UAB
Universitat Autònoma de Barcelona

Estudios del desempeño en prácticas científicas del alumnado mediante un proyecto de verificación de Apps

Mauricio Aguilera Tapia

Universitat Autònoma de Barcelona

Tesis doctoral

Estudios del desempeño en prácticas científicas del alumnado mediante un proyecto de verificación de Apps

Autor: Mauricio Aguilera Tapia

Director y tutor: Dr. Víctor López Simó

Doctorado en Educación

Departamento de Didáctica de la Matemática y las Ciencias Experimentales

Junio 2025

Índice

Índice de Figuras

Índice de Tablas

CAPÍTULO I

INTRODUCCIÓN

El Capítulo I presenta el propósito general de la tesis, que gira en torno a promover el pensamiento crítico en el aula de ciencias a través de la evaluación de la confiabilidad de aplicaciones móviles (Apps). Introduce el proyecto App Checkers como una propuesta que aprovecha el uso cotidiano de celulares para generar una experiencia basada en las prácticas científicas escolares. La propuesta se plantea como una manera de fortalecer competencias científicas en contextos escolares actuales.

El pensamiento crítico, entendido como un conjunto de disposiciones y habilidades que configuran la capacidad de los individuos para emitir juicios razonados en contexto (Facione, 1990; McPeck, 1981), se presenta como una herramienta clave frente a los desafíos epistémicos en el ámbito de la enseñanza de las ciencias. Esta competencia adquiere una relevancia particular, ya que se alinea con objetivos formativos como la alfabetización científica, la toma de decisiones basada en pruebas, y la participación ciudadana responsable (Puig y Uskola, 2021; Covitt y Anderson, 2022). Su desarrollo no puede ser abordado de forma puntual ni descontextualizada, sino que hace falta enseñarlo de manera continua y situada, que ponga en juego prácticas científicas genuinas como la indagación, la argumentación y la modelización (Vila et al., 2023). La ciencia escolar, lejos de reducirse a la transmisión de conocimientos, debe constituirse como un espacio para la problematización y el juicio crítico, promoviendo que el alumnado interprete, analice, evalúe y regule su propio pensamiento frente a fenómenos reales o simulados.

Una de las formas más potentes de vincular pensamiento crítico y enseñanza de las ciencias consiste en ofrecer a los estudiantes situaciones problemáticas que combinen el uso de tecnología cotidiana con la aplicación de criterios científicos. Una actividad que apunta en esta dirección es el análisis de la confiabilidad de aplicaciones móviles (Apps) que aseguran medir variables científicas o sociales. Estas Apps, ampliamente disponibles en tiendas digitales, plantean una apariencia de objetividad que rara vez es sometida a escrutinio. Sin embargo, cuando los estudiantes se enfrentan a ellas con una actitud crítica, pueden activarse algunas dimensiones del pensamiento crítico, desde el cuestionamiento de fuentes hasta el diseño de procedimientos de validación y fiabilidad. Este tipo de experiencias didácticas generan un conflicto cognitivo, ya que el alumnado se ve interpelado por herramientas que conoce o utiliza en su vida diaria, pero cuya legitimidad como instrumentos de medición es dudosa. Por tanto, se abre una oportunidad para que el aula de ciencias se convierta en un espacio de verificación epistémica, donde se promuevan habilidades como el diseño de experimentos, la recolección y análisis de datos, y la argumentación basada en pruebas (Torres Climent et al., 2017; Rodríguez-Arteche et al., 2024).

Este enfoque alcanza su mayor potencial cuando no solo se cuestiona el funcionamiento técnico de la App, sino que se diseñan actividades que permitan evaluar explícitamente su confiabilidad. En este marco, surge la necesidad de operar con una doble dimensión: por un lado, la fiabilidad, entendida como la consistencia y veracidad de las fuentes de información que alimentan la App; y por otro, la validez, concebida como el grado en que los resultados

producidos se ajustan a estándares de calidad, claridad, precisión y lógica o a los estándares considerados válidos por la comunidad científica en un contexto particular. La evaluación de Apps en el aula, por tanto, se puede plantear como una actividad a realizar por el alumnado que no se limite a un juicio intuitivo, sino que exija a los estudiantes formular criterios explícitos de evaluación, diseñar procedimientos experimentales, e interpretar resultados. Estas prácticas, cuando son guiadas, permiten un acercamiento auténtico a la forma de trabajar de la ciencia y al pensamiento crítico, promoviendo el desarrollo de competencias y su transferencia a otros ámbitos de la vida académica y social.

1.1 Como medir la confiabilidad de una App en una actividad escolar

Los profesionales de la educación científica enfrentan retos sociales relacionados con la propagación de bulos pseudocientíficos y el apoyo a teorías alternativas. Para abordar estos fenómenos, diversas iniciativas ciudadanas y periodísticas han desarrollado estrategias como los "fact-check", sistemas de verificación basados en pruebas que evalúan la confiabilidad de las afirmaciones difundidas en redes sociales o expresadas por figuras públicas y medios de comunicación (Graves, 2017). Como se mencionó, la enseñanza de la ciencia trasciende la simple acumulación de información, ya que debe desarrollar una alfabetización científica que capacite a los estudiantes para enfrentar problemas sociales relevantes desde una perspectiva científica. Esto requiere comprender los procesos y debates propios de la comunidad científica, evitando posturas simplistas como considerar que la ciencia ofrece soluciones infalibles o, por el contrario, que intenta engañar (Couso y Puig, 2021). Además, es esencial que el alumnado participe en actividades que integren aspectos procedimentales y epistémicos de la ciencia, lo que les permite analizar críticamente el conocimiento defectuoso o erróneo que circula, especialmente en internet. Dichos contenidos erróneos representan un riesgo significativo tanto a nivel individual como colectivo (Osborne y Pimentel, 2023). Al diseñar tareas que trasciendan la mera verificación de teorías en un laboratorio, se fomenta el desarrollo de competencias necesarias para enfrentar los desafíos del presente siglo, especialmente aquellos derivados del uso problemático de las nuevas tecnologías.

Inspirándose en esta lógica de la verificación, y tratando de buscar un contexto que cruce la desinformación, el auge de la tecnología en la era digital y el fenómeno de las creencias preexistentes con el uso del pensamiento crítico y de las habilidades científicas mediante las

prácticas científicas de indagación, modelización y argumentación, es que un grupo de innovación docente de profesores de física y química de Cataluña desarrollaron un proyecto de aula para estudiantes de secundaria llamado "App Checkers" (López-Simó, 2021), en el cual se propuso a estudiantes de ESO que verificaran la confiabilidad de una aplicación móvil comercial (App) descargable en celular, y que supuestamente midiera algo del exterior, ya fuera a través de la cámara del teléfono, de su micrófono o de cualquiera de sus sensores internos. El estudiante trabaja con una App para investigar, siempre que fuere de carácter gratuito y segura, que incluyera algún tipo de medición del exterior (fuera cierta o no), y que no infligiera ningún tipo de norma de convivencia y respeto a los demás.

