Evaluación de Política

Respaldo académico · marco teórico, fórmulas y bibliografía

Este documento sostiene metodológicamente las ocho mecánicas del módulo web de Evaluación (versión 2.0, con literatura 2020-2026 incorporada). Cubre la justificación de cada decisión de diseño, las fórmulas usadas y las referencias completas. Pensado para consultores y comités técnicos que quieran auditar el método.

1 · Marco general del módulo

El módulo asume el enfoque **theory-based evaluation** (Weiss 1995, Mayne 2008+, Funnell & Rogers 2011) como armazón conceptual, complementado con métodos contrafactuales modernos (Athey & Imbens 2017, Abadie 2010, Callaway-Sant'Anna 2021) y los criterios OCDE-DAC actualizados en 2019 como lenguaje de comunicación con organismos multilaterales y comités técnicos.

El módulo NO impone un método único. Diferentes preguntas evaluativas exigen diferentes métodos: una pregunta causal exige contrafáctico, una de valor exige métodos cualitativos o económicos, una de proceso exige observación participante. El selector del paso 4 explicita esa decisión y la deja documentada.

2 · Tipología de preguntas evaluativas (mecánica 1)

Patton (*Utilization-Focused Evaluation*, 4ª ed. 2008; 5ª ed. 2022) distingue cinco familias de preguntas que el módulo adopta:

- **Descripción** · ¿qué está pasando? Caracteriza magnitud, distribución, perfiles. No pretende atribuir causalidad.
- **Atribución causal** · ¿la política causó el cambio observado? Exige contrafáctico (RCT, DiD, RD, control sintético, matching).
- **Valor** · ¿vale la pena lo que cuesta? Juicio normativo: costo-beneficio social, value-for-money, equidad distributiva.
- **Proceso** · ¿cómo se está implementando? Fidelidad, calidad operativa, brechas diseño-ejecución.
- **Gestión** · ¿la organización está aprendiendo? Adaptación, decisiones gerenciales, uso de evidencia (Patton, *developmental evaluation*, 2011).

Rossi, Lipsey & Henry (*Evaluation: A Systematic Approach*, 8ª ed. 2018) complementan con cuatro alcances temporales que el módulo también registra: ex-ante, concurrente, ex-post y meta-evaluación.

3 · Teoría de cambio y marco lógico (mecánica 2)

El módulo adopta el marco lógico clásico de CEPAL/ILPES (Ortegón, Pacheco & Prieto, 2005) con cinco niveles canónicos: insumos → actividades → productos → resultados → impacto. Esta estructura es compatible con el Logical Framework de USAID y el formato de matriz de marco lógico exigido en formulación de proyectos por la mayoría de organismos multilaterales en América Latina.

La capa de **supuestos transversales** proviene de la tradición de *contribution analysis* (Mayne 2001, 2008+): para que una cadena causal funcione, deben mantenerse condiciones contextuales que la intervención no controla. Documentarlas explícitamente permite identificar dónde se rompe la cadena cuando algo falla.

Para problemas complejos, Pawson & Tilley (*Realistic Evaluation*, 1997) proponen formular la teoría como tríadas **contexto-mecanismo-outcome**. Esta refinación NO se exige en el módulo (introducirla agrega complejidad sin que el usuario promedio la aproveche), pero queda mencionada como ruta de profundización avanzada.

4 · Indicadores SMART y validación (mecánica 3)

El acrónimo SMART proviene de Doran (1981) y fue adoptado por la tradición de gestión por resultados. El módulo lo aplica con cinco criterios codificados en el validador automático:

Letra	Criterio	Validación en el módulo
S	Specific (específico)	Tiene nombre + definición operativa
M	Measurable (medible)	Tiene fórmula explícita
A	Achievable (alcanzable)	Tiene meta concreta
R	Relevant (relevante)	Tiene nivel asignado (teoría de cambio) + fuente verificable
T	Time-bound (temporal)	Tiene frecuencia definida

El chip SMART por fila muestra el score 0-5 y las letras faltantes (ej. *3/5 · falta M+R*). La validación es solo informativa: el módulo no bloquea indicadores parciales — un comité técnico puede aceptar indicadores en construcción si la línea base aún no se ha levantado.

