La formación en pruebas de software. Enseñanza del concepto de sesgo de confirmación desde la perspectiva del razonamiento pragmático.

Verónica D'Angelo.

La formación en pruebas de software. Enseñanza del concepto de sesgo de confirmación desde la perspectiva del razonamiento pragmático

Verónica D'Angelo

La siguiente es una versión traducida al castellano de un manuscrito publicado.
Cita del artículo original:

D'Angelo, V. (2024). Software testing education. Teaching the concept of confirmation bias from a pragmatic perspective of reasoning / La formación en pruebas de software. Enseñanza del concepto de sesgo de confirmación desde la perspectiva del razonamiento pragmático. *Studies in Psychology*, *45*(2-3), 297-337. https://doi.org/10.1177/02109395241274656

Resumen

La Tarea de Selección de Wason ha sido utilizada con frecuencia por investigadores en ciencias computacionales para destacar las supuestas deficiencias de los probadores de software desde una perspectiva logicista falsacionista, influyendo en los criterios y procesos de selección de personal en esta área. Sin embargo, un enfoque pragmático del razonamiento sugiere que la cognición humana está influenciada no solo por la lógica, sino también por factores contextuales. Esta perspectiva indica que las deficiencias de los probadores podrían surgir de priorizar la relevancia de la información sobre una estricta adherencia lógica. Abordar el sesgo de confirmación y las influencias contextuales es fundamental para una comprensión integral del problema y para desarrollar nuevas técnicas de prueba de software así como nuevas estrategias educativas. El Experimento 1 examinó dos versiones temáticas de la Tarea de Wason entre estudiantes de Informática, confirmando un rendimiento mejorado en línea, de acuerdo con la teoría de la relevancia. El Experimento 2 mostró la habilidad de los estudiantes para comprender conceptos relacionados con el falsacionismo y aplicarlos a tareas abstractas. Los resultados sugieren reutilizar el test de Wason, tradicionalmente empleado en la didáctica de la psicología, dentro de las ciencias computacionales, valorizando la diversidad temática y contextual como herramienta didáctica para enseñar lógica y metacognición de sesgos cognitivos, en lugar de limitarse únicamente a evaluar habilidades lógicas formales.

Palabras clave
Software testing; diversidad temática; formación en pruebas de software; psicología del razonamiento; tarea de selección de Wason; sesgo de confirmación; teoría de la relevancia

Introducción

La prueba de software no es una mera tarea técnica en el campo de la tecnología informática (TI) sino una actividad intelectualmente desafiante que depende en gran medida del comportamiento humano. No sólo requiere razonamiento hipotético deductivo, sino disposición psicológica, creatividad y experiencia práctica. Sin embargo, tanto las estrategias educativas diseñadas para formar a probadores de software como las investigaciones sobre su desempeño han adoptado mayoritariamente una perspectiva falsacionista, arraigada en un marco teórico que postula la razón como una capacidad general independiente del contexto.

Desde esta perspectiva, los errores realizados por los probadores de software suelen atribuirse a un fenómeno denominado sesgo de confirmación, que supuestamente influye en el proceso de adquisición de la información, llevándoles a favorecer aquella información que coincide con sus expectativas y a descartar la información que las contradice. En consecuencia, se han desarrollado métricas de sesgo para diferenciar entre razonadores buenos (con bajo nivel de sesgo) y malos (con alto nivel de sesgo). Por el contrario, entre las teorías pragmáticas del razonamiento, y en especial en la Teoría de la Relevancia de Sperber y Wilson, la razón se concibe como sensible al contexto, y el razonador prioriza la información en función de sus implicaciones prácticas y del esfuerzo cognitivo necesario para su procesamiento. Desde esta perspectiva, la cognición humana es notablemente eficiente y adaptable y los denominados sesgos pueden considerarse heurísticas adaptativas, es decir, un comportamiento esperado en condiciones atípicas.

A continuación describimos la emergencia de estos dos modelos cognitivos de la razón humana, el modelo intelectualista y el interaccionista, que luego serán puestos a prueba en la sección experimental. Describiremos también la relación de ambos modelos con el sesgo de confirmación según la interpretación de la investigación sobre las pruebas de software en el área informática.

Falsacionismo frente a verificacionismo

En primer lugar, debemos aclarar la importante diferencia entre *cálculo proposicional* y *razonamiento proposicional*. A pesar de que algunas perspectivas teóricas los han confundido, no son equivalentes.

El cálculo proposicional es un sistema formal de cálculo, compuesto por símbolos, reglas de formación y reglas de inferencia, cuya validez depende exclusivamente de su estructura formal y no del significado psicológico que pueda atribuirse a sus expresiones. Está basado en la estructura e inferencia de proposiciones utilizando operadores lógicos como AND, OR y NOT, y es muy utilizado en teoría y práctica de la computación.

En cambio, el *razonamiento proposicional*, es una actividad cognitiva que consiste en evaluar, interpretar y extraer conclusiones basadas en proposiciones, que pueden estar expresadas en lenguaje formal o natural, ya que las personas pueden utilizar el lenguaje natural y su capacidad de comprensión para inferir relaciones lógicas entre proposiciones y tomar decisiones informadas. El razonamiento proposicional es un componente fundamental del pensamiento humano y es referido en múltiples disciplinas, como en filosofía, en psicología y en la toma cotidiana de decisiones.

El cálculo proposicional "se parece" al razonamiento proposicional en diversos aspectos, pero las normas que lo rigen no son las mismas y no deberían confundirse con las que rigen el pensamiento humano. Sin embargo, la historia de la psicología muestra que algunos enfoques teóricos han caído en el equívoco de considerarlos equivalentes y las consecuencias de esta confusión se han trasladado tanto a algunas teorías psicológicas específicas como a sus proyecciones educacionales; en particular, en la enseñanza de las ciencias formales y especialmente en la enseñanza de las ciencias computacionales.

La presunción errónea de que las leyes de la lógica en la filosofía de la ciencia reflejan las leyes del razonamiento humano se remonta a las aportaciones de George Boole en su obra *Análisis matemático de la lógica* (*The Mathematical Analysis of Logic*, Boole, 1847). Boole sentó las bases para sus afirmaciones psicológicas en *Una investigación sobre las leyes del pensamiento* (*An Investigation on the Laws of Thought*, Boole, 1854). No solo resolvió el método para traducir la lógica en un cálculo práctico y eficiente, sino que agregó algo más: asumió que este cálculo tan conveniente y eficiente se basaba en los principios del pensamiento humano.

Dentro del cálculo proposicional, el uso del condicional y sus derivados para probar hipótesis ha sido de especial interés en la relación entre lógica y psicología. Dos conceptos desempeñan un papel clave en el estudio del razonamiento de las pruebas: el concepto de 'confirmación', tomado del positivismo lógico (Carnap, 1936, 1937), y el concepto de 'falsificación', tomado del racionalismo crítico de Popper (1959, 1962).

Karl Popper criticó firmemente el principio de verificación, alegando que conducía al estancamiento y al dogma en la investigación científica, puesto que los científicos tendían a buscar evidencias que confirmaran sus teorías preexistentes en lugar de tratar de refutarlas. A cambio, Popper propuso el principio del falsacionismo, que sostiene que una teoría científica solo puede considerarse científica si puede ser refutada mediante evidencia empírica. Esto significa que una teoría debe estar abierta al cuestionamiento y a ser potencialmente refutada por nuevos datos o evidencia, en lugar de simplemente buscar una confirmación continua de las creencias existentes. Mientras que el principio de verificación se consideró una estrategia conservadora que podría llevar a una confirmación acrítica de las creencias existentes, el principio de falsacionismo se presentó como una estrategia aperturista, ya que fomenta una actitud crítica y abierta hacia las teorías (Poletiek, 2001).

La teoría de Popper se convirtió en prescriptiva para la investigación psicológica y epistemológica relevante. Piaget dedicó una atención considerable en toda su obra a la relación entre lógica y psicología. Desde una perspectiva popperiana, Piaget equiparó el razonamiento humano al cálculo proposicional (Inhelder & Piaget, 1958) y afirmó que, al alcanzar la fase operativa formal, los adolescentes tienen la capacidad de razonamiento deductivo a través de reglas (Inhelder & Piaget, 1955).

La capacidad de encontrar contraejemplos (ejemplos que contradicen la propia hipótesis) se consideró una capacidad humana natural, adquirida durante el desarrollo del pensamiento formal. Según Piaget e Inhelder, frente a una situación causal compleja, el adolescente indagaría si el hecho *x* implica un hecho *y*, formulando una proposición de tipo '*p* implica *q*', y buscaría un ejemplo contrario para probarla; un caso en el cual *x* no implique *y* (Beth & Piaget, 1974). Por ejemplo, si la proposición es 'todas las tarjetas que son rojas de un lado, son verdes por el otro', la persona buscaría un caso en que una tarjeta roja no fuese verde del otro lado.

En esa misma década, Wason (1960) adoptó la perspectiva falsacionista de Popper para llevar a cabo el primer experimento psicológico sobre la comprobación de hipótesis, tratando de evaluar el comportamiento lógico formal de un grupo de adolescentes en una situación específica. Pero el resultado fue un alto porcentaje de errores en las soluciones de los adolescentes, que contradecía los postulados de Inhelder y Piaget (1955). Wason atribuyó estos supuestos errores a lo que denominó

sesgo de confirmación. A continuación explicamos en mayor detalle este experimento y el fenómeno del sesgo de confirmación.

Experimentos de Wason

Wason (1960, 1968) propuso dos experimentos dirigidos a investigar qué hacen las personas cuando tratan de comprobar sus propias hipótesis. Se preguntó si, en esas situaciones, las personas actúan como científicos según el principio de falsacionismo de Popper, es decir, buscando contraejemplos en lugar de buscar casos que confirmen su hipótesis. Sin embargo, en ambos experimentos, observó lo contrario: en la gran mayoría de los casos, las personas tendían a utilizar la estrategia del test positivo como heurística por defecto (Klayman & Ha, 1987). Sobre su primer experimento, con el que aparentemente logró demostrar que las personas adoptan un comportamiento de falsacionismo, la argumentación de Wason se consideró defectuosa puesto que no logró probar si los participantes buscaban ejemplos contrarios a la hipótesis o si el hallazgo de ejemplos contrarios fue una consecuencia involuntaria, como el propio autor reconoció (Wason, 1962). Dicho de otro modo, y volviendo a la distinción previa, debemos diferenciar entre el falsacionismo como producto y el falsacionismo como proceso psicológico. El primero es el resultado de una operación formal necesaria en las ciencias, porque permite descartar una hipótesis cuando existen casos que no cumplen la regla establecida. Sin embargo, eso no implica que el falsacionismo también exista como proceso psicológico. Todavía no hay consenso entre las teorías psicológicas en este campo. Hasta la fecha, los psicólogos han observado que las personas experimentan el falsacionismo como una estrategia imposible de ejecutar, como si no fuesen capaces plantear una hipótesis y cuestionarla al mismo tiempo (Poletiek, 2011). Cuando se obtienen resultados que refutan la hipótesis, estos parecen ser una consecuencia de la calidad de la hipótesis y no de un determinado comportamiento de comprobación (Poletiek, 1996).