Este proyecto se enmarca en una estrategia de Aprendizaje Basado en Proyectos, y se ha implementado en diversos grupos de estudiantes de distintos niveles, con el objetivo de introducir a los estudiantes en el diseño de investigaciones a partir de situaciones cotidianas y así promover, principalmente sus habilidades de indagación y uso de pruebas. Los estudiantes deben llevar a cabo un estudio científico sobre la confiabilidad de la App, independientemente de si la fiabilidad y validez se presumen de antemano. En el proyecto, la pregunta de investigación o problema de investigación es proporcionada a los estudiantes: ¿Qué tan confiable es la aplicación seleccionada?, donde por confiabilidad se entenderá la validez y fiabilidad de la App. En algunos casos, la App será un medidor concreto, mientras que, en otros, los estudiantes podrían no saber inicialmente qué mide la aplicación. Esta diversidad de enfoques busca dar a los estudiantes la posibilidad de explorar diferentes áreas de medición y poner a prueba sus habilidades de indagación científica y pensamiento crítico. El proceso incluye el diseño y la ejecución de experimentos, así como la evaluación de la fiabilidad y validez de las Apps seleccionadas, con el objetivo de fomentar la reflexión sobre la importancia de la verificación en la ciencia y la tecnología.

1.2 Definición del problema de investigación

El problema de investigación central se enfoca en comprender la interrelación entre las prácticas de indagación, argumentación y modelización que desarrollan los estudiantes, y cómo estas se ven influenciadas por el uso de Apps, específicamente en el contexto del proyecto App Checkers en las distintas versiones ha sido aplicado en el aula. La investigación busca analizar de qué manera los estudiantes aplican estrategias de indagación científica, argumentación estructurada y conceptos de validez y fiabilidad, y cómo estos procesos impactan en su desempeño dentro de escenarios científicos. De este modo, el problema se extiende al análisis

de cómo estas competencias se desarrollan y aplican mediante el soporte tecnológico y el andamiaje pedagógico provisto en el aula.

Como resultado de esta investigación, se propone la creación de una metodología que nos permita evaluar las prácticas científicas y niveles de desempeño que el proyecto promueve. Este sistema, que quedará como una herramienta para futuras evaluaciones, se basa en las estrategias que los estudiantes emplean para evaluar la validez y la fiabilidad en el uso de aplicaciones móviles. Su objetivo es proporcionar un marco práctico para que los docentes puedan evaluar el desempeño vinculado al uso de modelos de validación de las Apps empleadas durante el proceso de aprendizaje. El principal aporte de esta investigación a la didáctica radica en el desarrollo de un enfoque metodológico que integra el uso de aplicaciones móviles y herramientas de validación científica para fomentar el pensamiento crítico en el alumnado y busca conectar la enseñanza de la ciencia con la vida cotidiana de los estudiantes, mediante la creación de actividades que promueven la indagación y el análisis reflexivo. Asimismo, ofrece a los docentes herramientas prácticas para evaluar de manera efectiva el desarrollo de estas competencias en el aula, asegurando su aplicabilidad en diversos contextos educativos.

CAPÍTULO II

MARCO TEÓRICO

El Capítulo II presenta el marco teórico de la tesis en torno a tres prácticas científicas clave: indagación, argumentación y modelización. Se describen sus implicaciones en el aula de ciencias y se analizan estrategias para evaluar el desempeño del alumnado en cada una de ellas. Profundiza en marcos como los Concepts to Evidence (CoE) y el modelo de Toulmin, así como en nociones de fiabilidad, validez y modelos mentales. Todo ello conforma una base conceptual que guiará los análisis realizados en los estudios empíricos posteriores.

El aprendizaje científico cada vez ha ido evolucionando más de posturas centradas en los contenidos donde sólo se va acumulando información hacia posturas más centradas en el alumnado y en habilidades procedimentales y epistémicas donde importa cómo funciona la ciencia y el porqué de sus procesos. Es fundamental que el estudiantado se enfrente activamente a tareas que los involucren en estos aspectos de la ciencia. Esto les permite tener herramientas y abordar todo tipo de controversias socio-científicas con las que lidiaran en la comunidad. Al involucrar a los estudiantes en tareas de este tipo, en lugar de limitarse a la verificación de teorías en un laboratorio o ser observadores de experimentos demostrativos, se fomenta el desarrollo de habilidades que les permitirán abordar nuevos desafíos como los surgidos por el uso de nuevas tecnologías.

Existen pruebas de que la mejor manera de aprender ciencia escolar es practicándola (Jiménez-Liso, 2020), donde las principales prácticas de la ciencia son: la indagación, la modelización y la argumentación, las cuales “están profundamente relacionadas con el desarrollo del pensamiento crítico, ya que promueven habilidades que son fundamentales para la reflexión, el análisis y la toma de decisiones informadas” (Fussero y Occelli, 2024). Estas prácticas científicas, más recomendadas por la investigación didáctica, convergen con las necesidades de la era digital ya mencionadas, porque ayudan a desarrollar el pensamiento crítico, por ejemplo, en argumentación al “poner en duda cualquier afirmación que no esté apoyada en pruebas” (Jiménez-Liso, 2020) y "evaluar enunciados en base a evidencias" (Jiménez-Aleixandre, 2010); en modelización al desarrollar "conocimientos, metaconocimientos, destrezas y valores epistémicos" que les permiten a los estudiantes comprender cómo se genera el conocimiento científico (Oliva, 2019); y en indagación, para que los estudiantes "planifiquen y realicen investigaciones" y utilicen habilidades científicas y sociales para resolver problemas (Crujeiras-Pérez y Jiménez-Aleixandre, 2015).

Tal como se mencionó en la introducción, el pensamiento crítico desempeña un papel fundamental tanto en la ciencia escolar como en las prácticas científicas de indagación, modelización y argumentación, ya que estas contribuyen significativamente a su desarrollo. Por ello, en las secciones siguientes de este marco teórico nos centraremos en el análisis de dichas prácticas.

2.1 Prácticas Científicas

Según Reiser et al. (2012), una práctica científica es un conjunto de habilidades y procesos que los estudiantes deben desarrollar para comprender y aplicar las ideas centrales de la ciencia. Crujeiras-Pérez (2014) señala que el término práctica deriva de la concepción de la ciencia no solo como un conjunto de procesos, sino también como un producto de la interacción social y el discurso que acompaña la construcción del conocimiento científico (National Research Council [NRC], 2012). Estas prácticas no son objetivos de aprendizaje independientes, sino que enfatizan el razonamiento, el discurso y la aplicación en torno a las ideas científicas, y para comprenderlas los estudiantes deben experimentarlas por ellos mismos. Las prácticas mencionadas por el NRC (2012, p.41) son:

- Formular preguntas y definir problemas
- Desarrollar y utilizar modelos
- Planificar y llevar a cabo investigaciones
- Analizar e interpretar datos
- Utilizar las matemáticas y el pensamiento computacional
- Construir explicaciones y diseñar soluciones
- Participar en argumentos a partir de evidencia
- Obtener, evaluar y comunicar información

Estas derivan de aquellas prácticas en las que los científicos e ingenieros participan realmente en su trabajo y son las que el NRC (2012) considera esenciales para el currículo de ciencia e ingeniería en el nivel de enseñanza secundaria, y están diseñadas para guiar el aprendizaje en el aula.

El NRC (2012) también describe un marco conceptual con tres esferas principales de actividad; investigación, también conocida como indagación (en inglés, Inquiry), evaluación, también conocida como argumentación (en inglés, Argumentation), y desarrollo de explicaciones y/o soluciones, llamada modelización (en inglés, Modelling), tal como se muestra en la figura 2.1.A que es una adaptación del NRC (2012).

