PARTE III — CONSTRUIR PRODUCTOS DIGITALES · CAPÍTULO 09

La excelencia en ingeniería como decisión de producto

09

Cómo convertir las expectativas de una población en decisiones de ingeniería, experimentar con rigor y usar la evidencia para mejorar productos físicos, digitales y servicios.

3 nov 2026 · 23 min de lectura · Edición viva 0.9 · Estado: en progreso ·

Hay una parte de la ingeniería que me interesa especialmente: el ingenio que necesitamos para resolver una necesidad con los materiales, las capacidades y las restricciones que tenemos alrededor. Se vuelve visible cuando una silla se adapta al cuerpo, un beneficio bancario resulta comprensible o una aplicación permite terminar una tarea sin pedir ayuda. En todos esos casos, las decisiones técnicas terminan formando parte de lo que una persona siente al utilizar el producto.

A lo largo de mi carrera he encontrado una relación cada vez más estrecha entre esa capacidad de resolver y la calidad. Podemos disponer de buenos proveedores, especialistas y plataformas, pero todavía necesitamos saber qué esperamos conseguir y para quién. Una especificación puede cumplirse de manera impecable y entregar algo que el cliente no necesita. También podemos identificar una necesidad real y construir una solución incapaz de atenderla de forma consistente.

Por eso considero que la excelencia en ingeniería empieza en la definición del producto. Desde ahí decidimos a quién escuchar, qué medir, cuánto riesgo aceptar y qué evidencia nos permitirá afirmar que algo funciona. La estadística ayuda a sostener esas decisiones, siempre que recordemos que detrás de una media, una distribución o una tasa de abandono hay personas cuyas condiciones de uso pueden ser muy distintas.



FIGURA 9.1 Las expectativas se convierten en decisiones sobre mecanismos, materiales e interacciones.

Qué entiendo por calidad



Figura 9.2 — El usuario no conoce la especificación del tornillo; espera que la silla sea segura.

Mi forma de entender la calidad parte de cumplir con la expectativa del cliente dentro de una población determinada. Esa idea tiene antecedentes claros. Joseph M. Juran la relacionó con la adecuación al uso; Philip B. Crosby, con la conformidad con los requisitos. W. Edwards Deming insistió en comprender el sistema y su variación para mejorar de manera sostenida. No son definiciones idénticas, pero juntas permiten conectar la necesidad, la especificación y la consistencia de lo que entregamos.[1, 2]

Durante mucho tiempo he explicado mi concepto hablando de la expectativa del cliente en una media poblacional. La expresión necesita una precisión: una población no tiene una sola expectativa y su media no representa necesariamente a una persona real. Lo que buscamos es conocer la distribución de necesidades de quienes queremos atender, establecer compromisos y comprobar en qué medida los cumplimos. Mi definición es una síntesis personal de esas ideas, no una cita literal atribuible a uno de esos autores.

También conviene separar población, muestra y espacio muestral. La población es el conjunto sobre el que queremos aprender: por ejemplo, los titulares de una tarjeta que cumplen ciertas condiciones durante un periodo. La muestra es el subconjunto que observamos. En probabilidad, el espacio muestral es el conjunto de resultados posibles de un experimento; no es el nombre técnico de las personas entrevistadas. En investigación de clientes, además, necesitamos un marco muestral: la lista o mecanismo desde el cual podemos seleccionar participantes.

Con esa precisión, diría que calidad significa cumplir de manera consistente las expectativas y los requisitos relevantes de una población definida, con evidencia suficiente para conocer tanto el cumplimiento como sus límites. Esto incluye condiciones que el cliente quizá no sabe formular, como la seguridad de un mecanismo, la integridad de sus datos o la accesibilidad de una función. El usuario no tiene que conocer la especificación de un tornillo para esperar que la silla sea segura.

Si funciona unas veces y otras no, tendremos una promesa que depende de la suerte.

La consecuencia práctica es importante. Si el producto cumple una expectativa mal investigada, tendremos conformidad con una especificación equivocada. Si cumple únicamente para quienes se parecen al equipo que lo construyó, tendremos una cobertura limitada. Y si funciona unas veces y otras no, tendremos una promesa que depende de la suerte. Las tres situaciones requieren decisiones de producto, aunque se manifiesten como problemas de ingeniería.

El promedio y sus límites



Figura 9.3 — Una altura promedio no resuelve a la vez piernas cortas y piernas largas.

Pensemos en un ejemplo ilustrativo, ajeno a los resultados de mis proyectos. Cinco usuarios tardan 2, 3, 3, 4 y 18 minutos en completar un canje. La media es 6 minutos: sumamos los tiempos y los dividimos entre cinco. La mediana es 3: ordenamos los valores y tomamos el central. Ambas cifras son correctas, pero responden a preguntas diferentes.

La media ayuda a dimensionar el tiempo total consumido y puede servir para estimar capacidad operativa. La mediana describe mejor el centro de esta distribución, porque el caso de 18 minutos desplaza considerablemente el promedio. Sin embargo, escoger la mediana porque produce un resultado más agradable sería otra manera de ocultar el problema. Ese usuario también forma parte de la experiencia que entregamos.

Para decidir necesitamos observar la dispersión, las condiciones del caso extremo y, cuando el volumen lo permita, percentiles. Un percentil 95 de tiempo indica el valor que aproximadamente el 95 % de las observaciones no supera, según el método de cálculo utilizado. No promete que todos terminarán antes de ese límite. En una aplicación puede ayudar a ver esperas que la mediana disimula; en un servicio, a entender cuántos clientes enfrentan demoras particularmente largas.

La misma lógica aparece al diseñar una silla. Una altura promedio no resuelve simultáneamente la necesidad de alguien con piernas cortas y la de alguien con piernas largas. Tampoco una persona que está en el percentil 50 de estatura estará necesariamente en el mismo percentil de longitud de antebrazo o anchura de cadera. Por eso el diseño antropométrico requiere dimensiones pertinentes y combinaciones reales, además de mecanismos de ajuste.

