayor impulsividad. Esto sucede debido a que, si la diferencia entre dos valores de demora genera una disminución significativa en la elección de la recompensa demorada, la curva de descuento será más pronunciada, resultando en un mayor valor del parámetro estimado *k*. Por ello, a mayor valor de *k*, mayor impulsividad. Este modelo ofrece la ventaja de permitir predecir el punto en el que dos alternativas tienen el mismo valor subjetivo y, por lo tanto, el cambio en la preferencia de los sujetos entre ellas (Sjoberg y Johansen, 2018).

Sin embargo, el modelo de Mazur (1987) presenta una limitación importante debido a que no considera la magnitud de la recompensa como un factor relevante en la elección. Varios estudios han demostrado que las cantidades mayores tienden a descontarse de manera menos pronunciada que las cantidades más pequeñas (Green y Myerson, 1996; Mitchell, 2017; Raineri y Rachlin, 1993; Thaler, 1981). En otras palabras, el descuento por demora no solo depende de la demora, sino también del tamaño de la recompensa.

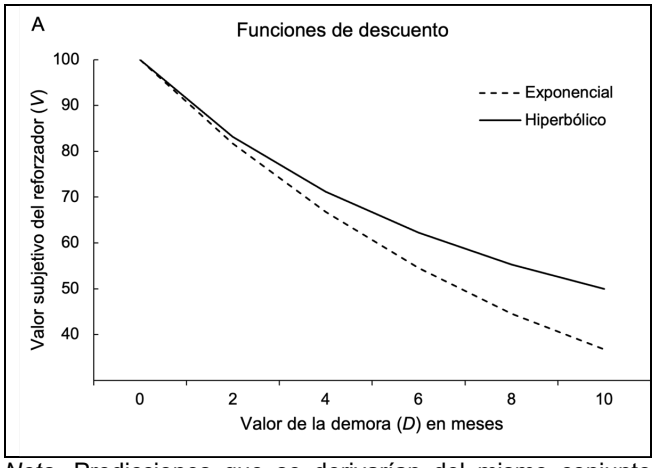
Para superar esta limitación, Myerson y Green (1995) propusieron el Modelo Hiperbólico Generalizado, que extiende el modelo de Mazur (1987) al incluir un parámetro adicional que captura la sensibilidad subjetiva al tiempo. La fórmula de este modelo es la siguiente:

$$V = \frac{A}{(1 + kD)^s},$$

donde, *s* es un parámetro estimado adicional que ajusta la sensibilidad al valor de la demora. Cuando *s* = 1, el modelo se reduce al de Mazur (1987), pero valores menores de *s* permiten representar una percepción más compleja del tiempo y de la magnitud de la recompensa. Este modelo asume que la percepción del tiempo y de la magnitud de la recompensa no es lineal, sino que varía de manera subjetiva según el individuo y el contexto. Al incluir *s*, el modelo se vuelve más flexible y puede capturar patrones de comportamiento que los modelos anteriores no lograban explicar, como la menor tasa de descuento observada con recompensas más grandes (Ballard et al., 2023; McKerchar et al., 2009).

**Figura 1**

*Comparación de las funciones matemáticas derivadas de los modelos Exponencial e Hiperbólico*




*Nota*. Predicciones que se derivarían del mismo conjunto hipotético de datos de descuento por demora.

La Figura 1 muestra una comparación de las funciones matemáticas derivadas de estos tipos modelos, ilustrando cómo los cambios en los supuestos teóricos afectan las predicciones sobre el comportamiento. Al analizar el comportamiento impulsivo, es esencial seleccionar el modelo adecuado, ya que cada uno ofrece diferentes conclusiones sobre los mecanismos que subyacen a las elecciones intertemporales.

A pesar de que el Modelo Hiperbólico Generalizado (Myerson y Green, 1995) ofrece importantes ventajas teóricas y empíricas, el modelo de Mazur (1987) sigue siendo el más ampliamente utilizado en Psicología, tanto en el ámbito experimental como en el clínico (Vanderveldt et

al., 2016). Esto se debe a que el modelo de Mazur logra un equilibrio óptimo entre precisión explicativa y parsimonia, lo que lo hace más accesible y práctico para su aplicación en una variedad de contextos psicológicos. Aunque los modelos más avanzados, como el Hiperbólico Generalizado, pueden proporcionar un mejor ajuste a los datos en ciertos casos, no siempre son necesarios para describir fenómenos básicos de descuento temporal (Green y Myerson, 2004). Esto refuerza la idea de que el modelo de Mazur (1987) es una herramienta altamente eficiente, especialmente en estudios centrados en análisis grupales o exploratorios. Su amplia aceptación refleja la preferencia de las personas psicólogas por modelos que ofrezcan una combinación equilibrada de simplicidad, facilidad de aplicación y capacidad predictiva, consolidándolo como una opción estándar en el estudio de las elecciones impulsivas. Sin embargo, siempre es recomendable elegir el modelo que describa mejor los datos en cuestión (Franck et al., 2015).

## Aplicaciones de los modelos matemáticos de descuento por demora fuera del laboratorio

Las condiciones controladas de laboratorio son ideales para construir y probar modelos matemáticos que den cuenta del comportamiento, ya que permiten aislar las variables relevantes, así como modificar los parámetros del modelo reduciendo potenciales fuentes de error. Sin embargo, la utilidad y poder explicativo de los modelos matemáticos en Psicología trasciende los contextos de laboratorio.

Para ilustrar la aplicación e interpretación del modelo hiperbólico de Mazur (1987), consideremos de manera hipotética su uso en una tarea de descuento por demora para analizar si los adolescentes con TDAH que consumen drogas son más impulsivos que aquellos que no lo hacen. Como resultado obtendríamos una k por cada sujeto o grupo. Si el grupo con consumo de drogas obtuviera una $k = 0.30$ frente a una $k = 0.15$ en el grupo sin consumo, se podría concluir que el primero muestra una mayor impulsividad, con una mayor sensibilidad a la demora, y, por lo tanto, un comportamiento más controlado por reforzadores inmediatos.

De forma similar, Reed y Martens (2011) demostraron cómo el modelo hiperbólico puede aplicarse en el ámbito educativo para evaluar la efectividad de distintos reforzadores en estudiantes durante tareas académicas. En su estudio, los participantes obtenían puntos por completar una tarea de matemáticas, los cuales podían ser canjeados inmediatamente o tras una demora de 24 horas. Los resultados revelaron que los reforzadores demorados eran menos efectivos en estudiantes con valores altos de *k*, destacando la necesidad de considerar la impulsividad al diseñar estrategias de intervención educativa.

Además, numerosos trabajos han trasladado estos modelos al ámbito aplicado, abordando problemáticas sociales y de salud. Estudios como los de Howatt et al. (2019) han destacado la relación entre la adherencia a hábitos saludables, como la dieta mediterránea, y el autocontrol, sugiriendo que las tasas de descuento podrían utilizarse para identificar poblaciones en riesgo que requieren intervenciones específicas. Del mismo modo, Albrecht e Iyengar (2021) demostraron que la comprensión de los mecanismos de descuento es clave para desarrollar estrategias de tratamiento y políticas públicas efectivas frente a problemas complejos como la obesidad infantil y el consumo de tabaco, conectando la Psicología, la economía y la salud pública. Finalmente, Berry et al. (2019) encontraron que pasar tiempo en entornos naturales en la ciudad reduce las tasas de descuento en decisiones relacionadas con la calidad del aire, sugiriendo que incorporar espacios verdes en áreas urbanas puede fomentar comportamientos más orientados al futuro relacionados con la salud ambiental y pública.

En conjunto, estos ejemplos demuestran la relevancia de trasladar conceptos y herramientas de la Psicología básica a aplicaciones prácticas, mostrando que los modelos matemáticos no solo permiten comprender fenómenos conductuales básicos, sino que también son una herramienta valiosa para diseñar intervenciones que impacten positivamente en la calidad de vida individual y social.

## Guía breve para ajustar el modelo de Mazur (1987) usando Excel

Aquí presentamos una breve guía sobre cómo ajustar un modelo por primera vez, utilizando la función *solver*[9] en Microsoft Excel. La guía está diseñada para aquellas personas que no están familiarizadas con el uso de modelos, pero que desean empezar a trabajar con ellos.

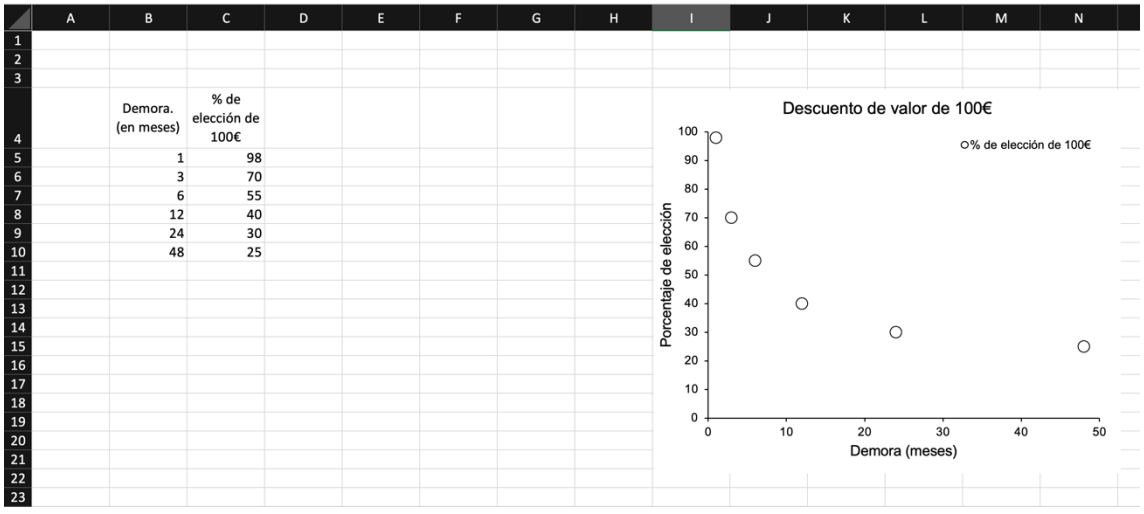
### *Preparando los datos*

Antes de proceder a ajustar un modelo es fundamental recordar que, si bien los modelos matemáticos son herramientas para mejorar la comprensión de los datos, no sustituyen a los mismos. Por ello, es indispensable contar con una base de datos recogidos con rigor científico, así como comprensión profunda de los mismos para interpretar adecuadamente los resultados del modelo. En el ejemplo desarrollado en esta guía, se trabajará con datos hipotéticos de descuento por demora, para aplicar el Modelo Hiperbólico de Mazur (1987), abordado en la sección anterior.

Imaginemos que se ha preguntado en repetidas ocasiones a una persona si preferiría 1 euro ahora o 100 euros después de un tiempo determinado. Los datos hipotéticos correspondientes se muestran en las columnas B y C de la Figura 2. Se puede observar que entre 1 euro ahora y 100 euros en 1 mes, la persona elegiría los 100 euros en el 98 % de las ocasiones. Sin embargo, ante la misma comparación, pero aumentando la demora a 48 meses para los 100 euros (1 euro ahora *vs.*, 100 euros en 48 meses) la persona elegiría los 100 euros en el 25 % de las ocasiones. Adicionalmente, en la Figura 2 (columnas I a N), los mismos datos hipotéticos han sido representados gráficamente (círculos blancos), lo que permite evaluar visualmente el ajuste del modelo. Se debe considerar que, dado que los círculos blancos representan los datos, estos deben mantenerse igual durante todo el ejercicio.

**Figura 2**

*Preparación de los datos en Excel*




*Nota.* Datos hipotéticos. Libro de Excel disponible en: https://osf.io/d2jgb/.

[9] Para activar la herramienta solver es recomendable consultar la página de soporte de Microsfot Office: https://support.microsoft.com/es-es/office/carga-del-complemento-solver-en-excel-2016-612926fc-d53b-46b4-872c-e24772f078ca#OfficeVersion=Windows.

## *Entendiendo el modelo*

Una vez organizados los datos en Excel, se deben identificar los parámetros del modelo y cómo se relacionan con los datos reales. En este ejemplo se usará el modelo hiperbólico de Mazur (1987), el cual, como se ha mencionado previamente, requiere de tres parámetros para calcular el valor subjetivo de una recompensa ($V$). El modelo se puede presentar así:

$$V = A/(1+kD).$$

En este caso, el valor de la recompensa ($V$) se medirá como el porcentaje de veces que la persona elige los 100 euros (columna C, Figura 2). Si el valor de 100 euros permanece intacto, debido a que se entrega de forma inmediata (demora, $D = 0$), esta opción debería ser elegida en todas las ocasiones (100 %). Por lo tanto, el parámetro A, que representa el máximo porcentaje de elección, será igual a 100 (%). Esto se puede comprobar realizando los siguientes cálculos:

$$V = 100/(1+k*0) = 100/1 = 100,$$

siendo en este caso $D = 0$ porque se entregó el reforzador de forma inmediata, lo cual implica que no es necesario conocer el valor de $k$ para calcular el denominador, ya que cualquier número multiplicado por cero resulta en cero, y 1 + 0 = 1; de forma que 100/1 = 100.

Una vez comprobado que los cálculos son adecuados, se procede a ajustar el modelo. Para determinar el valor hipotético de $V$, contamos con dos parámetros conocidos: $A$ y $D$. El parámetro $A$ corresponde siempre al máximo porcentaje de elección en ausencia de demora (100), mientras que $D$ corresponde a los datos en la columna B. Sin embargo, aún es necesario calcular el valor de $k$. Dado que es la primera vez que se trabaja con este modelo, y con un fin didáctico, inicialmente se explorará el valor de $k$ de manera manual, probando diferentes valores. Posteriormente, se automatizará el ajuste empleando la herramienta solver de Excel.