Este marco conceptual y las ocho prácticas están relacionadas de manera complementaria, la figura 2.1.A proporciona un marco general sobre las actividades fundamentales de científicos e ingenieros, mientras que las ocho prácticas desglosan estas actividades en competencias específicas que los estudiantes deben desarrollar en el aula. De manera que la esfera de la indagación agrupa prácticas como "Formular preguntas y definir

problemas", " Planificar y llevar a cabo investigaciones", y "Analizar e interpretar datos", prácticas fundamentales en la fase de investigación empírica, es decir la indagación. La modelización incluye prácticas como "Desarrollar y usar modelos", "Construir explicaciones y diseñar soluciones", y "Usar matemáticas y el pensamiento computacional". Finalmente, argumentación está relacionada con prácticas como "Participar en argumentación basada en evidencia" y "Obtener, evaluar y comunicar información". Cabe mencionar que en general los investigadores continentales (Ferrés, 2017; Crujeiras, 2014) llaman a las prácticas científicas: indagación modelización y argumentación, y a las ocho prácticas desglosadas, competencias. Las llamaremos así también desde ahora.

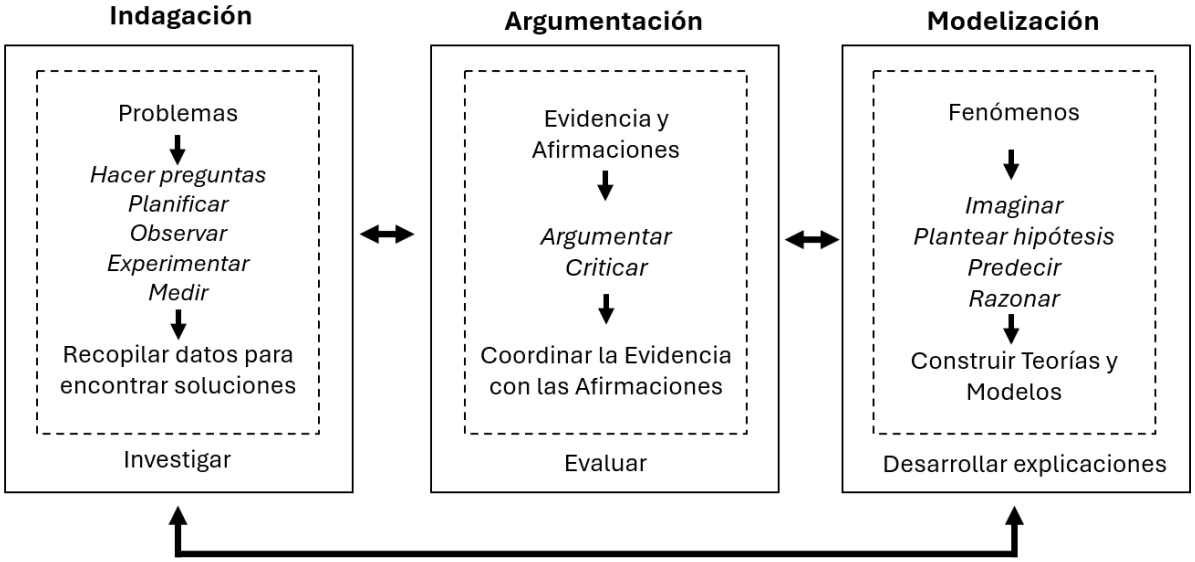

Figura 2.1.A - Modelo de las prácticas científicas publicado por en NRC (2012). Tomado de Jiménez-Aleixandre y Crujeiras-Pérez (2017, p.73).

En la izquierda en la figura 2.1.A se encuentra la indagación. Osborne (2014, p.181), explica que: “La ciencia comienza con una pregunta, que siempre es un subconjunto de la pregunta principal, ¿cómo es la naturaleza?”, por ello una de las prácticas clave en las que se involucran los científicos es la formulación de preguntas. Respecto al lado derecho, modelización, se menciona que: “Las observaciones a su vez generan una pregunta causal de ¿Por qué sucede? Tal pregunta suscita la imaginación de los científicos, la construcción de modelos y la producción de hipótesis explicativas” (Osborne, 2014, p.181). Finalmente, al centro se ubica la argumentación, su presencia en el modelo busca plasmar que la ciencia trata principalmente de ideas, no de datos, y que estas ideas deben ser argumentadas y criticadas para ser consensuadas por la comunidad científica: “¿Cómo sabemos?' y ¿Cómo podemos estar

seguros? Sin embargo, lograr consenso y establecer la validez de tales afirmaciones depende de los argumentos y la evaluación crítica de las pruebas". (Osborne, 2014, p.181).

2.2 La práctica científica de la indagación

La indagación constituye una práctica fundamental en la enseñanza y el aprendizaje de las ciencias escolares. Su implementación promueve un cambio en la forma de generar y validar el conocimiento, alejándose del sentido común y acercándose a prácticas más rigurosas basadas en pruebas. Este enfoque fomenta en el alumnado una actitud activa, autónoma y motivada hacia el aprendizaje científico, favoreciendo la construcción de explicaciones fundamentadas y el desarrollo del pensamiento crítico. Asimismo, la indagación no solo implica el desarrollo de habilidades procedimentales, sino también la comprensión de los métodos y procesos que sustentan el conocimiento científico. Tanto el profesorado como el alumnado deben recorrer este camino reconociendo el papel de las emociones, la importancia de las experiencias auténticas y la necesidad de construir argumentos sólidos basados en evidencias.

La indagación escolar no debe entenderse como una simple reproducción del "método científico" tradicional, ya que las científicas y los científicos no utilizan un único método para investigar. Cada disciplina científica adapta sus procedimientos de acuerdo con la naturaleza de los problemas que aborda. En este sentido, la indagación en el aula se centra en la búsqueda de pruebas, la obtención de datos coherentes y su uso para construir conclusiones fundamentadas, más que en seguir de manera rígida un esquema único de investigación.

Para que la indagación resulte efectiva, el profesorado debe recorrer primero el proceso que desea proponer al alumnado, diseñándolo y secuenciándolo cuidadosamente. Además, no es necesario que el alumnado alcance plena autonomía en todas las fases del proceso: en algunas ocasiones, puede proponer preguntas y diseños de investigación; en otras, puede trabajar a partir de propuestas elaboradas inicialmente por el profesorado. El grado de apertura dependerá de las experiencias previas de los estudiantes, del objetivo de aprendizaje y de la naturaleza del problema planteado. Así, la indagación en ciencias se configura como un proceso flexible, dinámico y adaptativo que articula de manera equilibrada la guía docente y la autonomía progresiva del alumnado.

2.2.1 La indagación en el aula de ciencias

Existen muchos investigadores educativos que han escrito acerca de la indagación desde varios puntos de vista, de manera que el término ha adquirido un uso polisémico en la ciencia escolar Couso, 2014). Existen tres maneras en los cuales se aborda la indagación (Couso, 2014, p.1):

- **Tipo a**: A las capacidades cognitivas que los estudiantes deben desarrollar: la capacidad de "indagar" o "investigar" científicamente.
- **Tipo b**: Lo que es necesario que el alumnado entienda sobre los métodos utilizados por los científicos para dar respuesta a sus preguntas: la naturaleza de la indagación científica.
- **Tipo c**: Una variedad de estrategias de enseñanza y aprendizaje que el profesorado debe desarrollar para que el alumnado aprenda capacidades de indagación (a) y sobre la indagación científica (b), así como para comprender y aprender conceptos científicos.

