IMPORTANCIA DEL RIGOR METODOLÓGICO EN INGENIERÍA DE SOFTWARE EMPÍRICA

Sira Vegas
Universidad Politécnica de Madrid

Jornadas SISTEDES, 12 de septiembre, 2023
Ciudad Real, España

Proyecto PID2022-137846NB-I00 financiado por:

CONTENIDOS

Importancia de la fiabilidad de los resultados

Caracterización de la validez

Estado actual de la metodología en IS empírica:

- Experimentos crossover
- Estudios MSR
- Experimentos DL
- Tipo participante

IMPORTANCIA DE LA FIABILIDAD DE LOS RESULTADOS

IMPORTANCIA DE LA FIABILIDAD DE LOS RESULTADOS

- Una cuestión fundamental en la Ingeniería de Software Empírica (ISE) es la fiabilidad de los resultados de los estudios empíricos
- **Fiabilidad:** Grado en el que un resultado empírico corresponde exactamente con la realidad
- Depende del grado de validez del estudio

¿SON FIABLES LOS RESULTADOS?

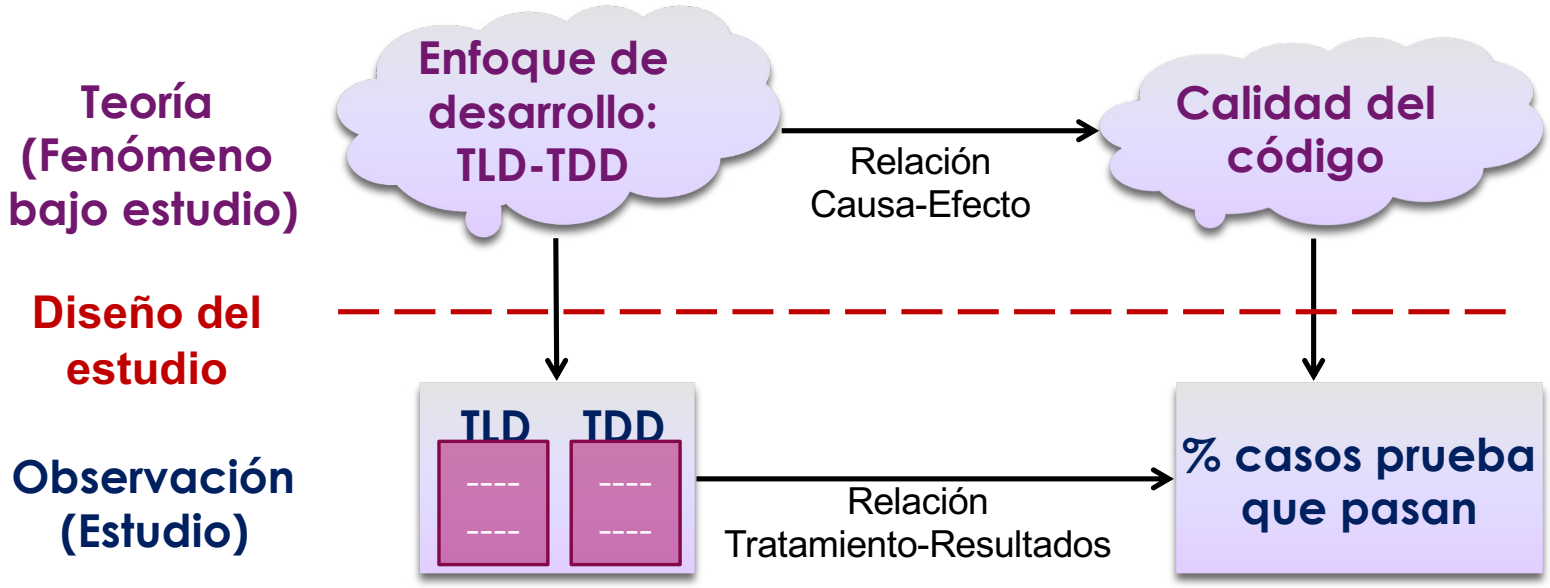
- En psicología se está experimentando una crisis relacionada con el "nivel de duda sin precedentes entre los profesionales sobre la confiabilidad de los hallazgos de la investigación en el campo" (Pashler y Wagenmakers, 2012)
- En IS hay estudios que arrojan preocupaciones similares (Sjoberg et al. 2005), (Shepperd et al., 2014)

CARACTERIZACIÓN DE LA VALIDEZ

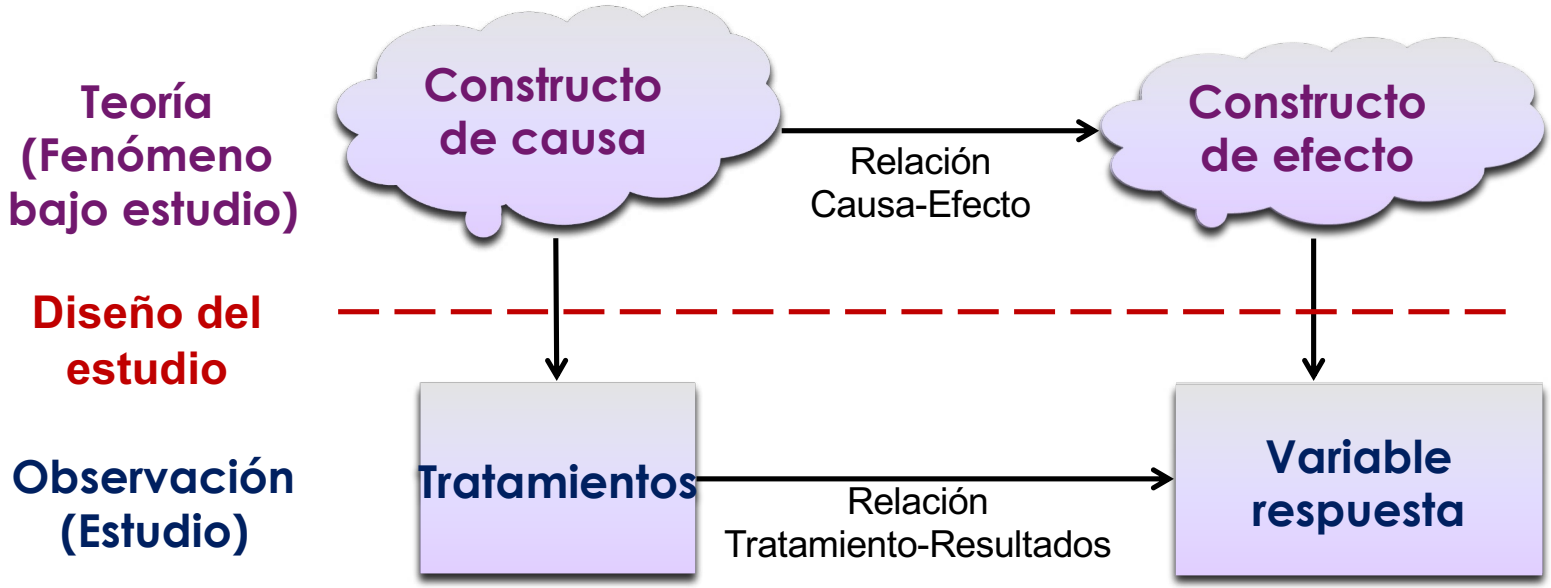
CARACTERIZACIÓN DE LA VALIDEZ

- Comúnmente se mide a diferentes niveles
- Posibles causas del desajuste entre resultados y realidad