Tanto el segundo experimento de Wason (1966), la Tarea de Selección, como un extenso número de investigaciones apoyan la opinión de que el razonamiento humano (al menos en la comprobación de hipótesis) no se guía por la lógica proposicional sino por un conjunto de reglas sensibles al contexto. Ese segundo experimento consistió en completar una tarea relacionada con el cumplimiento de una regla tipo $P \rightarrow Q$, que el participante tenía que verificar a partir de los datos presentados en cuatro tarjetas de las que solo era visible un lado. Cada tarjeta representa un ejemplo en el que la regla puede cumplirse o no. En su versión original, el contenido de las tarjetas eran letras y números (Figura 1) y las instrucciones explicaban que en una cara de la tarjeta había una letra y en la otra cara, un número, de forma que se cumplía esta regla:

> Si en una cara hay una E, entonces en la otra cara hay un 2. ¿Qué tarjetas deberías voltear para comprobar si la regla es verdadera para todas las tarjetas?

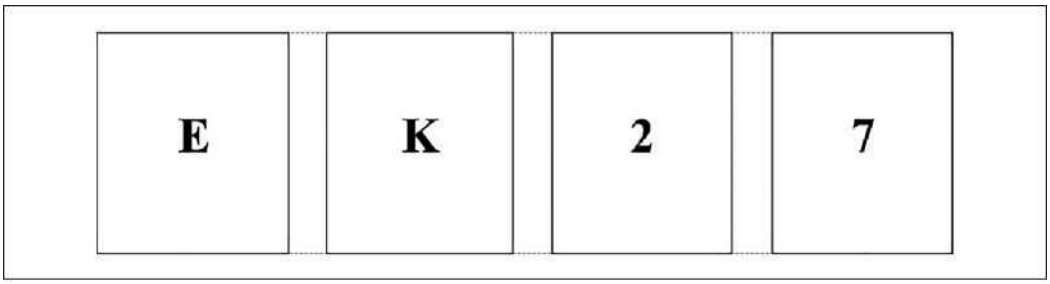

Figura 1. Tarjetas en la Tarea de Selección de Wason (1966).

Si las personas tendiesen de forma natural al pensamiento lógico normativo (falsacionista), los participantes tratarían de refutar la regla. La respuesta correcta sería E (*P*) y 7 (*no-Q*). Porque la tarjeta con una E podría tener un número distinto de 2 en la otra cara, y la tarjeta con un 7 podría tener una E en el anverso. En ambos casos, la regla quedaría refutada. Por otro lado, lo que podría aparecer en el reverso de la tarjeta con una *K* (¿qué pasa con *no-P*?) no importa, porque la regla no lo contempla, como tampoco importa el contenido detrás del 2. Si es P, confirmar únicamente un caso de la norma no es suficiente, puesto que solo se verifica completamente la norma cuando se comprueba que no existen casos que la refuten, no cuando existen casos que la verifican.

La mayoría de los participantes seleccionaron las tarjetas P y Q, o solo la tarjeta P. Solo en torno a un 10% de ellos indicaron la selección correcta, P y no-Q.

A principios de la década de 1970, dos variantes de esta tarea lograron obtener una mayoría de selecciones correctas según el pensamiento normativo: Evans (1972) introdujo una simple variación en la tarea estándar de selección, 'si P, entonces no Q'; es decir, la negación de la consecuencia de la condición. Otra variante de éxito fue transformar la regla en una regla deóntica, a diferencia de la tarea original, que era descriptiva. Por otro lado, Johnson-Laird et al. (1972) demostraron que cuando la regla es deóntica, es decir, cuando expresa un deber u obligación, la mayoría de los participantes

seleccionan correctamente P y *no-Q*. Inicialmente, esto se interpretó como evidencia de que un contenido realista y concreto, cercano a la experiencia de las personas, facilitaba el razonamiento. La versión deóntica de la tarea de selección de Griggs y Cox (1982) proponía a los participantes el siguiente escenario:

> Imagina que eres un policía de servicio. Tu función es garantizar que las personas cumplan ciertas normas. Las tarjetas que tienes delante contienen información sobre cuatro personas que están sentadas alrededor de una mesa. En un lado de la tarjeta se muestra la edad de la persona y en el otro lado se indica qué están bebiendo. Existe una regla: Si la persona bebe cerveza, debe ser mayor de 19 años. Selecciona la tarjeta o tarjetas necesarias para determinar si todos cumplen o no la regla (Figura 2).

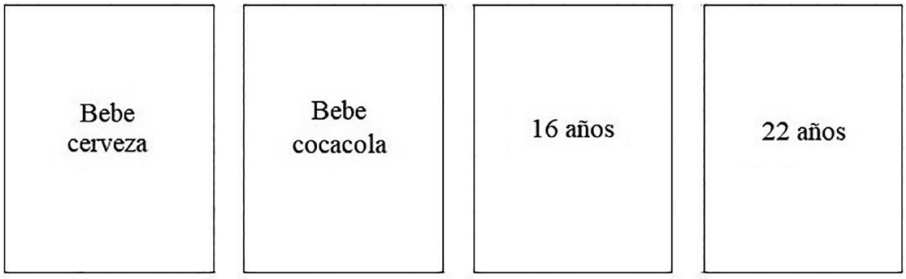

Figura 2. Tarjetas en la tarea de selección adaptada por Griggs y Cox (1982).

Sin embargo, otras pruebas con versiones familiares y concretas, pero no deónticas, no dieron los mismos resultados (Manktelow & Evans, 1979; Wason & Shapiro, 1971), mientras que otras versiones abstractas y deónticas, sí (Cheng & Holvoak, 1985, 1989; Cosmides, 1989; Girotto et al., 1989).

Cheng y Holyoak (1985, 1989) dedujeron que la relación de permiso puede representarse de forma abstracta como un esquema de tipo *'Si se realiza una acción A, debe cumplirse una condición previa P,* con lo que se consigue mejor rendimiento. Se demostró que el beneficio de las versiones temáticas no puede considerarse universal y que las versiones deónticas lógicamente difieren de otras versiones descriptivas de la tarea (Manktelow & Over, 1990, 1991; Mosconi & D'Urso, 1974; Yachanin & Tweney, 1982).

En versiones descriptivas, los participantes buscan evidencia para determinar la verdad o falsedad de una regla, mientras que en las versiones deónticas, los participantes buscan evidencia de una transgresión. Platt y Griggs (1993) demostraron un incremento sustancial en el número de respuestas correctas cuando se indicaba a los participantes que buscasen transgresiones de la regla en lugar de pedirles que revisasen la regla. Estos resultados fueron interpretados en el marco de la teoría de modelos mentales de Johnson-Laird y Byrne (1991) y del modelo de razonamiento de doble proceso de Evans (1993). El porcentaje de respuestas correctas en la versión abstracta de la tarea de Wason suele ser inferior a 10%, mientras que otras versiones temáticas serían más fáciles de resolver (Evans, 1982, 2013; Wason & Jolmson- Laird, 1972) y la facilidad se relaciona directamente con la relevancia (Sperber et al., 1995).

En la Teoría de la Relevancia, Liberman y Klar (1996), Love y Kessler (1995) y Sperber y Wilson (1986) coinciden en sus respectivas interpretaciones de la tarea de Wason, enfatizando los efectos positivos del contenido cuando se orienta a los participantes a descubrir '*P y no-Q*' (transgresión de la regla). Este foco puede lograrse poniendo de relieve la transgresión de la regla y demostrando la alta probabilidad de ocurrencias de P y *no-Q*. Sperber et al. (1995) propusieron que se logra facilitar la búsqueda de información para refutar la regla cuando esta información es *relevante* para el participante. A diferencia del concepto de relevancia de Evans (1989), la Teoría de la Relevancia (Sperber & Wilson, 1986, 2004), como explicamos a continuación, define la relevancia como una propiedad de cualquier input que interacciona con la información previa del participante, produciendo resultados inherentemente interesantes y efectos importantes a un costo razonable.

La teoría de la relevancia y la tarea de Wason

Según esta teoría, la eficiencia de la cognición humana, como otras formas de eficiencia, implica un equilibrio óptimo entre el esfuerzo mental ejercido y los efectos cognitivos resultantes. En términos evolutivos, la clave está en alcanzar ciertos objetivos cognitivos, como determinar si algo es comestible o identificar amenazas potenciales, como predadores. Esto significa asignar recursos de atención a aquellos inputs perceptuales y mnemónicos capaces de suscitar esfuerzos cognitivos significativos. Los efectos cognitivos incluyen diversas mejoras en la adquisición de conocimiento, como la obtención de más información, garantizando la veracidad y organización de los datos, y una acción orientadora. Esencialmente, la cognición actúa como una inversión en adquisición de conocimiento sobre el mundo exterior, que posteriormente informa a las acciones o a otros procesos cognitivos. El

foco está en la selección de inputs, que pueden incluir cualquier forma de representación derivada de la percepción o la memoria.

En su artículo 'Modularity and Relevance' (*Modularidad y relevancia*), Sperber (2005, p. 61) utiliza la analogía del comportamiento de una rana para esbozar la comparación entre la complejidad de la cognición humana y otros procesos más elementales. La rana espera a que aparezca una mosca en su campo de visión. Si esto sucede, se activa el proceso de cazar la mosca, dirigido por estímulos con operaciones asociadas casi obligatorias. Pero esa obligatoriedad depende de una lista de prioridades. El módulo de cazar no es el único en funcionamiento. Simultáneamente, también sigue funcionando el módulo para escapar de los predadores, una acción de supervivencia con mayor prioridad que la de comer. Si ambos estímulos (la mosca y el predador) apareciesen simultáneamente, el funcionamiento del módulo de caza de la mosca sería sustituido por el del módulo de huida del predador, que tiene una clara prioridad adaptativa sobre otros módulos. Sus acciones son totalmente obligatorias y su orden de preferencia está regulado por el sistema nervioso de la rana. Esta jerarquía de procesos con el sistema nervioso como el controlador supremo es viable porque la rana reconoce pocos estímulos y tiene procesos simples a modo de respuesta, que raramente compiten (es muy poco probable que se le presenten simultáneamente una presa potencial, un predador potencial y una pareja potencial). Por el contrario, para los humanos, la situación es muy distinta. La mente humana es enormemente modular y requiere una gran cantidad de información.