Para profundizar, Couso (2014, p.2) menciona que el *tipo a* se refiere a las destrezas indagativas (conocidas como *inquiry skills*) que según Arnold et al. (2021, p.3) inicialmente: "[…] eran simplemente un instrumento para ilustrar y consolidar el conocimiento del contenido, y que los alumnos deberían poder seguir e implementar como una serie de instrucciones similares a recetas para un propósito de investigación". Lo anterior, posteriormente sufrió una mejora cuando se agregó al marco el manejo de los conceptos en que se basa la investigación científica a partir de los trabajos de Roberts y Gott (1999); y Millar, Lubben, Gott y Duggan (1995), entre otros.

La segunda manera, *tipo b*, es de naturaleza epistémica, lo que es necesario que el alumnado entienda sobre los métodos utilizados por los científicos, es decir no importa solo lo procedimental sino también se requiere un entendimiento del porque se procede de una cierta forma, enfatizado la importancia de la comprensión procedimental (Arnold et al., 2021, p.3). La importancia de la indagación es que es un medio para: "[…] que el alumnado comprenda cual es la naturaleza de la ciencia (como es y cómo se hace) a partir de su participación en prácticas científicas lo más auténticas posible". Couso (2014) señala que el uso dependerá de la visión o escuela del autor, moviéndose entre dos extremos que pueden ir desde el llamado "método científico" a propuestas más actuales como la ciencia naturalizada, donde la actividad científica clave es "el desarrollo de explicaciones basadas en pruebas sobre cómo funciona el mundo" (Giere, 1991).

Finalmente, el *tipo c*, corresponde a una variedad de estrategias que el profesorado debe desarrollar para que el alumnado aprenda capacidades de indagación, sobre la naturaleza de la indagación, así como para el manejo de conceptos científicos. La literatura se refiere a este enfoque metodológico como "enseñanza de las ciencias como indagación" o "enseñanza de las ciencias centrada en la indagación", conocida como IBSE (*Inquiry-based Science Education*) por sus siglas en inglés (Couso, 2014, p.3), que se propone como alternativa a la escuela tradicional y asociada a otras metodologías como el aprendizaje basado en problemas.

En resumen, como mencionan Simarro et al. (2013, p.36) los *tipos a* y *b* se refieren a contenidos para hacer ciencias y la tercera *tipo c*, se refiere a un enfoque didáctico que permite trabajar y aprender contenidos de y sobre ciencias.

En el *tipo c*, la enseñanza de las ciencias centrada en la indagación se apoya en un ciclo estructurado que permite transformar las ideas iniciales del alumnado en conocimiento basado en pruebas. Las ideas alternativas del alumnado suelen ser de sentido común tendiendo a ser persistentes y, por tanto, el profesorado debe promover formas distintas de validar y generar conocimiento científico. Este enfoque no solo favorece la expresión de ideas personales, sino que guía a los estudiantes en la búsqueda de pruebas para contrastarlas.

La figura 2.2.A representa este ciclo de indagación, compuesto por seis fases que se desarrollan tanto desde la práctica de la indagación, a cargo del estudiante, como desde la secuencia instruccional que está a cargo del profesorado. El proceso se inicia con el reconocimiento de la necesidad de un modelo, donde el profesorado presenta un fenómeno y formula una pregunta guía que logre enganchar al alumnado. Estas preguntas deben ser investigables, no retóricas y relacionadas con fenómenos cercanos que despierten curiosidad. En esta fase inicial, el alumnado contextualiza y se "engancha" con el problema.

Posteriormente, el estudiante expresa y justifica ideas personales, generando una hipótesis o modelos iniciales a través de representaciones gráficas, frases o dibujos. El docente, en paralelo, fomenta esta expresión y facilita el marco para que esas ideas aparezcan con claridad.

En la tercera fase, el alumnado prioriza las pruebas que necesitará para sostener o refutar sus ideas. Aquí, el profesorado propone distintas formas de desarrollar un diseño de obtención de datos que puede ser cerrado, estructurado, guiado o abierto. Luego, en la cuarta fase, el estudiante recopila y expresa datos hipotéticos o empíricos, mientras que el docente proporciona los recursos necesarios y guía su representación y análisis.

Figura 2.2.A - Ciclo de indagación. Tomado de Jiménez-Liso (2020, p.55)

En la quinta fase, los estudiantes evalúan y conectan los datos obtenidos con sus ideas previas, reformulándolas o confirmándolas. El rol del docente consiste en ayudar a transformar esos datos en verdaderas pruebas, dotándolas de sentido dentro del marco de indagación. Finalmente, en la sexta fase, el alumnado construye un conocimiento descriptivo y lo comunica, tomando conciencia de lo aprendido. El profesorado debe facilitar la reflexión final y motivar la explicitación de lo construido.

Este ciclo no solo estructura el proceso de indagación, sino que articula una relación dinámica entre la acción del alumnado y la guía del profesorado. Como se observa en la figura 2.1.A, cada fase incorpora simultáneamente un objetivo didáctico (la práctica de la indagación por parte del alumno) y una fase de instrucción (orquestada por el profesor). Esta doble dimensión asegura que la indagación no sea espontánea ni desorganizada, sino acompañada y estructurada.

La práctica de la indagación en ciencias puede desarrollarse en distintos niveles de autonomía, transitando entre propuestas más guiadas y otras más abiertas. En la indagación guiada, el profesorado estructura el proceso proporcionando preguntas iniciales, orientaciones

para el diseño experimental y apoyos explícitos para la interpretación de resultados. Esta modalidad permite al alumnado concentrarse en aspectos específicos de la indagación científica, fortaleciendo habilidades y estrategias sin la exigencia de tomar decisiones complejas de manera autónoma. En cambio, la indagación abierta otorga mayor protagonismo al estudiante, quien debe formular sus propias preguntas, diseñar procedimientos, recoger datos, analizarlos y construir conclusiones de forma independiente. Esta forma de indagación, aunque más exigente, favorece el desarrollo de competencias científicas avanzadas y un entendimiento profundo de la naturaleza de la ciencia. La elección entre uno u otro enfoque dependerá de la experiencia previa del alumnado, los objetivos de aprendizaje y el grado de complejidad que se desee abordar en la enseñanza.

2.2.2 Estrategias para evaluar el desempeño del alumnado en indagación

En el estudio de la indagación desde el punto de vista *tipo b*, existe una línea de investigación, denominada de naturaleza epistémica. Crujeiras-Perez (2014) investiga operaciones epistémicas, centrándose en cómo los estudiantes desarrollan, refinan y mejoran progresivamente sus concepciones y competencias en indagación a lo largo del tiempo, particularmente a través de la participación en actividades de indagación en el laboratorio, donde evalúa la evolución de habilidades como la identificación de variables, el diseño de investigaciones y la interpretación de resultados, no desde un enfoque de ejecución técnica puntual, sino observando cómo estas operaciones contribuyen al progreso epistémico al permitir una comprensión más profunda y fundamentada del conocimiento científico. En su estudio Crujeiras-Perez (2014) usa como referente a Golding (2012), quien tiene la siguiente concepción acerca de progreso:

> El progreso es un constructo o categoría conceptual útil para comprender la indagación: podemos entender y evaluar una indagación considerando qué tipo de progreso se realiza (o no se realiza) y la magnitud de este progreso (Golding, 2012, p.678).