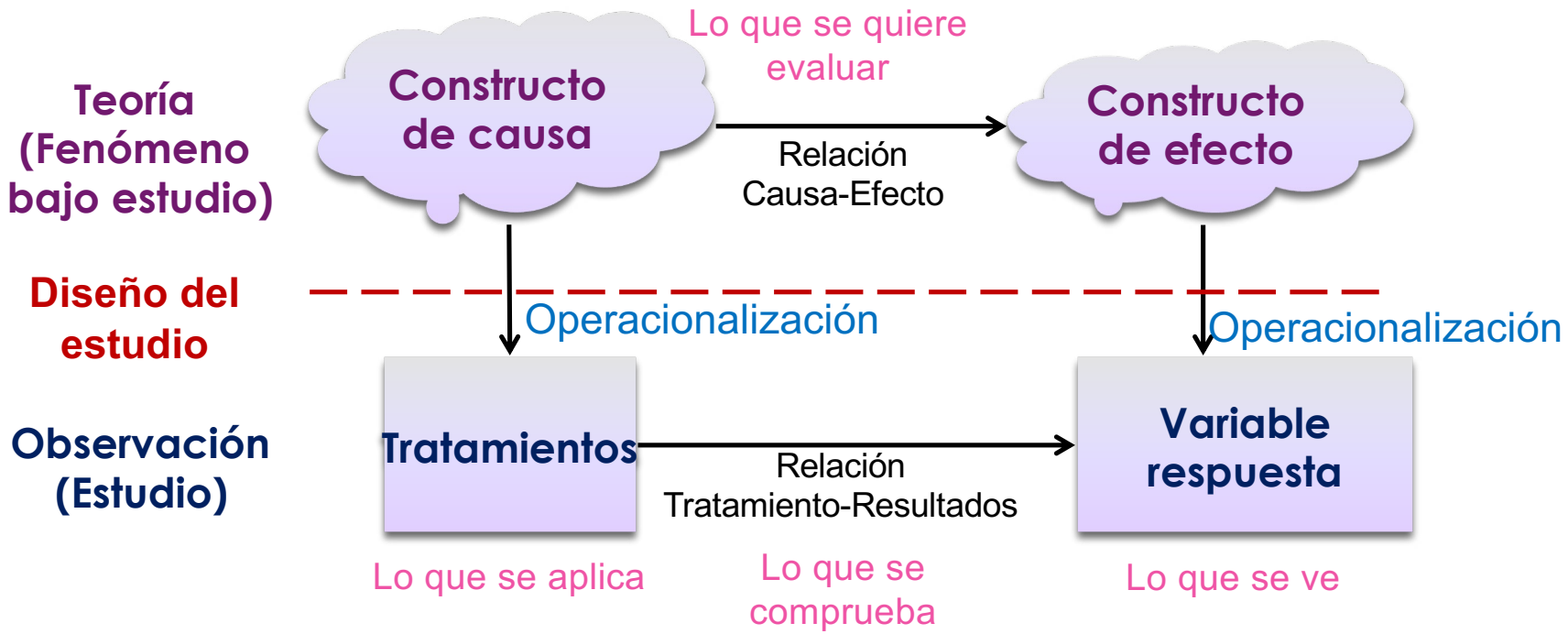
CARACTERIZACIÓN DE LA VALIDEZ



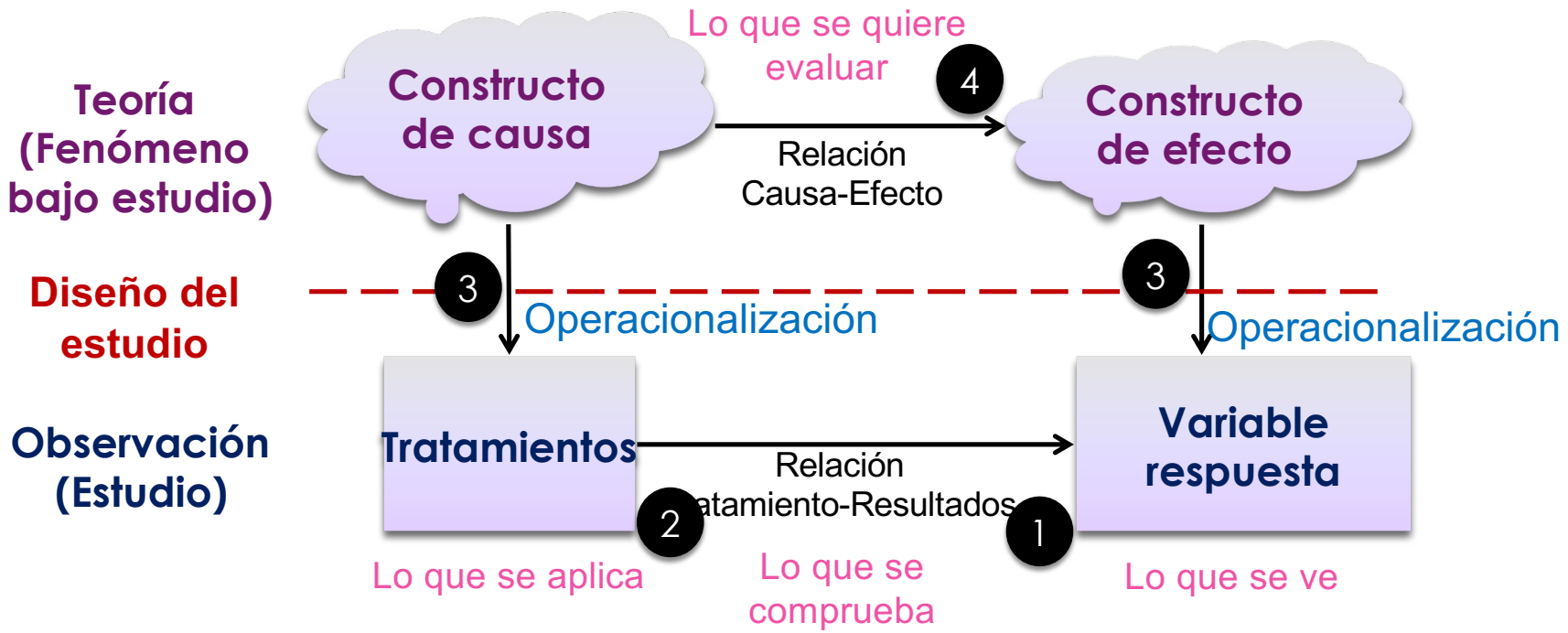
CARACTERIZACIÓN DE LA VALIDEZ



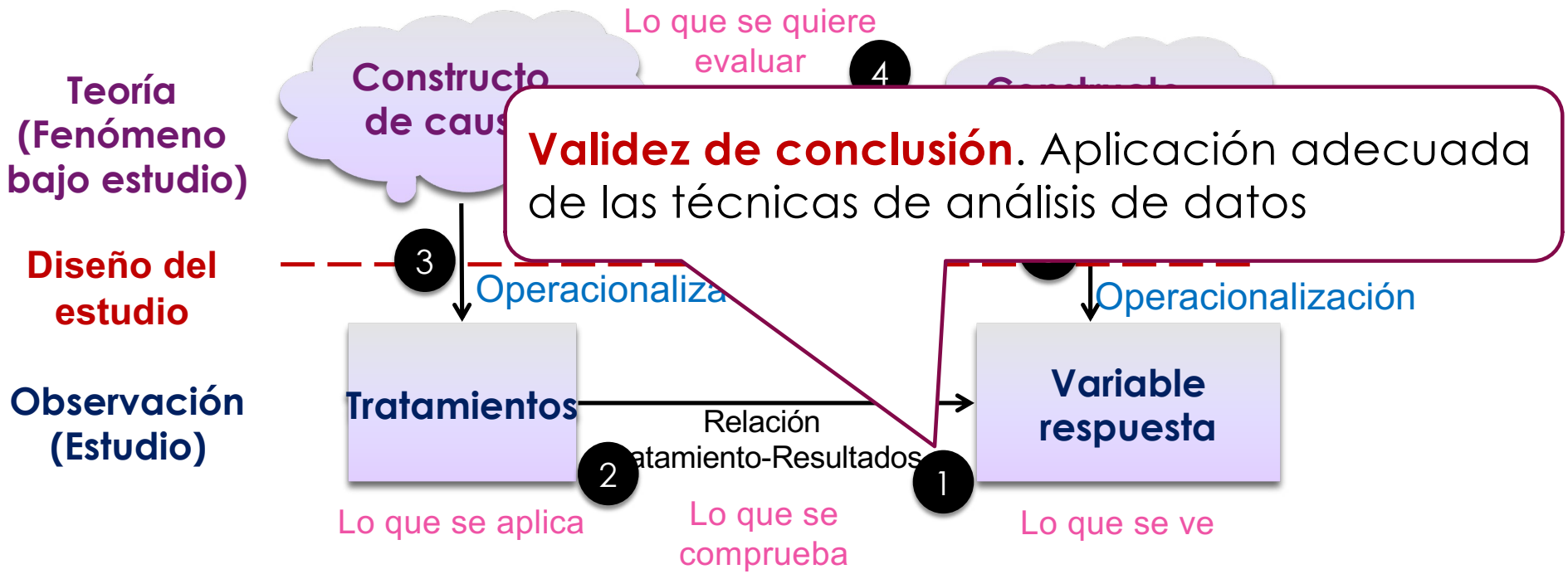
CARACTERIZACIÓN DE LA VALIDEZ



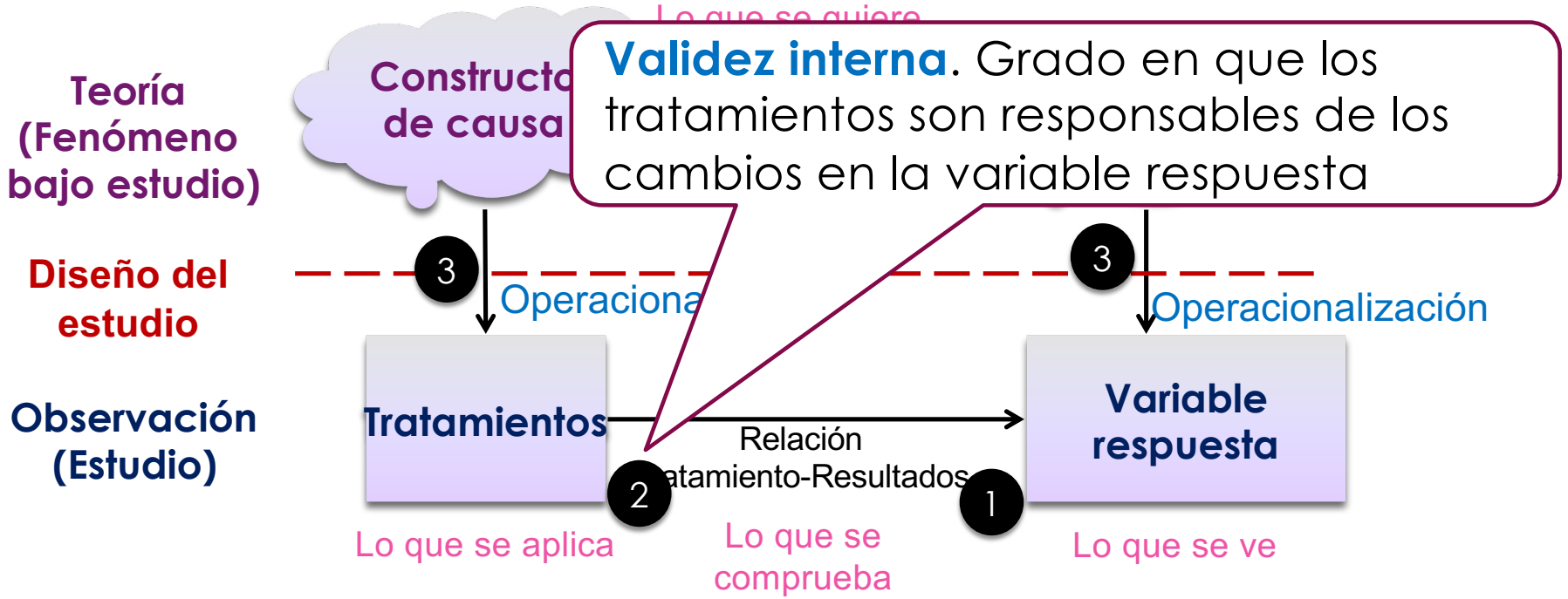
CARACTERIZACIÓN DE LA VALIDEZ



CARACTERIZACIÓN DE LA VALIDEZ

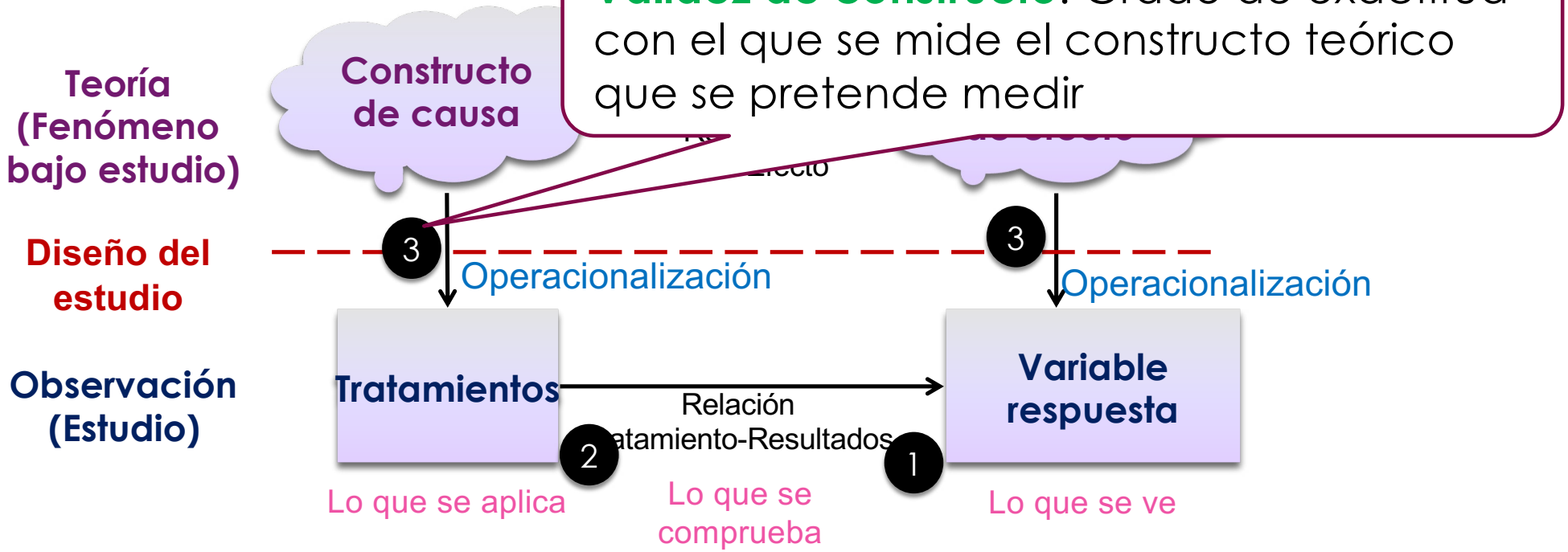