Comenzaremos con un valor inicial de $k = 0.02$ (celda "C13", Figura 3A), que al ser un valor muy pequeño debería generar una curva con una pendiente muy poco pronunciada. A continuación, se insertará una fórmula en la celda "D5", como se muestra en la Figura 3A: "= 100/(1+($C$13*B5))". En esta fórmula, los símbolos de dólar "$" indican que la fórmula debe usar siempre la celda C13, que corresponde a nuestra $k = 0.02$, la cual multiplicará a la demora registrada en la celda B5 (Figura 3A). Tras insertar esta fórmula, se procederá a arrastrarla hacia abajo, de forma que se calcule el mismo dato para las seis demoras, teniendo en cuenta que $k$ siempre va a estar en C13, pero D va a ir cambiando de fila (D6, D7…D10). El resultado debe verse como se muestra en la Figura 3B. Si se grafican estos datos (línea punteada, Fi-

**Figura 3**

*Ajuste del modelo con una k pequeña*

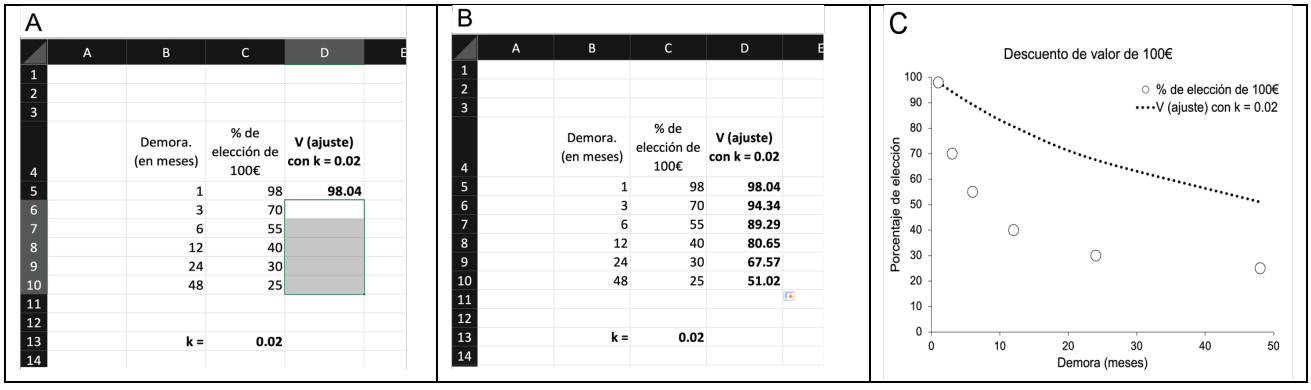
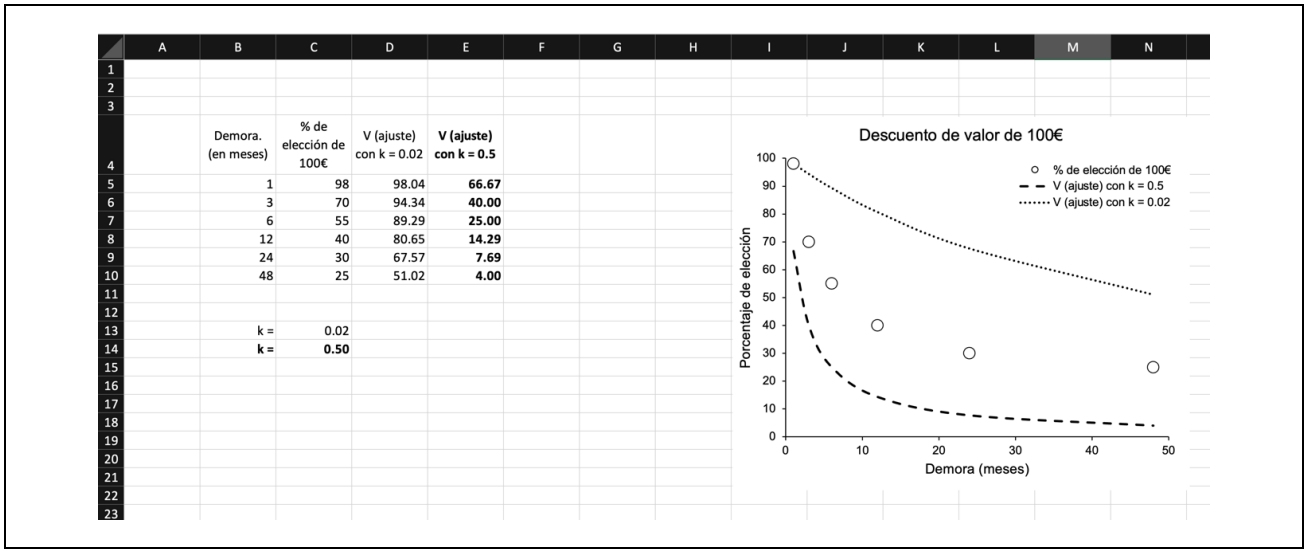


**Figura 4**

*Ajuste del modelo con una k grande*




| | Demora. (en meses) | % de elección de 100€ | V (ajuste) con k = 0.02 | **V (ajuste) con k = 0.5** |
|---|---|---|---|---|
| 5 | 1 | 98 | 98.04 | **66.67** |
| 6 | 3 | 70 | 94.34 | **40.00** |
| 7 | 6 | 55 | 89.29 | **25.00** |
| 8 | 12 | 40 | 80.65 | **14.29** |
| 9 | 24 | 30 | 67.57 | **7.69** |
| 10 | 48 | 25 | 51.02 | **4.00** |
| 13 | k = | 0.02 | | |
| 14 | **k =** | **0.50** | | |

gura 3C), se observará que la curva empieza en el mismo lugar que los datos reales, pero que no se sobrepone al resto de los datos. Esto indica que el modelo, cuando $k = 0.02$, no se ajusta adecuadamente a nuestros datos.

Ya se ha comprobado que $k = 0.02$ es demasiado pequeña, por lo que se utilizará un valor de k más grande, como por ejemplo $k = 0.5$ (celda C14, Figura 4). En este caso, se insertará una fórmula similar a la anterior, pero esta vez iniciando en la celda E5, y utilizando "$C$14" para incluir $k = 0.5$ en la fórmula. Una vez hecho esto, y habiendo arrastrado la fórmula hasta E10, se deberían obtener los datos mostrados en la Figura 4. En la gráfica de la Figura 4 (línea discontinua), se observa que el modelo tampoco logra un ajuste adecuado a los datos reales, ya que la curva generada se encuentra por debajo de estos.

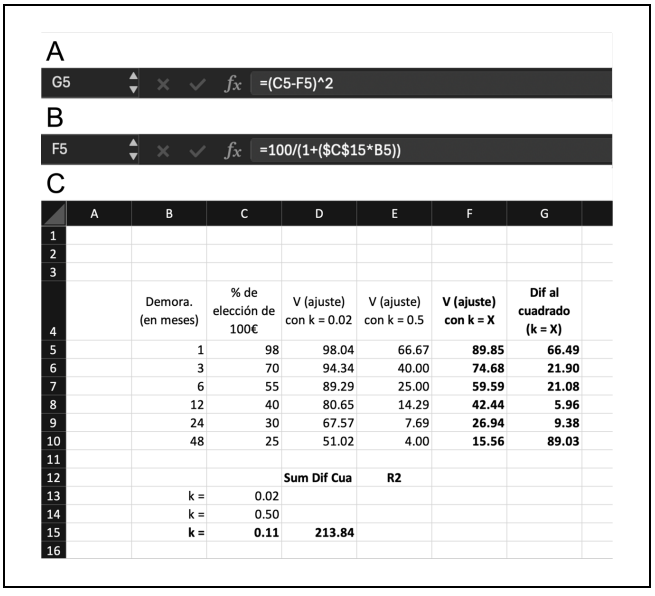
Con estos dos ejemplos, hemos ilustrado cómo la forma de la curva varía al modificar el valor de *k*; sin embargo, aún no hemos identificado un valor que proporcione un buen ajuste. Aunque sería posible encontrarlo mediante ensayo y error, este método resultaría demasiado tedioso. Por ello, utilizaremos la herramienta solver, disponible en el menú de Datos de Excel.

## *Ajuste del modelo usando la técnica de mínimos cuadrados y solver*

Como se ha mencionado previamente, ajustar un modelo implica que los valores obtenidos con el modelo sean lo más parecidos posible a los datos reales, y para ello utilizaremos la técnica estadística de los mínimos cuadrados (ver pp. 117-118 en Pardo y San Martín, 2004). Comenzaremos con un valor de k intermedio entre los dos probados previamente, por ejemplo, $k = 0.25$, (celda C15, Figura 5C). En la columna F, se insertará la fórmula que hemos usado anteriormente, pero usando C15 para el valor de *k* (Figura 5A). A continuación, se calcularán las diferencias al cuadrado. Para ello, en la columna G, se restará el porcentaje de elección (columna C), menos el valor de V obtenido con $k = 0.25$ (columna F), y el resultado se elevará al cuadrado: =(C5-F5)^2, como se muestra en la figura 5B. Por último, se sumarán esas diferencias al cuadrado y el resultado se colocará en la celda D15 (Figura 5C). La técnica de los mínimos cuadrados consiste en minimizar esta suma, buscando que el resultado sea lo más pequeño posible, lo que indicará un mejor ajuste del modelo a los datos.

**Figura 5**

*Ajuste del modelo y cálculo de la suma de la diferencia de cuadrados*



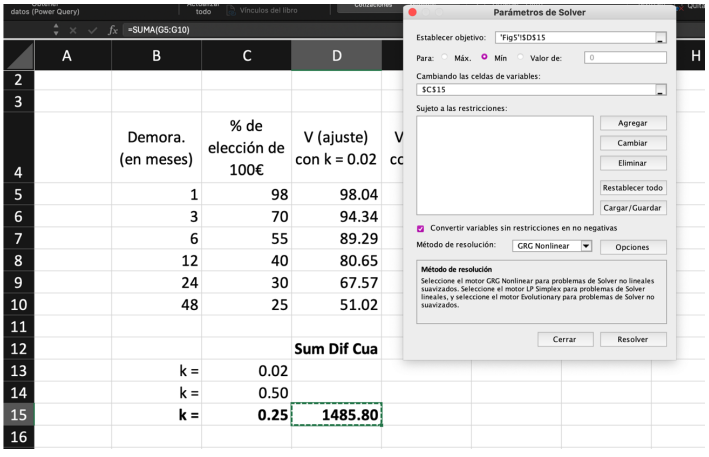

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | | | | | | |
| 3 | | | | | | | |
| 4 | | Demora. (en meses) | % de elección de 100€ | V (ajuste) con k = 0.02 | V (ajuste) con k = 0.5 | **V (ajuste) con k = X** | **Dif al cuadrado (k = X)** |
| 5 | | 1 | 98 | 98.04 | 66.67 | **89.85** | **66.49** |
| 6 | | 3 | 70 | 94.34 | 40.00 | **74.68** | **21.90** |
| 7 | | 6 | 55 | 89.29 | 25.00 | **59.59** | **21.08** |
| 8 | | 12 | 40 | 80.65 | 14.29 | **42.44** | **5.96** |
| 9 | | 24 | 30 | 67.57 | 7.69 | **26.94** | **9.38** |
| 10 | | 48 | 25 | 51.02 | 4.00 | **15.56** | **89.03** |
| 11 | | | | | | | |
| 12 | | | | **Sum Dif Cua** | **R2** | | |
| 13 | | k = | 0.02 | | | | |
| 14 | | k = | 0.50 | | | | |
| 15 | | **k =** | **0.11** | **213.84** | | | |
| 16 | | | | | | | |

Para realizar el análisis, se accederá al menú Datos y se seleccionará la herramienta solver, lo que abrirá la ventana mostrada en la Figura 6. Se establecerá como celda objetivo D15, donde se encuentra la suma de las diferencias de los cuadrados, y se seleccionará "Min", para indicar que se busque el mínimo valor posible que aparece en esa celda. En el campo "Cambiando las celdas de variables:", se seleccionará la celda C15, de manera que solver modifique el valor de esta celda. Finalmente, se hará clic en el botón "resolver", y tras aceptar la solución de solver, se observará como ahora el resultado es una $k = 0.11$. El procedimiento seguido por solver consiste en probar distintos valores de $k$ (C15) hasta encontrar aquel en el que el valor de D15 sea lo más pequeño posible, logrando así el mejor ajuste posible del modelo.

Al graficar los valores obtenidos en la columna F, que representan las predicciones del modelo cuando $k = 0.11$, se observa que se aproximan considerablemente a los datos reales (línea continua, Figura 7). Para evaluar con mayor precisión el grado de ajuste entre los datos observados y los predichos, se calcula el coeficiente de determinación ($R^2$) utilizando la fórmula "=COEFICIENTE.R2", en Excel). Este coeficiente indica la proporción de variabilidad de los datos que es explicada por el modelo. Un valor de $R^2 = 1$ implica que el modelo explica la totalidad de la variabilidad, por lo que cuanto más cercano sea el valor a 1, mejor será el ajuste del modelo. En este caso, se obtiene

**Figura 6**

*Ventana de* solver *en Excel*

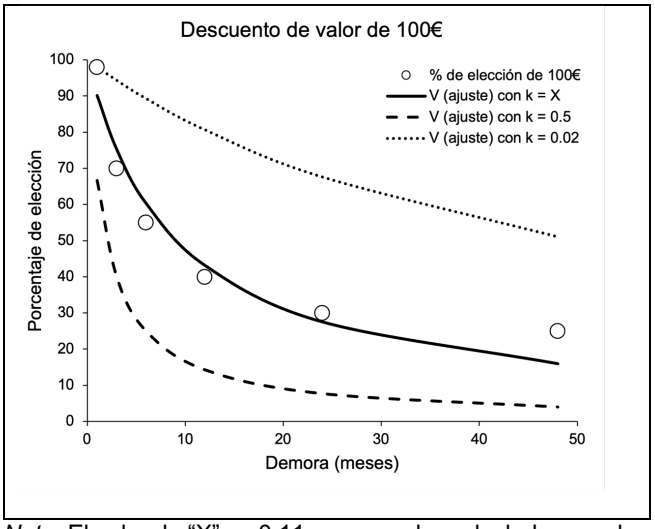



| | A | B | C | D |
|---|---|---|---|---|
| 2 | | | | |
| 3 | | | | |
| 4 | | Demora. (en meses) | % de elección de 100€ | V (ajuste) con k = 0.02 |
| 5 | | 1 | 98 | 98.04 |
| 6 | | 3 | 70 | 94.34 |
| 7 | | 6 | 55 | 89.29 |
| 8 | | 12 | 40 | 80.65 |
| 9 | | 24 | 30 | 67.57 |
| 10 | | 48 | 25 | 51.02 |
| 11 | | | | |
| 12 | | | | **Sum Dif Cua** |
| 13 | | k = | 0.02 | |
| 14 | | k = | 0.50 | |
| 15 | | **k =** | **0.25** | **1485.80** |
| 16 | | | | |

un $R^2 = 0.95$, lo que confirma que el modelo tiene un buen ajuste.

**Figura 7**

*Comparación de ajustes del modelo usando tres valores distintos de k*




*Nota.* El valor de "X" es 0,11, como se ha calculado usando *solver*.