Luego desarrolla la diferencia entre progreso procedimental, que sería del *tipo a*, y progreso epistémico, que sería del *tipo b*. Señala Golding (2012, p. 679) que al preguntarse "¿Qué tipos de cambio y mejora podrían indicar progreso en una indagación educativa?", los estudiantes pueden mejorar sus habilidades personales, como comparar y contrastar, o dar ejemplos; también pueden desarrollar habilidades grupales en una indagación dialógica, como escuchar con atención o parafrasear de manera más hábil. Por último, podrían avanzar en la

calidad de sus concepciones, respuestas o juicios, por ejemplo, al refinar una pregunta y clarificar posibles respuestas. Este último caso representa una mejora en el contenido sustantivo de una indagación y se diferencia del desarrollo de habilidades o métodos empleados, que Golding (2012) denomina progreso procedimental, que implica mejorar las habilidades y cualidades mentales necesarias para la indagación. Entonces, Golding (2012) establece que el progreso epistémico no es equivalente al progreso procedimental, aunque ambas formas de progreso tienen valor educativo. El trabajo en el aula del progreso epistémico le da mucha más legitimidad al proceso de indagación, ya que implica una mejora en las ideas o concepciones que resultan del proceso.

En su tesis, Crujeiras-Pérez (2014) evalúa operaciones epistémicas con un enfoque en el progreso del contenido epistémico, es decir, en la mejora de las ideas, concepciones y razonamientos a lo largo del tiempo. Esto implica que está interesada no solo en la ejecución técnica, sino en cómo las acciones contribuyen a un avance en la comprensión o en la calidad del conocimiento generado. Para hacerlo desarrolla un estudio longitudinal con tres pequeños grupos experimentales (Crujeiras-Pérez, 2014, p. 101), durante dos cursos académicos consecutivos, para observar cambios en el desempeño de los estudiantes entre 14 y 16 años. Realiza cinco actividades de laboratorio (Determinar qué pasta dental es más efectiva para prevenir caries, Diseñar cómo separar sustancias mezcladas, Identificar qué fábrica contaminaba un río, Establecer dónde disponer residuos químicos de manera segura, Resolver la autoría de una nota anónima) (Crujeiras-Pérez, 2014, p. 150), diseñadas para fomentar la indagación científica y evaluar competencias relacionadas. El progreso lo mide en términos de cómo los estudiantes refinen y mejoran sus concepciones y prácticas a medida que participan en actividades de indagación, si un grupo inicialmente identifica variables incorrectas o incompletas y, después de retroalimentación y práctica, logra seleccionar variables adecuadas y justificar su elección, se considera un progreso epistemológico. Utiliza análisis del discurso para rastrear la calidad y profundidad de las interacciones entre los estudiantes, por ejemplo, con el análisis del discurso puede mostrar si los estudiantes progresan de descripciones simples a explicaciones basadas en evidencia.

Así el progreso lo mide considerando:

- Idoneidad de las propuestas de diseño para resolver las tareas.
- Reducción de la dependencia del profesor durante el proceso.
- Evidencia de cambios cualitativos en las concepciones y estrategias empleadas.

Crujeiras-Pérez (2014, p. 172) elaboró una rúbrica, que se muestra en la tabla 2.2.A para analizar las operaciones epistémicas en tres dimensiones: producción, evaluación y comunicación del conocimiento, dentro de ellas se clasifican las operaciones epistémicas. Cada dimensión se divide en dos categorías que son: (a) Específico, operaciones características de los contextos de indagación, y (b) General, operaciones comunes a todos los contextos de argumentación, modelización e indagación. Aquí sucede que, aunque las operaciones están clasificadas en grupos, algunas de ellas se solapan, a veces para llevar a cabo una es necesario realizar primero otra de otro grupo.

Tabla 2.2.A - Rúbrica de Crujeiras-Pérez (2014) para estudiar el progreso epistémico en indagación

Práctica epistémica	Tipo	Operación
Producción de conocimiento	Específico	Propuesta de diseño
	General	Explicación científica
		Contextualización
		Propuesta de predicción
		Uso de representación
		Definición
		Ejemplificación
Evaluación del conocimiento	Específico	Identificación de variables
		Identificación de entidades problema
	General	Identificación/reformulación del objetivo de la tarea
		Argumentación
		Propuesta/uso de criterios
		Clasificación
		Evaluación de propuestas o productos
		Interpretación/discusión de resultados
		Establecimiento de conclusiones
		Legitimizaciones
Comunicación del conocimiento	Específico	Discusión de propuestas de diseño
		Redacción de propuestas de diseño
	General	Pregunta/respuesta de clarificación
		Traducción entre lenguajes
		Búsqueda de consenso

Esta rúbrica no es una rúbrica tradicional en el sentido de que no presenta cada dimensión con criterios organizados en niveles de desempeño como "bajo", "medio" o "alto". En lugar de ello, Crujeiras-Pérez (2014) utiliza esta rúbrica para identificar y clasificar las operaciones epistémicas realizadas por los estudiantes durante actividades de indagación, agrupándolas según su naturaleza específica (características del contexto de indagación) o general (comunes a múltiples contextos).

Por ejemplo, al analizar la operación epistémica de "identificación de variables", Crujeiras-Pérez (2014) emplea la rúbrica para registrar en qué episodios del discurso estudiantil se realizan acciones que reflejen esta operación, como cuando los estudiantes identifican magnitudes relevantes que cambian durante una investigación, como por ejemplo un grupo de estudiantes que, durante una de las tareas que les planteó, discutieron cómo identificar y controlar las variables, mostrando un proceso de razonamiento colaborativo.

Para evaluar el progreso en una operación epistémica como "identificación de variables", Crujeiras-Pérez (2014) analiza la evolución longitudinal de los episodios registrados. Esto incluye observar si los estudiantes, a lo largo de varias tareas, mejoran su capacidad para identificar variables relevantes de manera autónoma y consistente, reduciendo su dependencia del profesor y aumentando la coherencia y precisión en sus propuestas. Este enfoque permite valorar no solo si se realiza la operación, sino cómo cambia su calidad y complejidad en función de la experiencia acumulada por los estudiantes. De manera que el progreso en cada categoría de la rúbrica (tabla 2.2.A) se determina a lo largo del tiempo y los resultados finales muestran de forma cualitativas que grupos progresan más que otros.

En contraste con lo anterior, otra línea de investigación en indagación corresponde a la del *tipo a*, desarrollada por Ferrés (2017) en su tesis doctoral. Mientras que Crujeiras-Pérez (2014) realiza un estudio longitudinal, analizando el progreso del alumnado en el uso de operaciones epistémicas necesarias para la indagación científica, como la formulación de hipótesis y la evaluación de evidencia, Ferrés (2017) se centra en un análisis transversal, evaluando trabajos de investigación autónoma realizados por estudiantes de bachillerato en un momento determinado y estableciendo el desarrollo competencial en un momento particular. A través de su rúbrica, que llama NPTAI (New Práctical Test Assessment Inventory), identifica dificultades específicas y categoriza las competencias relacionadas con la indagación en dimensiones como la formulación de problemas, identificación de variables, diseño metodológico y análisis de datos.