En todo momento está abierta a innumerables inputs potenciales. Cada módulo tiene inputs disponibles y debe competir por la capacidad mental para procesarlos. Para identificar un input como tal, este debe ser más relevante que el resto. La decisión puede variar en función del contexto. Lo que podría ser un estímulo para alimentarse (la presencia de comida) en un contexto podría no serlo en otro (en el caso de una enfermedad digestiva o de seguir una dieta). Por tanto, un sistema de control general que asigne prioridades fijas como el sistema nervioso de la rana (ante la presencia de comida, activar el proceso de alimentación) no sería adaptativo porque la decisión de qué input procesar y qué acciones activar variará en función de las circunstancias. Además, una vez tomada la decisión sobre qué input procesar, el sistema debe acceder a la información a largo plazo, compararla con el input y realizar inferencias. Estas actividades mentales requieren esfuerzo, lo que supone un costo que solo debe asumirse con la expectativa de obtener un beneficio. Distintos inputs activarían distintas líneas de pensamiento que podrían implicar una asignación u otra de esfuerzo y produciría beneficios cognitivos muy distintos. Es necesario un proceso de asignación de energía flexible y sensible al contexto para contribuir a la eficacia del sistema cognitivo. La eficacia cognitiva es una cuestión de lograr invertir esfuerzo en el procesamiento de los inputs correctos, es decir, los más relevantes entre todos los inputs posibles en un momento dado.

Sperber y Wilson (1986) definen la relevancia como una propiedad característica de los inputs que la mente puede utilizar en un momento dado para seleccionarlos. Ese grado de relevancia será mayor 'cuanto mayores sean los efectos cognitivos conseguidos procesando un input determinado' y 'cuanto menor sea el esfuerzo de procesamiento necesario para conseguir esos efectos' (Sperber & Wilson, 2008). Asimismo, y esto es clave en el asunto que nos ocupa, en igualdad de condiciones, cuanto mayor sea el costo de procesamiento de un input, menor es su relevancia.

Tres ejemplos permiten ilustrar este concepto. El primero lo da Sperber (2005): si un doctor le dice a un paciente que está enfermo, esta información tiene cierto grado de relevancia. Pero si el doctor informa al paciente de que 'tiene una gripe', esa información es más relevante, no porque aporta más información sino porque tiene un mayor número de efectos cognitivos (dura en torno a una semana, requiere unos días de licencia en el trabajo, medicación, tratamiento, etc.). Ahora bien, si esa información se presenta de una forma que exija un alto costo de procesamiento, por ejemplo, si el doctor le dice que tiene una enfermedad cuyo nombre se deletrea con la séptima, decimoctava, novena, decimosexta, y quinta letra del abecedario (gripe) aunque esa información sea cierta y produjera, una vez procesada, el mismo número de efectos cognitivos (consecuencias), el esfuerzo de procesamiento es mucho mayor. Es decir, la información sobre la enfermedad sería menos relevante de esa manera que si el doctor hubiera dicho simplemente tiene gripe.

Como segundo ejemplo, suponga que un profesor indica a un estudiante que dibuje una figura a partir de un segmento de recta arbitrario, tal que en uno de sus extremos se erija una recta perpendicular, sobre la cual se marque un segundo segmento congruente al primero, repitiéndolo en un tercer segmento, cuyo extremo, en unión con el extremo libre del segmento inicial determine el cierre. Si el profesor le dijera al estudiante que dibuje un cuadrado, esta frase sería más relevante por el menor costo cognitivo, aunque tuviera menos información que la consigna geométrica.

Tomemos la prueba de Wason como tercer ejemplo. La versión abstracta original de la tarea, como se indica en la Figura 1, difiere en términos de relevancia de la tarea temática de Griggs y Cox (Figura 2). En este último caso, los participantes asumen el rol de búsqueda de transgresiones de la regla, de modo que la información sobre posibles transgresiones ('beber cerveza', 'tener menos de 16 años') es más relevante porque tiene mayores consecuencias cognitivas para un razonador que debe

hacer cumplir la ley. En este escenario, no se afirma que el razonador actúe como falsacionista de forma natural en el sentido que Piaget y Popper proponían, sino que su objetivo cambia hacia la búsqueda de transgresiones en un sentido de verificación. En la versión abstracta, sin embargo, esas mismas tarjetas tienen un costo cognitivo que es demasiado alto debido a su arbitrariedad, análogo a las letras de la palabra gripe, o a la consigna geométrica del cuadrado. Aparentemente, las tarjetas no guardan relación alguna con la regla excepto la que se menciona en la regla. Es de escasa relevancia, no debido a sus consecuencias sino debido al esfuerzo cognitivo necesario. El sistema cognitivo no parece querer invertir energía en este tipo de problemas.

La comprobación de hipótesis ha evolucionado porque la identificación de errores es esencial para la supervivencia de los razonadores. Por tanto, su éxito depende de que los participantes adopten el rol de buscadores de infracciones, como agentes de seguridad, revisores, científicos y probadores de software. De una forma similar a la obligación que un guardia de seguridad tiene de identificar infracciones de las normas, los probadores de software tratan de identificar errores en programas de terceras partes. Sin embargo, no es natural que los científicos apoyen una hipótesis y simultáneamente busquen sus flaquezas, igual que a los programadores les resulta complicado diseñar algoritmos al mismo tiempo que buscan errores en un ejercicio de resolución de problemas.

Si el contexto induce al razonador a adoptar el papel de buscador de errores en la tarea, seleccionarán 'P' y 'no-Q' con facilidad. No obstante, para que se detecten inputs falsacionistas, esta información (la evidencia de P y *no-Q*) debe ser igual o más fácil de representar que su alternativa P y Q, algo que suele suceder en situaciones coherentes con las experiencias previas del participante o sus roles como buscador de infracciones, como cuando se les pide que imaginen que son policías. Por tanto, es posible elaborar versiones simplificadas de la tarea de selección manipulando la relevancia para mejorar el rendimiento del participante, incorporando las funciones básicas P y Q de modo que P y no-Q sean más fáciles de representar que la alternativa P y Q. Estos escenarios también se discuten en la teoría de los modelos mentales como predictores de un falsacionismo preciso.

Por lo general, las personas no practican un pensamiento falsacionista a menos que asuman el papel de buscadores de infracciones, como en las profesiones vinculadas con la medicina, las ciencias o la seguridad. Sin embargo, incluso en estos roles, el acto de refutación suele implicar la búsqueda activa de estas infracciones. Por ejemplo, un guarda de seguridad busca activamente desviaciones de las normas estándar, de igual manera que un probador de software busca activamente errores en un programa desarrollado por otros. Estas acciones ocurren en contextos determinados, en los que la emergencia de posibles errores ocultos da lugar a respuestas diversas, ya sea un cambio en la perspectiva del programador o un cambio en el contexto global. Esta variabilidad de respuesta no es indicativa de la eliminación de sesgos, sino que refleja la relevancia cambiante de estos errores en la tarea que desarrolla el razonador.

El sesgo de confirmación en las pruebas de software

El uso de condicionales y derivados en el planteamiento y la comprobación de hipótesis constituye un tema de gran interés en el campo de la informática. De la misma forma que las personas tienden a evaluar más positivamente aquellos casos que más posiblemente confirmen sus creencias, en lugar de aquellos que podrían refutarlas — un fenómeno conocido como 'sesgo de confirmación' — en los profesionales de la informática se ha observado un comportamiento paralelo, tanto en programadores como en probadores de software.

Los científicos informáticos, en general, se caracterizan por un sólido razonamiento lógico y habilidades analíticas avanzadas. Existen semejanzas y diferencias notables entre las tareas realizadas por programadores y probadores. La principal tarea de un programador es la codificación, que implica escribir programas utilizando lenguajes artificiales para lograr funcionalidades concretas. En segundo lugar, la depuración de programas implica revisar y mejorar el código para optimizar su calidad o identificar y corregir errores para garantizar el funcionamiento del programa. Por otro lado, para un probador de software, la ejecución de pruebas es primordial. Los probadores deben diseñar pruebas que abarquen todos los aspectos del programa, garantizando que se verifiquen todas las funciones y vías de ejecución. Estas pruebas sistemáticas tienen por objeto verificar la funcionalidad, usabilidad, seguridad y rendimiento del programa. Además, los probadores son los responsables de detectar, grabar e informar de cualquier error o fallos identificados durante las pruebas.

Los programadores tienden a desarrollar un vínculo personal con el código que escriben y depuran, lo que puede dar lugar a un sesgo de confirmación, puesto que pueden estar más predispuestos a probar aquellos casos que confirman sus creencias iniciales sobre la corrección del programa. Esto hace que tiendan a centrarse más en resolver problemas específicos identificados durante el desarrollo, lo que podría influir en la selección de casos de prueba alineados con estas áreas de interés. Por tanto, en el desarrollo de software, la recomendación de no limitar la corrección del código al propio autor, sino de someterlo a revisión por otros programadores, se ha consolidado como una práctica estándar dentro de los procesos de aseguramiento de calidad. La revisión por pares

tiende a revelar un mayor número de errores que la autocorrección.

A diferencia de los programadores, los probadores de software, dado que no tienen una conexión personal con el código, podrían estar más abiertos a considerar una amplia variedad de casos de prueba, inclusive aquellos que podrían contradecir creencias existentes sobre la funcionalidad del programa. Su principal objetivo es identificar cualquier defecto o comportamiento inesperado, de modo que tienden a explorar escenarios que cuestionan las expectativas iniciales sobre el programa.

El sesgo en el comportamiento de prueba, como describen estudios como el de Katz y Anderson (1989) para los programadores y como se ha observado en los probadores de software en los trabajos de Leventhal et al. (1993), refleja una tendencia a realizar pruebas seleccionando aquellos casos que se anticipa verificar positivamente, lo que se conoce como 'sesgo positivo de las pruebas'. Este enfoque podría limitar la eficacia de las pruebas al no cuestionar adecuadamente el programa para descubrir posibles errores.

Myers (1979), en su libro clásico sobre las pruebas de software, fue el primer autor en el campo de la informática que subrayó el papel relevante de la psicología en las pruebas de software. Advirtió que las personas están muy orientadas a las metas: cuando asumen un rol, no son capaces de asumir de forma simultánea otro rol que esté en conflicto con el primero. Myers señaló que las pruebas de software constituyen el 50% del costo del proyecto y dedicó un capítulo a la psicología de las pruebas, destacando, entre otros principios clave, que una prueba satisfactoria es la que encuentra un error que todavía no había sido descubierto. Este concepto de las pruebas representó un cambio radical en la perspectiva de algunos desarrolladores de programas, para quienes una prueba satisfactoria era aquella en la que no se identificaba ningún error (Pressman, 2010). La expectativa de 'no encontrar errores' es un rasgo típico de la actitud de los programadores hacia su propia producción. Esta dificultad de los programadores para actuar como probadores de sus propios programas se ha atribuido a un 'sesgo de confirmación', considerado el error más significativo de inferencia lógica y, presumiblemente, un factor negativo en el rendimiento.