CARACTERIZACIÓN DE LA VALIDEZ

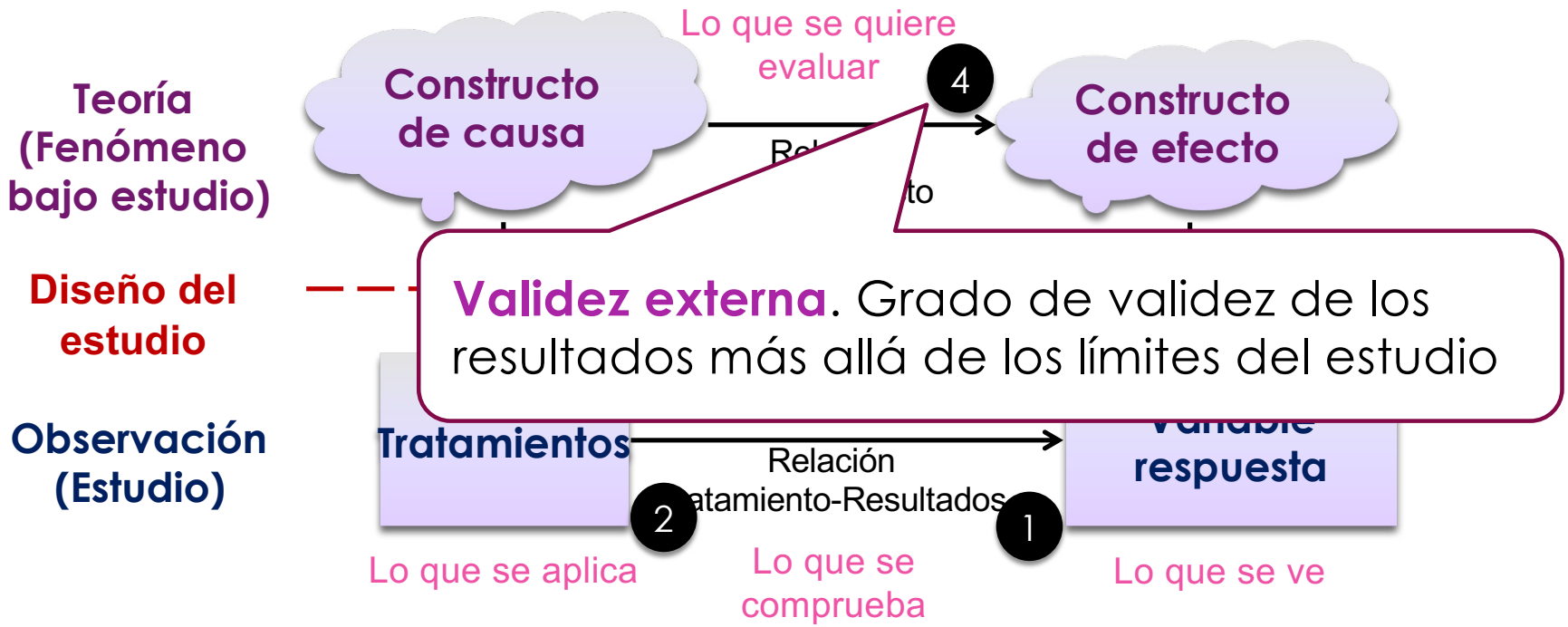


CARACTERIZACIÓN DE LA VALIDEZ

Validez de constructo. Grado de exactitud con el que se mide el constructo teórico que se pretende medir



CARACTERIZACIÓN DE LA VALIDEZ



CARACTERIZACIÓN DE LA VALIDEZ

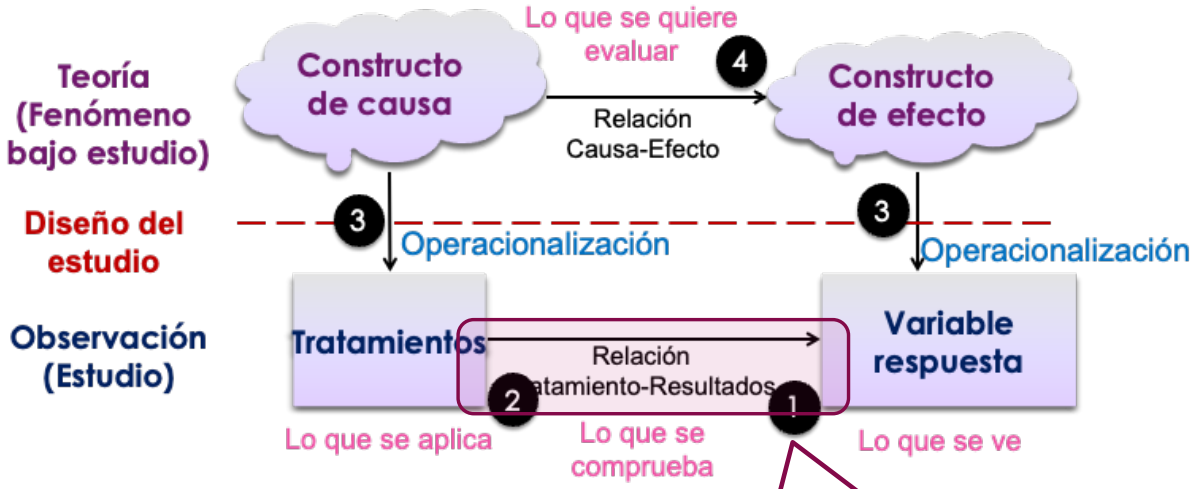
ESTADO ACTUAL DE LA METODOLOGÍA EN LA INVESTIGACIÓN EN IS EMPÍRICA

- Ejemplos:
  - **Validez de conclusión (estadística)**: Análisis de experimentos crossover
  - **Validez interna**: Estudios de Minería de Repositorios Software
  - **Validez de constructo**: Experimentos con sistemas basados en aprendizaje profundo
  - **Validez externa**: Estudiantes vs. profesionales