## Otros modelos matemáticos de interés

A lo largo de este artículo, hemos utilizado los modelos destinados a medir la elección impulsiva, como el de descuento temporal, para ilustrar la aplicación de modelos matemáticos en Psicología. No obstante, estos representan solo una parte de las herramientas disponibles, ya que la disciplina cuenta con una amplia variedad de modelos diseñados para explorar diferentes dimensiones del comportamiento.

Entre los muchos existentes, queremos destacar dos modelos de utilización actual: el modelo del valor esencial del reforzador (Hursh y Silberberg, 2008) y el de la ley de igualación temporalmente ponderada (TWML; Shahan, 2022). El primero, desarrollado en el marco de la economía conductual, evalúa la capacidad de un reforzador para elicitar y mantener una respuesta en función de su coste. Este enfoque utiliza curvas de demanda para modelar la relación entre el consumo del reforzador y el esfuerzo necesario para obtenerlo, determinando así su valor esencial (Strickland y Lacy, 2020). El modelo de demanda exponencial de Hursh y Silberberg (2008) cuantifica este valor mediante un parámetro estimado, donde menores valores reflejan una mayor demanda, independiente de la presencia de reforzadores alternativos. Ha sido aplicado en contextos como la comparación de la demanda de sustancias, por ejemplo, tabaco frente a sustitutos como parches de nicotina, contribuyendo significativamente al estudio del abuso de sustancias y otros problemas clínicos. Por su parte, la TWML explica y predice el estallido de extinción, un aumento temporal en la tasa de respuesta al inicio de un procedimiento de extinción antes de que el comportamiento disminuya. Este modelo amplía la ley de igualación clásica al dar mayor peso a las experiencias recientes, calculando el valor dinámico de los comportamientos que compiten por un reforzador. Este modelo muestra cómo la tasa de respuesta depende del valor del comportamiento objetivo y de factores que compiten por el tiempo del organismo. Aunque su validación empírica sigue en desarrollo, la TWML ofrece un marco prometedor para predecir conductas disruptivas en el contexto de intervenciones diseñadas para disminuir ciertos comportamientos (Fisher et al., 2022).

## Comentarios finales

El uso de modelos matemáticos en Psicología, a pesar de su larga historia y utilidad para analizar la influencia de variables críticas, no ha conseguido ser una estrategia dominante en el tratamiento de los datos en Psicología. Esto puede atribuirse a que todavía su uso se enfrenta a diversos retos relacionados con su ajuste y su implementación. Un desafío importante es la falta de bases de datos exhaustivas que permitan probar y validar modelos con un rango amplio de condiciones experimentales. Esto puede llevar al uso de valores hipotéticos (e.g., demoras en los modelos de descuento por demora) en lugar de datos reales, lo que, si bien facilita los ajustes iniciales, limita la generalización de los resultados. Asimismo, la elección del método de ajuste, como los mínimos cuadrados, presenta sus limitaciones. El método de mínimos cuadrados es el más ampliamente utilizado debido a que es más accesible y fácil de implementar con conocimientos básicos. Sin embargo,

su simplicidad estadística puede comprometer la robustez de los análisis, ya que no permite calcular intervalos de confianza con precisión (Busemeyer y Diederich, 2014; Myung, 2003). Por otro lado, diseñar experimentos para discriminar entre modelos similares puede resultar particularmente complejo, especialmente cuando pequeñas diferencias en los datos pueden favorecer un modelo sobre otro. De esta forma, resulta esencial entender que los modelos no ofrecen explicaciones perfectas, sino aproximaciones, por lo que las conclusiones deben fundamentarse en la teoría (Killeen, 2023). Por último, un desafío asociado a las competencias de las personas psicólogas radica en que la expansión y mejora de los modelos requiere tanto habilidades matemáticas avanzadas como acceso a recursos tecnológicos especializados, lo que limita su implementación a un porcentaje reducido de profesionales en el ámbito de la Psicología (Cavagnaro et al., 2013; Mazur, 2006).

Estas barreras dificultan la integración más amplia de los modelos matemáticos en la Psicología. No obstante, superarlas permite aprovechar plenamente su potencial como herramientas fundamentales para analizar variables críticas, validar teorías y operativizar conceptos psicológicos de manera cuantitativa, lo que facilita realizar predicciones precisas e incluso, en ocasiones, contraintuitivas. Un ejemplo destacado es el presentado en este artículo, el modelo de descuento por demora, que ilustra cómo estas herramientas pueden aplicarse tanto en la investigación básica como en áreas aplicadas, ofreciendo un marco robusto para comprender la impulsividad y la toma de decisiones. Este artículo busca contribuir a superar las barreras relacionadas con las competencias de los estudiantes y profesionales en Psicología, ya que ello no solo impulsaría la investigación hacia direcciones innovadoras, sino que también reforzaría el valor de los modelos matemáticos como un puente entre los datos empíricos, las teorías psicológicas y su aplicación en áreas aplicadas.

## Referencias

Albrecht, N. M. y Iyengar, B. S. (2021). Pediatric Obesity: An Economic Perspective. *Frontiers in Public Health, 8,* Artículo 619647. https://doi.org/10.3389/fpubh.2020.619647

Ballard, T., Luckman, A. y Konstantinidis, E. (2023). A Systematic Investigation into the Reliability of inter-Temporal Choice Model Parameters. *Psychonomic Bulletin & Review, 30*(4), 1294–1322. https://doi.org/10.3758/s13423-022-02241-7

Batchelder, W. H. (2010). Mathematical Psychology. *WIREs Cognitive Science, 1*(5), 759–765. https://doi.org/10.1002/wcs.46

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berry, M. S., Repke, M. A. y Conway, L. G. (2019). Visual Exposure to Natural Environments Decreases Delay Discounting of Improved Air Quality. *Frontiers in Public Health, 7*, Artículo 308. https://doi.org/10.3389/fpubh.2019.00308

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Busemeyer, J. R. y Diederich, A. (2014). Estimation and Testing of Computational Psychological Models. En P. W. Glimcher y E. Fehr (Eds.), *Neuroeconomics* (pp. 49–61). Elsevier. https://doi.org/10.1016/B978-0-12-416008-8.00004-8

Cavagnaro, D. R., Myung, J. I. y Pitt, M. A. (2013). Mathematical Modeling. En T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (Vol. 1): *Foundations* (pp. 438–453). Oxford University Press.

Evenden, J. L. (1999). Varieties of Impulsivity. *Psychopharmacology, 146*(4), 348–361. https://doi.org/10.1007/PL00005481

Fisher, W. W., Greer, B. D., Mitteer, D. R. y Fuhrman, A. M. (2022). Translating Quantitative Theories of Behavior into Improved Clinical Treatments for Problem Behavior. *Behavioural Processes, 198*, Artículo 104639. https://doi.org/10.1016/j.beproc.2022.104639

Franck, C. T., Koffarnus, M. N., House, L. L. y Bickel, W. K. (2015). Accurate Characterization of Delay Discounting: A multiple Model Approach Using Approximate Bayesian Model Selection and a Unified Discounting Measure. *Journal of the Experimental Analysis of Behavior, 103*(1), 218–233. https://doi.org/10.1002/jeab.128

Green, L. y Myerson, J. (1996). Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. *American Zoologist, 36*(4), 496–505. https://doi.org/10.1093/icb/36.4.496

Green, L. y Myerson, J. (2004). A Discounting Framework for Choice With Delayed and Probabilistic Rewards. *Psychological Bulletin, 130*(5), 769–792. https://doi.org/10.1037/0033-2909.130.5.769

Green, L. y Myerson, J. (2010). Experimental and Correlational Analyses of Delay and Probability Discounting. En G. J. Madden y W. K. Bickel (eds.), *Impulsivity: The Behavioral and Neurological Science of Discounting*. (pp. 67–92). American Psychological Association. https://doi.org/10.1037/12069-003

Howatt, B. C., Muñoz Torrecillas, M. J., Cruz Rambaud, S. yTakahashi, T. (2019). A New Analysis on Self-Control in Intertemporal Choice and Mediterranean Dietary Pattern. *Frontiers in Public Health, 7,* Artículo 165 https://doi.org/10.3389/fpubh.2019.00165

Hursh, S. R. y Silberberg, A. (2008). Economic Demand and Essential Value. *Psychological Review, 115*(1), 186–198. https://doi.org/10.1037/0033-295X.115.1.186

Killeen, P. R. (2023). Theory of Reinforcement Schedules. *Journal of the Experimental Analysis of Behavior, 120,* 289–319. https://doi.org/10.1002/jeab.880

Mazur, J. E. (1987). An Adjusting Procedure for Studying Delayed Reinforcement. En M. L. Commons, J. E. Mazur, J. A. Nevin y H. Rachlin (Eds.), *The Effect of Delay and of Intervening Events on Reinforcement Value* (pp. 55–73). Erlbaum.

Mazur, J. E. (2006). Mathematical Models and the Experimental Analysis of Behavior. *Journal of the Experimental Analysis of Behavior, 85*(2), 275–291. https://doi.org/10.1901/jeab.2006.65-05

McKerchar, T. L., Green, L., Myerson, J., Pickford, T. S., Hill, J. C. y Stout, S. C. (2009). A Comparison of Four Models of Delay Discounting in Humans. *Behavioural Processes, 81*(2), 256–259. https://doi.org/10.1016/j.beproc.2008.12.017

Mitchell, S. H. (2017). Devaluation of Outcomes Due to Their Cost: Extending Discounting Models Beyond Delay. En J. R. Stevens (Ed.), *Impulsivity: How Time and Risk Influence Decision Making* (pp. 145–161). Springer. https://doi.org/10.1007/978-3-319-51721-6_5

Myerson, J. y Green, L. (1995). Discounting of Delayed Rewards: Models of Individual Choice. *Journal of the Experimental Analysis of Behavior, 64*(3), 263–276. https://doi.org/10.1901/jeab.1995.64-263

Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. J*ournal of Mathematical Psychology, 47*(1), 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7

Pardo, A. y Castellanos, R. S. M. (2004). *Análisis de Datos en Psicología II* [Data Analysis in Psychology II]. Pirámide.

Raineri, A. y Rachlin, H. (1993). The Effect of Temporal Constraints on the Value of Money and Other Commodities. *Journal of Behavioral Decision Making, 6*(2), 77–94. https://doi.org/10.1002/bdm.3960060202

Reed, D. D. y Martens, B. K. (2011). Temporal Discounting Predicts Student Responsiveness to Exchange Delays in a Classroom Token System. *Journal of Applied Behavior Analysis, 44*(1), 1–18. https://doi.org/10.1901/jaba.2011.44-1

Renda, C. R., Rung, J. M., Peck, S. y Madden, G. J. (2021). Reducing Impulsive Choice VII: Effects of Duration of Delay-Exposure Training. *Animal Cognition, 24*(1), 11–21. https://doi.org/10.1007/s10071-020-01412-0

Rung, J. M. y Madden, G. J. (2018). Experimental Reductions of Delay Discounting and Impulsive Choice: A Systematic Review and Meta-Analysis. *Journal of Experimental Psychology: General, 147*(9), 1349–1381. https://doi.org/10.1037/xge0000462

Samuelson, P. A. (1937). A Note on Measurement of Utility. *The Review of Economic Studies, 4*(2), 155–161. https://doi.org/10.2307/2967612

Shahan, T. A. (2022). A Theory of the Extinction Burst. *Perspectives on Behavior Science, 45*(3), 495–519. https://doi.org/10.1007/s40614-022-00340-3

Sjoberg, E. A. y Johansen, E. B. (2018). Impulsivity or Sub-optimal Reward Maximization in Delay Discounting? A Critical Discussion. *Human Ethology Bulletin, 33*(2), 22–36. https://doi.org/10.22330/heb/332/022-036

Sosa, R. y dos Santos, C. V. (2019). Toward a Unifying Account of Impulsivity and the Development of Self-Control. *Perspectives on Behavior Science, 42*(2), 291–322. https://doi.org/10.1007/s40614-018-0135-z

Strickland, J. C. y Lacy, R. T. (2020). Behavioral Economic Demand as a Unifying Language for Addiction Science: Promoting Collaboration and Integration of Animal and Human Models. *Experimental and Clinical Psychopharmacology, 28*(4), 404–416. https://doi.org/10.1037/pha0000358

Thaler, R. (1981). Some Empirical Evidence on Dynamic Inconsistency. *Economics Letters, 8*(3), 201–207. https://doi.org/10.1016/0165-1765(81)90067-7

Vanderveldt, A., Oliveira, L. y Green, L. (2016). Delay Discounting: Pigeon, Rat, Human—does it Matter? *Journal of Experimental Psychology: Animal Learning and Cognition, 42*(2), 141–162. https://doi.org/10.1037/xan0000097

Wagenmakers, E. J., Wetzels, R., Borsboom, D. y van der Maas, H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. https://doi.org/10.1037/a0022790

# CREDIBILIDAD O BARBARIE: CÓMO LA CRISIS DE REPLICABILIDAD HA DESATADO UNA REVOLUCIÓN EN PSICOLOGÍA Y OTRAS CIENCIAS

# CREDIBILITY OR BARBARISM: HOW THE REPLICATION CRISIS SPARKED A REVOLUTION IN PSYCHOLOGY AND OTHER SCIENCES

ÓSCAR LECUONA[1], GUIDO CORRADI[2],
ARIADNA ANGULO-BRUNET[3] Y EDUARDO GARCÍA-GARZÓN[4]

Cómo referenciar este artículo/How to reference this article:

Lecuona, O., Corradi, G., Angulo-Brunet, A. y García-Garzón, E. (2025). Credibilidad o barbarie: Cómo la crisis de replicabilidad ha desatado una revolución en Psicología y otras ciencias [Credibility or Barbarism: How the Replication Crisis Sparked a Revolution in Psychology and Other Sciences]. *Acción Psicológica, 22*(1), 115–136. https://doi.org/10.5944/ap.22.1.43231

## Resumen

La ciencia actual vive tiempos críticos y revolucionarios. El surgimiento de la "crisis de replicación" ha supuesto un reto estructural histórico, junto con mala praxis científica y problemas derivados con una estructura de incentivos perversa en el sistema de publicaciones. Como respuesta, se han iniciado diversas reformas en la comunidad científica conocidas como la "revolución de la credibilidad". En este artículo revisamos estos sucesos, su cronología, características principales, y su relación entre ellos. De esta forma, buscamos divulgar y formar a los lectores en estos aspectos. Además, proponemos nuevas Buenas Prácticas de Investigación de la comunidad científica en Psicología y otras ciencias para producir y consumir una ciencia más replicable y creíble. En definitiva, concienciar y formar parte de una mejor comunidad científica para los retos del siglo XXI.