Durante décadas, los investigadores de las pruebas de software han mejorado las estrategias de las pruebas, han desarrollado programas de pruebas y han popularizado términos como 'romper el programa' o 'ataques' para referirse a los procesos de depuración del software, que son verdaderos 'ataques' en puntos del código considerados más débiles (Andrews & Whittaker, 2006; Whittaker, 2002).

En las primeras investigaciones empíricas sobre las métricas del sesgo de confirmación (Leventhal et al., 1993, 1994; Teasley et al., 1994), los probadores de software asumían la tarea de introducir valores para realizar pruebas, que podrían pertenecer bien a clases de equivalencia positiva (confirmatoria) o negativa (falsacio- nista).

Los investigadores observaron que, cuando las especificaciones del programa contenían una combinación equilibrada de información confirmatoria y falsacionista, el sesgo de confirmación tendía a disminuir. Esto se evidenció por un mayor ratio de valores falsacionistas relacionados con valores confirmatorios. Sorprendentemente, los participantes en estos estudios eran totalmente conscientes de su función como probadores de software. Estos resultados, anteriores a la teoría de la relevancia (Girotto et al., 1989; Sperber et al., 1995), son coherentes con resultados posteriores en el marco de esa teoría.

No obstante, los investigadores contemporáneos en el área informática han formulado métricas de sesgo, derivadas de la Tarea de Selección de Wason, con el objetivo de evaluar la capacidad de razonamiento de las personas. Los análisis de los datos indicaron un rendimiento superior de los estudiantes frente a los profesionales, lo que implicaría, para ellos, la necesidad de iniciativas más robustas de formación en lógica en los ámbitos profesionales para fortalecer sus habilidades de razonamiento falsacionista (Calikli & Bener, 2010, 2015; Çalıklı & Bener, 2013). Este enfoque logicista presupone la existencia de un razonamiento independiente del contexto y propone la Tarea de Selección de Wason como un método para diferenciar entre razonadores 'buenos' y 'malos'.

Hasta la fecha, no se han relevado estudios sobre el sesgo de confirmación, en el área informática, que hayan incorporado específicamente la perspectiva de las teorías pragmáticas del razonamiento a las que aquí haremos referencia.

Este enfoque sesgado de las teorías psicológicas lleva a los probadores profesionales a asumir responsabilidad por los niveles de productividad de sus empresas, sin que se defina claramente cuuál es el papel de la educación en este contexto.

Las pruebas de software ocupan un lugar secundario en el currículum de Informática, desde 1992 y hasta la actualidad. De acuerdo con la creciente complejidad de los sistemas informáticos y la creciente demanda de ingenieros de pruebas altamente cualificados, muchos educadores han explorado maneras de mejorar las estrategias pedagógicas en las pruebas de software aunque, hasta la

fecha, ninguno se ha dedicado a explorar con los estudiantes la cuestión del sesgo de confirmación como una temática psicológica y meta cognitiva.

Además, pese a que estudios recientes señalan la importancia de conceptualizar y utilizar teorías en la investigación pedagógica de las ciencias computacionales (S. Fincher & Petre, 2004; S. A. Fincher & Robins, 2019; Hannay et al., 2007; Nelson & Ko, 2018; Shull et al., 2007), una revisión sistemática de la literatura revela que solo 3.9% de los estudios sobre las pruebas de software se basa en teorías del aprendizaje debidamente fundamentadas (Garousi et al., 2020).

En resumen, primero cuestionamos la sostenibilidad de la investigación sobre el sesgo en el campo de la informática, que ha tendido a adoptar una perspectiva logicista limitada al no incorporar las teorías pragmáticas del razonamiento. Señalamos que el uso del Test de Wason como herramienta para medir el razonamiento puede ser cuestionable, del mismo modo que su eficacia ha sido cuestionada en la literatura científica. En segundo lugar, reconocemos la importancia de abordar el sesgo de confirmación de forma exhaustiva e imparcial en las ciencias computacionales, teniendo en cuenta todas las perspectivas psicológicas, incluidas aquellas más allá del enfoque falsacionista tradicional. En tercer lugar, subrayamos la necesidad de que los estudiantes comprendan el concepto de sesgo utilizando métodos pedagógicos fundamentados, basados en teorías psicológicas.

Dos modelos cognitivos del razonamiento lógico

A partir de lo expuesto en las secciones precedentes, podemos esbozar dos modelos cognitivos del razonamiento lógico. El campo de la psicología del razonamiento ha experimentado un extenso proceso iniciado por el experimento de Wason (1966, 1968), que puso de relieve una sorprendente 'incapacidad' para refutar condicionales y sentó las bases de dos perspectivas predominantes que tratan de explicar este fenómeno asumiendo dos hipótesis y modelos distintos del razonamiento humano. Por un lado, el enfoque tradicional logicista afirma que la razón es una habilidad cognitiva de uso general. Bajo esta perspectiva, los errores y sesgos del razonamiento se atribuyen a un sistema de *procesamiento dual* en el que un sistema intuitivo e irreflexivo, susceptible al sesgo, coexiste con un sistema analítico y racional (Evans & Over, 1996; Kahneman, 2003; Stanovich, 1999). Por otro lado, la hipótesis *interaccionista*, como propusieron Mercier y Sperber en 2017, basada en la teoría de la relevancia de Sperber y Wilson (1986), concibe la razón como uno de varios módulos inferenciales. Bajo esta perspectiva, el razonamiento humano no opera de forma aislada, sino que interactúa con otros procesos cognitivos y módulos mentales para generar inferencias y tomar decisiones. Este enfoque reconoce la complejidad de la mente humana y sugiere que el razonamiento no es más que una pare de un sistema más amplio de procesamiento de la información. Esta perspectiva sugiere que la intuición y la razón están entrelazadas en un proceso unificado, en el que inicialmente un único sistema genera intuiciones y después las racionaliza. Bajo esta perspectiva, el razonamiento no emerge como un mecanismo de resolución de problemas en busca de una verdad absoluta sino como una competencia social estrechamente vinculada a la argumentación en contextos sociales.

Desde esta segunda perspectiva, la cual adoptamos, si el rendimiento en la Tarea de Selección de Wason puede depender del grado de relevancia de cada input, es posible manipular la relevancia para mejorar el rendimiento en la tarea. Por lo tanto, nos postulamos en contra de las métricas de sesgo (Calikli & Bener, 2010, 2015) según las cuales, la ausencia de activación de procesos mentales falsacionistas no es un indicador válido del bajo nivel de rendimiento de los probadores de software sino que, más bien, podría deberse a que estos procesos no se dan en las personas. Asimismo, el éxito en la tarea y la producción de resultados falsacionistas no indican necesariamente la activación de un proceso mental de falsacionismo, aunque puede influir, como defiende la teoría de la relevancia (Sperber et al., 1995).

Desde esta perspectiva, resulta especialmente interesante demostrar que los estudiantes de informática son sensibles al contexto durante su razonamiento, y son capaces de resolver una tarea satisfactoriamente con una manipulación adecuada de los inputs (tarjetas). En segundo lugar, merece la pena investigar si la tarea de Wason podría utilizarse más eficazmente para incrementar el conocimiento de los estudiantes sobre el proceso de falsacionismo que sienta las bases para la comprensión del sesgo de confirmación, facilitando así un enfoque más constructivo y meta cognitivo de las prubas de software.

Método

Una estrategia pedagógica comúnmente utilizada con estudiantes de grado de psicología lleva a introducir el concepto de sesgo de confirmación. Inicialmente, se les presenta la versión abstracta de la tarea de Wason y los estudiantes tienen que resolverla. Después, se les facilita una versión temática, como el escenario de la edad y las bebidas alcohólicas, y de nuevo los estudiantes tienen que resolverlo (Figura 2). Por lo general, los estudiantes tienen menos dificultades con la versión temática, lo que suscita la exploración introspectiva de sus procesos de razonamiento.

En lugar de proponer cualquier versión temática, en este estudio adoptamos la perspectiva de la

teoría de la relevancia y diseñamos tareas temáticas específicamente pertinentes para los estudiantes de informática.

El objetivo de esta investigación era doble: por un lado, tratamos de comprobar si los resultados obtenidos en investigaciones previas en el marco de la teoría de la relevancia pueden ser replicados en esta población. Hipotetizamos que, si el diseño de la tarea era relevante para estos estudiantes y obtenían resultados similares o superiores a la media en comparación con la tarea abstracta, podrían comparar ambas situaciones y lograr un concepto más abstracto del sesgo de confirmación en la comprobación de hipótesis, mejorando así potencialmente su rendimiento incluso en la tarea abstracta.

Para maximizar la eficacia de las versiones temáticas, seleccionamos dos temas, ambos estrechamente relacionados con las experiencias de los estudiantes de informática. El primer tema giró en torno a una serie de televisión popular entre los estudiantes, mientas que el segundo tema abordó el rol del probador de software.

Experimento 1

En este experimento manipulamos la relevancia contextual de las tareas temáticas con el objeto de mejorar el interés y el rendimiento de los participantes en escenarios familiares en los que asumen el rol de controladores. Queríamos investigar las posibles diferencias en el rendimiento de los participantes entre las tareas abstractas y las temáticas, así como entre las distintas tareas temáticas. En particular, analizamos la influencia de un entorno cotidiano vinculado al ocio frente a un entorno académico y profesional, ambos familiares para los participantes. Para ello, diseñamos dos tareas temáticas específicas para comparar su rendimiento frente al obtenido en la versión abstracta. Una tarea temática versó sobre una popular serie de TV, mientras que la otra tarea estaba relacionada con entornos de programación y pruebas de software. Nuestro objetivo, en la misma línea que investigaciones anteriores, era mostrar que manipulando el contexto se puede mejorar el rendimiento en la tarea de Wason presentando entornos que resultan familiares a los estudiantes, suscitando así el cambio de una perspectiva verificadora a una perspectiva de búsqueda de transgresores de la regla. Además, exploramos posibles efectos de transferencia entre las versiones temáticas y las abstractas, analizando si la exposición inicial a las tareas temáticas influyó en el posterior rendimiento en las tareas abstractas.

Versión temática 1: vamos al cine. El primer problema, que denominamos 'Vamos al cine' (TemaTV) se inspiró en un episodio de la serie *The Big Bang Theory*, que la mayoría de los estudiantes conocen. El fragmento subido por un usuario en español (https://www.youtube.com/watch?v=Wduqd6oX7XI) pertenece a un episodio más extenso que no se proyectó ("The Financial Permeability" (2009). TBBT, temporada 2, episodio 14).