Antes de elegir un indicador conviene reconocer qué decisión dependerá de él. El promedio puede orientar una estimación de costo; la mediana, describir una experiencia central; los extremos, revelar exclusiones o riesgos. Y una escala de satisfacción necesita cuidado adicional: pasar de dos a tres estrellas no tiene por qué representar el mismo cambio subjetivo que pasar de cuatro a cinco. Ningún resumen sustituye la lectura de la distribución y de las respuestas que lo originaron.

Figura 9.4



La media y la mediana resumen aspectos diferentes de las mismas cinco experiencias.

El beneficio que no se entendía



Figura 9.5 — Clasificar las respuestas ayuda a distinguir interés, comprensión y obstáculos de uso.

Cuando trabajaba para un banco participamos en la definición de un programa de lealtad con puntos Payback. Había una intención comercial razonable: ofrecer algo a cambio del uso de la tarjeta y mantener una relación con el cliente. Existían convenios y recompensas, pero encontrábamos una distancia entre el interés que despertaba recibir un beneficio y el uso efectivo de lo que ofrecíamos.

La impresión que teníamos era que casi nadie entendía cómo funcionaba el programa. A las personas les interesaba recibir algo por sus compras, pero pocas accedían a los convenios que teníamos disponibles. Esa observación no

era todavía una explicación. Podíamos estar frente a un problema de comunicación, de relevancia del premio, de esfuerzo para canjear o de condiciones difíciles de cumplir. Cada posibilidad habría requerido una intervención distinta.

Empezamos por determinar la población que queríamos conocer y seleccionar una muestra. Después vinieron las entrevistas, las encuestas, la lectura de las respuestas, su clasificación y el análisis de resultados. La estrategia del plan de lealtad se construyó con esa información y posteriormente establecimos grupos de control dentro de la población seleccionada para medir el desempeño. No conservo aquí una cifra verificable de mejora, así que no tendría sentido convertir ese recuerdo en un porcentaje de éxito.

Disponer de un beneficio no equivale a entregar una experiencia de valor.

Lo que sí puedo recuperar es el aprendizaje: disponer de un beneficio no equivale a entregar una experiencia de valor. Entre ambos hay una secuencia que el cliente debe poder recorrer. Necesita enterarse, comprender las condiciones, identificar una recompensa que le importe, alcanzar los requisitos y recibir lo prometido. Cada transición tiene componentes de diseño, operación e ingeniería.

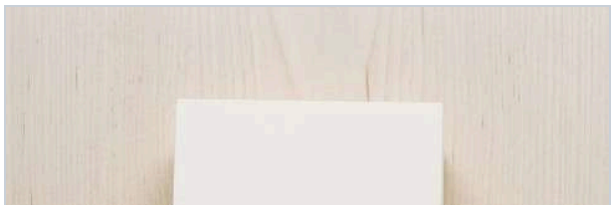
Si hoy tuviera que representar ese recorrido, distinguiría exposición al programa, comprensión, elegibilidad, intento de canje y recepción del beneficio. Son categorías propuestas para analizar el caso, no una reconstrucción de las mediciones exactas que hicimos. Nos ayudarían a evitar una conclusión apresurada: que un canje bajo significa necesariamente falta de interés. También puede indicar que el interés existe y el producto no consigue convertirlo en una acción posible.

Figura 9.6



Categorías propuestas para analizar el caso. No reconstruyen las mediciones del proyecto y no contienen datos.

Elegir a quién escuchar



Definir una población obliga a tomar decisiones incómodas. Si escuchamos únicamente a quienes ya canjean, aprenderemos mucho sobre usuarios

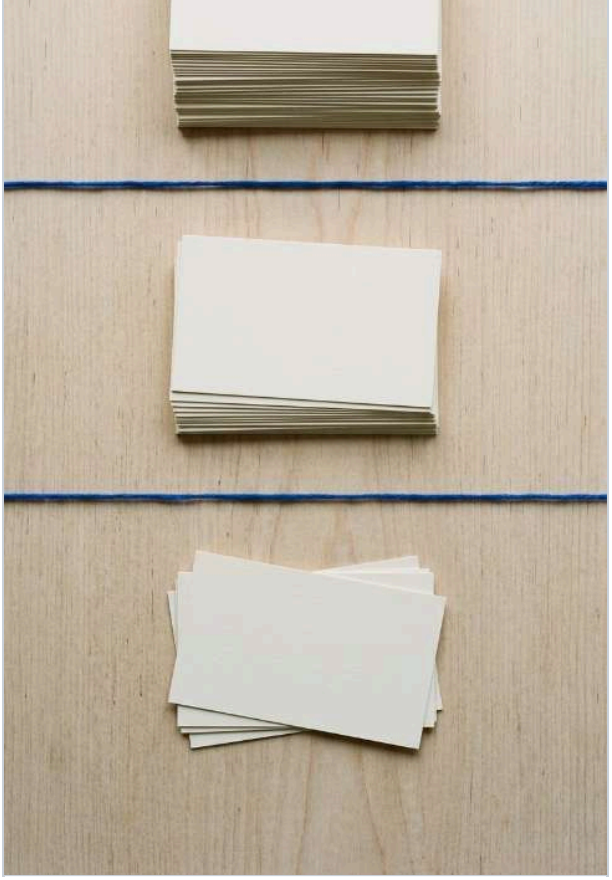


Figura 9.7 — Quién entra en la muestra decide lo que podremos aprender de ella.

que superaron las barreras y poco sobre quienes quedaron fuera. Si invitamos solo desde la aplicación, dejaremos de observar a quienes no entran en ella. Y si dependemos de respuestas voluntarias, es posible que participen con mayor frecuencia las personas más satisfechas o las más molestas.

En un programa como este podría ser útil distinguir clientes que conocen el beneficio y lo usan, quienes lo conocen pero no lo usan y quienes no lo conocen. También podrían importar la antigüedad, la frecuencia de uso, la ubicación o el canal habitual. Estas categorías deben responder a una hipótesis de producto, no a la facilidad de llenar una presentación con segmentos. Tener más grupos no garantiza entender mejor a las personas.