Existen diferencias entre las rúbricas de ambas autoras. La rúbrica de Crujeiras-Pérez (2014) está diseñada para identificar operaciones epistémicas desde el análisis del discurso. Por otro lado, la rúbrica de Ferrés (2017) tiene un enfoque más práctico y genérico, evaluando directamente el desempeño en investigaciones concretas realizadas por el alumnado y adaptándose a las características del entorno educativo. Mientras que Crujeiras-Pérez (2014) prioriza la comprensión de procesos cognitivos a lo largo del tiempo, Ferrés (2017) busca

evaluar y orientar prácticas de indagación mediante la evaluación de productos finales y el ajuste de las competencias a los estándares del bachillerato.

La rúbrica de Ferrés (2017) tiene una base teórica, principalmente del trabajo de Tamir et al. (1982) quienes desarrollaron el PTAI (Práctical Test Assessment Inventory) que es una rúbrica jerarquizada que consta de veintiuna categorías para evaluar habilidades científicas en el contexto de pruebas de laboratorio y exámenes de acceso a la universidad, que Ferrés (2017) adapta al contexto educativo de la indagación autónoma de estudiantes de bachillerato de Cataluña. Además, integra otros trabajos que describen y evalúan las competencias científicas, como el modelo de Mayer et al. (2008) que define cuatro sub-habilidades de conocimiento procedimental en la indagación científica, y el Framework de PISA (Organisation for Economic Co-operation and Development [OECD], 2013) que destaca las habilidades necesarias para formular preguntas investigables, diseñar investigaciones, interpretar resultados y evaluar conclusiones. La rúbrica NPTAI de Ferrés (2017) ha sido utilizada por diferentes investigadores en distintos contextos y tareas (Mariscal, 2015; Rosa, 2019; Pérez, 2019; Pozuelo y Cascarosa, 2018; Solé Llussà et al., 2017).

En otros trabajos, Crujeiras-Pérez y Cambeiro (2017) y, Crujeiras-Pérez y Cambeiro (2018) también miden habilidades procedimentales del alumnado en distintos contextos, en el mismo sentido que Ferrés (2017) es decir en un momento particular, y no para estudiar el progreso de operaciones epistémicas en un estudio longitudinal. En el primer caso, Crujeiras-Pérez y Cambeiro (2017) piden al alumnado que diseñen una investigación científica para determinar si el detergente Limpics es un fraude. Los estudiantes, organizados en pequeños grupos, deben planificar y poner en práctica una investigación en el laboratorio durante dos sesiones de 50 minutos. Para guiarlos, se les proporciona un conjunto de tarjetas de colores con preguntas relacionadas con aspectos del diseño de la investigación, como la identificación del problema, la formulación de hipótesis, la selección de materiales, el control de variables y el procedimiento. Con base en estas preguntas, los estudiantes elaboran sus diseños, los implementan en el laboratorio y redactan un informe con las pruebas obtenidas, con el objetivo de comprobar la eficacia del detergente y convencer a los consumidores de los resultados. La actividad se centra en introducirlos por primera vez en el proceso de planificación de investigaciones científicas a partir de una cuestión de la vida cotidiana. Además, construyen su rúbrica para analizar las producciones de los estudiantes primero fundamentada en literatura científica y luego refinándola con los datos empíricos obtenidos en su estudio.

Para construir una rúbrica y evaluar el trabajo de los estudiantes, usan tanto literatura teórica como los datos empíricos obtenidos de las respuestas de los estudiantes. El primer paso consistió en la definición de las dimensiones de la rúbrica, basándose en las operaciones de indagación que los estudiantes debían realizar en la actividad (identificación del problema, la formulación de hipótesis, la selección del criterio de medida, la selección de materiales y equipamiento, la identificación de variables, el control de variables, la propuesta de procedimiento y la repetitividad). Cada una de estas dimensiones se corresponde directamente con las preguntas planteadas en las tarjetas de colores utilizadas en la actividad. Posteriormente, desarrollaron categorías iniciales para cada dimensión. Estas categorías emergieron de las respuestas reales de los estudiantes y de la literatura previa. Por ejemplo, en la dimensión de identificación del problema, los desempeños incluyen "indica el problema de forma general" y "avanza un resultado sin identificar el problema"; en la formulación de hipótesis, las categorías consideran si las respuestas son incompletas o si los estudiantes avanzan un resultado en lugar de formular una hipótesis adecuada.

Las categorías iniciales fueron fundamentadas teóricamente en marcos conceptuales establecidos sobre competencias científicas e indagación. Entre las referencias utilizadas se encuentra la OECD (2013), que define la competencia para evaluar y diseñar investigaciones científicas; el trabajo de Jiménez-Aleixandre y Crujeiras-Pérez (2017), que analiza cómo los estudiantes planifican investigaciones; y los estudios de Zimmerman (2000) y Krajcik et al. (1998), que documentan dificultades comunes en el diseño de investigaciones, como la falta de detalle y precisión en las respuestas estudiantiles.

Una vez definidos los desempeños, la rúbrica fue validada y ajustada mediante un análisis empírico de las producciones escritas de los estudiantes. Aplicaron la rúbrica a las respuestas proporcionadas durante la actividad experimental para evaluar si estas encajaban en los desempeños iniciales. Este análisis permitió identificar nuevos desempeños que emergieron de las respuestas estudiantiles, como omisiones o conceptos malentendidos, lo que llevó a realizar ajustes en la rúbrica.

Como parte del proceso, incorporaron un modelo de referencia que define las respuestas ideales para cada dimensión. Por ejemplo, en la dimensión de selección del criterio de medida, el modelo de referencia describe cómo los estudiantes deben comparar la eficacia del detergente Limpics en distintos tejidos y manchas, considerando otras variables controladas. Este modelo sirve como estándar contra el cual se evalúan las respuestas de los estudiantes. Así, la rúbrica fue revisada y utilizada en un análisis multinivel. En el primer nivel, las

respuestas se clasificaron en categorías según su adecuación y nivel de detalle. En el segundo nivel, las respuestas se compararon con el modelo de referencia, evaluando si eran adecuadas, incompletas, inadecuadas o si no había respuesta. La rúbrica final es la de la tabla 2.2.B Una cuestión interesante es que, de las rúbricas estudiadas, esta es la única que considera como criterio la repetitividad de las lecturas hechas en una medición del mismo fenómeno.

Tabla 2.2.B - Rúbrica para evaluar el desempeño en la indagación científica en la tarea "¿Cómo podemos averiguar si Limpics es un fraude?" de Crujeiras-Pérez y Cambeiro (2017).