5 · Métodos evaluativos · frontera 2020-2026 (mecánica 4)

La versión 2 del módulo precarga **catorce métodos**, con seis estimadores frontera de la literatura 2018-2024 integrados al lado de los clásicos. La razón del salto: TWFE (DID clásico con múltiples períodos y rollout escalonado) produce sesgos serios — pesos negativos, signos invertidos — documentados por Goodman-Bacon (2021). En contextos colombianos con políticas que entran por fases (PND territorial, programas regionales), esto es regla, no excepción.

Métodos causales estado del arte (★)

- **DID escalonado · ATT(g,t)** (Callaway & Sant'Anna 2021; Sun & Abraham 2021; Borusyak-Jaravel-Spiess 2024). Estima un ATT por cada cohorte de inicio y período post-tratamiento; agrega vía exposure-weighted average. Evita la contaminación TWFE. Paquete did (R) o csdid (Stata).
- **Synthetic Control aumentado** (Ben-Michael, Feller & Rothstein 2021). Pesos de Abadie + ridge regression para desesgar el pre-ajuste imperfecto. Inferencia por placebos in-space (Abadie 2010) + p-values exactos. Estándar para 1 unidad tratada con N pequeño.
- **RDD moderno** (Cattaneo, Keele & Titiunik 2023, *A Practical Introduction*). Bandwidth óptimo MSE (Calonico-Cattaneo-Titiunik 2014) + bandas robustas + test McCrary de manipulación. Paquete rdrobust.

- **Double Machine Learning** (Chernozhukov et al. 2018, Econometrics Journal). Cross-fitting 5-fold. ML para funciones nuisance (propensity, outcome) preservando inferencia válida del parámetro causal. Paquete DoubleML (R/Python) · EconML (Microsoft).
- **Causal Forests** (Wager & Athey 2018; Athey-Tibshirani-Wager 2019). Honest splitting + cross-fitting para estimar efectos heterogéneos (CATE). N ≥ 5.000. Paquete grf.
- **Análisis de Contribución** (Mayne retrospectiva CJPE 2024; WB IEG Quality Guidance 2023). ToC explícita + identificación de riesgos por flecha causal + recolección mixta por riesgo. Narrativa de contribución auditable. Indicado cuando RCT/cuasi-experimental no es factible (reformas institucionales, programas multicomponente).

Métodos clásicos (siguen disponibles)

- **RCT.** Banerjee, Duflo & Kremer (Nobel 2019, J-PAL). Asignación aleatoria. Estándar de oro pero caro y a veces inviable.
- **DID clásico (2 períodos).** Card & Krueger 1994. Solo válido cuando todas las unidades se tratan al mismo tiempo. Si el rollout es escalonado, migrar a DID escalonado.
- **RD clásico.** Thistlethwaite-Campbell 1960 (original). Imbens-Lemieux 2008. Sensible a bandwidth ad-hoc. Cattaneo 2023 lo reemplaza en el estado del arte.
- **Control sintético clásico.** Abadie, Diamond & Hainmueller 2010. Mejor usar la versión aumentada de Ben-Michael 2021 si el pre-ajuste no es exacto.
- **Matching / PSM.** Rosenbaum & Rubin 1983. Para datos observacionales. Solo controla sesgos observables. Para >20 covariables, DML domina.
- **Cualitativo.** Patton 2022 (5ª ed.); Yin 2018 (6ª ed.). Esencial para preguntas de valor y proceso.
- **Mixto.** Creswell & Plano Clark 2017. QUANT + QUAL. Casi siempre el más defendible ante comités diversos.
- **Value-for-Money + MVPF.** HM Treasury *Green Book* 2022 + Hendren-Sprung-Keyser MVPF (NBER 2020). Ver sección 7.