Las entrevistas y las encuestas aportan evidencias diferentes. Una entrevista puede descubrir una dificultad que no habíamos considerado; una encuesta bien diseñada puede ayudarnos a estimar su frecuencia. Para pasar de respuestas abiertas a datos necesitamos criterios de clasificación, conservar el texto original y revisar los casos ambiguos. Una misma respuesta puede contener varias causas, y forzarla a entrar en una sola casilla puede distorsionar lo que el cliente dijo.

También hay que resistir la tentación de presentar toda muestra como representativa. El tamaño necesario depende de la precisión buscada, la variabilidad, el diseño de selección y la comparación que queremos realizar. Para un experimento importa, entre otras cosas, el efecto mínimo que necesitamos detectar. Una muestra

grande obtenida de un canal sesgado puede describir con mucha precisión a la población equivocada.

En un trabajo exploratorio pequeño podemos encontrar problemas valiosos sin estimar su prevalencia en el mercado. Eso fue relevante en mi experiencia con mobiliario: el presupuesto limitaba el número de participantes, pero sus observaciones podían cambiar decisiones concretas. El límite estaba en lo que podíamos afirmar después. Descubrir una dificultad en varias personas justifica investigarla; no demuestra automáticamente qué porcentaje de todos los compradores la experimentará.

Una historia anterior al software



FIGURA 9.8 Hogares, procesos, parcelas y ensayos: cuatro tradiciones de una disciplina compartida. Bodegón ilustrativo, no documentos históricos.

Cuando pienso en aquel programa bancario, reconozco una continuidad con prácticas que otras industrias habían desarrollado mucho antes. Procter & Gamble sitúa en 1924 el comienzo de su investigación formal de consumidores, con recolección de información y diferenciación por características de los hogares. Es un antecedente útil porque muestra que observar al comprador, clasificar respuestas y trasladarlas al producto no nació con la analítica digital.[3]

La historia industrial tiene otra línea complementaria. En mayo de 1924, Walter A. Shewhart presentó el esquema de una gráfica de control; su trabajo posterior ayudó a distinguir la variación habitual de un proceso de las señales que requerían investigación. El NIST recoge también las aportaciones de Harold Dodge y Harry Romig al muestreo para inspección. Aquí la pregunta era cómo producir con consistencia y cómo evaluar lotes sin depender exclusivamente de revisar cada pieza terminada.[4]

En paralelo, Ronald Fisher trabajó en Rothamsted desde 1919 y contribuyó al desarrollo del diseño de experimentos a partir de problemas agrícolas. Comparar tratamientos exigía considerar que dos parcelas podían diferir por motivos ajenos al fertilizante. Esa necesidad de separar el efecto que buscamos de las condiciones que lo rodean sigue siendo reconocible cuando comparamos dos experiencias digitales.[5]

La investigación clínica aportó otra tradición decisiva. El ensayo del Medical Research Council sobre estreptomicina y tuberculosis, publicado en 1948, es un hito por su diseño y por las medidas para evitar conocer de antemano la asignación de tratamientos. No fue el primer uso de comparación ni puede resumir por sí solo el nacimiento de la experimentación. Tampoco sería exacto afirmar que las farmacéuticas inventaron estos métodos y después los transfirieron a todas las demás industrias.[6]

La continuidad llega hasta compañías como Nestlé, que describe el uso de investigación sensorial y retroalimentación del consumidor para mejorar productos, y Unilever, que evalúa dimensiones como producto, empaque, propuesta y precio. Son aplicaciones distintas de una disciplina compartida: observar, comparar y ajustar.[7] En medicina, además, la eficacia clínica y la seguridad no se reducen a satisfacción comercial. Podemos aprender de sus métodos sin equiparar por completo a un paciente con un comprador ni convertir un ensayo clínico en un embudo de ventas.

1919
RONALD FISHER EMPIEZA EN ROTHAMSTED

1924
SHEWHART ESBOZA LA GRÁFICA DE CONTROL

1948
SE PUBLICA EL ENSAYO DEL MRC SOBRE ESTREPTOMICINA

La silla como laboratorio



Tuve la oportunidad de colaborar con compañeros dedicados a medir antropometría y estudiar su impacto en el diseño de sillas de oficina. El propósito incluía mejorar la postura y

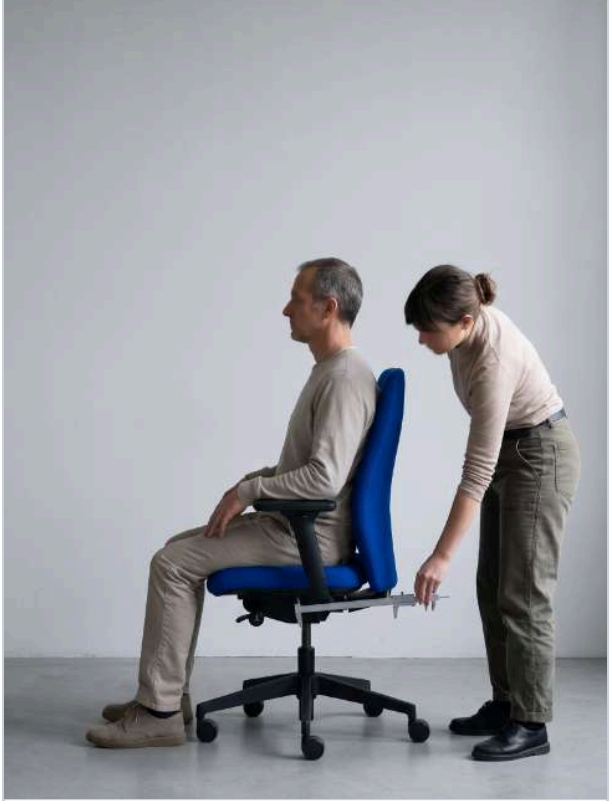


Figura 9.9 — Observar el cuerpo y escuchar al usuario orienta los ajustes que después deben poder fabricarse.

reducir condiciones que pudieran perjudicar las articulaciones. Al participar en ese proyecto encontré mucha experimentación, registro y escucha. Las respuestas de quienes probaban el producto se leían en relación con el mercado objetivo que buscábamos atender.