**Correspondence address [Dirección para correspondencia]**: Óscar Lecuona, Facultad de Psicología, Universidad Complutense de Madrid, España.
**Email**: olecuona@ucm.es

**ORCID**: Oscar Lecuona, Guido Corradi, Ariadna Angulo-Brunet y Eduardo García-Garzón.

[1] Universidad Complutense de Madrid, España.
[2] Universidad Villanueva, España.
[3] Universitat Oberta de Catalunya, España.
[4] Shakers, S.L.

**Agradecimientos**: Los autores extienden su agradecimiento a Pandelis Perakakis, por sus comentarios y aportaciones al manuscrito.

Recibido: 30 de mayo de 2024.
Aceptado: 28 de agosto de 2024.

**Palabras clave:** Crisis de replicación; Revolución de credibilidad; Prácticas cuestionables de investigación; Meta-ciencia; Ciencia abierta.

## Abstract

Current science is experiencing critical and revolutionary times. The emergence of the "replication crisis" has posed a historical structural challenge, alongside issues of scientific malpractice and problems stemming from a perverse incentive structure in scientific publishing. In response, various reforms have been engaged within the scientific community, known as the "credibility revolution". This article reviews these events, how they developed, their main characteristics, and their interconnections. In doing so, we aim to inform and educate readers about these aspects. In addition, we propose new Best Research Practices in the scientific community in psychology and other fields to produce and consume more replicable and credible science. Ultimately, we seek to raise awareness and encourage participation in a better scientific community to meet the challenges of the 21st century.

**Keywords:** Replication crisis; Credibility revolution; Questionable research practices; Meta-science; Open science.

## Introducción

Hoy en día la ciencia atraviesa una crisis sin precedentes, denominada «crisis de replicación». Mientras que la ciencia está evolucionando hacia un modelo más riguroso y exigente, están apareciendo preocupaciones sobre su credibilidad y, además, se están destapando casos de mala praxis científica (e.g., fraude o inventarse datos, plagio o autoplagio, entre otros; Van Noorden, 2023a; 2023b). Como señala *Retraction Watch* (2015), existen muchos casos de artículos retirados debido a malas praxis. Más allá de ser casos aislados, diversas fuentes señalan a universidades (e.g., Harvard) o países enteros (e.g., China, Rusia, Arabia Saudí o Irán) que han inflado artificialmente sus posiciones en *rankings* para obtener prestigio (e.g., Catazaro, 2023; Mole, 2024; The Economist, 2023; en prensa española, ver Ansede, 2024). Es decir, cuando la ciencia parece más necesaria (e.g., la pandemia de la COVID-19 o la emergencia climática) parecen surgir problemas que ponen en tela de juicio su integridad y credibilidad. Esta situación presenta un peligro adicional: Estos problemas pueden servir como ancla para que sectores extremistas aumenten su influencia y relevancia política. Algunos ejemplos son los movimientos terraplanistas, antivacunas, o diversos negacionismos (ver Materiales Suplementarios). Debido a la popularidad y alcance político de estas narrativas (e.g., ministerios en la presidencia de EE. UU.; BBC, 2024), esta situación pone directa e indirectamente en riesgo la vida de millones de personas.

### *Primeros indicios*

Para entender el momento actual, se puede empezar por sus antecedentes históricos. Por un lado, destaca el trabajo seminal de Rosenthal (1979) acuñando el concepto de «problema del archivador» (la no publicación de resultados no significativos, también llamado «sesgo de publicación»). Por otro lado, también cabe destacar el concepto de "*Cargo Cult Science*" (ciencia de culto a la carga) acuñado por Feynman (1998). Consiste en una metáfora sobre unas tribus que, a inicios del siglo XX, imitaba la tecnología occidental mediante madera y otros materiales. Esta imitación replicaba la forma, pero obviaba los mecanismos que los hacían funcionar. Feynman utilizó esta metáfora de «cultos a las naves de carga» para hablar de una "ciencia de culto a las naves de carga". Es decir, conocimiento que cumple con los rituales científicos pero que no provee resultados con calidad científica.

Todo esto llevó a la comunidad científica a revisar sus propios criterios para evaluar la calidad (la meta-ciencia). Así, por ejemplo, se ha señalado que la mayoría de la investigación (médica) no es replicable por una inflación de falsos positivos y una potencia estadística insuficiente (Ioannidis, 2005). Las causas incluirían tamaños muestrales bajos, selección a posteriori de hallazgos (*Cherry Picking*), sesgo de publicación, uso de $p < .05$ como criterio fijo y único, falta de estudios de replicación y compartir datos y análisis, tamaños del efecto pequeños o diseños poco rigurosos (Ioannidis, 2005, 2008). Lejos de reme-

diarse, 20 años después aún se encuentra una atenuación de efectos en los metaanálisis, incluso cuando los mismos están probablemente sobreestimados (e.g., Bartoš et al., 2023).

En Psicología, Simmons et al. (2011) ilustró la excesiva flexibilidad (también denominado «grados de libertad») que existe en los métodos estadísticos que se aplican para llevar a cabo los estudios, lo que podría conllevar a una inflación de falsos positivos (es decir, a encontrar efectos significativos que en realidad no existen). Frente a ello, se apostó por una mayor transparencia y replicabilidad a los estudios. La replicabilidad se puede definir como el grado en que se pueden obtener los mismos resultados de un estudio concreto en un nuevo contexto.

Continuando en el ámbito de la Psicología, se produjo un fracaso relevante al intentar replicar estudios clásicos de *priming* social (Doyen et al., 2012), suscitando un debate sobre la replicabilidad en la disciplina. No obstante, fueron los estudios sobre la percepción extrasensorial los que pusieron de relevancia estas preocupaciones. Bem (2011) publicó diversos experimentos siguiendo protocolos estándar y bajo revisión por pares, donde parecía evidenciarse la existencia de cierta precognición (e.g., los participantes parecían responder correctamente a una prueba antes de la presentación objetiva del estímulo). Aunque aparecieron refutaciones (Rouder y Morey, 2011; Wagenmakers et al., 2011), el propio Bem publicó un metaanálisis indicando la presencia de este efecto (Bem et al, 2015). Aparecieron dos interpretaciones en el debate: o bien los estudios son correctos y la percepción extrasensorial existe (Cardeña, 2018), o bien dichas percepciones no existen e incluso los métodos más exigentes son sensibles a falsos positivos (e.g., Lakens et al., 2016). Un estudio de replicación posterior concluyó la falta de replicabilidad de los resultados (Kekecs et al., 2023). Por tanto, la comunidad científica mayoritaria concluyó que era necesario reformular el sistema científico.

## *El detonante: la crisis de replicación*

El punto álgido de la preocupación sobre la ciencia psicológica llegó de la mano de los resultados de la *Open Science Collaboration* (2015). Esta colaboración realizó 100 estudios de replicación para efectos publicados en revistas de alto impacto en Psicología y otras disciplinas. Los hallazgos fueron claros: entre la mitad y dos tercios de los resultados no se replicaron. Esto confirmó que no eran casos o campos aislados, sino un problema estructural. Este suceso es la que se toma como inicio de la Crisis de la Replicabilidad donde se evidencia la sospecha, quizá fundada, de que muchos efectos que tomamos como relevantes, no se sostienen cuando se toman las salvaguardas metodológicas adecuadas. Los campos más afectados fueron la psicología social, cognitiva y del desarrollo (Youyou et al., 2023). Esto vino acompañado también de propuestas sobre problemas sistémicos en el desarrollo de instrumentos como el *α-hacking* (Flake y Fried, 2020; Hussey et al., 2024) y modelos teóricos (Eronen y Bringmann, 2021). Lejos de ser endémico de Psicología, estudios en otros campos indicaron resultados similares en medicina (Errington et al., 2021; Van Noorden, 2023a), economía (Bohannon, 2016; Camerer et al., 2016; Tsui, 2022; Wood et al., 2018), y gestión del agua (Stagge et al., 2019). Estudios interdisciplinares indican que el patrón parece consistente entre campos científicos (incluyendo física, química, e ingenierías; Baker, 2016; Gertler et al., 2018). Es decir, los problemas de replicabilidad están asociados a patrones sistémicos de la ciencia establecida.

## Diagnóstico: los hallazgos de la meta-ciencia

Entonces, ¿cuáles son las causas de este problema sistémico? Existen diversas propuestas, que se pueden agrupar como: (1) los grados de libertad de la persona investigadora y las prácticas cuestionables de investigación, y (2) el sistema de incentivos perverso de las editoriales e instituciones científicas.

### *Prácticas cuestionables de investigación*

Las prácticas cuestionables en investigación (QRP por sus siglas en inglés; Figura 1) son aquellas prácticas que, intencionadas o no, conducen a hacer inferencias incorrectas que no concuerdan con los objetivos, la metodología o

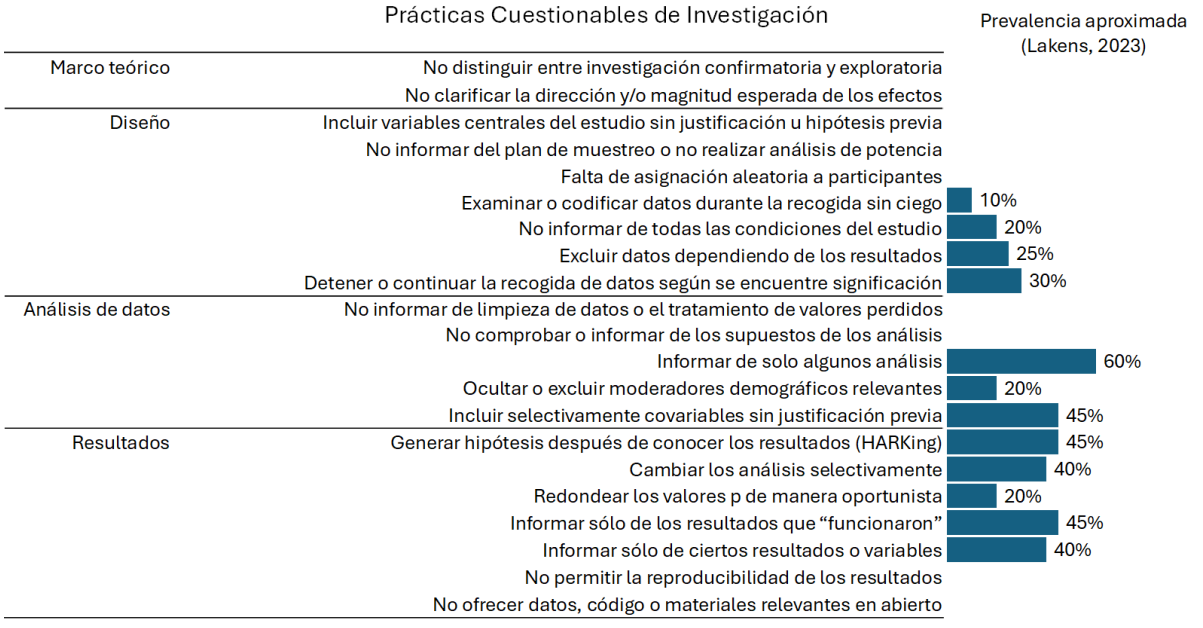
los resultados del estudio (Gerrits et al., 2019). Las QRP pueden cometerse por ignorancia o descuido, aunque también para provocar resultados estadísticamente significativos. Es importante separarlas del fraude y la mala praxis en general, debido a la falta de intencionalidad de engañar. Sin embargo, su mayor frecuencia podría hacerlas más dañinas. La Figura 1 muestra una lista con las principales QRPs.

Una de las QRP más conocidas es el *p-hacking* («piratear valores *p*»), que consiste en tomar decisiones y análisis estadísticos que obtengan resultados significativos al margen de los resultados reales (Stefan y Schönbrodt, 2022). Dentro del *p-hacking* pueden ocurrir muchos tipos distintos de QRP además de los ya indicados, como incluir o eliminar variables después de observar los resultados (o ítems de una escala), imputar valores perdidos con varios métodos sin justificación, buscar subgrupos sin planificación previa, o redondear los valores *p* (lista completa en Stefan y Schönbrodt, 2022). Incurrir en QRPs podría provocar una inflación de falsos positivos, y una infrarrepresentación de resultados disruptivos o contrarios a los marcos teóricos (Simonsohn et al., 2014; White, 2022). Además, las QRPs estarían directa y recíprocamente relacionadas con el sesgo de publicación descrito por Rosenthal (1976). Concretamente, llevarían a un aumento del sesgo de publicación, impactando sobre los tamaños del efecto y su heterogeneidad (Anderson y Liu, 2024). Y, de la misma forma, una literatura con mayor sesgo de publicación llevaría a una mayor proliferación de QRPs por (1) dirigir el trabajo científico hacia efectos endebles o inexistentes y tener presión por publicar resultados significativos, y (2) reproducir QRP por imitación al considerarlos estándares.

Las QRP pueden entenderse como un «problema del jardín de los senderos que se bifurcan» (*Forking Path Problem*; Nagy et al., 2024). Este concepto hace referencia al árbol de todas las posibles decisiones que se toman al aplicar métodos científicos. Cada decisión altera los resultados y normalmente solo se explora uno o pocos senderos,

**Figura 1**

*Prácticas Cuestionables de Investigación principales*




*Nota*. El valor de la prevalencia aproximada es de Lakens (2023). En los casos donde no existe valor es porque no se conoce la prevalencia de esa QRP.

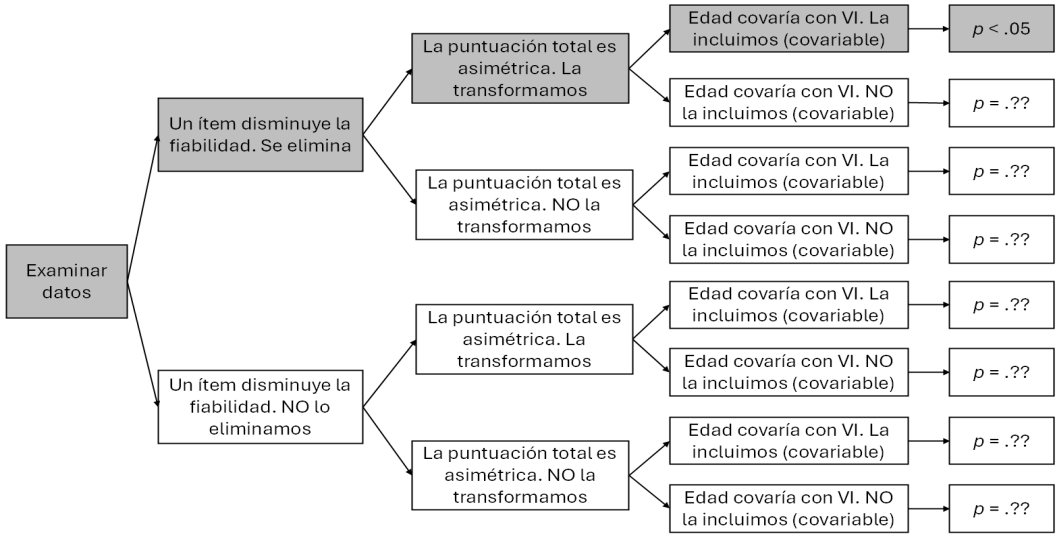
generalmente ignorando los otros. En la Figura 2 puede encontrarse un ejemplo aplicado en ciencia. El rango o amplitud de estos árboles de decisión puede conceptualizarse como los «grados de libertad de la persona investigadora». Esto puede llevar a la generación de sesgos y la proliferación de QRP. Idealmente, el rango de esos resultados debería describirse para estudiar su robustez (como en los «análisis multiverso», donde se construye un rango de una estimación estadística según se cambien partes del *forking path*; Parsons, 2022; Steegen et al., 2016).

