CET
Centro de Estudios Tributarios
Universidad de Chile
www.cetuchile.cl

LA EVALUACIÓN DE DISEÑOS EXPERIMENTALES EN EL ÁMBITO TRIBUTARIO.

Gustavo González

Licenciado en Economía,
Universidad de la República Uruguay.

RESUMEN:

Este artículo presenta una propuesta para integrar la práctica de evaluación de diseños experimentales a la planificación regular de acciones de tratamiento de las administraciones tributarias. La propuesta se justifica en que es la metodología más robusta para evaluar la eficacia de los tratamientos y la más informativa para contribuir a mejorar el diseño de éstos, así como su focalización.

El artículo está contextualizado para administraciones tributarias que desarrollan la gestión del cumplimiento tributario apoyadas en la metodología de la gestión de riesgos, para la cual la etapa de la evaluación es clave en el funcionamiento de un sistema de aprendizaje, que es imprescindible para gestionar la incertidumbre.

En este contexto, el objetivo de las evaluaciones no es solamente hacer un uso más eficaz y eficiente de los recursos, sino también asegurar un trato adecuado a los contribuyentes, con tratamientos que sean proporcionales al riesgo que presentan y a su disposición al cumplimiento.

La evaluación de la eficacia de los tratamientos es costosa bajo cualquier metodología, por eso es importante que los esfuerzos destinados a ésta den resultados certeros y aseguren un uso eficiente de los recursos disponibles. Sobre estas premisas, en este artículo se presenta una propuesta con consideraciones sobre cuándo y cómo evaluar, cómo integrar estas evaluaciones a la planificación, y qué requerimientos deben atenderse para esto. El artículo también presenta un ejemplo numérico que ayuda a ilustrar algunas de las características de esta propuesta.

1. INTRODUCCIÓN

La evaluación de diseños experimentales es una herramienta útil para generar conocimiento sobre los factores que influyen en el comportamiento de los contribuyentes y las posibles medidas para mejorar el cumplimiento tributario, así como para contribuir a un trato adecuado a los contribuyentes, en el sentido de que éste sea proporcional al nivel de riesgo que presentan y a su disposición a cumplir. Sin embargo, esta herramienta no ha sido aprovechada de forma sistemática por las administraciones tributarias de la región que, en general, se han limitado a realizar algunos experimentos aislados. En este ensayo, se analizan ventajas y desafíos de aplicar la evaluación de diseños experimentales en el ámbito tributario y se proponen algunas recomendaciones para incorporar esta práctica a la planificación de actividades de la administración tributaria.

La evaluación de diseños experimentales consiste en asignar aleatoriamente a los sujetos de estudio a diferentes grupos que reciben distintos tratamientos o intervenciones y comparar los resultados obtenidos entre los grupos. De esta forma, se puede aislar el efecto causal de los tratamientos o intervenciones sobre las variables de interés, controlando por otros factores que puedan influir. Esta metodología permite obtener evidencia empírica robusta y fiable sobre la efectividad de las políticas públicas y las acciones de gestión.

En el ámbito tributario, la evaluación de diseños experimentales puede servir para conocer mejor el comportamiento de los contribuyentes y los factores que lo determinan, tales como las normas sociales, las creencias, las actitudes, las preferencias, las motivaciones, los incentivos, las sanciones, la información, la educación, la asistencia, la confianza, la percepción de riesgo, la equidad, la reciprocidad, etc. Asimismo, puede servir para evaluar el impacto de diferentes medidas o estrategias para mejorar el cumplimiento tributario, tales como las campañas de comunicación, las cartas personalizadas, los mensajes persuasivos, los recordatorios, los "nudges", los premios, los sorteos, las auditorías, las multas, etc.

La literatura sobre la evaluación de diseños experimentales en el ámbito tributario es abundante y ha mostrado resultados interesantes y relevantes para la gestión tributaria. Por ejemplo, se ha encontrado que las cartas personalizadas que incluyen información sobre el uso de los impuestos, las normas sociales, las consecuencias del incumplimiento o la probabilidad de ser auditado pueden aumentar significativamente el pago voluntario de los impuestos[1]. También se ha encontrado que los premios o sorteos pueden incentivar la solicitud de facturas

1 Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. Journal of Public Economics, 148, 14-31

y el uso de medios electrónicos de pago[2]. Asimismo, se ha encontrado que las auditorías pueden tener un efecto disuasorio sobre el incumplimiento, pero también un efecto negativo sobre la confianza y la satisfacción de los contribuyentes[3].

La evaluación de diseños experimentales es una herramienta potente para ampliar el conocimiento que es relevante para las administraciones tributarias en su gestión del cumplimiento de las obligaciones tributarias. Este artículo propone algunas ideas y justificaciones para incorporar de forma sistemática esta práctica, de tal forma de aprovechar la oportunidad casi continua que tienen las administraciones tributarias de aprender a partir de la conducta de los contribuyentes, dado que permanentemente están realizando intervenciones sobre éstos.

En la sección 2 del artículo se presenta una reseña de la metodología de evaluación de diseños experimentales. La sección 3 es la parte central del artículo, en la que se hacen consideraciones sobre cómo integrar esta metodología a la práctica de planificación de acciones de tratamiento en las administraciones tributarias. Se discute sobre el objetivo y la oportunidad de la evaluación, la aplicación de los distintos pasos de la metodología en el contexto de la planificación, así como requerimientos específicos que pueden suponer desafíos de implementación para la administración tributaria. En la sección 4 se presenta un ejemplo numérico con datos ficticios, pero basado en estructuras de datos reales de las administraciones tributarias, que sirve para ilustrar lo descrito en la sección 3. Por último, en la sección 5 se presentan conclusiones.

2. RESEÑA METODOLÓGICA DE LA EVALUACIÓN DE DISEÑOS EXPERIMENTALES

Por evaluación de diseño experimental se hace referencia a la asignación aleatoria del tratamiento o programa que es objeto de la evaluación. También se lo suele denominar ensayo aleatorio controlado, evaluaciones aleatorias, evaluaciones experimentales, entre otras. En rigor, un experimento no tiene que identificar impactos mediante asignaciones aleatorias, pero en el ámbito de la evaluación usualmente se reserva el término "experimento" para referirse a evaluaciones querecurren a la asignación aleatoria.

Cuando se asigna de forma aleatoria a quienes serán sujetos de un tratamiento, es decir, mediante sorteo entre una población elegible numerosa, se puede generar una estimación robusta del contrafactual. Por contrafactual se hace referencia al

2 Mascagni, G., Nell, C., & Monkam, N. F. (2017). One size does not fit all: a field experiment on the drivers of tax compliance and delivery methods in Rwanda. ICTD Working Paper, 58

3 Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. American Economic Review, 105(8), 2539-69.

valor que habría tenido el resultado ("Y", la variable a ser medida) para quienes reciben el tratamiento si no lo hubieran recibido. Por definición, el contrafactual no es observable. Por lo tanto, debe estimarse utilizando un grupo de comparación, o también denominado grupo de control.

En la asignación aleatoria todas las unidades elegibles tienen la misma probabilidad de ser seleccionadas para recibir el tratamiento. Cuando se asignan unidades de manera aleatoria a los grupos de tratamiento y de control, ese proceso producirá dos grupos que tienen una alta probabilidad de ser estadísticamente idénticos, siempre que el número de unidades potenciales a las que se aplica el proceso sea suficientemente grande. En concreto, con un gran número de unidades el proceso de asignación aleatoria producirá grupos que tienen promedios estadísticamente equivalentes, así como distribuciones de valores muy semejantes, en todas sus características.

Con los datos de línea de base de la muestra de evaluación con la que se cuente se podrá comprobar empíricamente que los dos grupos (de tratamiento y de control) son estadísticamente equivalentes y, por lo tanto, no presentarán diferencias sistemáticas en ninguna de las características observables antes de la ejecución del tratamiento. Luego, si después de ejecutar el tratamiento se observan diferencias en los resultados entre los dos grupos, estas diferencias podrán atribuirse a la incidencia del tratamiento, dado que ambos grupos eran idénticos en la línea de base, antes de la ejecución, y que están expuestos a los mismos factores externos a lo largo del tiempo.

Esto último significa que si otros factores -distintos del tratamiento- pueden afectar al resultado de la variable de interés (la que es objeto de la evaluación, por ejemplo, el volumen de pagos de impuestos), ambos grupos (de tratamiento y de control) estarán expuestos de igual manera a estos factores. En otras palabras, el grupo de control funciona como contrafactual: el resultado promedio en la variable de interés observado en este grupo es informativo sobre lo que hubiera ocurrido con el grupo tratado si no hubiera recibido el tratamiento.

En este contexto de asignación aleatoria del tratamiento, el impacto de éste se mide como la diferencia entre el resultado observado bajo tratamiento (el resultado promedio del grupo tratado) y el resultado observado en el contrafactual (el resultado promedio del grupo de control). Este resultado se considera una estimación robusta del impacto del tratamiento, ya que, dadas las características del diseño, todos los demás factores (observados y no observados) que pueden afectar al resultado están controlados y afectan por igual a ambos grupos, por lo tanto, no inciden en la explicación de la diferencia en los resultados.

En cuanto a la validez de los resultados, debe distinguirse entre la validez interna y la validez externa. La validez interna significa que el impacto estimado del tratamiento está libre de todos los demás factores de confusión potenciales o, lo que es lo mismo, que el grupo de control represente una estimación precisa

del contrafactual. Como ya fue señalado, esto último está garantizado por la asignación aleatoria para integrar los grupos de tratamiento y control.

Por su parte, la validez externa quiere decir que la muestra de la evaluación representa con precisión a la población de unidades elegibles. Por consiguiente, los resultados de la evaluación se pueden extrapolar a la población de unidades elegibles. Esta validez se obtiene al utilizar muestreo aleatorio para extraer la muestra de la evaluación.

En definitiva, el proceso de evaluación de un diseño experimental persigue dos objetivos a través de selección aleatoria: seleccionar aleatoriamente una muestra para asegurar la validez externa, y asignar aleatoriamente el tratamiento como método de evaluación del impacto, para asegurar la validez interna. Una evaluación de impacto puede producir estimaciones internamente válidas del impacto mediante una asignación aleatoria del tratamiento; sin embargo, si la evaluación se lleva a cabo con una muestra no aleatoria de la población, puede que los impactos estimados no sean generalizables para el conjunto de unidades elegibles.