El presupuesto nos obligó a trabajar con una muestra reducida. Aun así, la información tuvo consecuencias importantes en la ingeniería: ayudó a definir la construcción, los sistemas de nivelación y los ajustes de postura. Después tuvimos que buscar proveedores capaces de responder al diseño y sus especificaciones técnicas. La investigación no terminaba en una recomendación; seguía hasta las condiciones necesarias para fabricar.

Para explicar el tipo de relación que aparece en esos proyectos, imaginemos que una persona no apoya bien los pies, que otra encuentra presión detrás de las rodillas y que una tercera no alcanza cómodamente el ajuste. Son ejemplos ilustrativos, no observaciones que esté atribuyendo al registro de aquel proyecto. Cada uno conduce a revisar variables diferentes: altura, profundidad, geometría o ubicación del mecanismo. La frase «no me resulta cómoda» necesita convertirse en preguntas que podamos investigar.

Además, comodidad percibida, postura observada y prevención de lesiones son resultados distintos. Una prueba breve puede ayudar a entender los dos primeros; demostrar reducción de lesiones requiere otro alcance, seguimiento y evidencia. La honestidad sobre ese límite forma parte de la calidad. No podemos

transformar un comentario favorable en una afirmación clínica sobre la protección de las articulaciones.

Lo que más me interesa de aquella experiencia es cómo la voz del usuario llegaba a una decisión material. Cambiar un rango de ajuste podía exigir otro mecanismo, alterar el costo, reducir el número de proveedores disponibles o modificar el mantenimiento. La ingeniería del entorno entraba así en la promesa del producto. Diseñar un ajuste excelente que no podemos fabricar de manera consistente sigue dejando la necesidad sin resolver.

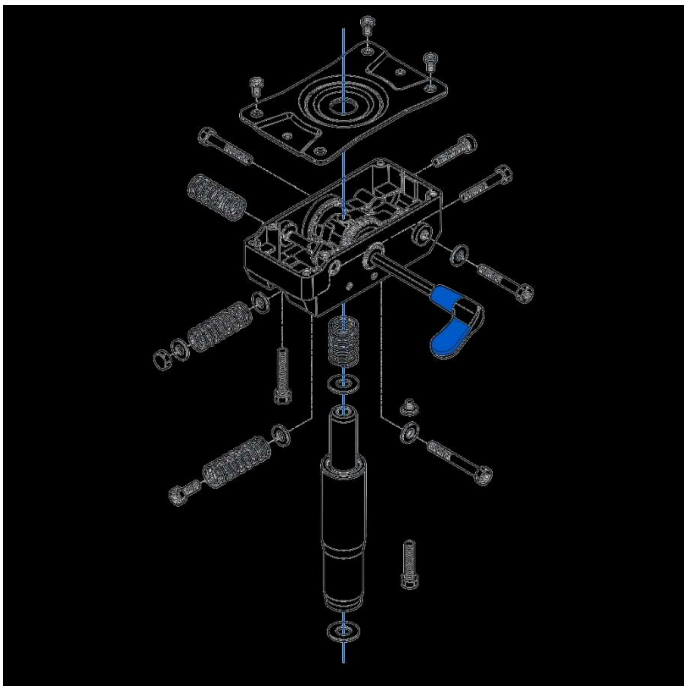


Figura 9.10 — Cambiar un rango de ajuste puede exigir otro mecanismo, otro costo y otro proveedor.
Vista conceptual, no el producto del proyecto.

Comparar para poder aprender



Figura 9.11 — Comparar exige definir antes qué mediremos, bajo qué condiciones y durante cuánto tiempo.

Si quisiéramos evaluar un nuevo mecanismo de silla, podríamos comparar un grupo que utiliza el modelo de referencia con otro que prueba la modificación. Tendríamos que definir antes qué mediremos, bajo qué condiciones y durante cuánto tiempo. La tarea, la duración de uso y las características de los participantes podrían influir tanto como el mecanismo. Una comparación útil necesita reconocer esos factores.

Cuando es viable, asignar al azar reduce el riesgo de que los grupos difieran sistemáticamente desde el inicio. Otra opción es que una persona pruebe ambas versiones, alternando el orden entre participantes. Ese diseño puede ser útil para ciertas medidas de uso, pero debe considerar aprendizaje, fatiga y efectos que persistan de una prueba a otra. No todo producto admite el mismo experimento.

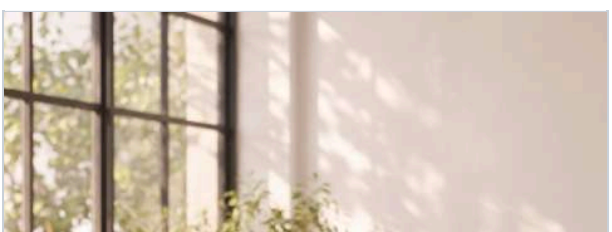
En el programa de lealtad, un grupo de control permite estimar qué habría ocurrido sin la intervención. Supongamos, solo como ejemplo, que una nueva explicación alcanza 12 % de canje y el control llega a 9 %. La diferencia es de tres puntos porcentuales, equivalente a un aumento relativo de aproximadamente 33.3 %. Esa aritmética no demuestra por sí misma un efecto: todavía necesitamos revisar el tamaño de los

grupos, la incertidumbre, la asignación y el periodo de observación.

Conviene fijar la métrica principal y las condiciones de decisión antes de mirar resultados. Si buscamos suficientes indicadores, siempre será más fácil encontrar alguno que parezca mejorar por azar. También debemos cuidar la unidad experimental: asignar por usuario evita ciertos cruces entre sesiones, pero quizá sea necesario asignar por hogar o sucursal cuando las personas comparten información y contaminan la comparación.