## *La perversión del sistema editorial*[1]

Los incentivos que impulsan la carrera investigadora son una de las principales razones por las que los científicos buscan obtener resultados significativos a toda costa. En general, el prestigio, promociones y acceso a financiación ha venido dado en las últimas décadas por el número de artículos publicados en revistas de alto impacto y el número de citas recibidas. En otras palabras, para tener una carrera científica exitosa, ha sido necesario publicar mucho y en revistas de calidad («publica o perece»; *publish or perish*). Esta presión lleva a hacer lo que sea necesario para publicar, incluyendo el *salami slicing* (dividir un trabajo en varios estudios publicables por separado), o ceñirse a narrativas complacientes con la audiencia (modificando resultados para que no contradigan ideas dominantes) o revistas (priorizando artículos novedosos para maximizar las citas y descargas de artículos). Aunque pueda parecer contradictorio, ambas estrategias responden al mismo patrón: Presentar hallazgos complacientes con la audiencia (por sorprendentes o continuistas) que generen citas (Giner-Sorolla et al., 2012; Nosek et al., 2012). Entre otras consecuencias, se ha encontrado que estudios genuinamente rompedores y valiosos (incluso descubrimientos laureados con un Nobel) tuvieron dificultades para ser publicados (Campanario, 2009). Más tarde, nuevas evidencias indicaron un declive en la innovación de las contribuciones científicas (Park et al., 2023).

**Figura 2**

Ejemplo de un "*forking path problem*"




[1] Usamos el adjetivo «perverso» no en su acepción popular (malvado, cruel, o sexualmente desviado), sino en su acepción en economía, es decir, a la producción de resultados indeseados o contraproducentes en una estructura de incentivos (Edwards y Roy, 2017).

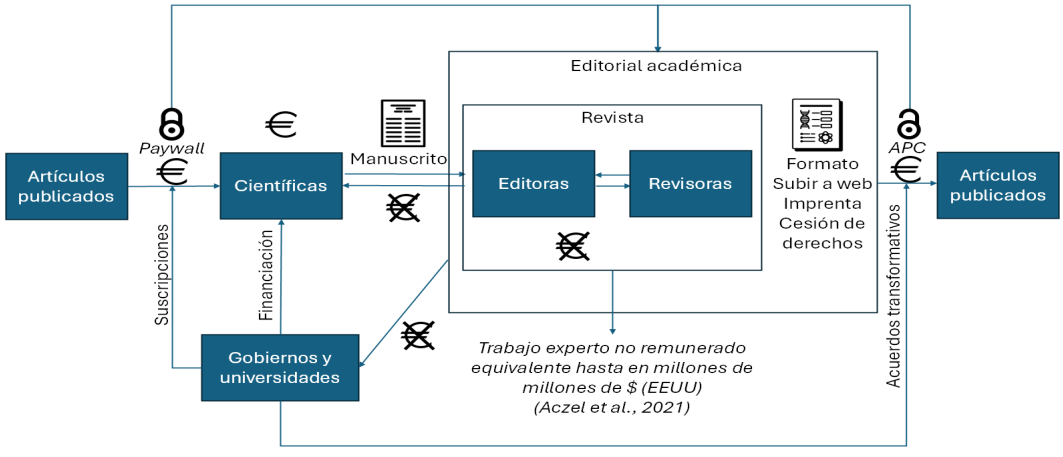
Además, los incentivos de las revistas y editoriales están más alineados con conseguir más citas que en favorecer la calidad de los trabajos que publican. Concretamente, los incentivos han dependido fuertemente del «factor de impacto», una métrica basada en citas que clasifica a las revistas en *rankings*. El problema es que dicha métrica no informa de la naturaleza de ese impacto en la literatura o sociedad, de si las citas respaldan o desacreditan el artículo, o de otros aspectos importantes. Es más, el factor de impacto y la calidad no parecen estar asociados (Brembs et al., 2013). En ese sentido, la estructura de incentivos resulta perversa por premiar aspectos contraproducentes o ajenos a los que debería.

Otro de los grandes problemas es el modelo económico que siguen las editoriales científicas. La Figura 3 ilustra dicho modelo (detalles en Materiales Suplementarios). Este modelo implica unos ingresos altos junto con gastos bajos, ya que la creación, revisión y edición de su producto la realizan gratuitamente las personas científicas, dejando los gastos en principalmente mantenimiento y *marketing* virtual, formateo de manuscritos e impresión física. Esto produce un margen de beneficios muy amplio (entre el 25 % y el 40 % de media) superior a multinacionales como Disney, McDonald's o Apple (Grossman y Brembs, 2021; Mayoni, 2024). En definitiva, las editoriales académicas se lucran de productos financiados con dinero público que no generan ni refinan, y los venden principalmente a aquellos que los producen por su acceso.

Debido a diversas críticas a este modelo, recientemente se implementó el modelo de acceso abierto. Actualmente el modelo más implementado es el *Gold Open Access* (ver Materiales Suplementarios): los autores de un artículo pagan los «gastos de procesamiento de artículo» (APC, por sus siglas en inglés), que varían entre los cientos a las decenas de miles de euros. Las instituciones y universidades han negociado con las editoriales los llamados «acuerdos transformativos», donde se otorgan un número concreto de artículos en acceso abierto al año en cada editorial, o descuentos sustanciales. Esto ha recibido críticas consistentes en que fomentan el oligopolio de las editoriales como una estrategia de *marketing* dejando el lucro intacto (Butler et al., 2023; Perakakis, 2021). De hecho, editoriales como *MDPI* o *Frontiers* han aprovechado este nuevo modelo para generar cantidades masivas de publicaciones rápidas a menudo con una calidad cuestionable, y siendo expulsadas algunas de sus revistas del *Journal Citation Reports*, el principal *ranking* por factor de impacto (Brainard, 2023). En otras palabras, la «solución» es que las editoriales ahora cobran al personal científico (o a las universidades donde trabajan)

**Figura 3**

*Modelo actual de publicación en revistas académicas*




*Nota*. APC = *Article processing charges* (cargo de procesamiento de artículos). El *paywall* y el *APC* son excluyentes.

por publicar una investigación financiada con dinero público, revisada y editada gratuitamente, en vez de cobrar por su acceso. Esta situación se agrava aún más cuando aparecen revistas «depredadoras», que prometen publicaciones rápidas de acceso abierto con estrategias de *marketing* (Cobey et al., 2018). La reacción más común entre personas ajenas a la ciencia es de incredulidad, o la opinión de que este modelo es una estafa.

En conjunto, la proliferación de revistas con *open access* lleva a que las personas con mayor acceso a recursos materiales y conceptuales tengan más publicaciones y citas y, por tanto, cuenten con más indicadores de prestigio para parte de la comunidad científica. Esta hipertrofia de las publicaciones (derivada del *publish or perish* más allá del *open access*) explicaría la inflación de retiradas de manuscritos (Van Noorden, 2023b), pero también la inabarcabilidad de la documentación y la dependencia de motores de búsqueda (como *Google Scholar*). Esto ha generado fenómenos complejos como la inflación de citas a artículos y autorías muy establecidas, o al aumento de desigualdades en la comunidad (Chu y Evans, 2021; Gomez et al., 2022; Nielsen y Andersen, 2021).

Finalmente, se han generado nuevas malas prácticas a raíz de este panorama (e.g., *paper mills*), junto con el surgimiento de la IA generativa en la actividad científica y el espionaje de revistas a usuarios mediante nuevas tecnologías (Brembs et al, 2023; ver Materiales Suplementarios). En conjunto, todas estas problemáticas (hipercrecimiento, afianzamiento de desigualdades, oligopolios, lucro por encima de la calidad) pueden describirse como una extensión del capitalismo al conocimiento científico, tratándolo como un producto de consumo más (Brembs et al., 2023).

## Soluciones: la revolución de credibilidad

Para poder superar estos problemas, se han propuesto reformas que han recibido el nombre de la «revolución de la credibilidad» (Vazire, 2018). Consisten en un mayor control de la arbitrariedad investigadora, el diseño de buenas prácticas (e.g., Tabla 1) y un fomento individual y comunitario de la replicabilidad (Figura 4; otras revisiones en Nelson et al., 2018; Nosek et al., 2022).

### *Pre-registros y la ciencia abierta*

Un pre-registro es esencialmente la publicación en abierto de un documento detallando el marco teórico, las hipótesis y los métodos de un estudio. Si este registro además se revisa en una revista con revisión por pares y se ejecuta el estudio según ese propio documento y se publica, se denomina «informe registrado» (*registered reports*). De esta forma, se limitan los grados de libertad las personas investigadoras y de la revista, ya que se comprometen a publicar los hallazgos sin importar los resultados (Hardwicke y Wagenmakers, 2023). Además, se busca prevenir las QRP, aunque aún no está claro si este objetivo se consigue (van der Akker, 2023). Aunque parezca inmovilista y perjudicial para la creatividad y serendipias, hay margen para desviarse del pre-registro mientras se justifique explícitamente (Lakens, 2024). Existen diversas páginas para realizar pre-registros (ver Tabla 1), mientras que los informes registrados se dan en las propias revistas.

Por otro lado, la ciencia abierta consiste en un movimiento heredero del *software* libre consistente en otorgar acceso libre y gratuito a los componentes de un estudio. Estas prácticas permiten describir el *forking path*, ya que se puede acceder a lo necesario para reproducir los hallazgos, o explorar otras alternativas. La vertiente más implementada es el acceso abierto, donde se garantiza el acceso libre al manuscrito (normalmente mediante pago previo de los autores a la revista). Sin embargo, otros componentes (e.g., datos, código, protocolos o materiales) no están garantizados y quedan a discreción de los autores. En muchos casos, esto resulta en la falta de acceso o en un acceso menos intuitivo a dichos elementos, lo que dificulta o incluso impide la reproducción de los resultados. Debido a los problemas de este modelo, han proliferado librerías piratas como *Sci-Hub* o *LibGen* (Maddi y Sapinho, 2023). Como alternativas, existen diversas revistas con acceso abierto sin APC (*Diamond Open Access*, ver Materiales Suplementarios; Bernal y Perakakis, 2023). El mantenimiento virtual de la revista estaría a cargo de instituciones académicas (e.g., bibliotecas de facultades, asociaciones científicas), mientras que el resto de los agentes seguirán ejerciendo sus funciones gratuitamente. Un ejemplo es *Psicológica* (https://psicologicajournal.com/), revista a cargo de la Universidad de València y otros organismos, e indexada en el JCR. Otros ejemplos son *PeerJ*, *Collabra*,

o *Psicothema* (un directorio de revistas en https://doaj.org/). Finalmente, también existen las pre-impresiones (*pre-prints*, manuscritos de artículos antes de enviarse a la revista) y post-impresiones (*post-prints*, manuscritos después de la revisión por pares sin formato de revista). Ambas garantizan el acceso libre, aunque los *pre-prints* deben advertir de su falta de revisión por pares (Wingen et al., 2022).

Además del acceso abierto, la ciencia abierta cuenta con datos, código, y materiales en abierto. Esto no solo permite la reproducibilidad de los resultados (esto es, el grado en que otra persona pueda obtener los mismos resultados con los mismos datos y análisis), la detección de errores en análisis y procesamiento de datos y la transparencia, sino que facilita la acumulación de conocimiento (e.g., mejores metaanálisis). Existen iniciativas para garantizar la confiabilidad de estos procesos, como las guías FAIR (acrónimo en inglés para datos localizables, accesibles, intercambiables y reutilizables; https://www.go-fair.org/). Finalmente, un campo en vías de implementación es la revisión por pares abierta (*open peer review*), el cual consiste en dar acceso libre a las identidades, contenidos y participación en la revisión (Walker et al., 2015; un ejemplo en https://peerj.com/). Esto permitiría una mayor calidad, por mayores incentivos y prevención del abuso, aunque existen críticas (Wolfram et al., 2020). Existen más ramas de la ciencia abierta (e.g., docencia, transferencia, etc.), que por razones de espacio no trataremos aquí (ver UNESCO, 2024).

Las prácticas de ciencia abierta pueden realizarse a través de diversas plataformas. Un ejemplo es la OSF, donde se pueden crear páginas y subir archivos gratuitamente con control de versiones (manuscritos, datos, código, etc.). Esto incluye publicar *pre-print* y *post-prints*, a través de servidores como *PsyArXiv*.

## *Estudiar y fomentar la replicabilidad*

La manera más inmediata de estudiar la replicabilidad han sido los estudios de replicación (e.g., Nosek et al., 2022). Estos estudios se centran en obtener la replicabilidad de uno o más hallazgos con procedimientos de alta calidad metodológica. Normalmente se distinguen tipos de replicaciones (ver Materiales Suplementarios), y es común que implementen otras prácticas destinadas para fomentar la replicabilidad: Muestras más grandes y planeadas, y estudios multi-equipo. Respecto a las muestras, la planificación a priori del tamaño y características de la muestra se recomiendan mediante el análisis de recursos disponibles y de la potencia estadística (Lakens, 2022; ver Materiales Suplementarios). Además, muestras más allá de países WEIRD (*Western, Educated, Industrialized, Rich and Democratic*) llevan normalmente a las colaboraciones multipaís (Jarke et al., 2022). Un gran ejemplo es la iniciativa *Many Labs*, entre otras (e.g., Klein et al., 2018; Ruggeri et al., 2021). Esto reduce las QRP al validarse entre equipos y aumenta la replicabilidad al contar con muestras más grandes y representativas. Contribuciones recientes señalan que parece suficiente con un solo estudio de replicación exitoso para concluir sobre si parece replicable o no, salvo que haya problemas metodológicos claros (e.g., muestra insuficiente, controles inadecuados, o diferencias sustantivas en instrumentos o procedimientos; Boyce et al., 2024). Esto no significa si el efecto existe en la población o no (e.g., podría ser un artefacto) o si es un efecto relevante. Además, es importante conocer el sesgo de publicación de la literatura que se pretende replicar, ya que puede inflar los falsos negativos en estudios de replicación (Nosek et al., 2022). Los metaanálisis se han propuesto como herramientas útiles (pero limitadas) para detectar e intentar corregir ese sesgo, como incluir literatura más allá de la revisada por pares, o datos de muestras (e.g., metaanálisis con datos de participantes individuales, o IPD por sus siglas en inglés; Wagner III, 2022; Wang et al., 2021). Finalmente, también existe el proyecto SCORE (Alipourfard et al., 2021), destinado a generar índices de confiabilidad a trabajos científicos mediante técnicas de aprendizaje automático y supervisión de expertos.