El diseño y la ejecución de un experimento aleatorio pueden enfrentar circunstancias que afecten la robustez de los resultados, en el sentido de que la diferencia en el valor de la variable objetivo entre el grupo tratado y el grupo de control no sea, como se esperaría, enteramente atribuible al efecto del tratamiento. La mayoría de estos problemas tiene solución a través del uso de metodologías complementarias y/o de ajustes al diseño. Este artículo no tiene como propósito ser una guía metodológica para la aplicación de la evaluación de diseños experimentales[4], pero, a los efectos de cumplir con los objetivos de este artículo, sí importa indicar algunos aspectos que deben ser considerados durante el diseño y la ejecución, particularmente en el marco de la planificación de acciones de tratamiento en una administración tributaria, con el propósito de minimizar la probabilidad de que se produzcan distorsiones que afecten negativamente a la robustez de los resultados.

La siguiente es una lista de verificaciones a considerar para asegurar la validez de los resultados. En la sección 3 se pondrá foco en los aspectos de la planificación de las acciones de tratamiento que contribuyen al cumplimentado de esta lista.

1. Deben compararse las características de la línea de base (antes de la ejecución del tratamiento) del grupo de tratamiento y del grupo de control. Debería encontrarse que las características en la línea de base están

4 La literatura sobre este tema es abundante. A vía de ejemplo, una guía en español de muy amigable lectura sobre metodologías de evaluación de impacto es Gertler, Paul J., Sebastián Martínez, Patrick Premand, Laura B. Rawlings y Christel M. J. Vermeersch. 2017. "La evaluación de impacto en la práctica, Segunda edición". Washington, DC: Banco Interamericano de Desarrollo y Banco Mundial. doi:10.1596/978-1-4648-0888-3.

equilibradas para las variables observables más relevantes, asegurando así la equivalencia estadística entre los dos grupos.

2. Se debe verificar si todas las unidades elegibles han recibido el tratamiento asignado y que no haya unidades no elegibles que hayan recibido tratamiento. En este sentido, en la sección 3 se hará referencia a algunas situaciones que pueden presentarse en la planificación de acciones de tratamiento y posibles soluciones.
3. Revisar la posible existencia de efectos indirectos del tratamiento en el grupo de control. El diseño estándar de un experimento aleatorio supone que las unidades que componen tanto el grupo de tratamiento como el grupo de control son independientes entre sí. Pero éstas pueden afectarse mutuamente a través de diversos canales de interacción. En el ámbito de la administración tributaria, esto podría darse, por ejemplo, entre contribuyentes que comparten a un mismo asesor tributario, o entre contribuyentes que mantienen relaciones comerciales recurrentes. Si se sospecha de la existencia de estos efectos, es conveniente ajustar el diseño para que éstos puedan ser identificados en la evaluación; en la sección 3 se aborda este problema.

3. PLANIFICAR LA EVALUACIÓN DE LAS ACCIONES DE TRATAMIENTO EN EL ÁMBITO DE LA ADMINISTRACIÓN TRIBUTARIA

3.1 Objetivo de la evaluación y alcance de este artículo

El proceso de gestión del cumplimiento tributario basado en la metodología de gestión de riesgos presupone que el conocimiento disponible en la administración tributaria es ampliamente aprovechado, al tiempo que éste crece y se perfecciona a partir del aprendizaje; el sistema de evaluación es clave para que esta función del proceso pueda desarrollarse.

Un sistema de evaluación apoyado en esta metodología debería comprender todas las fases del proceso de gestión de riesgos de cumplimiento tributario (en adelante, GRC), desde la identificación de los riesgos, su análisis y valoración, la priorización de riesgos, la asignación y ejecución de tratamientos, y hasta la propia evaluación, como fase integrante del proceso. No obstante, este artículo se enfoca exclusivamente en la evaluación de la eficacia de los tratamientos.

En este contexto, es esperable que el sistema de evaluación contribuya a:

- *Conocer qué factores determinan que un tratamiento sea más o menos eficaz para mejorar el cumplimiento tributario.* Este conocimiento

debería habilitar el rediseño de los tratamientos para aumentar su eficacia. Además, el tratamiento podría ser más eficaz para determinados perfiles de contribuyentes o en determinadas situaciones. Esta información también es relevante para optimizar su aplicación.

- *Una mayor comprensión de las causas del incumplimiento.* Los tratamientos se deberían diseñar en atención a estas causas, pero, generalmente éstas no son conocidas y se manejan en calidad de hipótesis de trabajo. La ejecución de las acciones de tratamiento articulada con un sistema de evaluación habilita un proceso de ensayo y error del cual se puede obtener información sistematizada. El mayor conocimiento de las causas puede habilitar la concepción de nuevos tratamientos.

- *Conocer fallas en la ejecución de los tratamientos.* Las acciones de tratamiento deben ejecutarse bajo cierto estándar, que garantice que todos los contribuyentes tratados reciban el mismo tratamiento. Deliberadamente se puede pretender que el tratamiento incluya algunas discriminaciones basadas en características del perfil del contribuyente; aún en este caso, el tratamiento sigue un estándar para cada perfil. La evaluación debería comprender verificaciones que certifiquen que este estándar haya sido cumplido; de no ocurrir, parte de los resultados observables podrían atribuirse a los desvíos encontrados. Asimismo, el análisis del diseño del tratamiento puede evidenciar errores a la luz de los resultados observables.

3.2 Oportunidad de la evaluación

Evaluar la eficacia de un tratamiento es una actividad costosa, que consume considerables recursos y ocasiona costos de oportunidad al vedar la intervención sobre ciertos contribuyentes que integran el universo de evaluación como grupo de comparación o de control. Además, no siempre es necesaria, en el sentido de que no siempre arrojará nueva información. Por ambas razones, que en definitiva hacen al uso eficiente de los recursos, es importante clarificar cuándo es oportuno evaluar.

En pocas palabras, podría decirse que es oportuno evaluar solamente si se presume que se pueden encontrar novedades, que pueden deberse a, por ejemplo:

- *No se ha evaluado el tratamiento.* Éste es el motivo más elemental; si el tratamiento no ha sido evaluado no conocemos su nivel de eficacia. Si es un tratamiento profusamente utilizado, su evaluación debería ser prioritaria.

- *El tratamiento es aplicado a un segmento de contribuyentes con un perfil diferente al de aplicaciones anteriores.* El tratamiento pudo haber

sido evaluado en el pasado, pero ahora se pretende aplicar sobre otros perfiles de contribuyentes, de tal manera que los resultados anteriores no necesariamente tienen validez como referencia para esta aplicación.

- *Hubo cambios en el diseño y/o en la ejecución del tratamiento.* El tratamiento ya ha sido evaluado, pero, a raíz de esta o por otros motivos, se dispusieron cambios en el diseño y/o en la ejecución. Es necesario conocer si estos cambios mejoran o empeoran la eficacia del tratamiento.

- *Se observan cambios estructurales en la economía y/o en la sociedad.* El tratamiento ya ha sido evaluado en el pasado, pero esto ocurrió hace cierto tiempo (años) y en el ínterin se han observado algunos cambios que podrían incidir en su eficacia. Por ejemplo, el crecimiento del comercio digital, o la relación de las personas con las tecnologías de la información y la comunicación, son ejemplos de cambios que podrían afectar, negativa o positivamente, la eficacia de algunos tratamientos.

3.3 Importancia de la evaluación para el diseño de tratamientos basados en la autorregulación

Un foco de interés específico refiere a los tratamientos basados en la autorregulación del contribuyente, es decir, cuando éste autorregula su comportamiento inducido por alguna acción de la administración tributaria. Los tratamientos que suponen autocorrección, por ejemplo, cuando se le indica al contribuyente que debe corregir una declaración de impuestos y se le dan instrucciones para completar esta corrección, constituyen una de las formas más conocidas de autorregulación, pero no es la única.

En términos generales, cualquier tratamiento que comprenda "nudges" (empujones) y/o un esquema de incentivos (pecuniarios y/o morales de premios y/o castigos) suele estar apoyado en la autorregulación del contribuyente. Es decir, tratamientos en los que se espera que el contribuyente adopte un cambio de comportamiento a partir del estímulo recibido. Esto supone, además, que el tratamiento incluye un sistema de seguimiento para verificar si el contribuyente cumple o no con la expectativa de cambio de comportamiento. Este cambio de comportamiento esperado puede ser observable a través de distintas variables asociadas al cumplimiento, lo cual dependerá de las características del tratamiento, por ejemplo: el valor del impuesto declarado, la frecuencia de presentación de declaraciones, el monto de pagos, el monto de créditos solicitados.

Las administraciones tributarias han diversificado su gama de tratamientos de esta clase en los últimos años, aprovechando el uso más intensivo de las tecnologías de la información y la comunicación. Una de sus ventajas por sobre los esquemas de control más tradicionales en las administraciones tributarias es la

capacidad de ampliar la cobertura de contribuyentes involucrados, ya que suelen presentar considerables economías de escala: la mayor parte de los costos de diseño, planificación, ejecución y seguimiento son fijos, es decir, no dependen de la cantidad de contribuyentes comprendidos en el plan de acción; por lo tanto, cuanto mayor es la cantidad de contribuyentes comprendida, mayor es el resultado esperado para la administración tributaria, a igual costo.

No obstante, el éxito de este tipo de tratamientos se juega, en buena medida, en la credibilidad de los mensajes que contiene. Usualmente, la inducción a un cambio de comportamiento se instrumenta a través de un mensaje que es enviado por diferentes vías y con muy variados contenidos. Sin perjuicio de esta heterogeneidad, el mensaje suele indicar: a) una expectativa de cumplimiento (qué se espera del contribuyente); b) lo que sucederá si el contribuyente no cumple ("amenaza").

Esta "amenaza" deber ser suficientemente conminativa para cambiar la conducta del contribuyente, pero, además, debe ser creíble. Concebir aquí el concepto de "amenaza" en un sentido amplio, como toda reacción de la administración tributaria en caso de que el contribuyente no cumpla[5]. En algunos de estos casos, cumplir con esta amenaza implica para la administración tributaria incurrir en una mayor utilización de recursos, por ejemplo, si aquella contiene la posibilidad de alguna acción de control de carácter presencial. En estas situaciones, entonces, la credibilidad descansa en la capacidad y determinación de la administración tributaria para comprometer esos recursos potencialmente requeridos.

En otras palabras, algunos tratamientos basados en la autorregulación pueden requerir un diseño de planificación más desafiante, ya que la cantidad de recursos que demandan dependerá de la respuesta de los contribuyentes frente a la acción inicial. En este contexto, poder determinar qué perfiles o características de los contribuyentes están asociados a una mayor probabilidad de eficacia del mensaje, o la tasa de respuesta positiva a la acción inicial discriminada por perfiles, son asuntos muy relevantes para una adecuada planificación de las acciones. Además, será relevante reforzar la credibilidad de los mensajes, lo cual se traduce en una mayor eficacia del tratamiento. En este contexto particular, la conveniencia de articular la planificación de las acciones de tratamiento con la evaluación de su eficacia utilizando la metodología de evaluación de diseños experimentales se vuelve más evidente.