El fragmento elegido describe un escenario de conflicto habitual entre Sheldon Cooper y sus amigos, originado por los comportamientos idiosincráticos de Sheldon. En última instancia, el conflicto se resuelve cuando los amigos deciden excluir a Sheldon e ir a cenar solos. Durante la serie, es evidente que el comportamiento del grupo tiende a ser flexible, y los amigos de Sheldon están dispuestos a ajustar sus preferencias para alcanzar un acuerdo mutuo. Por el contrario, el comportamiento de Sheldon es extremadamente inflexible; es incapaz de negociar o aceptar alternativas que se desvíen de su preferencia inicial.

Para completar el argumento iniciado por el video, se continúa desarrollando la situación. Además de la inflexibilidad de Sheldon y de su falta de empatía, surge una cuestión más seria. Sheldon no se percibe a sí mismo cometiendo un error o actuando de forma incorrecta.

Se invita a los participantes a adoptar el rol del psicólogo de Sheldon, que ha observado la resistencia de Sheldon a reconocer sus errores. El psicólgo teme que Sheldon se vea excluido de todos los grupos. En un esfuerzo por fomentar la autopercepción de Sheldon, el psicólogo propone un ejercicio y una regla. Sheldon tiene que registrar el resultado de cada situación social en tarjetas, representando de un lado de la tarjeta el comportamiento del grupo y del otro lado, el comportamiento de Sheldon –acuerdo (cara feliz) o desacuerdo (gesto de enojo). La regla del psicólogo dicta que cuando el grupo alcance un acuerdo, Sheldon debe estar de acuerdo también.

Dado que los participantes conocen al protagonista de la serie, están acostumbrados a su comportamiento y comprenden que está más predispuesto al desacuerdo y menos dispuesto a ceder. El psicólogo trata de ayudar a Sheldon a reconocer sus errores, fomentando su autopercepción y reconocimiento. En consecuencia, las tarjetas que representan el desacuerdo de Sheldon (no-Q) y el acuerdo unánime del grupo tienen un valor especial en este contexto, puesto que representan comportamientos que transgreden la regla de flexibilidad (P y no-Q).

Tras analizar las tarjetas en función de su relevancia, se observa que la tarjeta P, que representa el acuerdo del grupo, es un input que conlleva consecuencias significativas. Si Sheldon (representado en

la otra cara) no está de acuerdo, el resultado podría ser una nueva exclusión. Por el contrario, la tarjeta Q en la que Sheldon está de acuerdo, no tiene grandes implicaciones, puesto que el objetivo del psicólogo es subrayar los errores de Sheldon y no sus logros, pese a que la regla adopta la forma P → Q (el acuerdo del grupo implica el acuerdo de Sheldon).

Si el grupo no está de acuerdo (no-P), esto no significa la exclusión de Sheldon, incluso si él tampoco estuviese de acuerdo, lo que hace que este efecto sea menos significativo. Sin embargo, las consecuencias del desacuerdo de Sheldon (no-Q) cuando el grupo está de acuerdo en la otra cara son claras. Este input es relevante y suscita a levantar esa tarjeta.

Versión temática 2: probador de software. En este escenario, los participantes asumen el rol de probadores de software. Su tarea es evaluar si los programas creados por otros programadores contienen errores. Puesto que los participantes son estudiantes de informática, pese a ser estudiantes de primer curso, comprenden perfectamente lo que ello implica. El texto pone de relieve la gran responsabilidad de los probadores de software, ya que deben detectar errores de forma oportuna. Se les facilitó un nuevo programa en el que se había identificado una disfunción específica: la combinación de teclas *Ctrl*+*Mayús*+*Esc* debería finalizar la ejecución del programa y regresar al sistema operativo. Sin embargo, se dan casos en los que dicha combinación no cierra el programa como estaba previsto. El relato indica que un asistente de investigación dirigió las pruebas y documentó los resultados en tarjetas. En una cara de las tarjetas se registran las teclas pulsadas, mientras que en la otra cara se registra la respuesta del sistema.

Tras analizar las tarjetas resulta evidente que la tarjeta P, que indica la combinación de teclas *Ctrl*+*Mayús*+*Esc*, tiene consecuencias sustanciales. Si la respuesta del sistema reflejada en la otra cara de la tarjeta indica que el programa no finaliza, eso significa la detección de un error por parte del probador. Y al contrario, la tarjeta Q, que indica la finalización del programa, no tiene grandes implicaciones para el probador, incluso si se ha pulsado otra tecla. Además, si se pulsa una tecla distinta (no-P), no es de interés para el probador. Sin embargo, las implicaciones de que el programa no finalice (no-Q) son de gran relevancia para el probador.

Participantes y procedimiento. En el experimento participaron voluntariamente 96 estudiantes de primer curso de ingeniería electrónica de una universidad en la provincia de Buenos Aires: 10 mujeres y 86 hombres, con una media de edad de 20.4 (*DT* = 1.2). Los estudiantes cursaban el tercer mes de una materia dedicada a sistemas digitales, en la que estudiaban elementos lógicos como el álgebra booleana y las puertas lógicas. Aunque no se obtuvieron datos sobre la etapa de educación secundaria de los participantes, anticipábamos que un porcentaje significativo de ellos provenía de escuelas secundarias técnicas, donde habrían estado expuestos a los elementos fundamentales de la lógica. Dado que la versión abstracta de la tarea de Wason resulta difícil de resolver, incluso para académicos y profesionales, consideramos que los conocimientos incipientes de lógica de los estudiantes no deberían influir en su rendimiento en la tarea.

Los participantes fueron asignados aleatoriamente a una de cuatro condiciones con variaciones en el orden de presentación de los problemas: *TemaTV-Abstracta* (primero resolvieron una versión temática de la tarea de Wason contextuali- zada con el episodio de TV y, después, la versión abstracta), Abstracta-TemaTV (primero resolvieron la versión abstracta de la tarea de Wason y luego la versión temática contextualizada con el episodio de TV), *TemaST-Abstracta* (primero resolvieron una versión temática de la tarea de Wason contextualizada con el problema del probador de software y después resolvieron la versión abstracta) y *Abstracta-TemaST* (primero resolvieron la versión abstracta de la tarea de

Wason y, después, la versión temática contextuali- zada con el problema del probador de software).

No se compararon los cuatro grupos simultáneamente. Las comparaciones se realizaron entre subgrupos de 24 participantes cada uno, lo que dio lugar a una muestra de 48 participantes en cada análisis. Para pruebas estadísticas como la de chi cuadrado, utilizado en este estudio, asumimos un efecto de tamaño medio (*w* de Cohen ≈ .41) y un nivel de significatividad de $\alpha = .05$. Un análisis previo de potencia estadística utilizando el programa G*Power (Faul et al., 2009) sugirió que una muestra de al menos 47 participantes sería adecuada para alcanzar una potencia estadística de .8, con suficiente potencia estadística.

Procedimiento y materiales. Primero, los participantes de las cuatro condiciones experimentales completaron un breve cuestionario en papel sobre la prueba de Wason y el concepto de falsa- cionismo (refutación) de una hipótesis. Ninguno de los participantes declaró poseer conocimientos previos sobre estos conceptos.

Las condiciones *TemaTV-Abstracta* y *Abstracta- TemaTV* fueron evaluadas simultáneamente, en salas contiguas equipadas con un proyector cada una. Por otro lado, las condiciones *TemaST- Abstracta* y *Abstracta-TemaST* fueron evaluadas en la misma semana, en un intervalo de dos días, en otra sala sin proyector.

Cada uno de los participantes en las cuatro condiciones recibió un cuadernillo con dos problemas; cada problema ocupaba dos páginas. La primera página mostraba el primer enunciado, mientras que la página siguiente mostraba un ejercicio de selección de tarjetas, la tercera página mostraba el segundo enunciado del problema y la cuarta y última presentaba la selección de tarjetas correspondiente. Se indicó a los participantes que no procediesen a la página siguiente hasta haber leído completamente el problema.

Antes de leer el problema TemaTV, los participantes en las condiciones TemaTV-Abstracta y Abstracta-TemaTV vieron un videoclip de un minuto y 42 segundos del episodio mencionado 'Vamos al cine'. Este clip facilitaba el contexto para la posterior lectura del problema.

Debido al orden inverso en la presentación de la tarea temática y abstracta en ambas condiciones, se escalonó el tiempo de visionado del vídeo. En la condición Abstracta-TemaTV, los participantes completaron primero la tarea abstracta antes de ver el video. Y en orden inverso, la condición TemaTV-Abstracta. Por tanto, se organizaron en dos salas. Después de ver el videoclip, los participantes leyeron el problema *TemaTV*.

Tras leer el primer problema, los participantes de los cuatro grupos tenían que resolver la tarea de selección en la página siguiente. Lo mismo ocurrió con el problema siguiente en la página 3 y la consiguiente tarea en la página 4.

Versión Temática 1: TemaTV (Problema 'Vamos al Cine')

Cada una de las tarjetas en la tarea representaba una situación concreta; un lado reflejaba el comportamiento del grupo y el otro lado, el de Sheldon (véase Figura 3).

Figura 3. Las cuatro tarjetas del problema 'Vamos al Cine'.

Las expresiones recogidas en las tarjetas se utilizaron para ilustrar un estado de ánimo positivo o negativo. Las cuatro opciones de comportamiento se reflejaron en las tarjetas del siguiente modo: P — los amigos han llegado a un acuerdo, Q — Sheldon está de acuerdo, no-P — los amigos no están de acuerdo y no-Q — Sheldon no está de acuerdo (véase Figura 4).

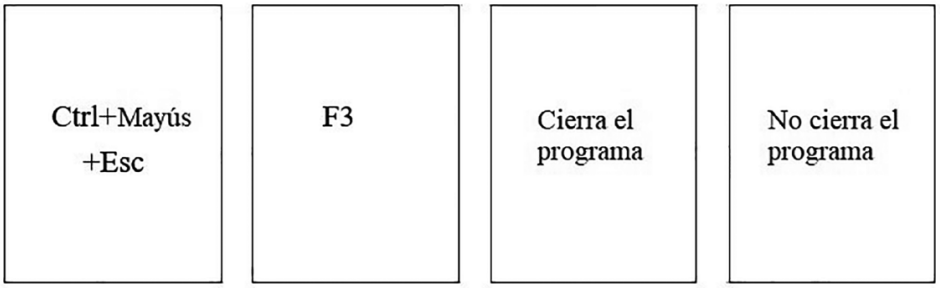

Figura 4. Tarjetas del problema del probador de software.