Desde el análisis de datos se han propuesto diversas reformas al contraste de hipótesis (NHST, por sus siglas en inglés) para prevenir sus malos usos. La reforma más sencilla han sido detectores de inconsistencias en resultados estadísticos, como *statcheck* (Nuijten y Wicherts, 2024). Otras propuestas proponen recuperar el uso más ortodoxo de los valores *p* (e.g., Haig, 2017). También proponen hacer más exigente el criterio de falsos positivos ($\alpha = .005$), o especificarlo en cada estudio siguiendo procedimientos concretos (e.g., Benjamin et al., 2018; Maier y Lakens, 2022). Otras propuestas apuestan por

expandir el NHST aportando, entre otros, tamaños del efecto, intervalos de confianza o bootstrap (Haig, 2017; Mayo, 2018). Finalmente, otras propuestas apoyan el abandono del NHST hacia paradigmas basados en cuantificar probabilidades sin puntos de corte (e.g., McShane et al., 2019). Aquí destaca la estimación Bayesiana (Wagenmakers et al., 2018; Kruschke, 2021). En resumen, las mejoras al NHST son más escalables, pero siguen expuestas a QRP. En cambio, su abandono provee paradigmas más robustos (e.g., Stefan y Schönbrodt, 2022), pero menos escalables y más complejos de aplicar. De igual manera, hay cierto consenso en que es necesario un mayor rigor y clarificación en los pasos efectuados en los análisis estadísticos (Devezer et al., 2021).

## *Educar a la comunidad en consumir y producir ciencia más creíble*

La última rama implica la formación de la comunidad científica en estas reformas. Un primer paso puede ser la «alfabetización» sobre los conceptos de este fenómeno a todo tipo de audiencias, con esfuerzos como el proyecto FORRT (https://forrt.org) y su glosario (Parsons et al., 2022). De manera más concreta, existen guías para personas editoras, y para la formación en campos de investigación (Schiavone y Vazire, 2023; Silverstein et al., 2024). Para el personal científico, los tutoriales sobre estas reformas pueden ser de gran ayuda (e.g., Hardwicke y Wagenmakers, 2023; ver Materiales Suplementarios), así como las guías para el uso de plataformas como la OSF (https://help.osf.io/). Además, el propio FORRT busca la construcción de comunidades de investigación en torno a estas reformas, que ayudarían a los estudios multi-país (Jarke et al., 2022). Para el estudiantado, la inclusión de estos aspectos en materias de grado y máster serían la principal medida (e.g., https://osf.io/cm4yb/; https://www.cos.io/events). Y finalmente, la divulgación científica de calidad resulta fundamental para la diseminación de resultados (e.g., Brembs et al., 2023, Figura 3; ver Material Suplementario). La Figura 4 resume todas estas reformas.

**Figura 4**

*Resumen de las reformas de la revolución de la credibilidad*

## ¿Qué tal está saliendo hasta ahora?

Por el momento, varios trabajos indican una mejora sistemática en las prácticas científicas (Forozish, 2024; Korbmacher et al., 2024; Nosek et al., 2022), tanto en general como en diversas ramas de conocimiento (e.g., Schiavone y Vazire, 2023). Es decir, se están encontrando cambios estructurales (como el incentivo a pre-registros y con prácticas de ciencia abierta, o revistas publicando artículos con *pre-prints*), procedimentales (plataformas como la *Open Science Framework*, *PsyArXiv*, o *Statcheck*), de comunidad (con una mayor concienciación e impulso de investigaciones más replicables y cuestionamiento de otras menos transparentes), y políticos (normas editoriales en revistas sobre incentivar pre-registros y prácticas de ciencia abierta, o requerir determinados grados de transparencia). Además, los pre-registros y *registered reports* parecen estar mejorando los métodos, se publican en revistas de mayor impacto, reciben más citas y no tardan más en publicarse (Lakens et al., 2024; van den Akker et al., 2023). Por último, parece que el *peer review* abierto también lleva a mayor número de citas, quizá indicando una mayor credibilidad (Zong et al., 2020).

Sin embargo, el panorama no parece tan simple. Hay algunas reformas que por sí solas no mejoran la credibilidad, o incluso la dañan. Por ejemplo, la población general no parece distinguir los *pre-prints* de artículos sometidos a *peer review*, incluso con explicaciones o avisos claros (Fleerackers et al., 2024; Wingen et al., 2022). Por otro lado, las políticas de datos abiertos no parecen mejorar la calidad (Berberi y Roche, 2022). Es decir, las reformas de la revolución de la credibilidad deben tomarse en conjunto y con una justificación sólida. Es importante recalcar que, además de mejorar la credibilidad de la ciencia, estas reformas son un ejercicio político dentro de la comunidad científica, con agentes y narrativas involucradas. Han de ser los agentes con mayor poder (revistas, agencias gubernamentales, entidades financiadoras, e investigadores senior) los que deben liderar y exigir estas reformas. Por ejemplo, aplicando modelos organizativos sin lucro económico (https://www.openscholar.info/; Brembs et al., 2023), políticas de replicabilidad y de ciencia abierta íntegra (e.g., los principios TOP y FAIR; Nosek et al., 2022), realizando pre-registros o enviando trabajos altamente citables a revistas ejemplares por su transparencia y rigor (e.g., con *diamond open access*). De hecho, dichos agentes están comenzando a adoptar estas reformas, como el incentivo hacia estudios pre-registrados o el cambio de incentivos de carrera investigadora más allá del número de citas (e.g., criterios DORA, https://sfdora.org/). Más información en materiales suplementarios.

La Tabla 1 resume estas recomendaciones como una lista de Buenas Prácticas Actuales de Investigación (BCPs, en contraste a las QRPs). De esta forma, el personal de investigación puede revisar su grado de adherencia y guiar su praxis científica.

## Conclusiones

A pesar de la crisis de replicación y sus retos, las reformas de la revolución de la credibilidad (pre-registros, ciencia abierta, y buenas prácticas de investigación) están siendo útiles para mejorar la ciencia. Con una comunidad comprometida y concienciada puede construirse una ciencia mejor. Y, de esa forma, escapar a la cargo *cult science* de la que nos quiso prevenir Feynman. Solo nos queda sumarnos a la revolución con un trabajo riguroso, honesto, y transparente.

**Tabla 1**

*Buenas Prácticas de Investigación (BRPs)*

| Etapa | Buenas Prácticas de Investigación | Plataformas o recursos |
|---|---|---|
| **Antes de la realización del estudio** | Pre-registrar el estudio (hipótesis, diseño, dirección y magnitud del efecto, métodos y análisis), o enviar manuscrito a una revista como propuesta para informe registrado | https://osf.io/registries<br>https://aspredicted.org/<br>https://clinicaltrials.gov/<br>https://osf.io/zab38/ |

| | | |
|---|---|---|
| | Declarar diseño exploratorio o confirmatorio | |
| | Declarar el uso y objetivo de cada manipulación e instrumento | |
| | Describir aleatorizaciones y manipulaciones llevadas a cabo | |
| | Análisis de potencia *a priori* para todos los análisis planeados, con tamaños del efecto de estudios previos | G*Power,<br>Kang (2021)<br>Jobst et al., (2024) |
| | Declarar criterios de inclusión y exclusión de participantes *a priori* | |
| | Usar medidas con pruebas de fiabilidad y validez para la población | |
| | Especificar las técnicas de análisis y contrastes estadísticos, así como si se optan por pruebas bilaterales o unilaterales | |
| | Especificar posibles datos perdidos y estrategias de gestión (eliminación o exclusión, imputación, etc.) | |
| **Durante la realización del estudio** | Si hay desvíos del pre-registro, describir y justificar dichas decisiones (e.g., eliminación de sujetos por criterios no anticipados) | Lakens (2024) |
| | Declarar los criterios de parada de recogida de muestra (ej. recursos económicos, tiempo). | Lakens (2024) |
| | Análisis de datos independiente o ciego (a marco teórico o conflictos de interés) | Nagy et al. (2024) |
| | Redactar una bitácora o registro de pasos en el procesamiento de la base de datos, con justificaciones (ej. recodificación de variables, transformación de datos, eliminación de sujetos) | |
| | Si se declaró un diseño exploratorio, realizar análisis exploratorios declarados como tales | |
| | Asegurar la invarianza de las mediciones entre condiciones o grupos | Leitgöb et al. (2023) |
| | Análisis de potencia *de sensibilidad* (si no se ha alcanzado el tamaño especificado *a priori*) | G*Power,<br>Kang (2021)<br>Jobst et al., (2024) |
| | Informar de contrastes sobre supuestos de las pruebas estadísticas. Si no se cumplen e influyen sobre estimaciones, seleccionar alternativas robustas, no paramétricas, o similares | |
| | Informar de tamaños del efecto e intervalos de confianza y credibilidad | Pek y Flora (2018) |
| | Informar de los valores *p* o factores de Bayes crudos (sin redondeos) y según el plan previsto | Greenland (2016) |
| **Después de la realización del estudio** | Compartir datos, análisis, materiales suplementarios y manuscritos en abierto (incluido un diccionario de variables con sus etiquetas y valores) | https://osf.io/ |
| | Informar de todos los análisis planificados | |
| | Seguir las guías FAIR para la disponibilidad de datos | https://www.go-fair.org/ |
| | Publicar *postprints* en plataformas abiertas | Wingen et al. (2022) |
| | Publicar *preprints* solo si es estrictamente necesario, y con avisos claro y visibles de que no tiene *peer review* (y redirigir si hay versión revisada) | https://osf.io/preprints/psyarxiv |
| | Establecer una lista de revistas para enviar el manuscrito con prácticas de ciencia abierta (e.g., *diamond open access*) | https://doaj.org/<br>https://jcr.clarivate.com/ |

## Referencias


Aczel, B., Szaszi, B. y Holcombe, A. O. (2021). A Billion-Dollar Donation: Estimating the Cost of Researchers' Time Spent on Peer Review. *Research Integrity and Peer Review, 6*(1) Artículo 14. https://doi.org/10.1186/s41073-021-00118-2

Alipourfard, N., Arendt, B., Benjamin, D. M., Benkler, N., Bishop, M., Burstein, M., Bush, M., Caverlee, J., Chen, Y., Clark, C., Dreber Almenberg, A.,

Errington, T. M., Fidler, F., Field, S., Fox, N., Frank, A., Fraser, H., Friedman, S., Gelman, B., Gentile, J., … Wu, J. (2021). Systematizing Confidence in Open Research and Evidence (SCORE). SocArXiv, 1–33. https://doi.org/10.31235/osf.io/46mnb

Al-Khatib, A. y Teixeira da Silva, J. A. (2016). Stings, Hoaxes and Irony Breach the Trust Inherent in Scientific Publishing. *Publishing Research Quarterly, 32*(3), 208–219. https://doi.org/10.1007/s12109-016-9473-4

Anderson, S. F. y Liu, X. (2023). Questionable Research Practices and Cumulative Science: The Consequences of Selective Reporting on Effect Size Bias and Heterogeneity. *Psychological Methods*. https://doi.org/10.1037/met0000572

Ansede, M. (2024, octubre 16). *La editorial Springer Nature retira 75 estudios del rector de Salamanca y sus colaboradores por prácticas fraudulentas*. El País. https://elpais.com/ciencia/2024-10-16/la-editorial-springer-nature-retira-75-estudios-del-rector-de-salamanca-y-sus-colaboradores-por-practicas-fraudulentas.html

Baker, M. (2016). 1,500 Scientists Lift the Lid on Reproducibility. *Nature, 533*(7604), 452–454. https://doi.org/10.1038/533452a

Bartoš, F., Maier, M., Shanks, D. R., Stanley, T. D., Sladekova, M. y Wagenmakers, E.-J. (2023). Meta-analyses in Psychology often Overestimate Evidence for and Size of Effects. *Royal Society Open Science, 10*(7), Artículo 230224. https://doi.org/10.1098/rsos.230224

BBC Mundo. (2024, Noviembre). Quién es Robert Kennedy Jr., el activista antivacunas y heredero de la dinastía Kennedy al que Trump elige para dirigir el Departamento de Salud. BBC Mundo. https://www.bbc.com/mundo/articles/c33e815jdpxo

Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. https://doi.org/10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., y Duggan, M. (2015). Feeling the Future: A Meta-Analysis of 90 Experiments on the Anomalous Anticipation of Random Future Events. *F1000Research, 4*, 1188. https://doi.org/10.12688/f1000research.7177.2

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., … Johnson, V. E. (2018). Redefine Statistical Significance. *Nature Human Behaviour, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berberi, I., y Roche, D. G. (2022). No Evidence that Mandatory Open Data Policies Increase Error Correction. *Nature Ecology & Evolution, 6*(11), 1630–1633. https://doi.org/10.1038/s41559-022-01879-9

Bernal, I. y Perakakis, P. (2023). No-pay Publishing: Use Institutional Repositories. *Nature, 619*(7971), 698–698. https://doi.org/10.1038/d41586-023-02315-z

Brainard, J. (2023). Fast-growing Open-Access Journals Stripped of coveted Impact Factors. *Science, 379*(6639), 1283–1284. https://doi.org/10.1126/science.adi0098

Brembs, B., Button, K., y Munafò, M. (2013). Deep Impact: Unintended Consequences of Journal Rank. *Frontiers in Human Neuroscience, 7,* Artículo 291. https://doi.org/10.3389/fnhum.2013.00291

Brembs, B., Huneman, P., Schönbrodt, F., Nilsonne, G., Susi, T., Siems, R., Perakakis, P., Trachana, V.,

Ma, L. y Rodriguez-Cuadrado, S. (2023). Replacing Academic Journals. *Royal Society Open Science, 10*(2), 230206. https://doi.org/10.1098/rsos.230206

Bohannon, J. (2016). About 40% of economics experiments fail replication survey. *Science, 3.* https://www.science.org/content/article/about-40-economics-experiments-fail-replication-survey