5 En el caso de incentivos positivos, por ejemplo, cuando se ofrece un plazo mayor para pagar impuestos, la amenaza está implícita por oposición: no es posible acceder al beneficio si no se cambia el comportamiento.

3.4 Aplicación de la metodología de evaluación de diseños experimentales en el marco de la planificación de las acciones de tratamiento

En esta sección se describen los detalles acerca de cómo se propone aplicar la metodología de evaluación de diseños experimentales en el marco de la planificación de acciones de tratamiento.

3.4.1 Determinación del universo de contribuyentes, selección y asignación aleatoria

El universo de contribuyentes que será objeto de la evaluación debería estar conformado por un conjunto de contribuyentes que cumplan con las reglas de asignación del tratamiento, es decir, aquellos para los cuales se considera, a priori, que el tratamiento es adecuado. En el marco de un modelo de gestión de cumplimiento basado en riesgos, se esperaría que este conjunto esté compuesto por contribuyentes que presentan cierto nivel de severidad en uno o más riesgos, además de otras características, por ejemplo, cierto nivel de predisposición al cumplimiento. Resulta obvio que esta definición conecta directamente la evaluación con la planificación de las acciones.

Siguiendo lo expuesto en la sección 2, el paso siguiente consiste en definir si se pretende que la validez de los resultados de la evaluación comprenda a la totalidad de esos contribuyentes elegibles para el tratamiento o solamente a un subgrupo. Si la pretensión es que los resultados sean válidos para todo el universo de elegibles (validez externa), entonces la selección muestral para integrar los grupos de evaluación (grupo de tratamiento y grupo de comparación o de control) debe ser enteramente aleatoria.

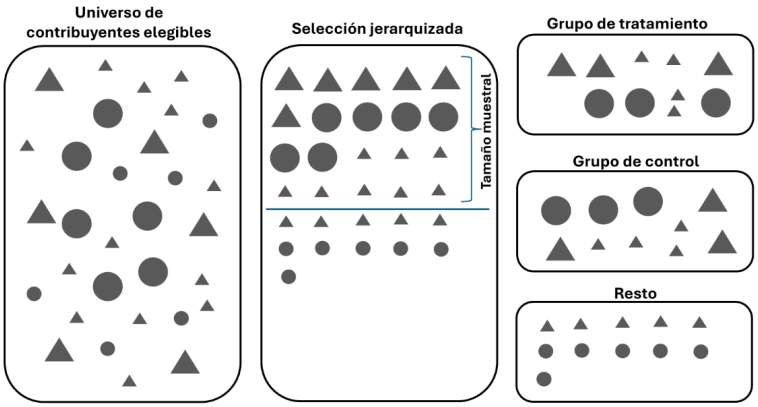
Si, en cambio, se pretende que los resultados sean solamente válidos internamente para el grupo de evaluación (tratamiento y control), entonces la selección muestral puede presentar algún sesgo. Típicamente, en el marco de la GRC y con el objetivo de hacer un uso eficiente de los recursos, este sesgo podría establecerse ordenando a los contribuyentes según el nivel de severidad que presentan en los riesgos valorados.

En cualquiera de las dos alternativas, con posterioridad a la selección de la muestra de evaluación, ya sea aleatoria o aplicando algún criterio jerárquico de selección, la asignación de cada contribuyente a los grupos de tratamiento y de control debe ser enteramente aleatoria para que los resultados tengan validez interna.

Estos procedimientos se ilustran en las siguientes figuras. La Figura 1 ilustra un diseño en el que la selección muestral es jerarquizada y la asignación a los grupos de tratamiento y control es aleatoria. Puede advertirse que los grupos de tratamiento y control presentan características semejantes (validez interna de

los resultados de la evaluación), pero el resto del universo que no conforma la muestra de evaluación presenta características distintas (no hay validez externa de los resultados). Esto se debe al sesgo introducido por los criterios de ordenación de los individuos al seleccionar la muestra.

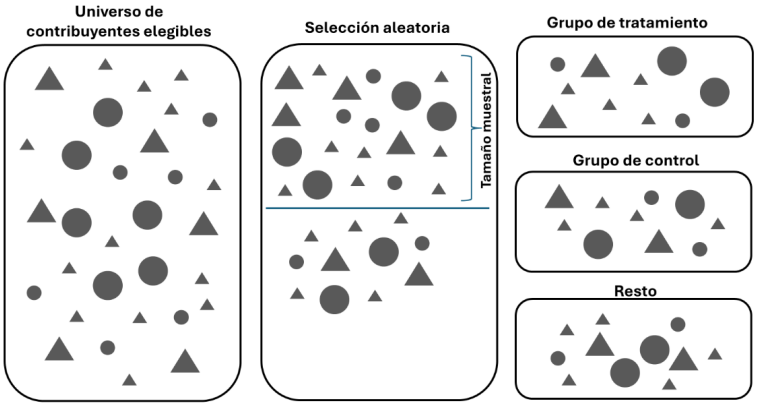
Figura 1. Selección muestral jerarquizada y asignación aleatoria



Fuente: elaboración propia

La Figura 2, por su parte, ilustra un diseño en el que tanto la selección muestral como la asignación a los grupos de tratamiento y control son aleatorias. En este caso puede advertirse que no solamente los grupos de tratamiento y control presentan características semejantes (validez interna), sino que además éstos son semejantes al grupo residual del universo que no integra la muestra (validez externa).

Figura 2. Selección muestral aleatoria y asignación aleatoria



Fuente: elaboración propia

3.4.2 Tamaño muestral

Dadas las características que acaban de reseñarse, el tamaño muestral estará condicionado por la capacidad de ejecución de la administración tributaria en el período planificado. Por ejemplo, si se establece una proporción 50/50 para los grupos de tratamiento y de control, entonces el tamaño muestral máximo será el equivalente a dos veces la capacidad de ejecución.

Si la capacidad de ejecución es muy elevada, eventualmente, el tamaño muestral podría ser menor a ese máximo, porque es posible que se consiga una representación estadística aceptable con tamaños menores.

La Tabla 1 muestra algunos ejemplos de resultados del Efecto Mínimo Detectable (EMD)[6] que corresponden a distintas características de los datos. El EMD es la diferencia mínima entre la media del grupo tratado y la media del grupo de control (para la variable objetivo) que es atribuible al tratamiento. La tabla expresa los valores de EMD para un valor estandarizado de la media muestral de la variable objetivo, igual a 1.

Por ejemplo, supóngase que la variable elegida para evaluar la eficacia de un tratamiento es el monto de pagos de los contribuyentes. El panel del medio de la Tabla 1 indica que, si el tamaño muestral es de 1000 individuos, el desvío estándar es 1,5 (es decir, 1,5 veces el valor de la media de pagos de toda la muestra) y las proporciones del grupo de tratamiento y de control son 50/50, entonces el EMD será 0,24. Con una media igual a 1, este resultado quiere decir que solamente si se encuentran diferencias entre las medias del grupo tratado y del grupo de control superiores a 24% podrán identificarse efectos atribuibles al tratamiento. Si lo que se observa es una diferencia menor, esto no significaría que el tratamiento no haya sido eficaz, sino que esta diferencia no es estadísticamente significativa, dado el tamaño muestral y la variabilidad que presentan los datos.

En los distintos paneles de la Tabla 1 puede advertirse que el valor del EMD será menor cuanto mayor sea el tamaño muestral y cuanto menor sea la variabilidad de los datos, expresada en el desvío estándar. Asimismo, puede advertirse que, dado un determinado tamaño muestral y un desvío estándar, el valor del EMD es menor para una proporción de 50% asignada al grupo de tratamiento. De hecho, puede demostrarse que la asignación de proporciones 50/50 minimiza el EMD.

6 En esta parte se utiliza como referencia a E. Duflo, R. Glennerster y M. Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit." Documento de discusión CEPR Núm. 6059. Londres: Center for Economic Policy Research. Disponible en www.cepr.org como Discussion Paper Nº 6059.

Tabla 1. EMD expresado en porcentaje de la media muestral para distintos tamaños muestrales y variabilidad

Tamaño muestral (N)	Proporción de la muestra asignada al grupo de tratamiento (P)		
	0,5	**0,25**	**0,1**
100	1,24	1,44	2,07
500	0,56	0,64	0,93
1000	0,39	0,45	0,66
5000	0,18	0,20	0,29
10000	0,12	0,14	0,21
Datos de la configuración:			
Power:	0,80		
Nivel de significación:	0,05		
Desvío estándar:	**2,50**		
Varianza:	**6,25**		

Tamaño muestral (N)	Proporción de la muestra asignada al grupo de tratamiento (P)		
	0,5	**0,25**	**0,1**
100	0,75	0,86	1,24
500	0,33	0,39	0,56
1000	0,24	0,27	0,39
5000	0,11	0,12	0,18
10000	0,07	0,09	0,12
Datos de la configuración:			
Power:	0,80		
Nivel de significación:	0,05		
Desvío estándar:	**1,50**		
Varianza	**2,25**		

Tamaño muestral (N)	Proporción de la muestra asignada al grupo de tratamiento (P)		
	0,5	**0,25**	**0,1**
100	1,99	2,30	3,32
500	0,89	1,03	1,48
1000	0,63	0,73	1,05
5000	0,28	0,32	0,47
10000	0,20	0,23	0,33
Datos de la configuración:			
Power:	0,80		
Nivel de significación:	0,05		
Desvío estándar:	**4,00**		
Varianza	**16,00**		

Fuente: elaboración propia

No obstante, esto no quiere decir, necesariamente, que siempre sea preferible una distribución de proporciones de este tipo. Esto dependerá del costo de ejecución y gestión de los grupos de tratamiento y de control. Si el costo unitario de gestión de los individuos del grupo de control es relativamente bajo, podría obtenerse un tamaño muestral grande aumentando el tamaño de este grupo, y así conseguir

un EMD relativamente bajo y acorde con las expectativas del experimento. En este sentido, Duflo et al (2007) proponen la siguiente regla de optimización para establecer las proporciones del grupo de tratamiento y de comparación:

$$\frac{P}{1-P} = \sqrt{\frac{c_c}{c_t}}$$

Donde c_c es el costo unitario por la gestión de cada individuo que integra el grupo de comparación (o control) y c_t es el costo unitario por la ejecución del tratamiento y gestión de cada individuo que integra el grupo de tratamiento. Como veremos en la sección siguiente, el costo del grupo de control puede ser nada desdeñable.