Los participantes afrontaron el siguiente problema:

En el episodio que has visto se describe un conflicto habitual entre Sheldon y sus amigos en el que, de nuevo, Sheldon muestra una falta extrema de flexibilidad. Finalmente, los amigos resuelven el problema excluyéndolo. El comportamiento de Sheldon es extremadamente obstinado; es incapaz de negociar o aceptar alternativas que lo desvíen de su preferencia inicial. Pero existe un problema más serio; es demasiado orgulloso y no reconoce sus errores.
Imagina que eres el psicólogo de Sheldon.
El objetivo del psicólogo es que Sheldon sea consciente de sus errores. Para ello, le propone un ejercicio y una regla. Le pide a Sheldon que registre el resultado de cada situación social en tarjetas, reflejando de un lado el comportamiento del grupo (acuerdo o desacuerdo) y del otro, el comportamiento de Sheldon.
El psicólogo le dice a Sheldon, esta es una regla que debes cumplir:
'Cada que el grupo completo esté de acuerdo, tú también lo estarás'.
Sospecha que Sheldon no será capaz de cumplir esta regla y quiere poner sus errores al descubierto exponiendo los resultados para que Sheldon los reconozca.
Al final del ejercicio, Sheldon se presenta con las cuatro tarjetas.
Si eres el psicólogo, ¿Qué tarjetas darás la vuelta como mínimo para descubrir si Sheldon cumplió la regla en

todas las situaciones?

Versión temática 2: TemaST (Problema del probador de software)

Cada una de las tarjetas en la tarea representaba una prueba realizada por el probador de software en ese programa. En un lado mostraba las teclas pulsadas y en el otro, la respuesta del programa (véase Figura 4).

Los participantes se enfrentaron al siguiente problema:

> Imagina que eres un probador de software recientemente contratado por una empresa de desarrollo de programas. Eres el responsable de detectar errores de forma oportuna. Se informa sobre un nuevo programa que presenta una disfunción: la combinación de teclas Ctrl+Mayús+Esc debería finalizar la ejecución del programa y regresar al sistema operativo. Sin embargo, en algunos casos, al pulsar esa combinación de teclas, no se produce el cierre del programa según lo previsto. Un asistente realiza las pruebas y documenta los resultados en tarjetas. Una cara muestra las teclas pulsadas, mientras que en la otra se refleja la respuesta del sistema.
>
> De las cuatro tarjetas siguientes, ¿a qué tarjetas darías la vuelta para comprobar si se cumple la
>
> regla de cerrar el programa al pulsar Ctrl+Mayús+Esc en todos los casos?

Versión abstracta

Se utilizó el mismo conjunto de cartas que en la versión clásica que se muestra en la Figura 1, acompañado de las siguientes instrucciones: 'Aquí hay cuatro tarjetas. Cada una de ellas tiene una letra en un lado y un número en el otro. Supuestamente, todas las tarjetas siguen la siguiente regla: "Si una tarjeta tiene una A en un lado, entonces tendrá un 4 en el otro lado". Tienes que decidir qué cartas tienes que voltear para comprobar si se confirma o se refuta la regla para todas las tarjetas'.

Resultados y discusión. Este estudio exploró el impacto de las versiones temáticas (TemaTV o TemaST) en comparación con la versión abstracta (Abstracta) en el rendimiento de los participantes en la tarea de Wason.

Cada grupo resolvió dos tareas. Para comparar el rendimiento de los participantes en la tarea temática frente a la tarea abstracta, se compararon los resultados de la tarea TemaTV en el grupo TemaTV-Abstracta con los resultados de la tarea Abstracta en el grupo Abstracta-TemaTV, dado que ninguna de estas tareas se vio influida por una tarea previa (Tabla 1). Asimismo, para comparar el rendimiento en la tarea TemaST frente a la tarea Abstracta, se compararon los resultados de la tarea TemaST en el grupo TemaST-Abstracta con los de la tarea Abstracta en el grupo Abstracta-TemaST (Tabla 2). Ambas comparaciones, reflejadas en la Tablas 1 y 2, contaron con una muestra total de 48 participantes (24 por grupo).

Tabla 1. Abstracta (AP) vs. Temática (TemaTV).

	P, no-Q	*P, Q*	otra
Abstracta en *Abstracta-TemaTV*	2	20	2
TemaTV en *TemaTV-Abstracta*	16	6	2

Tabla 2. Abstracta (AP) vs. Temática (TemaST).

	P, no-Q	*P, Q*	otra
Abstracta en Abstracta-	2	21	1

TemaST			
TemaST en TemaST-Abstracta	18	5	1

Tabla 3. Frecuencias de selección de la versión abstracta en TemaTV.

	Versión abstracta Respuestas		
	P, no *Q*	*P*, *Q*	otra
Grupo *Abstracta* en TemaTV-Abstracta	3	18	3
Abstracta- TemaTV	2	20	2

Tabla 4. Frecuencias de selección de la versión abstracta en TemaST.

	Versión Abstracta Respuestas		
	P, no *Q*	*P*, *Q*	otra
Abstracta en TemaST-Abstracta	4	20	0
Abstracta en Abstracta-TemaST	2	21	1
Total	11	79	6

La comparación reveló una diferencia significativa a favor de TemaTV (chi cuadrado = 18.2; $p < .05$). En la Tarea TemaST se observaron resultados igualmente significativos, como se muestra en la Tabla 2. La prueba chi cuadrado indicó una diferencia significativa (chi cuadrado = 22.64; $p < .05$) a favor de TemaST.

Además, para evaluar si la resolución de una tarea previa tiene algún impacto en la resolución de la tarea abstracta, se compararon los resultados de la tarea Abstracta realizada después de completar la tarea temática en el grupo TemaTV- Abstracta con los resultados de la tarea Abstracta, sin la influencia de una tarea previa, en el grupo Abstracta-Tema-TV (Tabla 3). Asimismo, la tarea Abstracta en el grupo TemaST-Abstracta se comparó con la tarea abstracta sin influencia de una tarea previa en el grupo Abstracta-TemaST (Tabla 4). Ambas comparaciones, de las Tablas 3 y 4, contaron con una muestra total de 48 participantes (24 por cada grupo).

En la tarea TemaST (Tabla 4) se observaron resultados similares, lo que indica un rendimiento constante de la versión abstracta con independencia del orden de presentación.

Estos resultados demuestran que los resultados de la teoría de la relevancia son aplicables a un grupo de estudiantes de informática. Las tareas temáticas, con un contenido familiar y una presentación equilibrada de información falsacio- nista y confirmatoria, mejora el rendimiento, lo que resulta en un mayor número de refutaciones (resultados P y no-Q), en comparación con las tareas abstractas tradicionales (Tablas 1 y 2). Sin embargo, no se observó ninguna relación significativa entre el orden de presentación y el rendimiento en la tarea abstracta. Independientemente de si la versión temática precede o sigue a la versión abstracta, no se observó un impacto discernible en el rendimiento de la tarea abstracta (Tablas 3 y 4). En general, la versión abstracta exhibió peor rendimiento de forma constante, independientemente de su posición relativa respecto a la versión temática.

Experimento 2

Desde una perspectiva pedagógica de la transferencia, para mejorar el conocimiento conceptual de los estudiantes sobre la comprobación de hipótesis, además de completar correctamente las versiones temáticas, es necesario comprender el concepto al enfrentarse a versiones abstractas. Para fomentar la transferencia del concepto de comprobación de hipótesis, los estudiantes tenían que comparar diversos casos estructuralmente comparables según la teoría de la proyección de la estructura (Gentner, 1983).

Los fundamentos de este enfoque surgen del reconocimiento de que la comparación de ejemplos concretos mejora la representación abstracta y la comprensión del concepto. La yuxtaposición de casos diversos facilita que los estudiantes identifiquen los patrones y principios subyacentes que transcienden el contexto específico, mejorando así su comprensión de la comprobación de hipótesis en distintos escenarios

La pregunta de investigación abordaba precisamente si el rendimiento en la tarea abstracta podría mejorar tras analizar ambos ejemplos temáticos e implicarse en procesos de comparación. En el contexto de las clases de programación, esto podría requerir facilitar la participación oral de los estudiantes. Hipotetizamos que la comparación de ejemplos fomenta la generalización, pero presentar ambos ejemplos seguidos sin incentivar la comparación o reflexión podría ser insuficiente para transferir un buen rendimiento a la tarea abstracta. Se comprobaron rigurosamente ambas condiciones para explorar esta hipótesis.

Método

Participantes. En este experimento participaron voluntariamente 120 estudiantes de primer curso de ingeniería de sistemas de información de una universidad pública de la provincia de Santa Fe: 65 mujeres y 55 hombres, con una media de edad de 19.50 (DT = 1.56). Los participantes cursaban una asignatura anual denominada Matemática discreta, en la que ya habían adquirido ciertos conocimientos de lógica proposicional.

Los participantes fueron asignados aleatoriamente a una de dos condiciones: *solo-ejemplos* (recibieron dos tareas temáticas para resolver de forma secuencial) y *comparación* (recibieron las mismas tareas temáticas junto a instrucciones para su comparación). Asumiendo un efecto de tamaño medio (w de Cohen ≈ .38) y un nivel de significatividad de $\alpha = .001$, el análisis previo de potencia utilizando el programa G*Power (Faul et al., 2009) sugiere que una muestra total de al menos 119 participantes será adecuada para alcanzar una potencia estadística de .8 con suficiente potencia estadística.

Procedimiento y materiales. Ninguno de los participantes en el Experimento 2 había participado en el Experimento 1. Todos ellos completaron un formulario inicial idéntico al del experimento previo para identificar y excluir a aquellos participantes familiarizados con la tarea de selección de Wason. Ninguno de ellos tenía conocimientos previos de la tarea. El experimento se dividió en dos fases: una fase de aprendizaje, en la que los participantes completaron tareas en su versión temática (de forma

secuencial o haciendo comparaciones) y una fase de aplicación, en la que completaron la versión abstracta. El objetivo era evaluar si los participantes que habían realizado procesos de comparación tenían mejor rendimiento en la tarea abstracta.

En la condición *solo-ejemplos*, los participantes afrontaron las mismas versiones temáticas utilizadas en el Experimento 1 (*TemaTV* y *TemaST*) durante la fase de aprendizaje. Los participantes tenían que resolver estas tareas de forma secuen- cial, sin la opción de revisar páginas anteriores, como se indicó en las instrucciones del Experimento 1. Tanto los grupos de comparación como los de solo-ejemplos recibieron versiones temáticas idénticas durante la fase de aprendizaje. La distinción entre los dos grupos reside en la condición de comparación, en la que los participantes recibieron instrucciones para comparar los ejemplos temáticos. Se buscaron principios comunes entre ambas situaciones y se presentaron las etiquetas ('P', 'no-P', 'Q', 'no-Q') para designar los contenidos que desempeñaban roles equivalentes en ambos escenarios. Por ejemplo, se motivó a los participantes con enunciados como 'Compara ambas historias. . . ¿Cuál es el principio común a ambas? ¿Cuál es la regla que Sheldon debe seguir? (Problema 1) y ¿Cuál es la regla que debe seguir el programa? (Problema 2)'.