Un grupo de control tampoco es lo mismo que una gráfica de control. El primero sirve para comparar una intervención; la segunda ayuda a vigilar un proceso a lo largo del tiempo. Y un experimento aleatorizado en una muestra sesgada puede estimar bien el efecto dentro de esa muestra sin justificar una conclusión para todo el mercado. La calidad de la comparación y la posibilidad de generalizar son preguntas relacionadas, pero distintas.

Lo que el software puede probar



En producto digital encuentro una afinidad con esa disciplina cuando hablamos de *Test-Driven Development*, o TDD. Su ciclo consiste en escribir

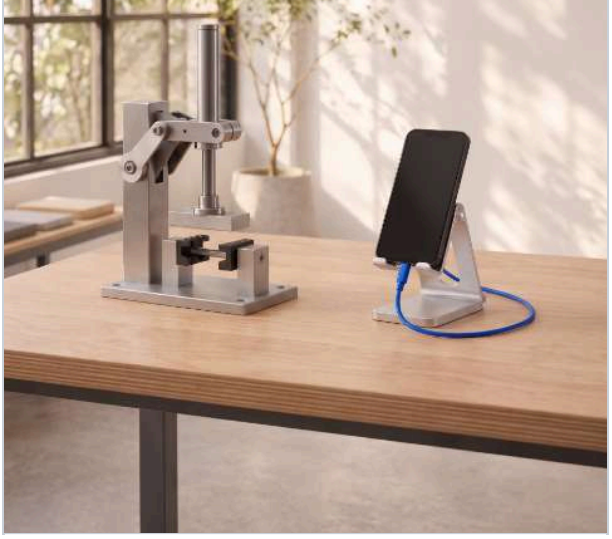


Figura 9.12 — Un banco de ensayo y una prueba automatizada verifican comportamientos definidos, no si el usuario los necesita.

una prueba automatizada que falla, implementar el código necesario para que pase y después mejorar la estructura del código conservando las pruebas satisfactorias. Es una práctica de desarrollo que obliga a explicitar comportamientos y proporciona retroalimentación rápida sobre ellos.[8]

La comparación con el trabajo físico resulta útil si respetamos sus límites. Una prueba de código puede verificar que un saldo insuficiente impida un canje o que un reintento no duplique una operación. No puede demostrar por sí sola que el cliente comprende los puntos, aprecia el beneficio o encuentra accesible el recorrido. TDD permite comprobar parte de la solución construida; investigar al usuario permite cuestionar si esa solución responde a la necesidad.

Por eso no describiría TDD, A/B, *soft launch* y *friends and family* como nombres intercambiables. Un A/B compara variantes mediante un diseño experimental; un lanzamiento limitado restringe la exposición inicial; una prueba con personas cercanas facilita observaciones tempranas, aunque introduce sesgos de selección y familiaridad. Podemos combinar estas prácticas, pero cada una produce evidencia diferente.

TABLA 9.1 — QUÉ PERMITE APRENDER CADA PRÁCTICA

PRÁCTICA	QUÉ PERMITE APRENDER	LÍMITE PRINCIPAL
TDD	Si el código cumple los comportamientos expresados en pruebas.	No valida por sí solo la necesidad ni la satisfacción.

PRÁCTICA	QUÉ PERMITE APRENDER	LÍMITE PRINCIPAL
Prueba de usabilidad	Cómo una persona entiende y realiza una tarea.	Un hallazgo cualitativo no estima automáticamente su frecuencia poblacional.
A/B con asignación adecuada	El efecto de una variante frente a otra en las condiciones estudiadas.	Necesita muestra, duración y métricas apropiadas.
Lanzamiento limitado	Cómo funciona una entrega con exposición acotada.	Limitar alcance no crea un control experimental.
Friends and family	Problemas tempranos observables con participantes cercanos.	La cercanía puede favorecer tolerancia, conocimiento previo y respuestas complacientes.

El vínculo valioso es que las observaciones de uso pueden alimentar requisitos y pruebas. Si aprendemos que el cliente teme perder puntos cuando se interrumpe la conexión, el equipo puede diseñar recuperación, confirmaciones y garantías de consistencia. Luego habrá que verificar el comportamiento técnico y volver a observar la experiencia. Una batería de pruebas completamente verde no cierra esa conversación; permite continuarla sobre una base más confiable.

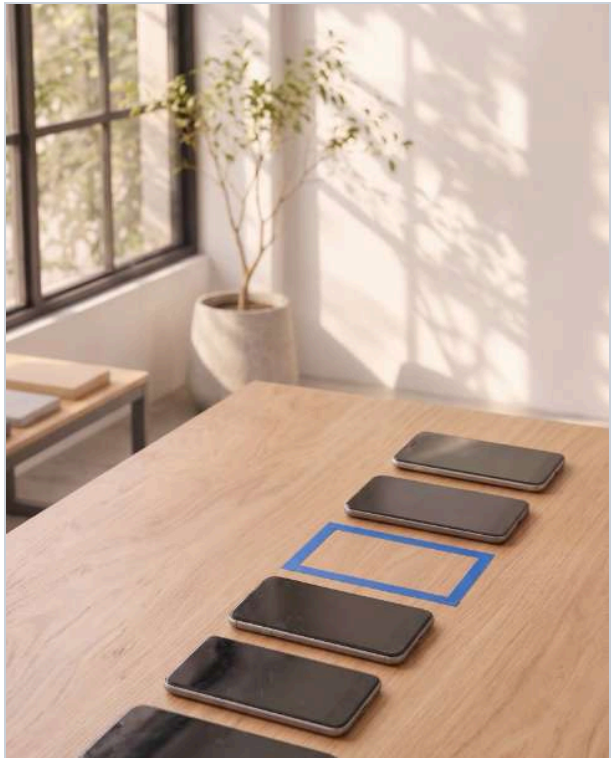
Figura 9.13



Las pruebas técnicas y la investigación de usuarios se complementan

La evidencia de uso modifica lo que pedimos al sistema; las pruebas verifican comportamientos definidos.