En la Tabla 1, por último, hay dos datos de la configuración que se establecieron como estándar: la potencia estadística, también llamado poder estadístico o *power*, y el nivel de significación.

El nivel de significación es la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera. Para el caso de estas pruebas, la hipótesis nula indicaría que el efecto del tratamiento es nulo, lo que en definitiva se traduciría en que no habría una diferencia estadísticamente significativa entre las medias del grupo tratado y el grupo de control para la variable objeto de evaluación. En la práctica de diseños experimentales y en pruebas estadísticas de hipótesis suele configurarse este nivel de significación en 5% y éste es el estándar aplicado en los resultados de la Tabla 1.

El poder estadístico puede interpretarse como la probabilidad de encontrar significancia estadística si la hipótesis alternativa de una prueba es verdadera. Para el caso que nos interesa, esta hipótesis alternativa indicaría que la media del grupo tratado es diferente de la media del grupo de control. Esto equivale a interpretar la potencia estadística como la probabilidad de rechazar una hipótesis nula falsa.

Configurar una baja potencia estadística en un experimento puede resultar problemático porque, en definitiva, implica una baja probabilidad de detectar un efecto verdadero, significativo en términos estadísticos. En otras palabras, significaría configurar un experimento con pocas posibilidades de obtener conclusiones, lo que podría suponer un despilfarro de recursos. En contrapartida, una potencia estadística muy elevada presenta el riesgo de que el experimento resulte muy sensible a pequeñas diferencias, es decir, que pequeñas diferencias entre el grupo tratado y el grupo de control resulten estadísticamente significativas, cuando en realidad podrían no ser relevantes.

Por lo tanto, es importante conservar un equilibrio entre ambos riesgos. En la práctica de diseños experimentales suele considerarse una potencia de 80% como una configuración estándar, y es ésta la que se ha establecido para la Tabla 1.

En suma, definir el tamaño muestral de un experimento que se coordina con la planificación de acciones de tratamiento de una administración tributaria está condicionado por la capacidad de ejecución del tratamiento (dotación de recursos dedicada a la ejecución del tratamiento) y dependerá de:

- El tamaño del efecto de tratamiento que se pretenda detectar.

- La dispersión de valores en la variable objetivo, resumida en el desvío estándar.

- Las proporciones muestrales que se asignen, respectivamente, al grupo de tratamiento y al grupo de control, teniendo presente que estas proporciones se pueden optimizar considerando los costos de ejecución y gestión unitarios de los individuos asignados a cada grupo (si esa información estuviera disponible).

- La potencia estadística y el nivel de significación que se establezcan para el experimento.

Para finalizar, se recomienda que una vez determinado el tamaño muestral, la selección efectiva sea por un número algo superior para contemplar eventuales reemplazos. En la sección 3.4.5 se advierte de las diferencias que pueden presentarse en la configuración de los grupos de tratamiento y de control entre el momento que éstos son conformados y el momento de ejecución del tratamiento. Las circunstancias que conducen a estas diferencias pueden requerir reemplazos, que deberían estar previamente contemplados en la selección. Esto es particularmente crítico si la selección es jerarquizada, es decir, si aplica algún tipo de ranking, ya que, de no preverse una porción extra, difícilmente se consigan reemplazos debido al sesgo de selección establecido por el ranking. En cambio, si la selección muestral es aleatoria, la posibilidad de obtener reemplazos es cierta. Pero, de todos modos, conviene que en ambos casos se seleccione un porcentaje extra (entre un 10 y un 20 por ciento del tamaño muestral) considerando los requerimientos de gestión de los grupos de tratamiento y de control (ver sección 3.4.5).

3.4.3 Estimación del efecto del tratamiento.

Una forma identificar el efecto del tratamiento es especificando un modelo de regresión lineal como el siguiente:

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

Donde Y_i es la variable objetivo cuyo desempeño se quiere evaluar, α es una constante estimada por el modelo T_i toma el valor 1 (uno) si el contribuyente i recibió el tratamiento o 0 (cero) si no lo recibió y ϵ_i es un término de error.

Adicionalmente, la especificación del modelo puede incorporar otras variables de control (X_i) que contribuyan a explicar el desempeño de la variable objetivo. Por ejemplo, si la variable objetivo es el monto de pagos de un período determinado, variables de control podrían ser el monto de pagos del período anterior, o algún indicador asociado con el volumen histórico de pagos del contribuyente.

El parámetro β es el efecto promedio del tratamiento. Si es estadísticamente significativo, su valor estará indicando, en promedio, cuánto es el efecto en la variable objetivo atribuible al tratamiento. La estimación de este parámetro en el modelo puede hacer por Mínimos Cuadrados Ordinarios (MCO).

3.4.4 Efectos indirectos del tratamiento.

Bajo este método de evaluación se asume que un tratamiento influye en los resultados de manera simple y lineal sobre todos los individuos tratados. Sin embargo, en el ámbito del cumplimiento tributario, es plausible que las acciones de tratamiento generen efectos en forma indirecta sobre contribuyentes no tratados. Algunos contribuyentes pueden optar por cambiar su comportamiento tributario solamente por acceder a información sobre la acción de tratamiento en curso, aun cuando ésta no los haya involucrado directamente. Esto dependerá, entre otros factores, de las características de la acción de tratamiento y de la reputación de la administración tributaria. Por supuesto, para que esto ocurra, la noticia de que el tratamiento se está ejecutando o se ejecutó tiene que llegar al conocimiento de esos contribuyentes. Esto puede ocurrir porque la administración tributaria les dio difusión pública a ciertas acciones de tratamiento, o por interacciones entre contribuyentes tratados y no tratados, especialmente cuando existe algún vínculo entre éstos.

La difusión pública de una acción de tratamiento puede interpretarse como un tratamiento en sí mismo, por lo tanto, sus efectos no son necesariamente en contribuyentes “no tratados”, en la medida que todos son receptores de esta difusión pública.

En cambio, cuando el efecto se da por interacciones entre contribuyentes vinculados, sí resulta más claro que estamos frente a efectos indirectos. Si un

contribuyente no tratado que cambia su comportamiento debido al tratamiento es integrante del grupo de control, el efecto indirecto estará violando un requisito básico del método, de que el resultado de una unidad (un contribuyente) no debería verse afectado por la asignación concreta del tratamiento a otras unidades. Este supuesto de estabilidad del valor de la unidad de tratamiento es necesario para asegurar que la asignación aleatoria produzca estimaciones no sesgadas. En otras palabras, si el grupo de control se ve indirectamente afectado por el tratamiento recibido por el grupo de tratamiento, la comparación no representa con precisión el contrafactual, es decir, qué habría ocurrido en el grupo de tratamiento en ausencia de tratamiento.

En tales casos, para que el grupo de control sea una medida precisa del contrafactual, será necesario que éste esté integrado por contribuyentes no expuestos a interacciones con contribuyentes del grupo de tratamiento. Ahora bien, si los efectos indirectos existen, constituyen un impacto real del tratamiento, y puede ser de interés su medición.

Si se sospecha que pueden producirse efectos indirectos del tratamiento (también llamados efectos de derrame), sería conveniente que la evaluación se proponga un objetivo doble: conocer tanto el efecto directo del tratamiento (sobre los individuos tratados) y el efecto indirecto (sobre individuos no tratados).

Este propósito podría demandar más información para diseñar la evaluación, ya que podría considerarse necesario disponer de datos sobre los vínculos entre los contribuyentes, por ejemplo, relaciones comerciales, societarias, o también identificar si se comparte un asesor tributario, que es un actor influyente en las decisiones de cumplimiento tributario. Con esa información se pueden establecer métricas que midan el grado de vinculación entre contribuyentes, particularmente, de cada contribuyente integrante de los grupos de evaluación (tratamiento y control) con otros integrantes de estos grupos.

La localización geográfica del contribuyente podría ser otro atributo de utilidad para discriminar si la probabilidad de interacción con otro contribuyente es alta o baja. Este atributo podría ser especialmente útil en el caso de personas físicas o pequeñas empresas (principalmente unipersonales).

El diseño muestral para este propósito se complejiza. Sería necesario establecer cuatro grupos:

- Grupo A: el grupo de tratamiento
- Grupo B: el grupo de comparación o control, cuyos integrantes deberían presentar bajo o nulo nivel de vinculación (según las métricas elaboradas) con los integrantes del grupo de tratamiento, para admitir que hay una baja probabilidad de interacción.
- Grupo C: un grupo de “afectados por el derrame”: contribuyentes con alto

grado de vinculación (según las métricas elaboradas) con los integrantes del grupo de tratamiento, que no fueron seleccionados para recibir el tratamiento.

- Grupo D: un grupo de control o comparación para el grupo de afectados por el derrame: también con bajo o nulo nivel de vinculación con el grupo de tratamiento, así como con el grupo de afectados por el derrame.

El efecto directo se podría medir a través de la estimación del modelo descrito en 4.4.3 considerando a los grupos A y B, y el efecto indirecto se estimaría configurando el mismo modelo, pero para los grupos C y D.

3.4.5 Gestión de los grupos que conforman la evaluación.

Definir a los grupos que conforman la evaluación (tratamiento y control) de la manera descrita supone que la evaluación de la efectividad de los tratamientos debe integrarse a la planificación de las acciones de tratamiento. Para que la evaluación de la eficacia de cada tratamiento tenga validez, el grupo de comparación o control no puede recibir ningún tratamiento durante el período de evaluación, a menos que sea un tratamiento que todos reciben, o que la intención sea evaluar el efecto diferencial de un tratamiento. Por razones semejantes, el grupo tratado no puede recibir otro tratamiento durante el período de evaluación, más que las acciones de tratamiento que serán objeto de evaluación.

Si estas condiciones no se cumplen, habrá otros factores incidiendo en la conducta de los contribuyentes, por lo que difícilmente se podrá aislar el efecto específico del tratamiento que se pretende evaluar.

Garantizar que las condiciones anteriores se cumplan requiere considerables esfuerzos de coordinación orientadas a gestionar a los grupos de control y de tratamiento, la cual se facilita si los grupos se diseñan en el contexto de la planificación de acciones de tratamiento.

Durante el período de evaluación, los contribuyentes que integren los grupos de control y de tratamiento no pueden ser considerados para recibir otras acciones de tratamiento incluidas en la planificación de acciones operativas.

Hay dos maneras de conseguir esto, y ambas pueden utilizarse en forma complementaria. Por un lado, si la planificación incluye acciones que ya tienen nóminas de contribuyentes, se puede hacer un cruce de éstas con las nóminas de los grupos de control y de tratamiento y, los casos de coincidencia se pueden quitar de estos dos últimos y se buscan reemplazos.