Dimensión	Categoría
Identificación del problema	- Indica el problema a investigar de forma general - Avanza un resultado sin identificar el problema - Asocia el título de la tarea al problema a investigar
Formulación de hipótesis	- Indica una de las hipótesis posibles, pero incompleta - Avanza un resultado en vez de establecer la hipótesis
Selección del criterio de medida	- Propone analizar la efectividad del detergente en distintos tejidos, sin considerar distintas manchas - Propone analizar la efectividad con distintas manchas, pero no con diferentes tejidos - Propone un criterio general sin indicar cómo llevarlo a cabo
Selección de materiales y equipamiento	- Proporciona algunos materiales e instrumentos, pero sin utilizar terminología científica - No indica materiales e instrumentos
Identificación de variables	- Considera variables aquellas a mantener constantes - Considera variables aquellas a mantener constantes e incluye otras no relacionadas con la investigación - Sin respuesta
Control de variables	- Señala 3 de las 4 variables a mantener constantes - Indica solo una variable - La respuesta no guarda relación con la investigación - Sin respuesta
Propuesta de procedimiento	- Indica algunos pasos a seguir, pero no todos los necesarios para completar el proceso - Considera las operaciones generales de la metodología científica como procedimiento a seguir - La respuesta no guarda relación con la pregunta
Repetitividad	- Indica un mínimo de 3 repeticiones - No especifica un número concreto de repeticiones

En el segundo caso, Crujeiras-Pérez y Cambeiro (2018) diseñaron una actividad basada en la indagación cooperativa para estudiantes de 4º de ESO. A lo largo de esta experiencia, dividen la tarea en cuatro fases, orientadas a promover destrezas científicas y fomentar la colaboración en pequeños grupos. En la fase de preparación, los estudiantes responden preguntas de apoyo para identificar propiedades de las pelotas y planificar el procedimiento experimental, acordando criterios comunes mediante discusiones grupales. Durante la experimentación, realizan mediciones (altura del bote, número de botes, masa y volumen) con

herramientas como probetas y grabaciones a cámara lenta, repitiendo los procedimientos para garantizar la fiabilidad de los datos, que registran en tablas para compartir. En la fase de comunicación, preparan presentaciones orales y digitales para compartir sus hallazgos con la clase, sin una guía para estructurarlas. Finalmente, analizan los datos obtenidos por todos los grupos para justificar cuál es la mejor pelota saltarina, aunque no todos logran establecer conclusiones fundamentadas.

Para determinar los niveles de desempeño en la tarea de cuatro fases mediante una rúbrica que evalúa cinco dimensiones clave: preparación, experimentación y toma de datos, comunicación de resultados, análisis y establecimiento de conclusiones, y trabajo cooperativo. Para cada dimensión establecen tres niveles de desempeño (0, 1 y 2), donde el nivel 0 representa la ausencia o insuficiencia de la acción esperada, el nivel 1 indica un desempeño parcial o con errores, y el nivel 2 corresponde a un desempeño adecuado y completo. Los niveles se asignan en función de criterios específicos, como la capacidad de planificar la investigación, registrar datos de forma precisa, utilizar lenguaje científico en la comunicación, justificar conclusiones basadas en evidencias y colaborar de manera efectiva en los grupos. Este sistema permite evaluar las acciones realizadas por los estudiantes en cada fase de la tarea, identificando tanto fortalezas como áreas de mejora. La rúbrica de Crujeiras-Pérez y Cambeiro (2018) se muestra en la tabla 2.2.C.

Tabla 2.2.C - Rúbrica para evaluar el desempeño en la tarea de indagación cooperativa en estudiantes de 4º de ESO, basada en Crujeiras-Pérez y Cambeiro (2018).

Dimensión	**2**	**1**	**0**
Preparación	Planifica la investigación, pero tiene dificultades para identificar las variables	Planifica la investigación, pero tiene dificultades para identificar las variables y establecer un procedimiento	No planifica la investigación
Experimentación y toma de datos	Registra las observaciones realizadas y toma los datos pertinentes para la investigación de forma adecuada	Registra las observaciones realizadas, pero toma los datos de forma poco adecuada	No registra las observaciones, los datos o ambos
Comunicación de resultados	Utiliza un tono y lenguaje científico correcto y la presentación digital contiene la información adecuada	Presenta un discurso rápido y un lenguaje científico no adecuado. La presentación digital no sigue un orden lógico	Presenta un discurso rápido y un lenguaje científico no adecuado. La presentación digital no es adecuada

Análisis y establecimiento de conclusiones	Analiza los datos y establece una conclusión justificada en base a los resultados obtenidos	Analiza los datos, pero no los utiliza para justificar su conclusión	No establece ninguna conclusión
Trabajo cooperativo	Buena integración en la dinámica de grupos y propone soluciones a los problemas que surgen durante la investigación	Buena integración, pero no propone soluciones a los problemas	Mala integración, y no propone soluciones a los problemas

Un trabajo que explícitamente presenta una distinción entre lo que es habilidades procedimentales y conceptuales es el de Knaggs y Schneider (2012). Ellos las llaman habilidades de proceso (*process skills*) y habilidades conceptuales (*concept skills*). A diferencia del resto explicita con mayor énfasis la diferencia al evaluar ambas por medio de un instrumento.

Knaggs y Schneider (2012) incorporan dos rúbricas que, a primera vista, podrían parecer redundantes debido a la similitud de los criterios utilizados, como "pregunta central", "hipótesis", "procedimiento", "datos y análisis" o "conclusión". Sin embargo, esta duplicación responde a una razón metodológica: la evaluación de dos productos diferentes generados por los estudiantes durante una secuencia didáctica basada en indagación. En concreto, se trató de evaluar por separado el desempeño del alumnado al utilizar un organizador gráfico, la V de Gowin (Novak y Gowin 1984; Novak 1990), y al redactar un informe de investigación formal al finalizar la actividad experimental. La V Gowin también es conocida como diagrama V o Vee heurístico de Gowin, y fue propuesta como una herramienta para ayudar a los estudiantes a organizar y articular el conocimiento científico en procesos de indagación y producción de conocimiento (Novak 1990).

Knaggs y Schneider (2012) explican que su objetivo fue analizar cómo se expresaban las habilidades científicas tanto en el momento de realizar la investigación como en el momento posterior de comunicarla por escrito. Para ello, adaptaron una misma estructura de rúbrica a dos contextos distintos. La primera rúbrica, mostrada en la tabla 2.2.D, está dirigida a evaluar la V de Gowin, donde esta última es entendida como un andamiaje visual y conceptual utilizado durante el diseño y la ejecución del experimento. En ella se recogen elementos como la pregunta central que guía la investigación, el mapa conceptual donde los estudiantes conectan sus conocimientos previos, la formulación de hipótesis y el procedimiento propuesto. También se contempla la organización de los datos recogidos y el valor otorgado a la experiencia

investigativa. Este instrumento evalúa cómo los estudiantes integran saberes conceptuales con el proceso metodológico mientras trabajan, reflexionan y toman decisiones en tiempo real.

En cambio la segunda rúbrica, mostrada en la tabla 2.2.E, se utiliza para valorar el informe de laboratorio escrito tras la experiencia, es decir, la producción textual que sistematiza y comunica los resultados obtenidos. Aunque los criterios son equivalentes en contenido, de hecho, repiten formulaciones similares o idénticas en los distintos niveles de logro, se aplican en una situación comunicativa distinta, en la cual se espera del estudiante una articulación más elaborada y formal de sus comprensiones. La formulación de la pregunta científica, la descripción del procedimiento, la representación de los datos y las conclusiones aparecen aquí en un "formato narrado", con el objetivo de comunicar a un lector externo lo que se hizo y lo que se aprendió.

Ambas rúbricas se derivan del marco teórico "Scientific Discovery as Dual Search" (SDDS), que plantea que la indagación científica combina tres procesos cognitivos: la generación de hipótesis, la planificación experimental y la evaluación de la evidencia. La V de Gowin representa este marco visualmente al dividir el conocimiento en dos lados: el lado conceptual ("knowing") y el lado metodológico ("doing"), lo que permite trabajar tanto los conceptos científicos como las habilidades del proceso de investigación. Así, la rúbrica que evalúa la V de Gowin se centra en cómo los estudiantes articulan estos dos planos durante la planificación y ejecución, mientras que la rúbrica del informe evalúa cómo se sintetizan y comunican esos elementos una vez finalizada la experiencia.