Quién quedó fuera de la muestra



La historia de la investigación médica muestra que disponer de recursos y rigor técnico no elimina los sesgos de selección. En Estados Unidos, una guía de la FDA de 1977 recomendó excluir a mujeres con posibilidad de embarazo de la fase I y de las primeras etapas de fase II, por preocupación ante riesgos reproductivos y del desarrollo fetal. La restricción se retiró en 1993. En 2026 hablamos de una política originada hace 49 años y revisada hace 33, no simplemente de que «hace cuarenta años solo se probaba en hombres».[9]

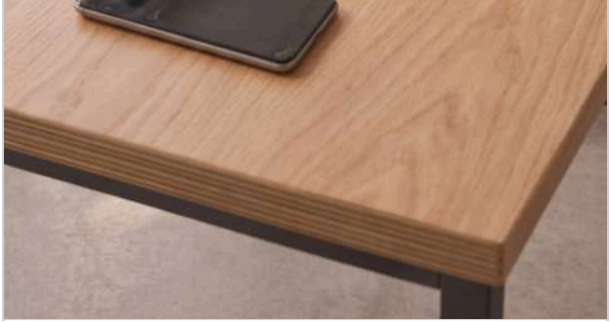


Figura 9.14 — Lo que no entra en la prueba tampoco entra en la conclusión.

La diferencia importa: no se excluía a todas las mujeres de todos los ensayos y de todas las fases. El problema incluía una representación insuficiente y la pérdida de oportunidades para conocer diferencias de respuesta desde etapas tempranas. En 1993, además, la NIH Revitalization Act incorporó requisitos de inclusión de mujeres y minorías en investigación clínica financiada por los NIH, con condiciones y excepciones definidas. Su alcance no debe confundirse con el de toda investigación realizada en cualquier país.[10]

También debemos cuidar la interpretación de las advertencias sobre embarazo. Algunas reflejan falta de datos suficientes; otras responden a riesgos identificados mediante evidencia humana, animal o farmacológica. No es correcto asumir que una advertencia existe únicamente porque el medicamento no se probó en embarazadas. La FDA plantea un etiquetado que distingue resumen de riesgos, consideraciones clínicas y datos disponibles precisamente porque esas situaciones no son equivalentes.[11]

La lección para producto digital es reconocer a quiénes excluye nuestro mecanismo de observación. Podemos probar en teléfonos recientes, redes estables y participantes familiarizados con tecnología, y después presentar el resultado como una propiedad universal. Podemos escuchar solo a clientes activos y omitir a quienes abandonaron durante el registro. Incluso podemos confundir una ausencia de quejas con satisfacción cuando el canal para quejarse resulta inaccesible.

Etiquetar a las personas no equivale a comprenderlas.

El problema también alcanza a las categorías. Sexo, género, edad, discapacidad y pertenencia étnica no son variables intercambiables ni explicaciones automáticas. En un estudio debemos justificar qué información necesitamos, cómo se obtiene y qué hipótesis permite evaluar. Etiquetar a las personas no equivale a comprenderlas. Una segmentación puede revelar una barrera, pero también puede consolidar prejuicios si interpretamos la etiqueta como la causa.

Medir el recorrido completo



Hoy podemos observar mapas de calor, botones utilizados, funcionalidades adoptadas y tableros consultados. Esa capacidad es valiosa, aunque genera una tentación: medir lo que la herramienta entrega con facilidad y después buscarle un significado. Un botón muy pulsado puede ser útil, confuso o incapaz de responder al primer intento. Una página poco visitada puede



Figura 9.15 — El clic vive en la interfaz; el cumplimiento llega después.

ser irrelevante o resolver una necesidad excepcional de enorme importancia.

Para evaluar calidad prefiero conectar las señales con una tarea y su resultado. En un proceso de canje distinguiría visualizar el beneficio, iniciar la solicitud, completarla en el sistema y recibir efectivamente la recompensa. El clic vive en la interfaz; el cumplimiento puede depender de un servicio externo y de una operación posterior. Si solo observamos el primero, podemos celebrar una mejora que el cliente nunca recibió.

La ingeniería de medición requiere acordar nombres de eventos, identificadores, marcas de tiempo y reglas para evitar duplicados. También necesita diferenciar intención de confirmación. Un evento emitido al presionar «canjear» no debe llamarse «canje completado» si todavía falta la respuesta del sistema. Las fallas de instrumentación pueden producir conclusiones tan equivocadas como las fallas del producto.

En una herramienta empresarial sucede algo parecido. Que un tablero sea el más visto no demuestra que mejore una decisión. Podemos explorar si reduce el tiempo necesario para resolver una tarea, si sus datos son confiables y si evita trabajo posterior. En un servicio, una atención más corta podría indicar eficiencia o una conversación interrumpida antes de resolver el problema. La medición necesita seguir al cliente hasta un resultado significativo.

Finalmente, una asociación entre acciones no basta para establecer causalidad. Los clientes más comprometidos podrían usar más una función y renovar más, sin que la función

explique por sí sola la renovación. Ese hallazgo justifica una hipótesis y quizá un experimento. Saltarse ese paso convierte una correlación prometedora en una prioridad de inversión que todavía no hemos demostrado.

Plataformas con trabajos diferentes



Figura 9.16 — Cada plataforma tiene su trabajo; ninguna resuelve automáticamente el de las demás.

En las conversaciones sobre datos de clientes suelen mezclarse CRM, CDP, analítica de producto y gestión de ingresos. Conviene ordenar sus papeles porque una plataforma no resuelve automáticamente el trabajo de las demás. Aquí uso RGM en el sentido habitual de *Revenue Growth Management*: una disciplina para gestionar crecimiento mediante precio, promociones, mezcla y presentación de la oferta, que puede apoyarse en distintas soluciones tecnológicas.[12]

Un CRM organiza la relación comercial y de servicio. HubSpot reúne capacidades de gestión de contactos, oportunidades, marketing y atención; su utilidad para producto está en relacionar promesas comerciales, fricciones y continuidad de la relación. Treasure Data es una CDP orientada a integrar datos y construir perfiles y segmentos. Mixpanel trabaja con eventos para analizar recorridos, conversiones y retención. Las capacidades disponibles

dependen de la configuración y contratación de cada producto.[13, 14, 15]