EJEMPLO: ANÁLISIS DE EXPERIMENTOS CROSSOVER

S. Vegas, C. Apa, N. Juristo, Crossover Designs in Software Engineering Experiments: Benefits and Perils, *IEEE Transactions on Software Engineering*, 42(2):120-135, 2016

VALIDEZ DE CONCLUSIÓN (ESTADÍSTICA)



Se debe aplicar un **análisis estadístico adecuado** a las observaciones (datos brutos) recopilados en el experimento para poder obtener conclusiones correctas sobre las relaciones entre los tratamientos y la variable respuesta

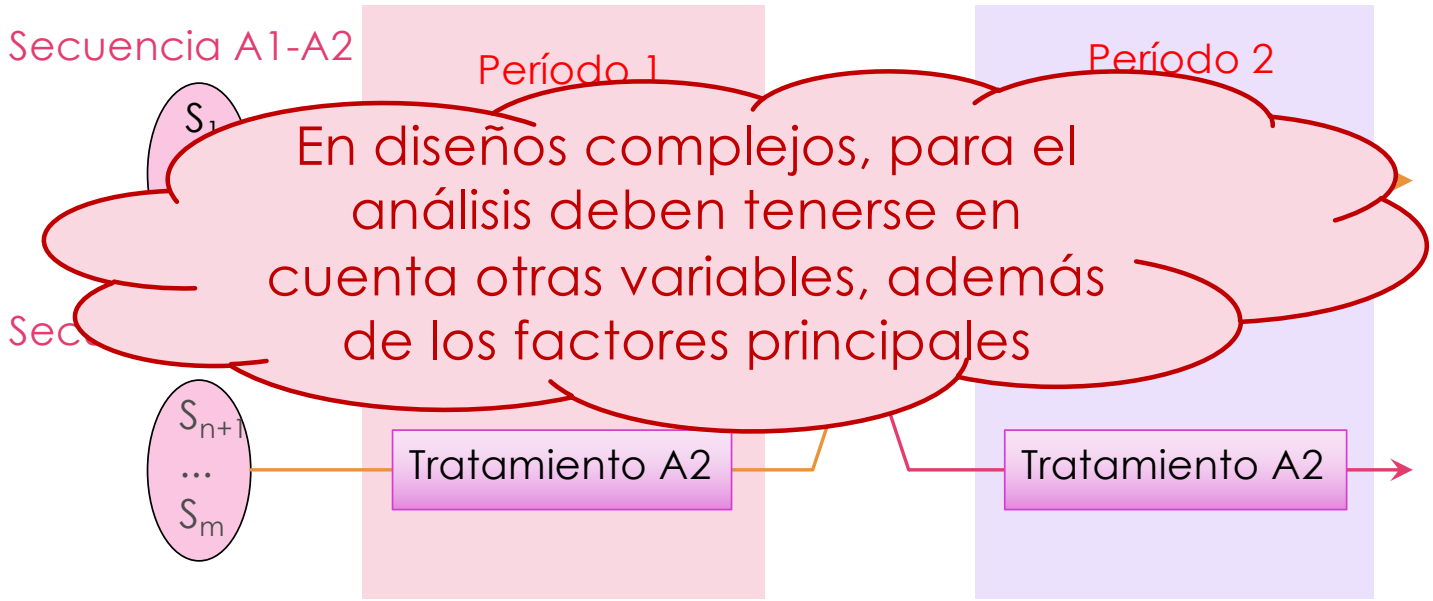
ELECCIÓN DEL TEST ESTADÍSTICO ADECUADO

N. Factores /tratamientos	Sujetos	Escala intervalo o Ratio	Al menos Ordinal
Un factor: 2 tratamientos	Independientes	Prueba t independiente	Prueba Mann-Whitney
	Relacionados	Prueba t pareada	Prueba Wilcoxon
Un factor: > 2 tratamientos	Independientes	ANOVA 1 factor	Prueba Kruskal-Wallis
	Relacionados	ANOVA medidas repetidas	Prueba Friedman
(Fraccional) Factorial	Independientes	ANOVA factorial	-
	Relacionados	ANOVA medidas repetidas	-

Pruebas paramétricas

Pruebas no paramétricas

DISEÑO CROSSOVER (CRUZADO)



DISEÑOS CROSSOVER

veces se realiza la tarea experimental

Orden de los tratamientos

	Período 1	Período 2	Período 3
Secuencia 1	Tratamiento A →	Tratamiento B →	Tratamiento C
Secuencia 2	Tratamiento A →	Tratamiento C →	Tratamiento B
Secuencia 3	Tratamiento B →	Tratamiento A →	Tratamiento C
Secuencia 4	Tratamiento B →	Tratamiento C →	Tratamiento A
Secuencia 5	Tratamiento C →	Tratamiento A →	Tratamiento B
Secuencia 6	Tratamiento C →	Tratamiento B →	Tratamiento A

- Caso particular de tratamiento*período
- Efecto físico/psicológico

UN EXPERIMENTO CROSSOVER REAL

Prueba t pareada

Source	Asymptotic Sig. (2-sided)
Técnica	0.133

Modelo lineal mixto

Source	Sig.
Secuencia	0.774
Período	0.080
Técnica	0.041

El nivel de significación se ha reducido

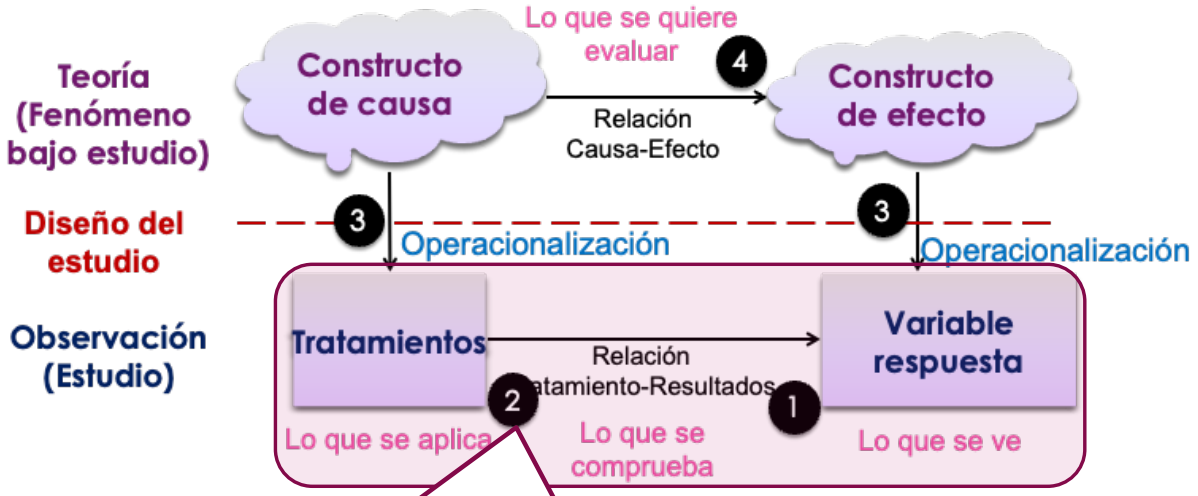
REVISIÓN DE LA LITERATURA

- Publicaciones relevantes (2012-2014):
  - Revistas: TSE, EMSE, TOSEM
  - Congresos: ICSE, ESEC/FSE, ESEM
- 40.2% de las publicaciones (39 de 82) realizan experimentos crossover
- 54.8% experimentos (68 de 124) realizan experimentos crossover
- 0% analizan sus experimentos crossover apropiadamente