Boyce, V., Prystawski, B., Abutto, A. B., Chen, E. M., Chen, Z., Chiu, H., Ergin, I., Gupta, A., Hu, C., Kemmann, B., Klevak, N., Lua, V. Y. Q., Mazzaferro, M. M., Mon, K., Ogunbamowo, D., Pereira, A., Troutman, J., Tung, S., Uricher, R. y Frank, M. C. (2024). Estimating the Replicability of Psychology Experiments After an Initial Failure to Replicate. *Collabra: Psychology, 10*(1), Artículo 125685. https://doi.org/10.1525/collabra.125685

Butler, L.-A., Matthias, L., Simard, M.-A., Mongeon, P. y Haustein, S. (2023). The Oligopoly's Shift to Open Access: How the Big Five Academic Publishers Profit from Article Processing Charges. *Quantitative Science Studies, 4*(4), 778–799. https://doi.org/10.1162/qss_a_00272

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. y Wu, H. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Cardeña, E. (2018). The Experimental Evidence for Parapsychological Phenomena: A Review. *American Psychologist, 73*(5), 663–677. https://doi.org/10.1037/amp0000236

Catazaro, M. (2023). Saudi Universities Entice Top Scientists to Switch Affiliations—Sometimes with Cash. *Nature, 617*(7961), 446–447. https://doi.org/10.1038/d41586-023-01523-x

Chu, J. S. G. y Evans, J. A. (2021). Slowed Canonical Progress in large Fields of Science. *Proceedings of the National Academy of Sciences, 118*(41), Articulo e2021636118. https://doi.org/10.1073/pnas.2021636118

Cobey, K. D., Lalu, M. M., Skidmore, B., Ahmadzai, N., Grudniewicz, A. y Moher, D. (2018). What is a Predatory Journal? A Scoping Review. *F1000Research, 7.* https://doi.org/10.12688/f1000research.15256.2

Doyen, S., Klein, O., Pichon, C. L. y Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PloS one, 7*(1), Artículo e29081. https://doi.org/10.1371/journal.pone.0029081

Edwards, M. A. y Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science, 34*(1), 51–61. https://doi.org/10.1089/ees.2016.0223

Fleerackers, A., Ratcliff, C. L., Wicke, R., King, A. J. y Jensen, J. D. (2024). Public Understanding of Preprints: How Audiences Make Sense of Unreviewed Research in the News. *Public Understanding of Science, 34*(2), 154–171. https://doi.org/10.1177/09636625241268881

Enserink, M. (2012). Final Report: Stapel Affair Points to Bigger Problems in Social Psychology. *Science*. https://www.science.org/content/article/final-report-stapel-affair-points-bigger-problems-social-psychology

Eronen, M. I. y Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science, 16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. y Nosek, B. A (2021). Investigating the Replicability of Preclinical

Cancer Biology. *eLife, 10*, Article 71601. https://doi.org/10.7554/eLife.71601

Feynman, R. P. (1998). Cargo Cult Science*. En J. Williams (Ed.), *The Art and Science of Analog Circuit Design* (pp. 55–61). Newnes. https://doi.org/10.1016/B978-075067062-3/50008-X

Flake, J. K., y Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. A*dvances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Forozish, A. O. (2024). How the Credibility Revolution Created a Paradigm Shift. Available at *SSRN* 4744474. https://doi.org/10.2139/ssrn.4744474

Gertler, P., Galiani, S. y Romero, M. (2018). How to Make Replication the Norm. *Nature, 554*(7693), 417–419. https://doi.org/10.1038/d41586-018-02108-9

Gerrits, R. G., Jansen, T., Mulyanto, J., van den Berg, M. J., Klazinga, N. S. y Kringos, D. S. (2019). Occurrence and Nature of Questionable Research Practices in the Reporting of Messages and Conclusions in International Scientific Health Services Research Publications: A Structured Assessment of Publications Authored by Researchers in the Netherlands. *BMJ Open, 9*(5), Artículo e027903. https://doi.org/10.1136/bmjopen-2018-027903

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science, 7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Gomez, C. J., Herman, A. C. y Parigi, P. (2022). Leading Countries in Global Science Increasingly Receive more Citations than other Countries Doing Similar Research. *Nature Human Behaviour, 6*(7), 919–929. https://doi.org/10.1038/s41562-022-01351-5

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. y Altman, D. G. (2016). Statistical Tests, P Values, Confidence Intervals, and Power: A guide to Misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grossmann, A. y Brembs, B. (2021). Current Market Rates for Scholarly Publishing Services. *F1000Research, 10*, 20. https://doi.org/10.12688/f1000research.27468.2

Hardwicke, T. E. y 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

Haig, B. D. (2017). Tests of Statistical Significance Made Sound. *Educational and Psychological Measurement, 77*(3), 489–506. https://doi.org/10.1177/0013164416667981

Hussey, I., Alsalti, T., Bosco, F., Elson, M. y Arslan, R. (2025). An Aberrant Abundance of Cronbach's Alpha Values at. 70. *Advances in Methods and Practices in Psychological Science, 8*(1). https://doi.org/10.1177/25152459241287123

Ioannidis, J. P. A. (2005). Why Most Published Research Findings are False. *PLOS Medicine, 2*(8), Artículo e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations are Inflated. *Epidemiology, 19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Jarke, H., Anand-Vembar, S., Alzahawi, S., Andersen, T. L., Bojanić, L., Carstensen, A., Feldman, G., Garcia-Garzon, E., Kapoor, H., Lewis, S., Todsen, A. L., Većkalov, B., Zickfeld, J. H. y Geiger, S. J.

(2022). A Roadmap to Large-Scale Multi-Country *Replications in Psychology. Collabra: Psychology, 8*(1), Artículo 57538. https://doi.org/10.1525/collabra.57538

Jobst, L. J., Bader, M. y Moshagen, M. (2023). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods, 28*(1), 207–221. https://doi.org/10.1037/met0000423

Kang, H. (2021). Sample size determination and power analysis using the G*Power software. *Journal of Educational Evaluation for Health Professions, 18*, 1–12. https://doi.org/10.3352/jeehp.2021.18.17

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., Bakos, B. E., Cousineau, D., Tressoldi, P., Schmidt, K., Grassi, M., Evans, T. R., Yamada, Y., Aczel, B., Adam-Troian, J., Albers, C. J., Alfano, M., Alicke, M. D., Alister, C., … Nosek, B. A. (2023). Raising the Value of Research Studies in Psychological Science by Increasing the Credibility of Research Reports: The Transparent Psi Project. *Royal Society Open Science, 10*(2), Artículo 191375. https://doi.org/10.1098/rsos.191375

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., Cabak Rédei, A., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø.-S., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., Alzahawi, S., … Evans, T. (2023). The Replication Crisis has led to Positive Structural, Procedural, and Community Changes. *Communications Psychology, 1*(1), 1–13. https://doi.org/10.1038/s44271-023-00003-2

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour, 5*(10), 1282-1291. https://doi.org/10.1038/s41562-021-01177-7

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology, 8*(1), Artículo 33267. https://doi.org/10.1525/collabra.33267

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology, 10*(1), Artículo 117094. https://doi.org/10.1525/collabra.117094

Lakens, D., Hilgard, J. y Staaks, J. (2016). On the Reproducibility of Meta-Analyses: Six Practical Recommendations. *BMC Psychology, 4*(1), Artículo 24. https://doi.org/10.1186/s40359-016-0126-3

Lakens, D., Mesquida, C., Rasti, S., y Ditroilo, M. (2024). The benefits of preregistration and registered reports. *Evidence-Based Toxicology, 2*(1), 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P. y van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research, 110,* Artículo 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Liverpool, L. (2023). AI Intensifies Fight against 'Paper Mills' that churn out Fake Research. *Nature, 618*(7964), 222–223. https://doi.org/10.1038/d41586-023-01780-w

Maddi, A. y Sapinho, D. (2023). On the Culture of Open Access: The Sci-hub paradox. *Scientometrics,*

*128*(10), 5647–5658. https://doi.org/10.1007/s11192-023-04792-5

Maier, M., y Lakens, D. (2022). Justify your Alpha: A Primer on two Practical Approaches. *Advances in Methods and Practices in Psychological Science, 5*(2), 25152459221080396. https://doi.org/10.1177/25152459221080396

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to get Beyond the Statistics Wars*. Cambridge University Press.

Mayoni, S. (2022, diciembre 12). *Scientific publishers are reaping huge profits from the work of researchers, and the universities are paying for it. University Post* – Independent of management. https://uniavisen.dk/en/scientific-publishers-are-reaping-huge-profits-from-the-work-of-researchers-and-the-universities-are-paying-for-it/

McShane, B. B., Gal, D., Gelman, A., Robert, C. y Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician, 73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Mole, B. (2024, enero 22). *Top Harvard cancer researchers accused of scientific fraud; 37 studies affected. Ars Technica*. https://arstechnica.com/science/2024/01/top-harvard-cancer-researchers-accused-of-scientific-fraud-37-studies-affected/

Nagy, T., Hergert, J., Elsherif, M., Wallrich, L., Schmidt, K., Waltzer, T., Payne, J. W., Gjoneska, B., Seetahul, Y., Wang, Y. A., Scharfenberg, D., Tyson, G., Yang, Y.-F., Skvortsova, A., Alarie, S., Graves, K. A., Sotola, L. K., Moreau, D. y Rubínová, E. (2024). *Bestiary of Questionable Research Practices in Psychology*. https://osf.io/preprints/psyarxiv/fhk98Nelson

Nelson, L. D., Simmons, J. y Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nuijten, M. B. y Wicherts, J. M. (2024). Implementing Statcheck during Peer Review is Related to a Steep Decline in Statistical-Reporting Inconsistencies. *Advances in Methods and Practices in Psychological Science, 7*(2), 1–14. https://doi.org/10.1177/2515245924125894

Nielsen, M. W. y Andersen, J. P. (2021). Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences of the United States of America, 118*(7), 1–10. https://doi.org/10.1073/pnas.2012208118

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D. y Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology, 73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nosek, B. A., Spies, J. R. y Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/174569161245905

Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. *Science, 349*(6251), Artículo aac4716 https://doi.org/10.1126/science.aac4716

Park, M., Leahey, E. y Funk, R. J. (2023). Papers and Patents are Becoming Less Disruptive over time. *Nature, 613*(7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x

Parsons, S. (2022). Exploring Reliability Heterogeneity with Multiverse Analyses: Data Processing Decisions Unpredictably Influence Measurement

Reliability. *Meta-Psychology, 6,* 1–22. https://doi.org/10.15626/MP.2020.2577

Parsons, J. A., Alperin, J. P., Bishop, D. V. M., Bowman, T. D., Crick, T., de Rijcke, S., Fortunato, S., Frassl, M. A., Guggenheim, C., Hahnel, M., Heise, C., Kramer, B., Labib, K., Loizides, F., Madan, C. R., Moore, S., O'Donnell, D. P., Rice, D. B., Ross-Hellauer, T., ... Sugimoto, C. R. (2022). A Community-Sourced Glossary of open Scholarship Terms. Nature Human Behaviour, 6(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pek, J. y Flora, D. B. (2018). Reporting Effect Sizes in Original Psychological Research: A Discussion and Tutorial. *Psychological Methods, 23*(2), 208–225. https://doi.org/10.1037/met0000126

Perakakis, P. (2021, mayo 3). *¿Qué son los «transformative agreements» y qué necesitamos saber antes de utilizarlos?* Pandelis Perakakis. https://pandelisperakakis.info/2021/05/03/que-son-los-transformative-agreements-y-que-necesitamos-saber-antes-de-utilizarlos/

Retraction Watch (2015, junio 16) *The Retraction Watch Leaderboard.* https://retractionwatch.com/the-retraction-watch-leaderboard/. Accedido el 21/10/2024

Retraction Watch. (2024, marzo 18). *Papers and peer reviews with evidence of ChatGPT writing*. Retraction Watch. https://retractionwatch.com/papers-and-peer-reviews-with-evidence-of-chatgpt-writing/. Recuperado el 21/10/2024.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., Bursalıoğlu, A., Bühler, V., Čadek, M., Çetinçelik, M., Clay, G., Cortijos-Bernabeu, A., Damnjanović, K., Dugue, T. M., Esberg, M., Esteban-Serna, C., Felder, E. N., ... Folke, T. (2021). The General Fault in our Fault Lines. *Nature Human Behaviour, 5*(10), 1369–1380. https://doi.org/10.1038/s41562-021-01092-x

Schiavone, S. R., y Vazire, S. (2023). Reckoning with our Crisis: An agenda for the Field of Social and Personality Psychology. *Perspectives on Psychological Science, 18*(3), 710–722. https://doi.org/10.1177/17456916221101060

Silverstein, P., Elman, C., Montoya, A., McGillivray, B., Pennington, C. R., Harrison, C. H., Steltenpohl, C. N., Röer, J. P., Corker, K. S., Charron, L. M., Elsherif, M., Malicki, M., Hayes-Harb, R., Grinschgl, S., Neal, T., Evans, T. R., Karhulahti, V.-M., Krenzer, W. L. D., Belaus, A., Moreau, D., ... Syed, M. (2024). A guide for Social Science Journal Editors on Easing into Open Science. *Research Integrity and Peer Review, 9*(1), Artículo 2. https://doi.org/10.1186/s41073-023-00141-5

Simmons, J. P., Nelson, L. D. y Simonsohn, U. (2011). False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22(*11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D. y Simmons, J. P. (2014). P-curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Nelson, L. y Simmons, J. (2023, septiembre 1). [113] *Data Litigada: Thank You* (And An Update). Data Colada. https://datacolada.org/113

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A. y James, R. (2019). Assessing Data Availability and Research Reproducibility in Hydrology and Water Resources. *Scientific Data, 6,* Artículo 190030. https://doi.org/10.1038/sdata.2019.30

Steegen, S., Tuerlinckx, F., Gelman, A. y Vanpaemel, W. (2016). Increasing Transparency through a Multiverse Analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stokel-Walker, C. (2023). ChatGPT listed as author on Research Papers: Many Scientists Disapprove. *Nature, 613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z

The Economist. (2023) *There is a Worrying amount of Fraud in Medical Research*. https://www.economist.com/science-and-technology/2023/02/22/there-is-a-worrying-amount-of-fraud-in-medical-research

Weissgerber, T. L., Milic, N. M., Winham, S. J. y Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology, 13*(4), Artículo e1002128. https://doi.org/10.1371/journal.pbio.1002128

Tsui, A. S. (2022). From Traditional Research to Responsible Research: The Necessity of Scientific Freedom and Scientific Responsibility for better Societies. *Annual Review of Organizational Psychology and Organizational Behavior, 9*, 1–32. https://doi.org/10.1146/annurev-orgpsych-062021-021303