Si la planificación incluye acciones sin nóminas, dado que se prevé generarlas en una fecha más cercana a su ejecución, el cruce solamente se podrá hacer una

vez que la nómina esté lista, y llegado el caso es probable que el tratamiento que es objeto de la evaluación ya esté en curso. Por lo tanto, en estas situaciones, si hubiera coincidencia de casos entre la nómina en cuestión con integrantes de grupos de control o de tratamiento, lo que correspondería es que el caso se planifique para ser ejecutado después que finalice el período de evaluación.

Se hace referencia al "grupo de tratamiento" como aquel que está integrado por los contribuyentes elegidos para recibir el tratamiento. En cambio, el "grupo tratado" está compuesto por quienes efectivamente reciben el tratamiento. Es probable que se presenten circunstancias en el tiempo transcurrido entre la definición del grupo de tratamiento (planificación) y la ejecución del tratamiento, que generen como resultado diferentes composiciones entre un grupo y otro. Estas circunstancias podrían estar vinculadas a, por ejemplo, la imposibilidad de contactar al contribuyente, o a una información tributaria que se actualiza y da como resultado que el contribuyente ya no es elegible para el tratamiento (esta segunda circunstancia también puede afectar a los integrantes del grupo de control).

La forma más satisfactoria de resolver estas diferencias es buscando reemplazos antes de que la ejecución del tratamiento dé inicio. En la medida que el tamaño de la muestra de evaluación sea inferior al tamaño del universo de contribuyentes elegibles, esta reposición de casos será posible, aplicando el mismo procedimiento ya descrito en la sección 3.4.1 y siguiendo las sugerencias indicadas sobre el final de la sección 3.4.2.

Si esta reposición en la muestra no se aplica oportunamente, o bien no es posible aplicarla porque no existen reemplazos (el tamaño muestral es igual al tamaño del universo), entonces será necesario ajustar la composición de los grupos de evaluación (tratado y control) antes de proceder a estimar el efecto del tratamiento. Existen técnicas que permiten ajustar esta composición cuando se presentan diferencias entre el diseño muestral y lo efectivamente ejecutado. Para una explicación detallada véase, por ejemplo, Duflo y otros (2007).

3.4.6 Período de evaluación.

Por "período de evaluación" se hace referencia al tiempo que transcurre entre la ejecución de las acciones del tratamiento evaluado y el momento en el que se recogen los resultados para ser evaluados. Como fue señalado, los contribuyentes que integran el grupo de control no deberían recibir tratamiento alguno durante el período de evaluación, y los que integran el grupo tratado ningún otro tratamiento más que el que está siendo objeto de la evaluación.

Por este motivo, además del costo de coordinación que supone gestionar los grupos de tratamiento y de control para que sigan siendo tales, con la evaluación del tratamiento se produce un costo de oportunidad para la administración tributaria derivado de la "veda" de tratamientos que se establece sobre los contribuyentes que integran estos grupos.

Como consecuencia de esto, durante el período de evaluación la administración tributaria estará resignando un beneficio potencial, el que se obtendría de intervenir a estos contribuyentes con otros tratamientos. Si se sigue el diseño propuesto en la sección 3.4.1 para la selección muestral, los contribuyentes seleccionados presentarán cierto interés fiscal para la administración tributaria, por lo que este costo de oportunidad podría ser, en algunos casos, elevado.

En definitiva, hay dos costos a considerar como restricción al momento de establecer el período de evaluación: los costos de gestión de los grupos y el costo de oportunidad por la veda de tratamientos sobre estos grupos[7]. Ambos costos se incrementan con la duración del período de evaluación.

Por otro lado, hay que considerar las características del tratamiento y de la variable que es objeto de evaluación para establecer cuánto es el tiempo mínimo necesario para observar resultados que puedan atribuirse al efecto del tratamiento.

Los tratamientos basados en la autorregulación del contribuyente prevén, en algunos casos, una secuencia de intervenciones en función de la respuesta del contribuyente, por lo que pueden requerir de más tiempo para evaluar sus resultados.

Por último, la extensión del período de evaluación también debe contemplar si existe interés en evaluar la duración del efecto de un tratamiento, es decir, si el cambio en el comportamiento del contribuyente es permanente o se diluye en el tiempo.

En definitiva, la duración del período de evaluación puede ser muy variable, ya que son muchas las condiciones que influyen en esta definición. No obstante, en el contexto de una coordinación de la evaluación con la planificación de las acciones de tratamiento, podría considerarse, como referencia general, una duración de 6 meses para el período de evaluación. Ésta constituye una duración aceptable para evaluar a la mayoría de los tratamientos, y resulta compatible con la lógica de planificación anual de las operaciones, que suele ser la más aplicada en las administraciones tributarias.

3.4.7 Otros requerimientos: estandarización y comunicaciones internas.

Evaluar un tratamiento debería presuponer que éste se ejecuta en forma estandarizada. Si cada caso individual que recibe un tratamiento (o una serie de acciones de tratamiento) sigue el mismo estándar, entonces los resultados de la evaluación serán más precisos. Si las acciones individuales de tratamiento no siguen un estándar, cabe preguntarse cuáles serían las características concretas del tratamiento a ser evaluado, ya que éste dependería de cómo es ejecutado

7 El costo de oportunidad es mayor cuando la selección muestral es enteramente aleatoria, porque significa que no todos los contribuyentes seleccionados para el tratamiento son los que presentan las situaciones más severas en los riesgos analizados que dieron lugar a éste. Cuando la selección muestral es jerarquizada por el nivel de riesgo, en cambio, hay un costo de oportunidad menor.

en cada caso individual. Una pregunta semejante podría formularse sobre la efectividad: ¿Cuál sería la forma específica del tratamiento que resulta eficaz? Será muy difícil contestar a una pregunta así, si no existe una definición y un control específicos sobre cómo el tratamiento es ejecutado.

La ejecución de un tratamiento puede involucrar a muchas personas en una organización como la administración tributaria, a su vez, el contribuyente que recibe el tratamiento puede interactuar, directa o indirectamente, con diferentes personas o áreas de la organización. De manera que la estandarización del proceso puede comprender diversos aspectos, como los servicios de consulta que se ponen a disposición para aclarar dudas o inquietudes sobre el tratamiento, la depuración de casos previo a la ejecución de las acciones, la ejecución misma, el seguimiento de casos, las acciones que se desplieguen en respuesta a la conducta del contribuyente, entre otros.

Por otra parte, la "veda de tratamientos" para los grupos de tratamiento y de control demanda que todas las áreas que interactúan con los contribuyentes deben estar en condiciones de poder verificar, antes de iniciar cualquier acción que involucre a un contribuyente, si éste integra una nómina de grupo de control o de grupo de tratamiento. Como se mencionó en la sección 3.4.5, se esperaría que estas verificaciones y controles cruzados se efectúen durante la planificación de las acciones operativas. No obstante, en esa sección también se mencionó que algunas acciones planificadas podrían no contener una nómina disponible al momento de la planificación. Adicionalmente, podrían desplegarse acciones por fuera de lo planificado. Por lo tanto, es relevante que la posibilidad de verificación esté disponible en todo momento.

Por razones de seguridad y de confidencialidad de los datos, no se esperaría que las áreas que ejecutan otras acciones tengan acceso a las nóminas de los grupos de evaluación, pero sí la posibilidad de consultar y acceder a una respuesta inmediata, de tal manera que no les suponga un entorpecimiento de su operativa. Por ejemplo, a través de un servicio web corporativo, en el que pueda indicarse el número de identificación tributario y el servicio les devuelva si el contribuyente está en "veda" por integrar un grupo de control o de tratamiento que está siendo evaluado, y el plazo que rige para esta veda.

Si la acción a ser ejecutada por el área es de carácter indiscriminado, es decir, se aplica a todos los contribuyentes o a una gran masa de contribuyentes, por ejemplo, comunicaciones con información general (por ejemplo, sobre la disponibilidad de un nuevo servicio) o recordatorios de vencimientos de plazos para presentación de declaraciones que se envíen en forma general, no sería necesario efectuar una consulta de este tipo. En la medida que se trate de una acción que involucre a todos los contribuyentes (y más en particular, a todo el grupo de tratamiento y a todo el grupo de control), la intervención no genera diferenciaciones, por lo que no es necesario interrumpirla o suspenderla.

Pero si se trata de acciones de carácter más individual, que afectan a algunos contribuyentes en particular, con determinadas características, entonces sería conveniente que estas acciones sean suspendidas por el término del período de evaluación.

Si el área involucrada considera que, debido a la gravedad del caso que está analizando y por razones de oportunidad, es conveniente que el caso sea tratado sin dilaciones (por ejemplo, el establecimiento de un embargo sobre un contribuyente con una deuda morosa significativa que tiene un alto riesgo de no pago), el caso podría ser comunicado para que el área encargada de la evaluación del tratamiento lo analice y proponga un caso de reemplazo.

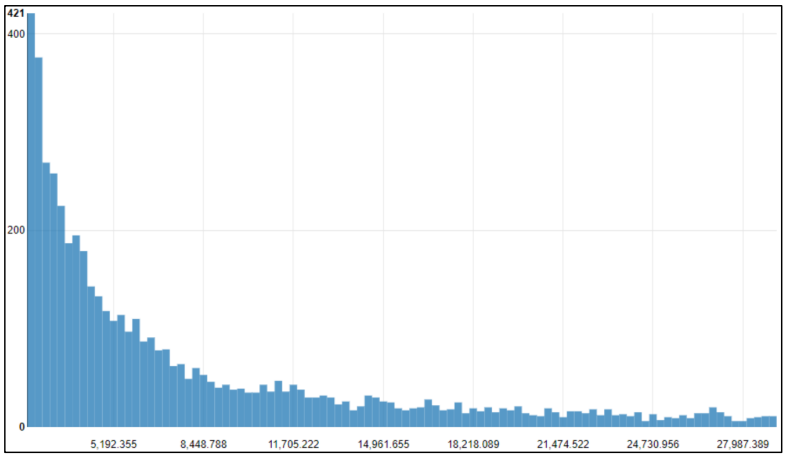
4. UN EJEMPLO NUMÉRICO

En esta sección se describe un ejemplo numérico para ilustrar la práctica que ha sido recomendada en este artículo. El ejemplo contiene datos ficticios, pero la estructura de estos datos está basada en casos de la realidad de las administraciones tributarias.