EJEMPLO: ESTUDIOS DE MINERÍA DE REPOSITORIOS SOFTWARE

N. Saarimäki, V. Lenarduzzi, S. Vegas, N. Juristo, D. Taibi. Cohort Studies in Software Engineering : A Vision of the Future. In Proceedings of the International Symposium on Empirical Software Engineering and Measurement, 2020.

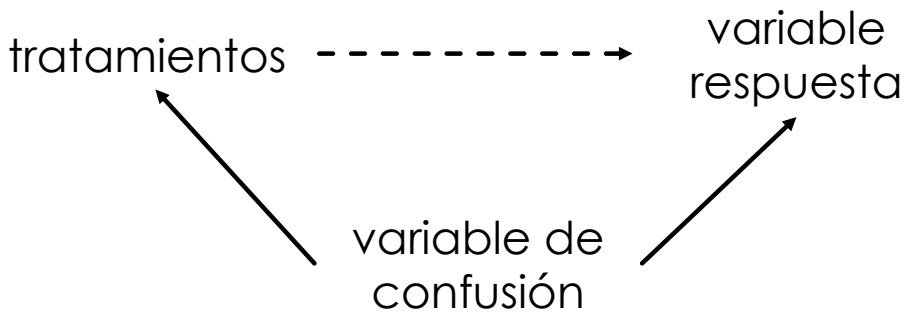
VALIDEZ INTERNA



El diseño del estudio asegura que las variaciones observadas en la variable respuesta se deben a los tratamientos y no son causadas por otras variables. Así, la relación detectada entre tratamientos y variable respuesta es una **verdadera relación causa-efecto**

CONTROL DE TERCERAS VARIABLES: PRINCIPIO DE CONFUSIÓN

- Distorsión que modifica la asociación entre los tratamientos y la variable respuesta porque existe una tercera variable (de confusión)



CONTROL DE TERCERAS VARIABLES: PRINCIPIO DE CONFUSIÓN

- No controlar las variables de confusión puede:
  - Hacer que parezca que existe una asociación entre la exposición y el resultado cuando no la hay
  - Enmascarar una verdadera asociación

PRINCIPIO DE CONFUSIÓN: UN EJEMPLO

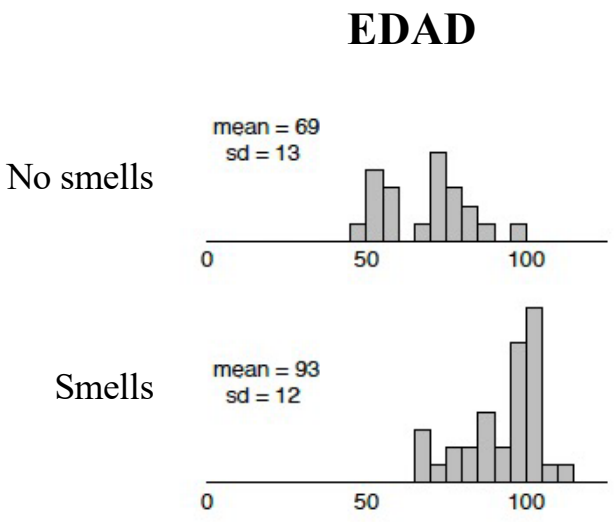
- Afecta el código "smell" a la complejidad del código Java?
- Clases elegidas :
  - Un Proyecto de Apache Software Foundation
  - Versiones 1.0.0 (agosto 2015) a 1.2.0 (Mar 2016)

Source	Sig.
Código smell	0.036

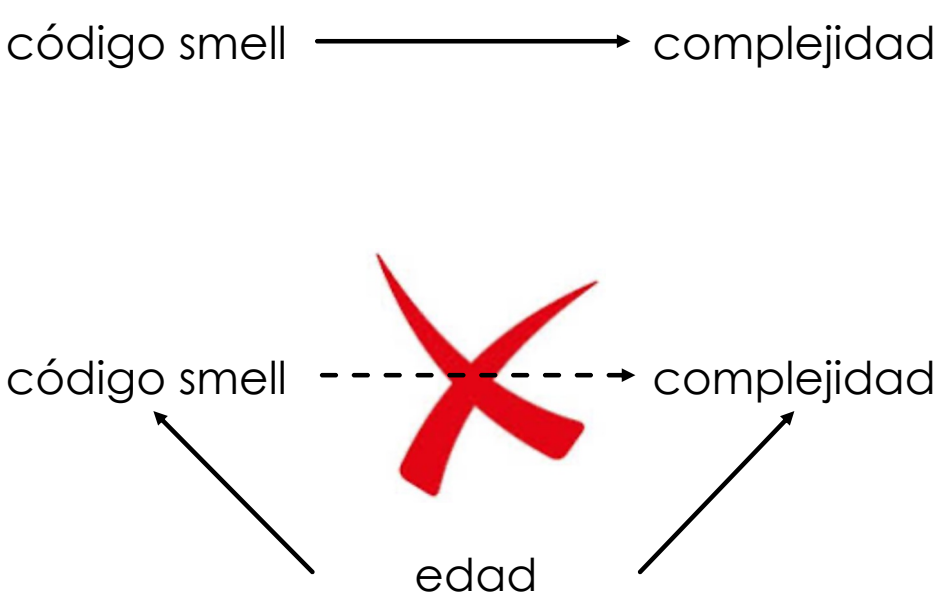
PRINCIPIO DE CONFUSIÓN: UN EJEMPLO

- Variable de confusión: Edad
- Las clases más viejas:
  - Tienen código "smell"
  - Son más complejas

Source	Sig.
Código smell	0.47
Edad	0.042



PRINCIPIO DE CONFUSIÓN: UN EJEMPLO



REVISIÓN DE LA LITERATURA

- Artículos publicados en el technical track de la conferencia Mining Software Repositories (MSR) 2020-2022