UNESCO. (2022). *Understanding open science*—UNESCO Biblioteca Digital (UNESCO open science toolkit, p. 6) [Factsheet]. https://doi.org/10.54677/UTCD9302

Van den Akker, O. R., van Assen, M. A. L. M., Bakker, M., Elsherif, M., Wong, T. K. y Wicherts, J. M. (2024). Preregistration in Practice: A Comparison of Preregistered and Non-Preregistered Studies in Psychology. *Behavior Research Methods, 56*(6), 5424–5433. https://doi.org/10.3758/s13428-023-02277-0

Van Noorden, R. (2023a). Medicine is Plagued by Untrustworthy Clinical Trials. How many Studies are Faked or Flawed? *Nature, 619*(7970), 454–458. https://doi.org/10.1038/d41586-023-02299-w

Van Noorden, R. (2023b). More than 10,000 research papers were retracted in 2023—A new record. *Nature, 624*(7992), 479–481. https://doi.org/10.1038/d41586-023-03974-8

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science, 13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Walker, R. y Rocha da Silva, P. (2015). Emerging Trends in Peer Review—A Survey. *Frontiers in Neuroscience, 9,* Artículo 169. https://doi.org/10.3389/fnins.2015.00169

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N. y Morey, R. D. (2018). Bayesian Inference for Psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagner III, J. A. (2022). The Influence of Unpublished Studies on Results of Recent Meta-Analyses: Publication Bias, the File Drawer Problem, and Implications for the Replication Crisis. *International Journal of Social Research Methodology, 25*(5), 639–644. https://doi.org/10.1080/13645579.2021.1922805

Wang, H., Chen, Y., Lin, Y., Abesig, J., Wu, I. X. y Tam, W. (2021). The Methodological Quality of Individual Participant Data Meta-Analysis on Intervention Effects: Systematic Review. *BMJ, 372,* Artículo 736. https://doi.org/10.1136/bmj.n736

White, N., Parsons, R., Collins, G. y Barnett, A. (2023). Evidence of Questionable Research Practices in Clinical Prediction Models. *BMC Medicine, 21*(1), Artículo 339. https://doi.org/10.1186/s12916-023-03048-6

Wingen, T., Berkessel, J. B. y Dohle, S. (2022). Caution, preprint! Brief Explanations Allow Nonscientists to Differentiate between Preprints and Peer-Reviewed Journal Articles. *Advances in Methods and Practices in Psychological Science, 5*(1), 1–15. https://doi.org/10.1177/25152459211070559

Wolfram, D., Wang, P., Hembree, A. y Park, H. (2020). Open Peer Review: Promoting Transparency in Open Science. *Scientometrics, 125*(2), 1033–1051. https://doi.org/10.1007/s11192-020-03488-4

Wood, B. D. K., Müller, R. y Brown, A. N. (2018). Push Button Replication: Is Impact Evaluation Evidence for International Development Verifiable? *PLOS ONE, 13*(12), Artículo e0209416. https://doi.org/10.1371/journal.pone.0209416

Youyou, W., Yang, Y. y Uzzi, B. (2023). A Discipline-wide Investigation of the Replicability of Psychology Papers over the past Two Decades. *Proceedings of the National Academy of Sciences, 120*(6), Artículo e2208863120. https://doi.org/10.1073/pnas.2208863120

Zong, Q., Xie, Y. y Liang, J. (2020). Does open Peer Review Improve Citation Count? Evidence from a Propensity Score Matching Analysis of PeerJ. *Scientometrics, 125*(1), 607–623. https://doi.org/10.1007/s11192-020-03545-y

## Materiales Suplemantarios

Existen diversas fuentes documentando los negacionismos y fenómenos parecidos como el terraplanismo (ver Bauer, 2014; Björnberg et al., 2017; Schmid y Betsch, 2019).

**Divulgación en redes sociales**

La divulgación a través de redes sociales y en formato audiovisual es de especial relevancia en los medios actuales, donde destacamos canales de *Youtube* como *QuantumFracture* en castellano, (https://www.youtube.com/@QuantumFracture), y *Kurzgesagt* en inglés (https://www.youtube.com/@kurzgesagt) por su rigor, transparencia y calidad de producción.

**Etapas de implementación de reformas**

Aquí Nosek et al. (2022) proponen una secuencia de implementación consistente en infraestructura («hacerlo posible»), usuarios («hacerlo fácil»), comunidades («hacerlo norma»), agencias de incentivos («hacerlo deseable»), y legislación («hacerlo un requisito»), donde consideramos que nos encontramos entre las comunidades y las agencias de incentivos. Esto implicaría que estamos pasando de los primeros usuarios (o *early adopters*) a la primera masificación (o *early majority*), con el objetivo de llegar a una masificación total (o *late majority*) gracias a legislaciones y agencias de incentivos.

**Tabla S1**

*Glosario con Términos Básicos Sobre la Crisis de Replicación y revolución de la credibilidad*

| **Clasificación** | **Término español *(traducción)*** | ***Definición*** |
|---|---|---|
| **Conceptos generales** | Crisis de replicación (*replication crisis*) | Crisis metodológica actual donde muchos estudios científicos tomados como confiables no pudieron ser replicados, dañando la credibilidad de la comunidad científica. |

| | | |
|---|---|---|
| | Meta-ciencia (*meta-science*) | Disciplina científica que estudia cómo la comunidad científica desarrolla su actividad, como sus sistemas de prácticas y métodos, productos, control de calidad y diseminación. Asociada a la cienciometría y la bibliometría. |
| | Replicabilidad (*replicability*) | Grado en el que un hallazgo científico puede volver a obtenerse cambiando aspectos no esenciales del contexto en el que fue obtenido (equipo investigador, muestras, procedimientos, o instrumentos) |
| | Reproducibilidad (*reproducibility*) | Grado en el que un estudio científico para re-generar los hallazgos usando los datos, análisis y materiales empleados por los investigadores originales |
| | Revolución de la credibilidad (*credibility revolution*) | Conjunto de reformas estructurales de la investigación científica para aumentar la replicabilidad y robustez de los hallazgos científicos, normalmente consistentes en prácticas de ciencia abierta, evitar prácticas cuestionables y fomentar el análisis y aumento de métodos robustos (e.g., muestras más grandes o pre-registros) |
| **Diagnóstico** | Estructura de incentivos (*incentive structure*) | Conjunto de reglas de recompensa explícitas e implícitas dentro del sistema académico, como la ponderación de méritos en contratos y promoción de puestos académicos o prestigio en investigación, docencia y gestión (e.g., número de publicaciones en revistas de impacto, presentaciones en congresos, u horas de docencia). Son comúnmente contempladas como contraproducentes con la calidad científica y académica, ya que desincentivan el rigor |
| | Grados de libertad del investigador (*researcher degrees of freedom*) | grado de amplitud del "jardín de senderos que se bifurcan", dando mayor arbitrariedad a un trabajo académico o científico y dañando su replicabilidad |
| | Problema del jardín de senderos que se bifurcan (*forking path problem*) | Árbol de decisiones tomadas por investigadores para realizar un análisis estadístico, normalmente ignorando el impacto que ha podido tener en el resultado obtenido, y expuesto a inflación de falsos positivos y por tanto a una pérdida de replicabilidad |
| | Mercado de autorías / artículos (*author / article marketplace*) | Conjunto de organizaciones e investigadores que ofrecen artículos rápidos a cambio de coautorías o viceversa, normalmente también con pagos o cargos económicos |
| | Molino de papers (*paper mill*) | Organización involucrada en mala praxis científica, produciendo mútliples artículos con datos fraudulentos, texto plagiado, o manipulación de figuras. |
| | Pagar para publicar (*pay per publish*) | Aforismo que describe prácticas de negocio en editoriales académicas donde la calidad o rigor de la revisión por pares es despriorizada (o directamente anulada) para obtener pagos económicos a los autores de las publicaciones. Comúnmente empleadas por revistas predatorias |
| | Prácticas Cuestionables de Investigación (PCI) (*Questionnable Research Practices* | Grupo de actividades que intencionada o inintencionadamente distorsionan los datos para favorecer las hipótesis de los investigadores (e.g., inclusión interesada de datos, selección de variables en el artículo, o p-hacking). Ver Figura 1 para una lista detallada. |

| | | |
|---|---|---|
| | *(QRPs)*) | |
| | Publicar o perecer (*publish or perish*) | Aforismo que describe la cultura de presión de los académicos por publicar con regularidad en revistas especializadas sus investigaciones si desean progresar en su carrera |
| | Revista predatoria (*predatory journal*) | Revista científica que sigue prácticas de negocio donde las editoriales buscan obtener pagos del mayor número de autores posibles a costa de una menor calidad o rigor (e.g., con una revisión por pares deficiente, o ninguna en absoluto) |
| | Ciencia salami (*salami science*) | Estrategia de publicación consistente en dividir un estudio científico en varias partes para maximizar la cantidad de publicaciones a costa de la calidad de su trabajo. Concepto adaptado de la "táctica del salami" (*salami slicing*) |
| **Soluciones** | Análisis de potencia (*power analysis*) | Conjunto de técnicas estadísticas para estimar la potencia estadística de un contraste estadístico concreto (es decir, la probabilidad de rechazar una hipótesis nula falsa, o poder detectar correctamente la presencia de un efecto), normalmente especificando el tamaño del efecto, confianza, potencia deseada y otros parámetros |
| | Ciencia abierta / academia abierta (*open science / open scholarhip*) | Conjunto de prácticas académicas para fomentar la transparencia de trabajos científicos y académicos. Incluye el acceso libre y gratuito a artículos, datos, análisis, protocolos y otros materiales, pero también puede incluir informes de revisión por pares o recursos docentes |
| | Estudio de replicación (*replication study*) | Estudio científico dedicado a evaluar la replicabilidad de uno o más estudios científicos buscando un alto grado de garantías (e.g., muestras altas, análisis robustos, pre-registros) |
| | Informe registrado (*registered report*) | Publicación similar al pre-registro, pero en una revista científica y tras una revisión por pares. Así se fomenta el compromiso de autores y revista de publicar los hallazgos sin importar los resultados |
| | Muestras WEIRD (*WEIRD samples*) | Muestras obtenidas de países occidentales, con niveles altos de educación, industrialización, riqueza y políticas democráticas |
| | Pre-registros (*preregistation*) | Publicación científica donde se especifica el marco teórico, hipótesis y métodos de un estudio antes de la recogida de datos. Así se fomenta el compromiso de los autores de publicar los hallazgos sin importar los resultados |
| | Acceso abierto verde (*green open access*) | Modalidad de acceso abierto donde los autores de trabajos científicos cuelgan *pre-prints, post-prints*, o los propios artículos en repositorios abiertos como forma de auto-archivar sus publicaciones. |
| | Acceso abierto oro (*gold open access*) | Modalidad de acceso abierto donde los artículos son de acceso abierto desde las revistas mediante el pago de los APC. |
| | Acceso abierto diamante (*diamond open access*) | Modalidad de acceso abierto donde los artículos son de acceso abierto desde las revistas sin pago de APC, también denominado *no-pay publishing*. |
| | Pre-impresión (*pre-* | Manuscrito académico publicado en un repositorio antes de ser enviado a la |

| | |
|---|---|
| *print*) | revista, normalmente sin haber pasado por revisión por pares. Es necesario tomar la información que contienen con cautela y deben indicar que no están revisados por pares. |
| Post-impresión (*post-print*) | Manuscrito académico publicado en un repositorio después de ser revisado y publicado por una revista, normalmente tras una revisión por pares. Esto permite el acceso libre al manuscrito. |
| Replicación conceptual (*conceptual replication*) | Hallazgo científico que ha sido replicado siguiendo la idea pero con diferencias sustanciales en los métodos empleados en el estudio original. Esto permite contrastar la replicabilidad |
| Replicación directa o exacta (*direct / exact replication*) | Hallazgo científico que ha sido replicado siguiendo exactamente o con un grado muy alto los métodos empleados en el estudio original. Esto permite contrastar la replicabilidad del estudio primario con altas garantías. |
| Replicación sistemática o experimental (*systematic / experimental replication*) | Hallazgo científico que ha sido replicado siguiendo en gran medida los métodos empleados en el estudio original salvo desviaciones justificadas en el procedimiento (e.g., una operacionalización distinta de las variables, instrumentos distintos, o muestras con demografía distinta). Esto permite contrastar la replicabilidad del estudio primario y estudios similares. |

**Análisis de potencia estadística**

Estos análisis requieren especificar niveles de potencia, confianza y tamaños del efecto para un análisis concreto, obteniendo un tamaño muestral mínimo para garantizar esas condiciones (Kang, 2021; Jobst et al., 2023). Un software muy conocido es *G*Power*, aunque para técnicas complejas es necesario usar simulaciones (Kang et al., 2021; Kyriazos, 2018). Existen distintas perspectivas, como los análisis de precisión y secuenciales.

**El uso de Inteligencia Artificial Generativa**

Paralelamente, el uso de IAs en la actividad científica ha cobrado relevancia, ya que permitiría en principio una mayor eficiencia. Sin embargo, su uso indebido está comenzando a ocurrir a gran escala y es preocupante, como la escritura de secciones de artículos sin el debido reconocimiento, errores, o textos con poco valor añadido (Retraction Watch, 2024). Esto se enlaza con una tradición de «timos» (*stings*) satíricos al sistema editorial, donde hay contenidos generados por IA confusos o totalmente aleatorios que consiguieron publicarse (Al-Khatib et al., 2016; una lista disponible en Wikipedia). El mensaje de estos stings parece estar cumpliéndose en forma de artículos fraudulentos generados por IA (Liverpool, 2023). Posibles soluciones podrían ser recoger a las IA en la autoría (Stokel-Walker, 2023), declarar el uso de IA en el artículo, o limitarlas a la mejora de expresión escrita o tareas repetitivas con una estructura clara.

**Explicación de Figura 3**

Para poder realizar un artículo científico, será necesario acceder a artículos de pago, constituyendo un muro de pago (*paywall*) al conocimiento científico. Una vez realizado, se envía a una revista donde un grupo de revisores y editores sin remuneración revisan el artículo (en un proceso que puede durar meses o años, bajo el incentivo de que ser revisor y editor es un mérito en la carrera investigadora). Dicho trabajo gratuito está estimado en hasta millones de millones de dólares para países como EE. UU. (Aczel et al., 2021). Si resulta aceptado, la editorial se encargará de formatearlo, colgarlo en su página web y (solo en ciertos casos) imprimirlo físicamente. Esto va acompañado de la cesión de todos los derechos de reproducción asociados al artículo.