Tras completar la fase de aprendizaje, se retiraron los formularios y se guardó una separación contextual de 15 minutos, durante la que los participantes completaron el test de reflexión cogni- tiva (TRC). A continuación, los participantes pasaron a la fase de aplicación, en la que recibieron la versión abstracta de la tarea. Esta versión incorporaba etiquetas tales como 'P', 'no-P', 'Q' y

Tabla 5. Respuestas abstractas. Solo-ejemplos vs. Comparación.

	Respuestas abstractas		
	P, no-Q	*P, Q*	other
Solo-ejemplos	12 (20%)	42 (70%)	6 (10%)
Comparación	42 (71.1%)	17 (28.8%)	0 (0%)

Tabla 6. Frecuencias de selección Temática y Abstracta.

	Solo ejemplos		
	P, no-Q	*P, Q*	otro
TemaTV	38 (63,3%)	15 (25%)	7 (11,6%)
TemaST	45 (75%)	9 (15%)	6 (10%)
AP	12 (20%)	42 (70%)	6 (10%)
	Comparación		
	P, no-Q	*P, Q*	otro
TemaTV	37 (62.7%)	15 (25.4%)	7 (11.8%)
TemaST	46 (77.9%)	7 (11.8%)	6 (10.1%)
AP	42 (71.1%)	17 (28.8%)	0 (0%)

'no-Q', ausentes en las anteriores iteraciones del experimento. Estas etiquetas correspondían a las utilizadas en las instrucciones para la comparación. Dado que son de uso común para representar operaciones de lógica proposicional, no se facilitaron más aclaraciones a los participantes en el grupo control. La pregunta que se formuló a los participantes en ambos experimentos se mantuvo constante: '¿Qué tarjetas deberías voltear para comprobar si se cumple la regla? Si hay una E en un lado, entonces debería haber un 2 en el otro lado'.

Resultados y discusión. Los resultados del experimento arrojan luz sobre la eficacia de distintos enfoques de aprendizaje en la mejora del rendimiento en las tareas abstractas. En la Tabla 5 se muestran las frecuencias de selección en la tarea abstracta, que revelan disparidades entre los grupos de comparación y los de solo-ejemplos. En particular, los participantes en la condición de

solo-ejemplos mostraron un índice inferior de selección de la respuesta correcta; únicamente un 20% optó por 'P, no-Q'. Por el contrario, el grupo de comparación mostró un índice de precisión superior, 71%, lo que indica una clara ventaja relacionada con los procesos de comparación ($\chi^2 = 33.25$; $p <$ .001).

Más allá de la tarea abstracta, en la Tabla 6 se muestran las frecuencias de selección y los porcentajes tanto para las tareas temáticas (TemaTV, TemaST) como para la tarea abstracta (Abstracta) en ambos grupos. Aunque los participantes en la condición solo-ejemplos exhibieron un rendimiento elogiable en las tareas temáticas (63.3% y 75% en TemaTV y TemaST, respectivamente), este no se transfirió en un rendimiento igualmente satisfactorio en la tarea Abstracta. Pese a observarse un rendimiento comparativamente superior (20%) en relación con la presentación de un único ejemplo concreto en el Experimento 1 (11.6% y 10%), los participantes siguieron exhibiendo una laguna considerable a la hora de resolver la versión abstracta en el mismo grupo. Esto sugiere que, aunque la exposición a ejemplos temáticos mejora la comprensión de la tarea, puede no facilitar intrínsecamente la transferencia de habilidades a escenarios abstractos.

No obstante, los participantes en la condición de comparación mostraron una habilidad notable para generalizar su aprendizaje de contextos temáticos a contextos abstractos. Mediante un proceso guiado de comparación, exhibieron un conocimiento más exhaustivo, reflejado en un índice superior de respuestas correctas en la tarea abstracta (71.1%). Este fenómeno subraya la importancia de los procesos de comparación estructurados para facilitar la transferencia de conocimiento y habilidades entre distintas disciplinas.

Estos resultados ponen de relieve la potencial eficacia de las comparaciones guiadas entre casos de pruebas estructuralmente análogas como estrategia pedagógica. Al incentivar a los estudiantes a identificar principios y patrones comunes entre distintos contextos, los educadores pueden fomentar el desarrollo de representaciones abstractas y objetivas de conceptos complejos tales como la comprobación de hipótesis y el sesgo de confirmación. Este enfoque no solo mejora el rendimiento específico de la tarea sino que además contribuye a desarrollar la capacidad de aplicar los principios aprendidos a situaciones noveles, una habilidad clave para fomentar un conocimiento más profundo y las habilidades de pensamiento crítico de los estudiantes.

Discusión

Retomando la problemática planteada en la introducción, algunas investigaciones en el campo de la informática han recurrido a la psicología para diseñar métricas para el sesgo de confirmación basadas en la Tarea de Selección de Wason. Estos estudios se han desarrollado bajo un enfoque logicista falsacionista tradicional, postulando que si los probadores de software no muestran comportamientos afines a elementos falsacionistas, estos se consideran inherentemente sesgados (Calikli & Bener, 2010, 2015). En estas configuraciones experimentales, los participantes o razonadores no suelen ser

conscientes de estar siendo evaluados como probadores de software, por lo que no encarnan ese papel. En particular, el estudio revela niveles inferiores de rendimiento entre los evaluadores experimentados que entre los novicios. Nuestra investigación pretendía evaluar el rendimiento de un grupo de estudiantes que no habían recibido educación formal en pruebas de software. Hicimos uso del test de Wason tanto en su versión temática como abstracta, fundamentado en teorías de razonamiento pragmático. Se recurrió a la manipulación del contexto para mejorar el rendimiento. Los resultados empíricos indican que, aunque el rendimiento disminuye en la tarea abstracta, este mejora cuando la tarea se sitúa en un contexto familiar, caracterizado por especificaciones exhaustivas y equilibrio entre informaciones positivas y negativas. Nuestros resultados confirman los de investigaciones previas en torno al sesgo de confirmación (Leventhal et al., 1994) y la psicología del razonamiento (Sperber et al., 1995). La habilidad de transferir el rendimiento a tareas noveles depende de la comparación deliberada entre distintos ejemplos. Sostenemos que caracterizar el rendimiento de los razonadores en la versión abstracta como un 'fracaso' es inexacto; más bien, estos se adhieren a un conjunto de reglas distinto.

Durante décadas, la evidencia empírica ha demostrado que el test de Wason abstracto no cumple con la finalidad asignada, lo que pone en duda la existencia de un razonamiento humano independiente del contexto disciplinar. Por el contrario, las versiones temáticas del test revelan la existencia de reglas sensibles al contexto que pueden influir en el rendimiento de los probadores.

Nuestros resultados ponen de relieve la importancia de la adaptación contextual del razonamiento en la comprobación de hipótesis, lo que invita al análisis del contexto en el área de las pruebas de software. En este ámbito, el contexto está compuesto por toda la información sobre la tarea que el programador tiene en mente (lo que está en su memoria operativa).

El contexto del *programador* está compuesto por las especificaciones del programa orientadas al objetivo del programa (lo que *hace* el código). En cambio, el contexto del *probador de software* está formado por todos los posibles fallos que se pueden tener en cuenta al escribir el test.

La integración del desarrollo de software no implica la integración psicológica. Los desarrolladores que se encuentran en nuevos entornos de desarrollo (por ejemplo, el desarrollo basado en pruebas o TDD) realizan una especie de "multitarea" infructuosa, alternando constantemente entre los dos roles. Por ejemplo, si el desarrollador pretende escribir un código limpio y comprobable, sigue ciertas premisas como tratar de no utilizar métodos largos, diseñar métodos estáticos o no contar con un estado global. Pero no debería tener en cuenta lo que hace el programa porque ese es un objetivo del programador, no del probador.

Motivados por la creciente demanda de ingenieros altamente cualificados, muchos educadores están de acuerdo en que la prueba de software debe ser un tema ineludible en el currículum de Informática desde el principio, cuando se aprende a programar, puesto que los estudiantes deben beneficiarse de probar sus primeros casos (Edwards, 2003, 2004; Goldwasser, 2002; Jones, 2000; Leska, 2004), pero algunos estudios empíricos, más centrados en las percepciones de los estudiantes, plantean la necesidad de abordar las dificultades (Scatalon et al., 2017) y prestar atención a las percepciones negativas (Baldassarre et al., 2022), aunque estas necesidades no han sido abordadas desde un enfoque psicológico.

Cuando los estudiantes noveles se ven obligados a escribir pruebas antes de escribir código, deben alternar entre el rol del programador y el rol del probador. Los estudiantes de primer curso comienzan a desarrollar algoritmos para resolver problemas con una actitud positiva hacia sus producciones, lo cual es incompatible con la autocrítica e incluso puede llegar a ser desmotivador. Si las pruebas de software se presentasen como un problema conceptual antes de progresar a una práctica repetitiva, podría introducirse este concepto sin comprometer el entusiasmo por las primeras experiencias de programación.

Conclusiones

Históricamente, los educadores han interpretado las pruebas de software como una labor técnica, mientras que los investigadores informáticos las han considerado simplemente una tarea sesgada.

Esta subestimación de las teorías psicológicas del razonamiento pragmático está aceptada e instalada en las ciencias computacionales y termina por excluir los factores contextuales de los modelos mentales posibles.

Nuestro estudio aporta nuevos conocimientos teóricos y evidencia empírica para ampliar estos modelos del razonamiento humano, el comportamiento de pruebas y el sesgo de confirmación en las pruebas de software. La teoría de la relevancia defiende, en la misma línea que los primeros estudios empíricos en el campo de la informática, que el equilibrio entre las informaciones verificacionistas y falsacionistas (pruebas positivas y negativas) mejora el rendimiento de los probadores.

En lo que respecta a la formación profesional, es esencial que los probadores de software reciban formación sobre el sesgo de confirmación y sus efectos. Al ser conscientes de este sesgo, los probadores pueden ser más críticos y reflexionar sobre su enfoque hacia las pruebas, buscando activamente tanto funcionalidades de verificación como de refutación.

Si bien las computadoras operan con una lógica binaria, el razonamiento humano se rige por reglas sensibles al contexto. Dado que los errores de software tienen su origen en interacciones humano-computadora, el refuerzo de la educación formal en lógica desde una perspectiva logicista puede no ser suficiente para subsanar el rendimiento deficitario de los probadores.

Sugerimos firmemente que las versiones temáticas del test de Wason, así como otros ejercicios que incorporen diversidad temática al razonamiento lógico —el cual suele enseñarse de manera exclusivamente sintáctica y descontextualizada en las ciencias computacionales—, sean incluidos en el currículo universitario. Esto permitiría introducir las dificultades conceptuales asociadas con la comprobación de hipótesis y los aspectos metacognitivos del sesgo de confirmación, antes de abordar los aspectos prácticos de las pruebas de software en los cursos introductorios de las carreras de informática. Se fomentaría también una comprensión más profunda de la importancia del contexto en el estudio y aplicación de la lógica.