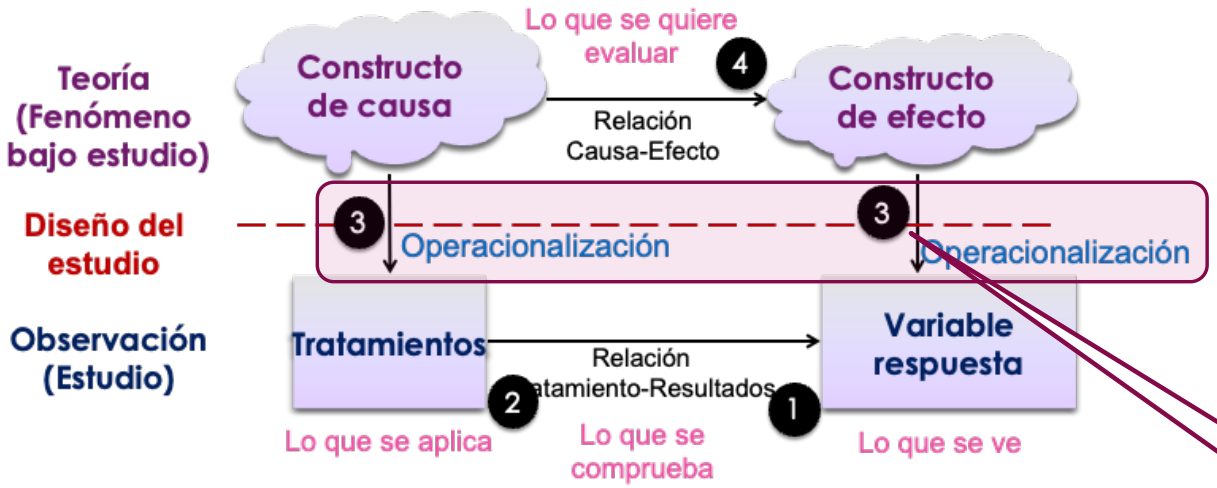
	Papers			RQs		
	Si	**No**	**Parcial**	**Sí**	**No**	**Parcial**
Control de terceras variables	23%	77%	-	25%	75%	-

EJEMPLO: EXPERIMENTOS CON SISTEMAS BASADOS EN APRENDIZAJE PROFUNDO

S. Vegas, S. Elbaum. Pitfalls in Experiments with DNN4SE: An Analysis of the State of the Practice. ESEC/FSE 2023.

https://arxiv.org/abs/2305.11556

VALIDEZ DE CONSTRUCTO



Los tratamientos y medición de las variables respuesta son instancias válidas de los constructos teóricos de la causa y el efecto respectivamente.

TIPOS DE CONSTRUCTOS

- **Causa**: Se refieren a los tratamientos
- **Efecto**: Se refieren a las variables respuesta

CONSTRUCTOS vs VARIABLES

- Constructos:
  - Son conceptos o temas amplios para un estudio
  - Se conceptualizan en el plano teórico (abstracto)
  - Pueden ser abstractos
  - No necesariamente tienen que ser directamente observable

Desarrollo dirigido por modelos

Experiencia en programación

Calidad del código

CONSTRUCTOS vs VARIABLES

- Variables:
  - Se crean desarrollando el constructo en una forma mensurable
  - Se operacionalizan y miden en el plano empírico (observacional)

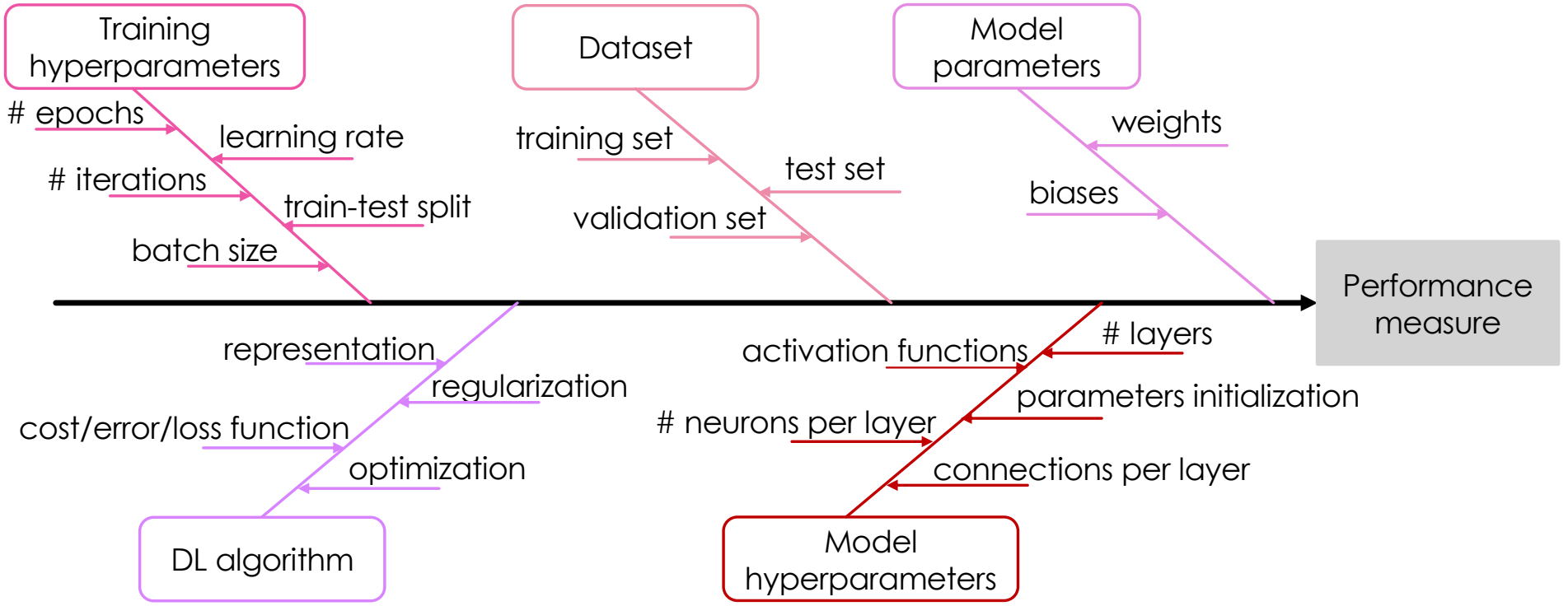
REDES NEURONALES PROFUNDAS EN IS

- El desarrollo de software de calidad a tiempo requiere la mayor automatización posible de las tareas de IS
- Los investigadores en IS están explorando el potencial de los enfoques de aprendizaje automático
- Las redes neuronales (NN) se pueden utilizar para crear herramientas para automatizar tareas de desarrollo y mantenimiento de software

USO DE REDES NEURONALES PROFUNDAS EN IS

- Requisitos:
  - Extracción de requisitos a partir de texto en lenguaje natural
- Diseño:
  - Identificación de patrones de diseño
- Codificación:
  - Reparación de código
  - Sugerencias de código
- Pruebas:
  - Predicción de defectos
  - Estimación del esfuerzo de las pruebas
- Mantenimiento
  - Detección de malware
  - Localización de errores
  - Detección de clones
- Gestión
  - Estimación del esfuerzo
  - Coste de desarrollo
  - Clasificación del código fuente