TABLA 9.2 — PLATAFORMAS Y CAPACIDADES CON TRABAJOS DIFERENTES

CAPACIDAD O PLATAFORMA	DATOS QUE PUEDE REUNIR O ANALIZAR	INDICADORES POSIBLES	APLICACIÓN A LA MEJORA
RGM	Precios, unidades, descuentos, canales y mezcla de productos.	Margen, ingreso por unidad, respuesta a promociones y elasticidad estimada.	Revisar si la oferta es atractiva y económicamente sostenible.
CRM / HubSpot	Contactos, oportunidades, interacciones, tickets y encuestas configuradas.	Conversión comercial, tiempos de atención, recurrencia de problemas y satisfacción reportada.	Conectar expectativas de compra con dificultades posteriores.
CDP / Treasure Data	Perfiles e información integrada de canales y transacciones.	Cobertura de identidad, actividad y conversión por segmento, según los datos disponibles.	Unificar observaciones y activar experiencias pertinentes para cada grupo.
Analítica / Mixpanel	Eventos, propiedades, usuarios y secuencias instrumentadas.	Conversión por etapa, retención, frecuencia y adopción de funciones.	Localizar fricciones y seguir la evolución del comportamiento.

Estos indicadores no aparecen necesariamente listos para usar ni tienen una definición universal. La retención depende del evento que consideramos valioso y del intervalo elegido. El margen requiere costos confiables; la elasticidad exige un análisis que separe otros cambios; la satisfacción de quienes responden puede diferir de la de quienes guardan silencio. Comprar la plataforma no sustituye ese trabajo.

Aplicado al ejemplo de lealtad, el CRM podría mostrar consultas repetidas sobre puntos; la CDP, ayudar a vincular esas consultas con perfiles y transacciones; la analítica, localizar una caída antes del canje; y RGM, evaluar el costo de modificar el beneficio. Juntas pueden sostener una investigación. Ninguna autoriza por sí sola a concluir cuál es la causa de la caída, y no siempre necesitamos desplegarlas todas para tomar una primera decisión útil.

La IA dentro de la investigación

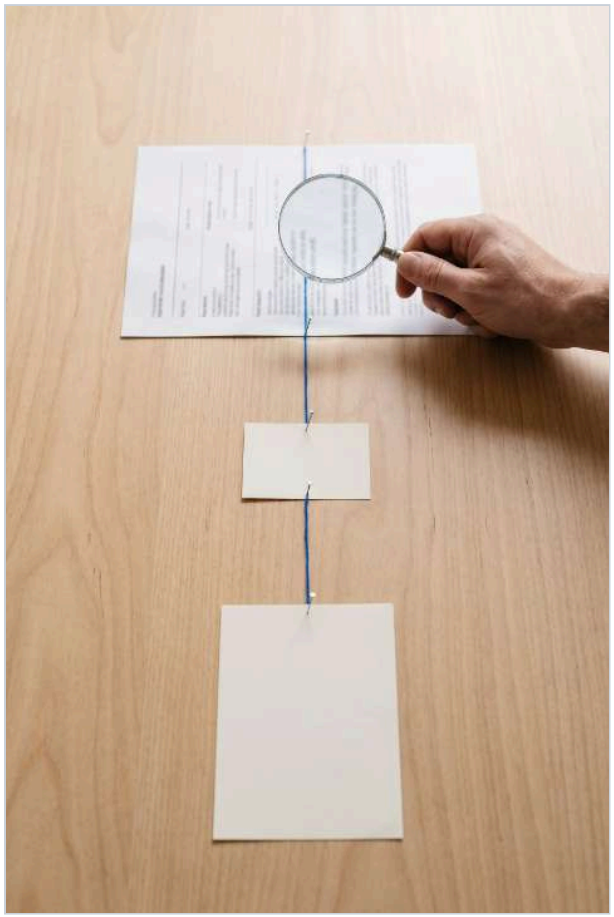


Figura 9.17 — Conservar el hilo entre respuesta, categoría y conclusión permite comprobar lo que hizo el modelo.

La inteligencia artificial puede ayudar a clasificar respuestas abiertas, proponer temas de entrevista, resumir incidencias o explorar agrupaciones de comportamiento. En una investigación con cientos de comentarios, una aplicación razonable sería usarla para sugerir categorías y después revisar una muestra de las clasificaciones contra los textos originales. Así podemos aprovechar velocidad sin perder la posibilidad de comprobar qué ocurrió con cada respuesta.

El riesgo aparece cuando confundimos fluidez con fidelidad. Un resumen puede borrar contradicciones, una clasificación puede fallar más en cierto vocabulario y un segmento puede reflejar principalmente las variables que decidimos capturar. Si una población no participó en el estudio, generar respuestas sintéticas con IA no la incorpora a la evidencia. Puede servir para explorar hipótesis, pero no reemplaza observar a esas personas.

Por eso conservaría la trazabilidad entre respuesta, categoría y conclusión; revisaría errores por grupo y documentaría cambios en el modelo o en las instrucciones de clasificación. También limitaría los datos personales a los necesarios para la pregunta. La capacidad de unir fuentes no justifica hacerlo indiscriminadamente. La investigación debe

permitir aprender sin convertir cada interacción del usuario en información disponible para cualquier propósito.

La oportunidad más interesante está en acortar el recorrido entre observación y decisión. Un patrón detectado puede convertirse en una hipótesis, esa hipótesis en un experimento y su resultado en una modificación. La mejora continua ocurre cuando completamos ese recorrido y volvemos a medir. Clasificar más información sin cambiar nada produce archivos más ordenados, pero no necesariamente un producto mejor.

Decidir cómo se sostiene la promesa



FIGURA 9.18 La mejora ocurre cuando la información modifica una decisión y volvemos a comprobar su efecto.

Todo lo anterior termina en elecciones concretas de ingeniería. En la silla, la evidencia puede exigir un mecanismo con otro rango, tolerancias diferentes o un proveedor capaz de sostener una especificación. En el programa de lealtad puede llevar a cambiar reglas, simplificar explicaciones, confirmar transacciones o revisar la integración con quien entrega el beneficio. En ambos casos estamos decidiendo cómo construir una promesa que pueda cumplirse.