Referencias

Andrews, M., & Whittaker, J. A. (2006). *How to break web software: Functional and security testing of web applications and web services*. Addison-Wesley Professional.

Baldassarre, M. T., Caivano, D., Fucci, D., Romano, S., & Scanniello, G. (2022). Affective reactions and test-driven development: Results from three experiments and a survey. *Journal of Systems and Software*, *185*, 111154. https://doi.org/10.1016/j. jss.2021.111154

Beth, E. W., & Piaget, J. (1974). Strict demonstration and heuristic procedures. In E. W. Beth & J. Piaget (Eds.), *Mathematical epistemology and psychology* (pp.86–100). Springer. https://doi.org/10.1007/978- 94-017-2193-6_4

Boole, G. (1847). *The mathematical analysis of logic*. Philosophical Library.

Boole, G. (1854). *An investigation of the laws of thought: On which are founded the mathematical theories of logic and probabilities*. Walton and Maberly.

Calikli, G., & Bener, A. (2010, September 12–13). *Empirical analyses of the factors affecting confirmation bias and the effects of confirmation bias on software developer/tester performance* [Conference session]. Proceedings of the 6th International Conference on Predictive Models in Software Engineering, Timişoara, Romania (pp. 1–11). Association for Computing Machinery.

Calikli, G., & Bener, A. (2015). Empirical analysis of factors affecting confirmation bias levels of software

engineers. *Software Quality Journal*, *23*(4), 695–722. https://doi.org/10.1007/s11219-014- 9250-6

Çalıklı, G., & Bener, A. B. (2013). Influence of confirmation biases of developers on software quality: An empirical study. *Software Quality Journal,* 21, 377–416.

Carnap, R. (1936). Testability and meaning. *Philosophy of Science*, *3*(4), 419–471.

Carnap, R. (1937). Testability and meaning — Continued. *Philosophy of Science*, *4*(1), 1–40.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416. https://doi.org/10.1016/0010-0285(85)90014-3

Cheng, P. W., & Holyoak, K. J. (1989). On the natural selection of reasoning theories. *Cognition*, *33*(3), 285–313. https://doi.org/10.1016/0010- 0277(89)90031-0

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010- 0277(89)90023-1

Edwards, S. H. (2003, July 31–August 2). *Using test- driven development in the classroom: Providing students with automatic, concrete feedback on performance* [Conference session]. Proceedings of the International Conference on Education and Information Systems: Technologies and Applications EISTA, Orlando, FL (*Vol. 3*).

Edwards, S. H. (2004, December 5–8). *Using software testing to move students from trial-and-error to reflectionin-action* [Conference session]. Proceedings of the 35th SIGCSE Technical Symposium on

Computer Science Education, Norfolk, VA (pp. 26–30). Association for Computing Machinery. https://doi.org/10.1145/971300.971312

Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology*, *24*(2), 193–199. https:// doi.org/10.1080/00335557243000067

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. Taylor & Francis Group.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences* (pp. ix, 145). Lawrence Erlbaum Associates, Inc.

Evans, J. St. B. T. (1993). Bias and rationality. In K.

I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 6–30). Taylor & Frances/Routledge.

Evans, J. St. B. T. (2013). *The psychology of deductive reasoning (Psychology revivals)*. Psychology Press.

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*(2), 356–363. https://psycnet.apa.org/record/1996-01742-008

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fincher, S., & Petre, M. (2004). *Computer science education research*. CRC Press.

Fincher, S. A., & Robins, A. V. (2019). *The Cambridge handbook of computing education research*. Cambridge University Press.

Garousi, V., Rainer, A., Lauvås, P., & Arcuri, A. (2020). Software-testing education: A systematic literature mapping. *Journal of Systems and Software*, *165*, 110570. https://doi.org/10.1016/j. jss.2020.110570

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155– 170. https://doi.org/10.1016/S0364-0213(83)80 009-3

Girotto, V., Gilly, M., Blaye, A., & Light, P. (1989). Children's performance in the selection task: Plausibility and familiarity. *British Journal of Psychology*, *80*(1), 79–95. https://doi. org/10.1111/j.2044-8295.1989.tb02304.x

Goldwasser, M. H. (2002). A gimmick to integrate software testing throughout the curriculum. *ACM SIGCSE Bulletin*, *34*(1), 271–275. https://doi. org/10.1145/563517.563446

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic- materials effect in Wason's selection task. *British Journal of Psychology*, *73*(3), 407–420. https://doi. org/10.1111/j.2044-8295.1982.tb01823.x

Hannay, J. E., Sjoberg, D. I. K., & Dyba, T. (2007). A systematic review of theory use in software engineering experiments. *IEEE Transactions on Software Engineering*, *33*(2), 87–107. https://doi. org/10.1109/TSE.2007.12

Inhelder, B., & Piaget, J. (1955). *De la logique de l'enfant à la logique de l'adolescent* [From the logic of the child to the logic of the adolescent] (p. 314). Presses Universitaires de France.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Psychology Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction* (pp. xii, 243). Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, *63*(3), 395–400. https://doi. org/10.1111/j.2044-8295.1972.tb01287.x

Jones, E. L. (2000, December 1). *Software testing in the computer science curriculum–a holistic approach* [Conference session]. Proceedings of the Australasian Conference on COMPUTING Education, Melbourne, Australia (pp. 153–157). Association for Computing Machinery.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475. https://doi. org/10.1257/000282803322655392

Katz, I. R., & Anderson, J. R. (1989). Debugging: An analysis of bug-location strategies (Abstract only). *ACM SIGCHI Bulletin*, *21*(1), 123. https://doi. org/10.1145/67880.1046598

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi. org/10.1037/0033-295X.94.2.211

Leska, C. (2004). Testing across the curriculum: Square one! *Journal of Computing Sciences in Colleges*, *19*(5), 163–169.

Leventhal, L. M., Teasley, B., & Rohlman, D. (1994). Analyses of factors related to positive test bias in software testing. *International Journal of HumanComputer Studies*, *41*, 717–749. https://doi. org/10.1006/ijhc.1994.1079

Leventhal, L. M., Teasley, B. M., Rohlman, D. S., & Instone, K. (1993). Positive test bias in software testing among professionals: A review. In L. J. Bass, J. Gornostaev & C. Unger (Eds.),

Human-computer interaction (pp. 210–218). Springer. https://doi.org/10.1007/3-540-57433-6_50

Liberman, N., & Klar, Y. (1996). Hypothesis testing in Wason's selection task: Social exchange cheating detection or task understanding. *Cognition*, *58*(1), 127–156. https://doi.org/10.1016/ 0010-0277(95)00677-X

Love, R. E., & Kessler, C. M. (1995). Focusing in wason's selection task: Content and instruction effects. *Thinking & Reasoning*, *1*(2), 153–182. https://doi.org/10.1080/13546789508251502

Manktelow, K. I., & Evans, J. S. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology*, *70*(4), 477–488. https:// doi.org/10.1111/j.2044-8295.1979.tb01720.x

Manktelow, K. I., & Over, D. E. (1990). *Inference and understanding: A philosophical and psychological perspective* (pp. vii, 194). Taylor & Frances/ Routledge.

Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, *39*(2), 85–105. https://doi. org/10.1016/0010-0277(91)90039-7

Mosconi, G., & D'Urso, V. (1974, April). *The selection task from the standpoint of the theory of double code* [Conference session]. Conference on The Selection Task, University of Trento, Trento, Italy.

Myers, G. J. (1979). *The art of software testing*. Wiley. http://archive.org/details/artofsoftwaretes00 00myer

Nelson, G. L., & Ko, A. J. (2018, August 13–15). *On use of theory in computing education research* [Conference session]. Proceedings of the 2018 ACM Conference on International Computing Education Research, Espoo, Finland (pp. 31–39). Association for Computing Machinery. https://doi. org/10.1145/3230977.3230992

Platt, R. D., & Griggs, R. A. (1993). Facilitation in the abstract selection task: The effects of attentional and instructional factors. *The Quarterly Journal of Experimental Psychology Section A*, *46*(4), 591–613.

https://doi.org/10.1080/14640749308401029

Poletiek, F. H. (1996). Paradoxes of falsification. *The Quarterly Journal of Experimental Psychology Section A*, *49*(2), 447–462. https://doi.org/10.1080/713755628

Poletiek, F. H. (2001). *Hypothesis-testing behaviour* (pp. x, 172). Psychology Press.

Poletiek, F. H. (2011). You can't have your hypothesis and test it: The importance of utilities in theories of reasoning. *The Behavioral and Brain Sciences*, *34*, 87–88. https://doi.org/10.1017/S0140525X10002980

Popper, K. R. (1959). *The logic of scientific discovery*.

Hutchinson. http://archive.org/details/logicof- scientifi00popp

Popper, K. R. (1962). *Conjectures and refutations, the growth of scientific knowledge*. Basic Books. http://archive.org/details/conjecturesrefut0000popp

Pressman, R. S. (2010). A practitioner's approach. *Software Engineering*, *2*, 41–42.

Scatalon, L. P., Barbosa, E. F., & Garcia, R. E. (2017, October 18–21). *Challenges to integrate software testing into introductory programming courses* [Conference session]. IEEE Frontiers in Education Conference (FIE), Indianapolis, IN (pp. 1–9). IEEE. https:// doi.org/10.1109/FIE.2017.8190557

Shull, F., Singer, J., & Sjøberg, D. I. K. (2007). *Guide to advanced empirical software engineering*. Springer Science & Business Media.

Sperber, D. (2005). Modularity and relevance. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind: Structure and contents*(*Vol. 1*, pp. 53–68). Oxford University Press.

Sperber, D.,Cara,F., & Girotto,V.(1995). Relevancetheory explains the selection task. *Cognition*, *57*(1), 31–95. https://doi.org/10.1016/0010-0277(95)00666-m

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. *142*). Harvard University Press.

Sperber, D., & Wilson, D. (2004). Relevance theory. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 607–632). Blackwell.

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. In R. W. Gibbs, Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 84–105). Cambridge University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Psychology Press. https://www.taylorfrancis.com/books/ mono/10.4324/9781410603432/rational-keith- stanovich

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https:// doi.org/10.1080/17470216008416717

Wason, P. C. (1962). Reply to wetherick. *Quarterly Journal of Experimental Psychology*, *14*(4), 250–250. https://doi.org/10.1080/17470216208416543

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 106–137). Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content* (*Vol. 86*). Harvard University Press.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*(1), 63–71. https://doi.org/10.1080/00335557143000068

Whittaker, J. A. (2002). *How to break software: A practical guide to testing with Cdrom*. Addison-Wesley Longman Publishing Co., Inc. https://dl.acm.org/ doi/abs/10.5555/515499

Yachanin, S. A., & Tweney, R. D. (1982). The effect of thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychonomic Society*, *19*(2), 87–90. https://doi.org/10.3758/ BF03330048