4.1 Características generales.

La estructura de datos comprende a 5142 contribuyentes ficticios que, para cierto riesgo de cumplimiento valorado, presentan un nivel de severidad (probabilidad de ocurrencia y consecuencia estimadas) dentro de cierto rango para el cual se ha dispuesto un tratamiento específico. La Figura 3 muestra el histograma con la distribución de la severidad que presentan estos contribuyentes, expresadas en unidades monetarias de la moneda local (ficticia).

Figura 3. Histograma del nivel de severidad en el riesgo valorado (en unidades monetarias)



Fuente: elaboración propia con base en un ejercicio de simulación

Por su parte, la Tabla 2 muestra estadísticas básicas tanto de la severidad del riesgo valorado y considerado para la asignación del tratamiento a ser evaluado como del volumen de pagos de estos contribuyentes. El valor del volumen de pagos considerado en la tabla corresponde al promedio de los últimos 4 períodos calculado para cada contribuyente. La severidad está calculada para un período, por lo tanto, el valor de los pagos es comparable con el de la severidad. Así, por ejemplo, el valor medio de severidad (7.910) es 1,55 veces el valor medio de los pagos (5.117).

Tabla 2: Estadísticas básicas de severidad en el riesgo valorado y de volumen de pagos (promedio de los últimos 4 períodos) – Datos por segmento de contribuyentes

Segmento	Cantidad de contribuyentes	Severidad en el riesgo valorado			Volumen de pagos		
		Media	Desviación estándar	Desv. Std. / Media	Media	Desviación estándar	Desv. Std. / Media
A	1800	7.671	6.325	0,82	5.273	20.279	3,85
B	2354	8.342	6.872	0,82	4.776	18.677	3,91
C	988	7.316	6.119	0,84	5.646	22.032	3,90
Total	**5142**	**7.910**	**6.555**	**0,83**	**5.117**	**19.920**	**3,89**

Fuente: elaboración propia con base en un ejercicio de simulación

Los datos de la Tabla 2 se muestran abiertos por segmento de contribuyentes. La categorización por segmento es aquí abstracta y meramente referencial; podría corresponder a regiones geográficas, atributos económicos de los contribuyentes, personerías jurídicas, etcétera. Su presentación es al solo efecto de tener una referencia de desagregación, para posteriormente analizar resultados de la evaluación por segmento.

4.2 Selección de la muestra de evaluación.

Como fue descrito en la sección 3.4.1, si se pretende evaluar la eficacia de este tratamiento, la selección muestral puede ser aleatoria o jerarquizada. En el primer caso, los resultados de la evaluación tendrán validez para el conjunto de los 5142 contribuyentes, mientras que en el segundo caso la validez se restringirá al conjunto de contribuyentes comprendidos en los grupos de evaluación (grupo de tratamiento y grupo de comparación o control).

Por su parte, como fue descrito en la sección 3.4.2, el tamaño muestral dependerá del nivel de dispersión que presenten los datos y del tamaño del efecto derivado del tratamiento que pretenda detectarse. En tal sentido, la Tabla 2 muestra que la dispersión de los datos, sintetizada en el cociente entre la desviación estándar y la media, depende de la variable que se proponga como objeto de estudio para la evaluación.

En este ejemplo, si la variable objetivo fuera la severidad del riesgo, es decir, comparar la severidad del riesgo después del tratamiento entre el grupo de tratamiento y el grupo de control, el grado de dispersión resultaría muy bajo: la desviación estándar es 0,83 veces el valor medio. En cambio, si la variable objetivo fuera el volumen de pagos (comparar el desempeño en pagos después del tratamiento entre el grupo de tratamiento y el grupo de control) el grado de dispersión es considerablemente más alto: 3,89.

Con estos valores de dispersión, si, por ejemplo, la muestra tuviera un tamaño de 4000 individuos y se dividiera en partes iguales entre el grupo de tratamiento y el grupo de control, el EMD sería 0,07 (7% del valor medio) si la variable objetivo fuera la severidad del riesgo y 0,31 (31% del valor medio) si la variable objetivo fuera el volumen de pagos.

Asúmase que éstos valores de EMD son aceptables para los objetivos de la evaluación y que la configuración antedicha es la utilizada para el diseño muestral. A continuación, con este ejemplo se ilustra el proceso de diseño muestral bajo las dos alternativas de selección, aleatoria y jerarquizada, y su posterior asignación aleatoria para la conformación de los grupos de evaluación.

4.2.1 Selección y asignación aleatorias.

La Tabla 3 muestra el resultado de seleccionar aleatoriamente 4000 individuos para la muestra de evaluación, para posteriormente asignar aleatoriamente una mitad de estos individuos al grupo de tratamiento y la otra mitad al grupo de control. Puede advertirse que el valor medio de la severidad en el riesgo es muy semejante entre los dos grupos, pero también lo es con el resto de los individuos no seleccionados para la muestra (en la tabla, la fila “Resto”).

Las pruebas típicas de igualdad de medias y varianzas (por ejemplo, prueba *F* y test de Levene) con estos datos muestran que la hipótesis nula de igualdad de medias y varianzas no puede descartarse, lo que a su vez se refleja en que los intervalos de confianza al 95% para el valor de la media de los tres grupos estén mayoritariamente superpuestos.

En otras palabras, los dos grupos que conforman la evaluación tienen igual distribución de valores de la severidad del riesgo y, adicionalmente, la distribución de valores del grupo residual también es igual. Por lo tanto, los resultados que se obtengan al evaluar el desempeño comparado del grupo de tratamiento y del grupo de control serán válidos no solamente para estos dos grupos, sino también para todo el universo de 5142 contribuyentes considerados.

Tabla 3. Selección y asignación aleatorias: comparación de medias y varianzas entre los grupos de evaluación

Grupo	Cantidad de contribuyentes	Media	Desviación estándar	Intervalo de confianza para la media	
				Límite inferior	Límite superior
Tratamiento	2.000	7.990	6.601	7.700	8.279
Control	2.000	7.923	6.552	7.636	8.210
Resto	1.142	7.747	6.485	7.370	8.124
Total	5.142	7.910	6.555	7.731	8.089

Fuente: elaboración propia con base en un ejercicio de simulación

Como resultado directo de la aleatoriedad aplicada en el proceso, el análisis de cualquier característica de los contribuyentes, analizada estadísticamente, no debería mostrar diferencias entre los grupos. A vía de ejemplo, la Tabla 4 muestra la distribución de los contribuyentes en cada grupo por segmento, donde puede advertirse que las composiciones de cada grupo son muy semejantes.

Tabla 4. Selección y asignación aleatorias: distribución por segmento

Grupo	Distribución por segmento		
	Segmento	Cantidad de contribuyentes	Participación
Tratamiento	A	693	35%
	B	907	45%
	C	400	20%
Control	A	703	35%
	B	933	47%
	C	364	18%
Resto	A	404	35%
	B	514	45%
	C	224	20%

Fuente: elaboración propia con base en un ejercicio de simulación

4.2.2 Selección jerarquizada y asignación aleatoria

En este caso, la selección muestral es el resultado de ordenar a los contribuyentes a través de algún criterio. La Tabla 5 muestra el resultado de ordenar a los contribuyentes en forma descendente según su nivel de severidad en el riesgo valorado, para luego elegir a los primeros 4000 para conformar la muestra de evaluación. La asignación de los contribuyentes a los grupos de tratamiento y de control es completamente aleatoria, por lo que se espera que sus medias y varianzas sean iguales, algo notorio en la Tabla 5.

Tabla 5. Selección jerarquizada y asignación aleatoria: comparación de medias y varianzas entre los grupos de evaluación

Grupo	Cantidad de contribuyentes	Media	Desviación estándar	Intervalo de confianza para la media	
				Límite inferior	Límite superior
Tratamiento	2.000	9.520	6.650	9.228	9.811
Control	2.000	9.411	6.668	9.119	9.703
Total	4.000	9.465	6.658	9.259	9.672

Fuente: elaboración propia con base en un ejercicio de simulación

La validez interna de los resultados está asegurada por la asignación aleatoria, y también se puede observar al analizar cualquier característica. Otra vez, a vía de ejemplo, la Tabla 6 muestra la composición de cada grupo por segmento.

Lo que también es notorio al comparar los resultados de esta tabla con los de la Tabla 3, es que la media de los grupos (9.465 en la Tabla 5) es superior a la media de todo el universo (7.910 en la Tabla 3). Esto es el resultado directo de la selección jerarquizada: se elige a los primeros 4000 del ranking de severidad, por lo tanto, su valor medio en severidad es superior. Por consiguiente, los resultados que se obtengan con esta evaluación serán válidos al interior del subgrupo de 4000 contribuyentes, pero no para todo el universo de 5142.

Tabla 6. Selección jerarquizada y asignación aleatoria: distribución por segmento

Grupo	Distribución por segmento		
	Segmento	Cantidad de contribuyentes	Participación
Tratamiento	A	689	34%
	B	934	47%
	C	377	19%
Control	A	684	34%
	B	940	47%
	C	376	19%

Fuente: elaboración propia con base en un ejercicio de simulación

4.3 Simulación de resultados del tratamiento

Como parte de este ejemplo numérico se han simulado resultados del tratamiento a ser evaluado. En aras de la brevedad, la simulación se focaliza en una selección

jerarquizada con asignación aleatoria (sección 4.2.2), pero las características del análisis son semejantes si se tratara de una selección y asignación aleatorias (sección 4.2.1).

Supóngase que la variable objetivo para la evaluación es el volumen de pagos a ser observado en el período posterior a la ejecución del tratamiento, al que denominaremos genéricamente "t+1". Para los 4000 contribuyentes que integran los grupos de evaluación (tratamiento y control) de la sección 4.2.2 la media y desviación estándar del volumen promedio de pagos de los 4 períodos anteriores al tratamiento son las que se muestran en la Tabla 7.

Tabla 7. Selección jerarquizada y asignación aleatoria: comparación de medias y varianzas entre los grupos de evaluación para el promedio de pagos

Grupo	Cantidad de contribuyentes	Media	Desviación estándar	Intervalo de		Desv. Std. / Media
				Límite inferior	Límite superior	
Tratamiento	2.000	2.898	9.062	2.495	3.301	3,13
Control	2.000	3.180	9.536	2.754	3.605	3,00
Total	4.000	3.038	9.301	2.746	3.331	3,06

Fuente: elaboración propia con base en un ejercicio de simulación

Si bien la media de pagos entre los dos grupos no es idéntica, las pruebas típicas no permiten rechas la hipótesis nula de igualdad de medias y varianzas, lo que se refleja en la superposición de los intervalos de confianza para la media.