*El selector del módulo detecta automáticamente cuando el tratamiento es **escalonado** (toggle del paso 4) y emite warning con redirección al DID escalonado si el usuario tenía DID clásico seleccionado. Esta lógica implementa el consenso post-2020 sobre la inferencia causal aplicada.*

6 · Corrección por hipótesis múltiples (MHT)

Cuando una evaluación reporta efectos sobre múltiples outcomes primarios, la probabilidad de encontrar al menos un falso positivo crece rápidamente. Con $\alpha = 0.05$ y k outcomes independientes, P(al menos un falso positivo) = 1 − 0.95^k. Para k = 5 outcomes, esa probabilidad es 23%; para k = 20, supera 64%. El módulo aplica una regla automática según el número de outcomes primarios pre-registrados:

k primarios	Corrección recomendada	Justificación
1	No requerida	Un solo outcome → no hay inflación FWER
2-3	Bonferroni (α / k)	Control FWER conservador y simple
4-8	Holm (1979) o Romano-Wolf (2005)	Holm domina a Bonferroni; RW captura correlaciones
≥9	Benjamini-Hochberg FDR ≤ 0.10	Control FDR razonable para muchos outcomes

Referencias clave: Anderson (2008, JASA) sobre Romano-Wolf; List, Shaikh & Xu (2019, Experimental Economics) sobre buenas prácticas MHT en RCT; Benjamini & Hochberg (1995, JRSS-B) sobre FDR. La corrección y la justificación quedan registradas en el Pre-Analysis Plan exportable.

7 · Análisis económico · CBA · MVPF · CEA (mecánica 6)

La versión 2 incorpora un paso opcional con tres calculadoras económicas convivientes. La distinción importa porque cada una pregunta algo diferente y exige supuestos distintos:

CBA · Cost-Benefit Analysis (Green Book HM Treasury 2022)

VPN = Σ_t (B_t – C_t) / (1 + r)^t, con t ∈ [1, h]. El módulo permite configurar r (DNP: 9%; Green Book: 3.5% ajustado) y h ∈ [1, 50] años. Reporta VPN en pesos colombianos y B/C como métrica auxiliar. El Green Book 2022 incorpora *weights distribucionales* para pesar más los beneficios sobre poblaciones vulnerables.

MVPF · Marginal Value of Public Funds (Hendren-Sprung-Keyser 2020)

MVPF = WTP_recipients / Net cost to government. WTP_recipients = valor monetario que los beneficiarios darían a la política. Net cost = costo presupuestal ± efectos fiscales (ahorros por menores beneficios futuros, mayores impuestos por más empleo, etc.). MVPF > 1 → política Pareto-superior. El propósito original (Hendren, NBER WP 26144) es comparar programas heterogéneos (transferencias, becas, capacitación, seguros) en un solo número, evitando la falsa precisión de monetizar todos los beneficios sociales.

CEA · Cost-Effectiveness (J-PAL)

CEA = Costo total / Outcome total en unidad natural (años adicionales de escolaridad, vidas salvadas, casos prevenidos, kilogramos de CO■ evitados). Útil cuando monetizar el beneficio es éticamente controvertido o técnicamente imposible. J-PAL publica un repositorio público de CEAs comparables (povertyactionlab.org/cea).