OPERACIONALIZACIÓN DE UNA RED NEURONAL



REVISIÓN DE LA LITERATURA

- Publicaciones relevantes (2018-2021):
  - Revistas: TSE
  - Congresos: ICSE, ESEC/FSE
- 76 artículos proponen técnicas usando DNNs para automatizar tareas del desarrollo de software
- Realizan la operacionalización:

	Completa	Parcial	Ausente
Factores y tratamientos	14%	**82%**	4%
Variables respuesta	**76%**	18%	6%

REVISIÓN DE LA LITERATURA

- La operacionalización del constructo de causa es problemática
- En concreto:

	Completa	Parcial	Ausente
Model hyperparameters	7%	**85%**	8%
Model parameters	2%	0%	**98%**
DL algorithm	26%	**72%**	2%
Training hyperparameters	19%	**73%**	8%
Training data	**69%**	27%	4%

REVISIÓN DE LA LITERATURA

- El 82% de los experimentos definen los factores adecuadamente, pero las definiciones de los tratamientos son incompletas
  - La lista de hiperparámetros explorados es completa, pero no se especifica el rango de valores
  - Los tratamientos se definen a nivel arquitectónico (LSTM, CNN, transformer, etc.). Sin embargo, se comparan implementaciones específicas de estas arquitecturas. Otras variables no especificadas (hiperparámetros, el algoritmo de DL o la representación de los datos podría ser las causas subyacentes de la diferencia, y no la arquitectura

REVISIÓN DE LA LITERATURA

- Sólo el 4% de los experimentos tienen una definición de factores incompleta:
  - La lista de hiperparámetros explorados no es completa, por lo que no se sabe qué factores se exploraron

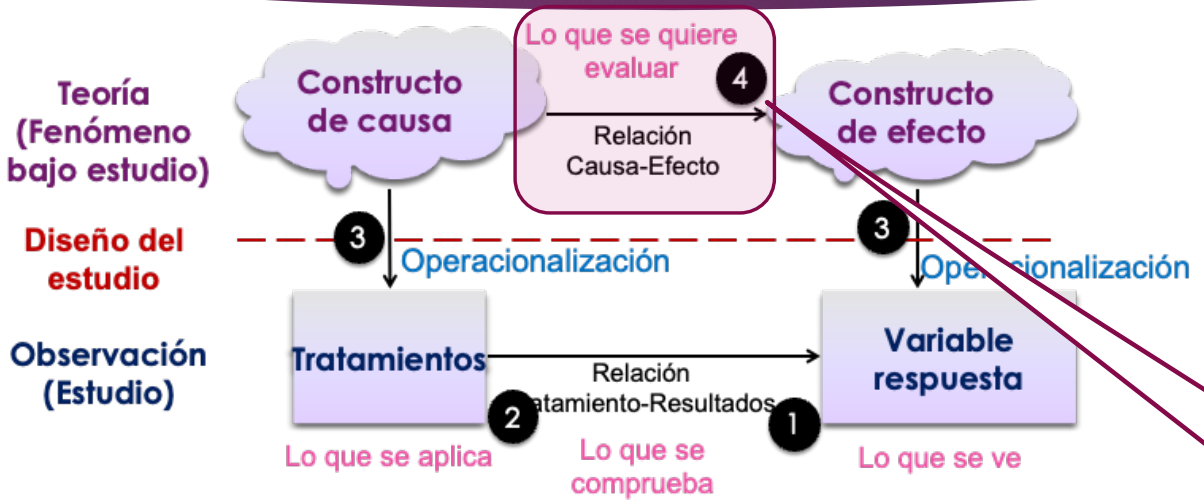
EFECTO DE PROPAGACIÓN: UN EJEMPLO

- Se compara la solución propuesta frente a otros tres enfoques de última generación
- Ninguno de ellos está realmente disponible:
  - En un caso, existe una versión estable en Github, pero puede ser diferente de la que está en el artículo (según lo mencionado anteriormente).
  - En otro caso, los autores tuvieron que re-implementar el enfoque siguiendo el artículo original donde se propone, que podría implementar la solución de otra forma
  - En otro caso, se utilizan los resultados reportados en el artículo en el que se propone la solución

EJEMPLO: ESTUDIANTES VS PROFESIONALES

P. Riofrío, S. Vegas, N. Juristo. Does Subject Type Influence Software Engineering Experiment Results? International Conference on Software Engineering- Poster track, 2017

VALIDEZ EXTERNA



Grado de generalización que se puede hacer de la relación causal detectada más allá del experimento

GENERALIZANDO LOS RESULTADOS

- La validez externa se refiere a variaciones en:
  - Participantes
  - Entorno
  - Tratamientos
  - Variables respuesta
- Principalmente en **aquellos que no estaban en el estudio**

EJEMPLO REAL

- Una característica típica de los experimentos con humanos en IS es el uso de estudiantes en lugar de profesionales.
- ¿Se pueden generalizar a los profesionales los resultados de los experimentos con estudiantes?
- Exploramos esto en el contexto de un experimento que examina la calidad del código generado (% caso de prueba que se pasan) utilizando TDD frente a ITL

EJEMPLO REAL

Tratamiento	Grupo	Media estimada	95% CI
ITL	Estudiantes	40,47	(29,19-51,75)
	Profesionales	49,53	(31,66-67,41)
TDD	Estudiantes	35,61	(30,37-40,85)
	Profesionales	38,26	(28,33-48,18)

Los profesionales logran una calidad ligeramente superior a la de los estudiantes tanto con TDD como con ITL

La caída del rendimiento de los profesionales que aplican TDD respecto a ITL casi duplica al de los estudiantes

EJEMPLO REAL

Profesionales	• Más experiencia beneficia a ITL
Estudiantes	• Más experiencia en programación ***tiende*** a beneficiar a TDD • Más experiencia en testing *tiende* a beneficiar a ITL

REVISIÓN DE LA LITERATURA

- Publicaciones relevantes (2014-2016):
  - Revistas: TSE, EMSE, TOSEM
  - Conferencias: ICSE, ESEC/FSE, ESEM
- 93 artículos reportan experimentos con humanos
  - En 11 de ellos se puede estudiar el impacto del tipo de sujeto
- Resultados no concluyentes:

No influye	38%
Influye	54%
Contradictorios	8%

CONCLUSIONES

- Se cumplen 42 años desde que se publicó el primer experimento en IS
- La IS finalmente ha abrazado el empirismo
- Pero corremos el riesgo de tener una crisis similar a la experimentada en psicología (Pashler y Wagenmakers, 2012), refiriéndose al "nivel de duda sin precedentes entre los profesionales sobre la confiabilidad de los hallazgos de la investigación en el campo"
- Es cierto que el estudio perfecto no existe
- Pero **la metodología importa** y deberíamos aspirar a mejorar nuestros estudios empíricos

¡GRACIAS!

Proyecto PID2022-137846NB-I00 financiado por:

MINISTERIO DE CIENCIA E INNOVACIÓN

UNIÓN EUROPEA

FONDO EUROPEO DE DESARROLLO REGIONAL

"Una manera de hacer Europa"

AGENCIA ESTATAL DE INVESTIGACIÓN