Nótese, por un lado, que el valor de la media de pagos (3.038) es inferior al de todo el universo (5.117 en la Tabla 2). Es decir que este grupo seleccionado a través de un criterio jerárquico presenta mayor nivel de riesgo (comparación entre Tablas 5 y 3) y menor volumen de pagos de impuesto; esto es un hecho frecuente en los análisis de cumplimiento tributario y de los riesgos asociados a éste.

Por otro lado, nótese también que el nivel de dispersión en el volumen promedio de pagos, sintetizado en el cociente entre la desviación estándar y la media (3,06) es algo inferior al observado para todo el universo (3,89 en la Tabla 2). Esto determina que, para el tamaño y diseño muestral elegidos (4000 dividido en proporciones 50/50 entre tratamiento y control) el EMD sea algo inferior: 0,24, cuando es 0,31 para todo el universo. O sea que sería posible detectar efectos del tratamiento algo menores, en comparación con las posibilidades de detección cuando se trabaja con una muestra seleccionada en forma aleatoria. Éste es la diferencia principal entre los análisis con uno y otro diseño muestral.

El modelo postulado para detectar el efecto del tratamiento sigue lo descrito en la sección 3.4.3. El modelo se estima para cada período, utilizando como variables de control al volumen de pagos del año anterior y a un indicador de tamaño del contribuyente. En los períodos previos al tratamiento, no es esperable encontrar un valor estadísticamente significativo para β, el parámetro que mide el efecto medio del tratamiento, por la simple razón de que el tratamiento no ha sido aplicado. En cambio, en el período posterior al tratamiento sí se esperaría un valor estadísticamente significativo para este parámetro, pero esto dependerá del tamaño del efecto, ya que, como fue señalado, no todos los efectos son detectables para este tamaño y diseño muestrales.

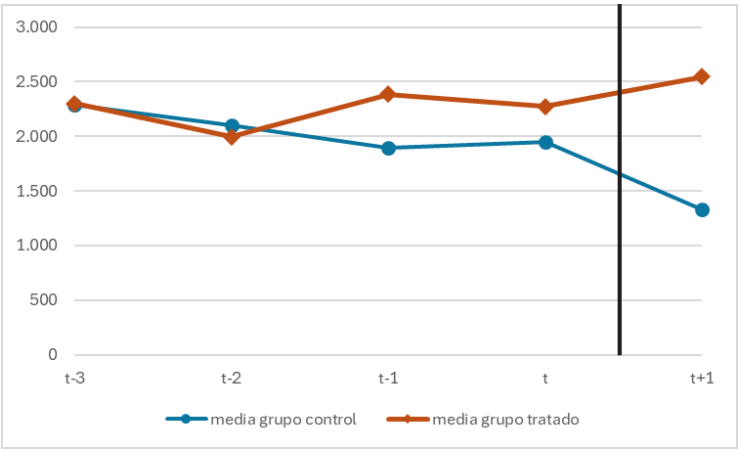
Como parte de este ejercicio se simulan dos escenarios, en el primero se suponen efectos más grandes y, por lo tanto, más fáciles de detectar; en el segundo los efectos son más pequeños.

4.3.1 Escenario 1.

La Figura 4 muestra los resultados simulados del tratamiento para este escenario 1, de efectos más grandes. La gráfica muestra la evolución del valor medio de pagos para el grupo tratado y el grupo de control, entre el período “t-3” y “t”, antes del tratamiento, así como en el período “t+1”, después del tratamiento. La línea vertical trazada indica el momento de ejecución del tratamiento, es decir, entre “t” y “t+1”, o lo que lo mismo, durante “t+1”, pero antes de su finalización.

El efecto simulado es considerablemente grande, ya que mientras el volumen medio de los pagos del grupo de comparación desciende entre “t” y “t+1” (la variación es -32%), el volumen medio de pagos del grupo tratado aumenta 12% en igual período; hay una diferencia de 44 puntos porcentuales. Como fue señalado, el EMD para este tamaño y diseño muestrales es 0,24, por lo que con una diferencia media en los pagos de 0,44 es altamente probable que se pueda detectar un efecto del tratamiento estadísticamente significativo.

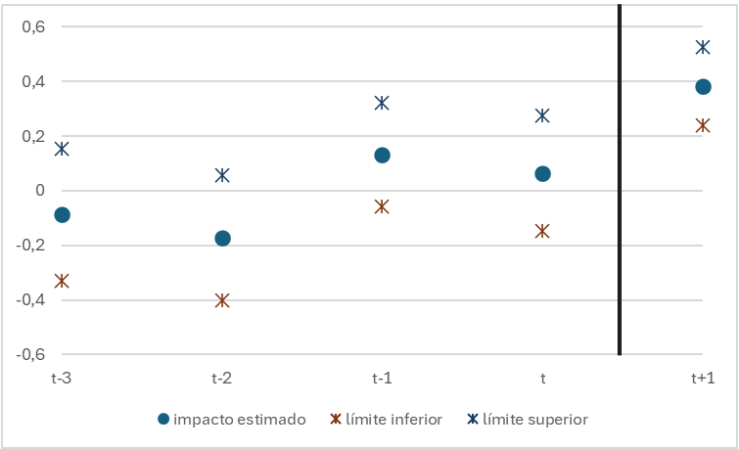
Figura 4. Efecto simulado del tratamiento en el escenario 1: media de volumen de pagos en cada período



Fuente: elaboración propia con base en un ejercicio de simulación

Por su parte, la Figura 5 muestra el valor estimado al configurar el modelo de la sección 3.4.3 para cada período. El círculo indica el valor del parámetro β estimado en cada período expresado en porcentaje del promedio de pagos del grupo tratado correspondiente a cada período. Los asteriscos indican los límites inferior y superior del intervalo de confianza al 95% del parámetro.

Figura 5. Efecto simulado del tratamiento en el escenario 1: impacto medio (en porcentaje) estimado para cada período



Fuente: elaboración propia con base en un ejercicio de simulación

Puede advertirse que en todos los períodos previos a la ejecución del tratamiento los intervalos de confianza contienen al cero, indicando que el parámetro no es estadísticamente significativo. Esto es esperable porque se trata de dos grupos con medias y varianzas del volumen de pagos estadísticamente iguales (Tabla 7) y con evoluciones del volumen de pagos semejantes; antes del tratamiento, no hay razones para esperar un desempeño distinto en los pagos.

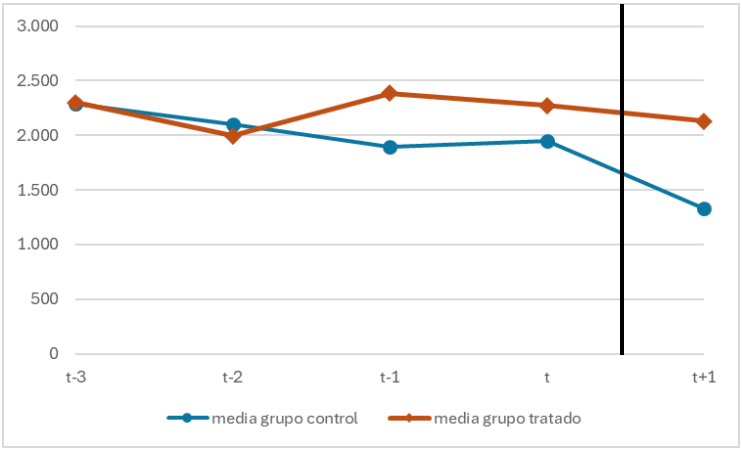
En cambio, en "t+1" se observa que el intervalo de confianza está por encima del cero, indicando que el efecto del tratamiento es estadísticamente significativo. En concreto, el valor promedio estimado es 0,38, lo que quiere decir que, en promedio, los contribuyentes tratados pagan un 38% más que su contrafactual (el grupo de control) como consecuencia del tratamiento. El intervalo de confianza es [0,24; 0,53], por lo que el efecto es estadísticamente significativo.

4.3.2 Escenario 2.

La Figura 6 muestra los resultados simulados del tratamiento para el escenario 2, de efectos más pequeños. La gráfica es idéntica a la Figura 4 en los períodos "t-3" a "t", antes del tratamiento, ya que la simulación solamente afecta al período "t+1", después del tratamiento.

El efecto simulado es más pequeño, pero aún positivo. La variación del volumen medio de los pagos del grupo de control es igual al mostrado en la Figura 4 (-32%), y la variación del volumen medio de pagos del grupo tratado también es negativa, pero menor en términos absolutos: -6%. La diferencia entre ambos es de 25 puntos porcentuales. Con un EMD para este tamaño y diseño muestrales de 0,24, una diferencia media en los pagos de 0,25 puede permitir detectar un efecto del tratamiento estadísticamente significativo, pero con menor probabilidad.

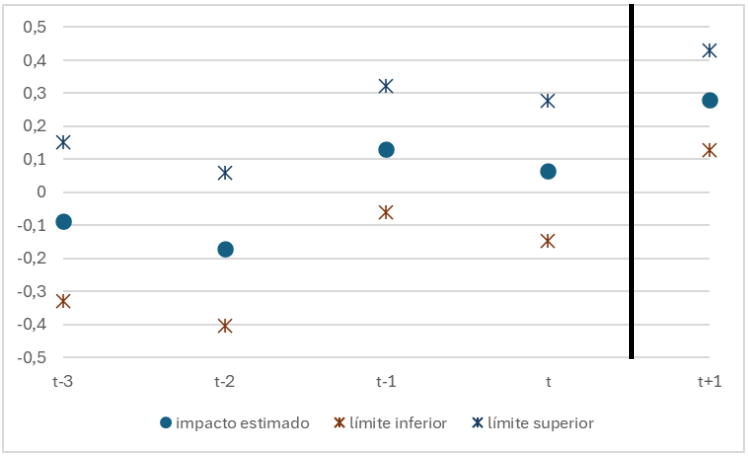
Figura 6. Efecto simulado del tratamiento en el escenario 2: media de volumen de pagos en cada período



Fuente: elaboración propia con base en un ejercicio de simulación

Análogamente a lo presentado con la Figura 5, la Figura 7 muestra el valor estimado al configurar el modelo de la sección 3.4.3 para cada período, pero en este caso para el escenario 2, de efectos pequeños. De nuevo, los resultados para los períodos "t-3" a "t" son idénticos a los observados en la Figura 5; la diferencia está en el período posterior al tratamiento, "t+1".

Figura 7. Efecto simulado del tratamiento en el escenario 2: impacto medio (en porcentaje) estimado para cada período



Fuente: elaboración propia con base en un ejercicio de simulación

El valor promedio estimado del efecto del tratamiento es 0,28, lo que quiere decir que, en promedio, los contribuyentes tratados pagan un 28% más que su contrafactual (el grupo de control) como consecuencia del tratamiento. El intervalo de confianza es [0,13; 0,43], por lo que el efecto es estadísticamente significativo. Nótese que el efecto promedio es inclusive mayor a la diferencia entre las variaciones de la media de pagos de ambos grupos. Esto es factible y depende de la composición de los resultados.