> *El PAP exportable incluye los tres números calculados, un disclaimer sobre el análisis de sensibilidad esperado, y la recomendación de reportar tornado diagram con los 5 parámetros más sensibles.*

8 · Criterios OCDE-DAC en detalle (mecánica 5)

Los criterios DAC fueron definidos por el Development Assistance Committee de la OCDE en 1991 y actualizados sustancialmente en 2019 (comunicación oficial OCDE/DAC/STAT(2019)16) para alinear con la Agenda 2030 de ODS. Estado actual de los seis:

Criterio (EN/ES)	Pregunta canónica
Relevance / Relevancia	¿La intervención está haciendo lo correcto?
Coherence / Coherencia	¿Encaja con otras intervenciones, políticas y prioridades?
Effectiveness / Efectividad	¿Está logrando sus objetivos?
Efficiency / Eficiencia	¿Hace buen uso de los recursos?
Impact / Impacto	¿Qué diferencia hace? (incluye contrafáctico)
Sustainability / Sostenibilidad	¿Los beneficios netos durarán en el tiempo?

Coherence es el criterio agregado en la actualización 2019 (antes solo eran 5). Mide compatibilidad interna (con el resto del portafolio de la organización ejecutora) y externa (con políticas nacionales y otros donantes/actores).

> *El módulo permite marcar un criterio como no aplica con justificación. Esto es válido y consistente con la guía OCDE-DAC: no toda evaluación necesita cubrir los 6.*

9 · Pre-Analysis Plan exportable

El módulo trata el plan de evaluación como el equivalente público a un **Pre-Analysis Plan** (PAP) en investigación experimental. Los PAP fueron impulsados por el AEA RCT Registry (Olken 2015 JEP) y la comunidad de pre-registration en ciencias sociales (Nosek et al. 2018 PNAS). Su lógica:

- **Antes** de recolectar datos finales, se registra qué se va a hacer, cómo y con qué criterios. Timestamp público en socialscienceregistry.org o osf.io.
- **Durante** el análisis, se mantiene el plan; cualquier desviación se documenta en addendum fechado *antes* del unblinding.
- **Después**, se reportan hallazgos contra el plan; anything no pre-registrado se reporta como exploratorio.

Esto blinda al evaluador contra el *p-hacking* (Simmons, Nelson & Simonsohn 2011 Psych Sci) y el *HARKing* (Hypothesizing After Results are Known, Kerr 1998), las dos fallas más extendidas en evaluación de política pública aplicada.

El módulo exporta un PAP estructurado en 13 secciones (research question, hipótesis primarias/secundarias, outcomes pre-registrados, teoría de cambio, identificación + especificación econométrica, MHT correction, heterogeneidad pre-especificada, cálculo de poder, recolección de datos, análisis económico opcional, OCDE-DAC, limitaciones + protocolo de desviaciones, referencias) listo para subir a AEA o OSF.

10 · Sistemas oficiales de monitoreo y evaluación · Colombia

El módulo está pensado para ser compatible con tres sistemas institucionales colombianos:

- **SINERGIA** · Sistema Nacional de Evaluación de Gestión y Resultados, DNP. CONPES 3134/2001 (creación). Hoy gestiona el monitoreo del PND y evaluaciones de política pública. La **tipología de evaluaciones DNP** (ejecutiva, operaciones, resultados, impacto, institucional, mapas de evidencia) está integrada al paso 1 del módulo desde la versión 2. La matriz de indicadores .csv del módulo es importable a SINERGIA con ajuste mínimo de columnas.
- **SUIFP** · Sistema Unificado de Inversiones y Finanzas Públicas. Para proyectos de inversión pública formulados con marco lógico. La teoría de cambio del módulo se mapea directamente a la estructura SUIFP.
- **SINERGIA-Seguimiento** · módulo de seguimiento a metas del PND con visualización pública. Indicadores SMART exportados por este módulo son compatibles con el formato esperado.

Como referente internacional, **Ivàlua** (Institut Català d'Avaluació de Polítiques Públiques) es la institución pública de referencia en habla hispana con guías metodológicas abiertas que son consistentes con la arquitectura de este módulo.