El entorno importa porque no diseñamos en el vacío. Tenemos determinadas competencias, procesos, presupuestos y dependencias. Una solución que el equipo no puede operar introduce fragilidad; un proveedor sin capacidad de reposición compromete el mantenimiento; una integración sin confirmaciones confiables dificulta saber si el cliente recibió lo suyo. Estas condiciones necesitan estar presentes desde la definición del producto y en la planeación del proyecto.

Yo organizaría ese trabajo en cuatro movimientos: definir la población y la expectativa; traducirlas en requisitos observables; comparar alternativas con un método apropiado; y revisar el resultado durante la operación. El responsable de producto sostiene la relación entre necesidad y decisión, los especialistas diseñan y verifican la solución, y la gestión del proyecto asegura que investigación, pruebas y correcciones tengan tiempo y recursos. Si solo presupuestamos la construcción, estamos suponiendo que ya sabemos todo lo necesario.

Cuatro movimientos

Definir, traducir, comparar y revisar: del entendimiento a una promesa que se pueda cumplir.

01

Definir

La población y la expectativa.

02

Traducir

En requisitos observables.

03

Comparar

Alternativas con un método apropiado.

04

Revisar

El resultado durante la operación.



Figura 9.19 — Cuando algo está bien hecho, se siente.

La decisión de ampliar una entrega debería considerar la mejora buscada y los límites que no aceptamos deteriorar. Un canje más frecuente puede ser valioso, pero hay que revisar fallas, reclamaciones, costos y capacidad de entrega. Una silla puede mejorar su ajuste y volverse más difícil de reparar. La calidad necesita una lectura conjunta porque el cliente recibe el producto completo, no la métrica que cada área escogió para demostrar su avance.

Al final regreso a lo que me interesa de la ingeniería: su capacidad de convertir comprensión en una solución que funciona. Escuchar a una persona, ubicar su experiencia dentro de una población y reconocer los límites de nuestra evidencia cambia lo que construimos. La excelencia aparece cuando esas decisiones permiten cumplir de forma consistente lo que

prometimos y seguir aprendiendo de quienes usan el producto. Cuando algo está bien hecho, se siente; nuestro trabajo es entender qué lo hace posible y conseguir que esa experiencia se repita.

REFERENCIAS Y NOTAS

Las experiencias del banco y del proyecto de sillas proceden del relato del autor. Los tiempos, porcentajes y diseños experimentales propuestos son ejemplos didácticos, no resultados históricos de esos proyectos. Las referencias externas se consultaron el 3 de noviembre de 2026. La figura 9.8 es un bodegón ilustrativo, no una reproducción de documentos históricos.

1. American Society for Quality, Quality Glossary, entrada «Quality»: atribuciones de adecuación al uso a Juran y conformidad con requisitos a Crosby. Véase también Joseph M. Juran. La definición desarrollada en el capítulo es una síntesis del autor. ↩
2. The W. Edwards Deming Institute, Understanding Variation and Appreciating Variety. Contexto para variación, sistemas y aprendizaje mediante iteración. ↩
3. Procter & Gamble, P&G History, entrada de 1924. Se utiliza para documentar la historia de investigación de consumidores de la compañía, sin adoptar su afirmación de primacía sobre toda la investigación de mercados. ↩
4. NIST/SEMATECH, How did Statistical Quality Control Begin?. Shewhart, memorando del 16 de mayo de 1924, y contribuciones de Dodge y Romig al muestreo de inspección. ↩
5. Rothamsted Research, Surprising birthplace of modern data analysis celebrates 100 years. Inicio del departamento de estadística en 1919 y desarrollo de métodos vinculados a experimentos agrícolas. ↩
6. Medical Research Council (1948), registro del informe original y materiales históricos en The James Lind Library. Se cita como hito metodológico, no como primer experimento controlado de la historia. ↩
7. Nestlé, Nutrition and pleasure; Unilever, Why invest in Unilever. Descripciones corporativas de investigación sensorial, retroalimentación y evaluación del producto; no son evaluaciones independientes de eficacia. ↩
8. Agile Alliance, Test-Driven Development; Government Digital Service, Test-driven development — The GDS Way. Ciclo de prueba fallida, implementación y refactorización. ↩
9. FDA, Guideline for the Study and Evaluation of Gender Differences in the Clinical Evaluation of Drugs, Federal Register, 22 de julio de 1993; FDA, Evaluation of Sex Differences in Clinical Investigations. Documentan el alcance de la guía de 1977 y su revisión de 1993. ↩

10. NIH Office of Research on Women's Health, History of Women's Participation in Clinical Research; NIH, Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research. Alcance en investigación financiada por los NIH. ↩

11. FDA, Pregnancy and Lactation Labeling Final Rule. Regla publicada en 2014 con entrada en vigor en 2015; estructura de información sobre riesgos, consideraciones clínicas y evidencia disponible. La discusión histórica no constituye orientación para usar un medicamento. ↩

12. Kantar, Revenue Growth Management. RGM se utiliza en este capítulo como disciplina de gestión de ingresos y no como nombre de una plataforma específica. ↩

13. HubSpot, Products, Features, and Benefits. Capacidades generales de CRM, marketing, ventas y servicio. La tabla propone usos analíticos, no garantiza disponibilidad en todos los planes. ↩

14. Treasure Data, Complete CDP y Segments Overview. Integración de datos, perfiles y segmentación. ↩

15. Mixpanel, Product analytics capabilities. Referencia a informes de eventos, embudos, retención y recorridos; no se hacen recomendaciones de compra ni afirmaciones de precio. ↩

← CAPÍTULO ANTERIOR

08 — Qué se rompe al escalar un producto

SIGUIENTE CAPÍTULO →

10 — Agentes, automatización y la nueva línea de producción — en desarrollo

© 2026 J. Angel Sosa. *Pensar en Productos, Construir Sistemas*.

Privacidad · Contacto