Otro aspecto destacable de estos resultados simulados es que la variación en los pagos del grupo tratado es también negativa, pero, de todos modos, es posible (si la diferencia es lo suficientemente grande) detectar un efecto positivo del tratamiento. El contrafactual está indicando que, en ausencia del tratamiento, el volumen promedio de pagos del grupo tratado hubiera disminuido aún más.

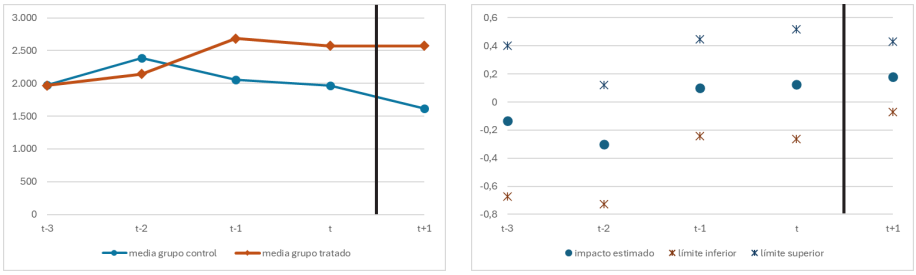
4.3.3 Efectos por segmento.

La Figura 8 muestra los efectos del tratamiento en cada uno de los segmentos de contribuyentes bajo la simulación del escenario 2. Los tres segmentos presentan evoluciones diferentes en el volumen medio de pagos, y también en la respuesta al tratamiento.

En el segmento A, por ejemplo, la variación de los pagos es nula después del tratamiento, y la variación en el grupo de comparación es -18%. Esta diferencia no es suficiente para identificar un efecto positivo derivado del tratamiento. Como se observa en la gráfica de la derecha correspondiente a este segmento, el intervalo de confianza para el parámetro β en el período posterior al tratamiento contiene al cero, por lo que no es estadísticamente significativo. Esto no quiere decir que no exista un efecto positivo del tratamiento en este segmento, sino que éste no puede detectarse para este tamaño y diseño muestrales.

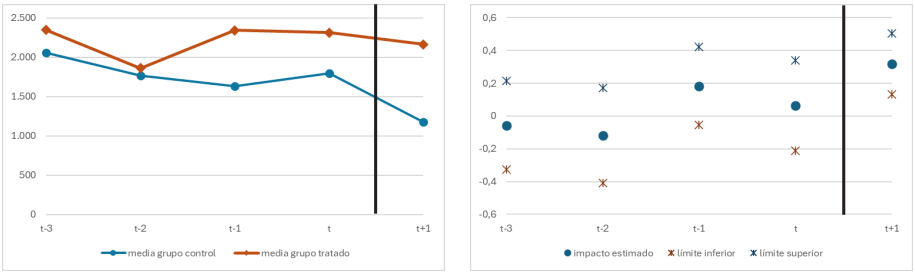
Figura 8. Efectos del tratamiento por segmento en la simulación del escenario 2

Segmento A



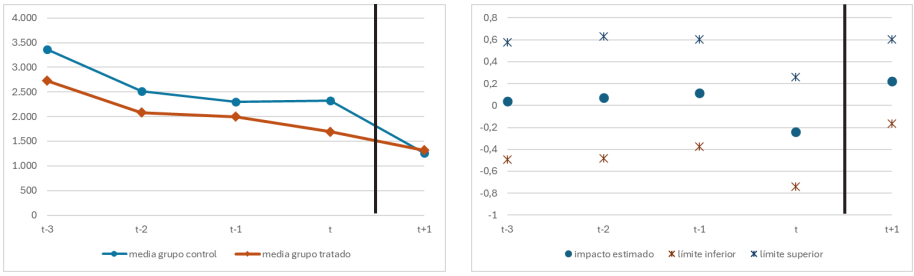
Variación en el volumen medio de pagos entre "t+1" y "t": grupo tratado, 0%; grupo de control, -18%

Segmento B



Variación en el volumen medio de pagos entre "t+1" y "t": grupo tratado, -6%; grupo de control, -34%

Segmento C



Variación en el volumen medio de pagos entre "t+1" y "t": grupo tratado, -22%; grupo de control, -46%

Fuente: elaboración propia con base en un ejercicio de simulación

En el segmento B, en cambio, la variación de los pagos en el grupo tratado es negativa (-6%) pero con una diferencia de 28 puntos porcentuales respectos del grupo de comparación (la variación en éste es -34%). La diferencia es suficientemente grande para detectar un efecto positivo del tratamiento, tal como lo muestra la gráfica de la derecha correspondiente a este segmento: el intervalo de confianza para el parámetro β está en su totalidad por encima del cero, indicando que es estadísticamente significativo.

Por último, en el segmento C las variaciones son considerablemente negativas en ambos grupos y, si bien el desempeño es mejor en el grupo tratado, la diferencia en medias, que es de 24 puntos porcentuales, no resulta suficiente para identificar un efecto positivo del tratamiento.

En suma, si bien es posible que existan efectos positivos del tratamiento en los tres segmentos, solamente en uno (el segmento B) éstos son detectables, en el sentido de que resultan estadísticamente significativos. Para estimar estos efectos en cada segmento es necesario configurar el modelo en cada período y en cada submuestra correspondiente a cada segmento. Por lo tanto, los tamaños muestrales son más pequeños, lo que exige que las diferencias sean más grandes para poder ser detectables como efectos del tratamiento.

En cualquier caso, la idea de presentar estos efectos diferenciados por segmento es ilustrar que la evaluación de diseños experimentales permite este tipo de análisis, es decir, identificar si existe algún efecto diferencial al considerar alguna característica de los individuos, en el ejemplo, la pertenencia a un segmento.

Este tipo de análisis puede contribuir a entender mejor los procesos por los cuales el tratamiento es o no eficaz, o también, puede contribuir a profundizar en el análisis de causas del incumplimiento. Por ejemplo, cabe preguntarse qué características asociadas a cada segmento pueden asociarse con estos resultados diferenciales.

5. CONCLUSIONES

En el marco de la GRC, la realización de evaluaciones de impacto de los tratamientos es clave para obtener elementos informativos que ayuden a mejorar el diseño de éstos, así como su focalización en los segmentos o perfiles de contribuyentes para los que resulta más eficaz. Asimismo, los resultados de las evaluaciones pueden ser de utilidad para arrojar luz sobre las causas del incumplimiento.

La práctica de la evaluación contribuye a aumentar el conocimiento de la administración tributaria orientado a asignar en forma más eficaz y eficiente los recursos, y a un trato más adecuado sobre los contribuyentes, haciéndolo proporcional al nivel de riesgo y a su disposición a cumplir.

La evaluación de diseños experimentales es reconocida como la metodología más robusta para este propósito, porque al basarse en asignaciones aleatorias asegura una estimación precisa del contrafactual (lo que sucedería en ausencia del tratamiento). En este artículo se ha procurado mostrar cómo la planificación de acciones de tratamiento que regularmente diseñan las administraciones tributarias constituye una oportunidad para la inclusión de estos diseños experimentales. De esta forma, además, se conecta en forma directa la operativa de la administración tributaria con su sistema de aprendizaje.

Por último, los requerimientos de estos diseños experimentales constituyen también una oportunidad para sistematizar procedimientos y así asegurar un estándar de calidad en las acciones de tratamiento, resultado que también contribuye a mejorar las interacciones con los contribuyentes.

6. REFERENCIAS

Bérgolo, M., Ceni, R., Cruces, G., Giaccobasso, M., Perez-Truglia, R. 2023. "Tax Audits as Scarecrows: Evidence from a Large-Scale Field Experiment," American Economic Journal: Economic Policy, vol 15(1), pages 110-153.

Bérgolo, M., Ceni, R., Cruces, G., Giaccobasso, M., Perez-Truglia, R. 2018. "Misperceptions about Tax Audits." AEA Papers and Proceedings, 108: 83-87.DOI: 10.1257/pandp.20181039

Brondolo, J., Chooi, A., Schloss, T., Siouclis, A. 2022. "Compliance Risk Management: Developing Compliance Improvement Plans". International Monetary Fund. Technicals Notes. https://www.imf.org/en/Publications/TNM/Issues/2022/03/18/Compliance-Risk-Management-Developing-Compliance-Improvement-Plans-515263

Centro Interamericano de Administraciones Tributarias (CIAT) – Servicios de Impuestos Internos de Chile (SII) – Fondo Monetario Internacional (FMI). 2020. "Manual sobre Gestión de Riesgos de Incumplimiento para Administraciones Tributarias". www.ciat.org

Duflo, E., Glennerster, R., Kremer, M. 2007. "Using Randomization in Development Economics Research: A Toolkit." Documento de discusión CEPR Núm. 6059. Londres: Center for Economic Policy Research. www.cepr.org - Discussion Paper Nº 6059.

Gertler, Paul J., Sebastián Martínez, Patrick Premand, Laura B. Rawlings y Christel M. J. Vermeersch. 2017. "La evaluación de impacto en la práctica, Segunda edición". Washington, DC: Banco Interamericano de Desarrollo y Banco Mundial. doi:10.1596/978-1-4648-0888-3.

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. 2017. "The behavioralist as tax collector: Using natural field experiments to enhance tax compliance". Journal of Public Economics, 148, 14-31,
14-31.
Mascagni, G., Nell, C., & Monkam, N. F. 2017. "One size does not fit all: a field experiment on the drivers of tax compliance and delivery methods in Rwanda. ICTD Working Paper, 58.

Organisation for Economic Co-operation and Development (OECD). 2014. "Measures of Tax Compliance Outcomes: A Practical Guide." Paris: OECD Publishing.https://www.oecd.org/ctp/administration/measures-of-tax-compliance-outcomes-9789264223233-en.htm.

Organisation for Economic Co-operation and Development (OECD). 2010. "Overview Note: Evaluating the effectiveness of compliance risk treatment strategies." Centre for Tax Policy and Administration. Paris: OECD Publishing. https://www.oecd.org/ctp/administration/46274397.pdf

Organisation for Economic Co-operation and Development (OECD). 2004. "Guidance Note: Compliance Risk Management; Managing and Improving Tax Compliance". Paris: OECD Publishing. https://www.oecd.org/tax/administration/33818656.pdf.

Pomeranz, D. 2015. "No taxation without information: Deterrence and self-enforcement in the value added tax." American Economic Review, 105(8), 2539-69.