11 · Bibliografía

- Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2), 391-425.
- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490), 493-505.
- Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484), 1481-1495.
- Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2), 3-32.
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *Annals of Statistics*, 47(2), 1148-1178.
- Banerjee, A., & Duflo, E. (2009). The Experimental Approach to Development Economics. *Annual Review of Economics*, 1, 151-178.
- Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The Augmented Synthetic Control Method. *Journal of the American Statistical Association*, 116(536), 1789-1803.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289-300.
- Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting Event Study Designs: Robust and Efficient Estimation. *Review of Economic Studies*, 91(6), 3253-3285.
- Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2), 200-230.
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6), 2295-2326.
- Card, D., & Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772-793.
- Cattaneo, M. D., Keele, L., & Titiunik, R. (2023). *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge Elements in Quantitative and Computational Methods for the Social Sciences.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1), C1-C68.
- CONPES 3134 de 2001. *Lineamientos para el Sistema Nacional de Evaluación de Resultados*. DNP, Bogotá.
- Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and Conducting Mixed Methods Research* (3ª ed.). SAGE.
- De Chaisemartin, C., & D'Haultfœuille, X. (2022). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. *The Econometrics Journal*, 26(3), C1-C30.
- DNP (2014). *Guía metodológica para el seguimiento y evaluación de políticas públicas*. Dirección de Seguimiento y Evaluación de Políticas Públicas, Bogotá.
- Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review*, 70(11), 35-36.

- Funnell, S. C., & Rogers, P. J. (2011). *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. Jossey-Bass.
- Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics*, 225(2), 254-277.
- Hendren, N., & Sprung-Keyser, B. (2020). A Unified Welfare Analysis of Government Policies. *Quarterly Journal of Economics*, 135(3), 1209-1318. NBER Working Paper 26144.
- HM Treasury (2022). *The Green Book: Central Government Guidance on Appraisal and Evaluation*. London: HMSO.
- Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2), 615-635.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- J-PAL (2024). *Cost-Effectiveness Analysis: Methodology and Applications in Education and Health*. Abdul Latif Jameel Poverty Action Lab.
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196-217.
- List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple Hypothesis Testing in Experimental Economics. *Experimental Economics*, 22(4), 773-793.
- Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. *The Canadian Journal of Program Evaluation*, 16(1), 1-24.
- Mayne, J. (2008). *Contribution analysis: An approach to exploring cause and effect*. ILAC Brief 16.
- Mayne, J. (2024). Contribution Analysis: A Retrospective. *The Canadian Journal of Program Evaluation*, 38(2), 246-260.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS*, 115(11), 2600-2606.
- OCDE-DAC (2019). *Better Criteria for Better Evaluation: Revised Evaluation Criteria – Definitions and Principles for Use*. OECD/DAC Network on Development Evaluation.
- Olken, B. A. (2015). Promises and Perils of Pre-Analysis Plans. *Journal of Economic Perspectives*, 29(3), 61-80.
- Ortegón, E., Pacheco, J. F., & Prieto, A. (2005). *Metodología del marco lógico para la planificación, el seguimiento y la evaluación de proyectos y programas*. Serie Manuales 42, CEPAL/ILPES, Santiago.
- Patton, M. Q. (2011). *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. Guilford Press.
- Patton, M. Q. (2022). *Qualitative Research & Evaluation Methods* (5ª ed.). SAGE.
- Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. SAGE.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2ª ed.). Cambridge University Press.
- Romano, J. P., & Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, 73(4), 1237-1282.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
- Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2018). *Evaluation: A Systematic Approach* (8ª ed.). SAGE.

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, 22(11), 1359-1366.
- Sun, L., & Abraham, S. (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, 225(2), 175-199.
- Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), 309-317.
- Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.
- Weiss, C. H. (1995). Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families. En J. Connell et al. (eds.), *New Approaches to Evaluating Community Initiatives*. Aspen Institute.
- World Bank IEG (2023). *Quality Guidance for Contribution Analysis*. Independent Evaluation Group, Washington DC.
- Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods* (6ª ed.). SAGE.