Por lo tanto, la presencia de dos rúbricas busca determinar cómo los estudiantes comprenden y aplican tanto las habilidades del proceso como los conceptos científicos involucrados, y cómo esas comprensiones se manifiestan en diferentes contextos comunicativos. Por lo anterior, podemos inferir que en este sentido es un marco alternativo para evaluar lo que Crujeiras-Pérez (2014) llama conocimiento epistémico.

Ahora, Knaggs y Schneider (2012), por cada una de las dos rúbricas, evalúan por separado los que llaman habilidades de proceso y habilidades conceptuales. Estas son dos dimensiones complementarias pero distintas de las competencias científicas que se espera que los estudiantes desarrollen y demuestren durante experiencias de indagación. Las habilidades de proceso son las asociadas con el hacer ciencia. Es decir, se relacionan con el diseño y ejecución de investigaciones científicas, serían el análogo de lo que Ferrés (2017) llama

habilidades procedimentales y Crujeiras-Pérez (2014) conocimiento procedimental. Knaggs y Schneider (2012) las describen explícitamente como aquellas involucradas en:

- Formular preguntas científicas.
- Desarrollar hipótesis comprobables.
- Diseñar experimentos controlados.
- Identificar y manipular variables (dependientes, independientes y de control).
- Registrar, analizar e interpretar datos experimentales.
- Evaluar evidencia.

Estas habilidades son metodológicas y procedimentales, es decir, constituyen el componente operativo de la indagación. Son las acciones que permiten poner en marcha el proceso de generación de conocimiento científico. En la rúbrica de Knaggs y Schneider (2012), estos elementos aparecen subrayados, porque forman parte del componente "doing" del V-map (lado de la acción) que los autores usan y es una adaptación de la "V de Gowin.

Las habilidades conceptuales están relacionadas con el conocimiento de los conceptos científicos y la capacidad de conectarlos entre sí para dar sentido a un fenómeno. Son las que Crujeiras-Pérez (2014) llama epistémicas. Incluyen según Knaggs y Schneider (2012):

- Reconocer, seleccionar y emplear adecuadamente conceptos científicos relevantes.
- Elaborar mapas conceptuales que estructuren y conecten ideas científicas.
- Utilizar conocimiento previo para interpretar los resultados de una investigación.
- Comunicar la importancia del experimento dentro de un marco conceptual más amplio.

Estas habilidades están más ligadas a la comprensión sustantiva de la ciencia, a lo que Knaggs y Schneider (2012) llama el lado "knowing" del V-map. No son habilidades experimentales en sí mismas, sino capacidades cognitivas que permiten construir sentido a partir de la experiencia de laboratorio.

Tabla 2.2.D - Rúbrica para evaluar el organizador gráfico V-map, tomada de Knaggs y Schneider (2012, p.619).

Habilidades de proceso (procedimentales)
Pregunta central
La pregunta es clara, precisa y comprobable mediante experimentación.
La pregunta es poco clara o vaga, pero comprobable mediante experimentación.
La pregunta es poco clara y no es comprobable.
Hipótesis
Responde claramente a la pregunta planteada.

Responde parcialmente a la pregunta. La hipótesis no es comprobable.
Procedimiento
Las variables y el control están identificados y utilizados correctamente en el experimento. La descripción del experimento es clara. El experimento responderá la pregunta central.
Las variables o el control pueden no estar bien definidos. La descripción es imprecisa; algunos pasos son vagos (puede que no se proporcione información cuantitativa específica). El experimento responderá la pregunta central.
Las variables y el control no están bien definidos o están ausentes. La descripción es incompleta, con pasos faltantes. El experimento no responderá la pregunta central.
Datos y análisis
Tabla de datos clara y bien organizada. El gráfico es claro y los ejes están etiquetados correctamente. Los datos están descritos de manera clara, concisa y precisa.
Hay una tabla de datos presente, pero puede ser poco clara o incompleta. El gráfico está presente, pero puede carecer de etiquetas. Los datos están descritos mínimamente.
La tabla de datos es incorrecta o falta. El gráfico es poco claro, incompleto o falta. Los datos no están descritos o están descritos incorrectamente.
Conclusiones
La hipótesis es evaluada con base en los datos. La pregunta central se responde directamente.
La hipótesis es evaluada con base en los datos. La pregunta central puede no abordarse directamente, pero es respondida.
Puede que la hipótesis no se evalúe con base en los datos. La pregunta central no está correctamente respondida.
Habilidades conceptuales (epistémicas)
Mapa conceptual (Concept map)
Lista completa de palabras clave relevantes utilizadas (6–10+ palabras utilizadas correctamente). Las palabras están correctamente vinculadas. Los niveles y enlaces cruzados están presentes**.**
Algunas palabras utilizadas son relevantes para el laboratorio (3–5 palabras utilizadas de forma correcta). Algunos enlaces son incorrectos. Los niveles están presentes, pero no hay enlaces cruzados.
Lista escasa de palabras utilizadas (0–2 palabras utilizadas correctamente). El mapa conceptual está incompleto o poco claro.
Hipótesis
Está apoyada por conceptos científicos utilizados en el mapa - justificada usando conocimiento científico
Hay evidencia, pero no está conectada con los conceptos presentes en el mapa.
La hipótesis no es comprobable y puede no estar respaldada por evidencia.
Conclusiones
Se hacen conexiones claras entre las afirmaciones conclusivas y los conceptos científicos en el mapa original. El valor del experimento está relacionado con conceptos, eventos o problemas fuera del aula.
Hay algunas conexiones entre conceptos científicos. Se intenta conectar el valor del experimento con conceptos externos, aunque dicha conexión puede no estar clara.
Las conexiones que se hacen entre las conclusiones y los conceptos son escasas o incorrectas.

El valor del experimento no se aborda o no está claro.

Tabla 2.2.E - Rúbrica para evaluar informes de laboratorio como producto final, tomada de Knaggs y Schneider (2012, p.619).

Habilidades de proceso (procedimentales)
Pregunta científica
La pregunta es clara, precisa y comprobable mediante experimentación.
La pregunta es poco clara o vaga, pero comprobable mediante experimentación.
La pregunta es poco clara y no es comprobable.
Hipótesis
Responde claramente a la pregunta planteada.
Responde parcialmente a la pregunta.
Hay evidencia, pero no está conectada con los conceptos previamente mencionados.
Procedimiento
Las variables y el control están identificados y utilizados correctamente en el experimento. La descripción del experimento es clara. El experimento responderá la pregunta central.
Las variables o el control pueden no estar bien definidos. La descripción es imprecisa; algunos pasos son vagos (puede que no se proporcione información cuantitativa específica). El experimento responderá la pregunta central.
Las variables y el control no están bien definidos o están ausentes. La descripción es incompleta, con pasos faltantes. El experimento no responderá la pregunta central.
Datos y análisis
Tabla de datos clara y bien organizada. El gráfico es claro y los ejes están etiquetados correctamente. Los datos están descritos de manera clara, concisa y precisa.
Tabla de datos presente, pero puede ser poco clara o incompleta. El gráfico está presente pero puede carecer de etiquetas. Los datos están descritos mínimamente.
La tabla de datos es incorrecta o falta. El gráfico es poco claro, incompleto o falta. Los datos no están descritos o están descritos incorrectamente.
Conclusión
La hipótesis es evaluada con base en los datos. Se responde directamente la pregunta central, si es posible.
La hipótesis es evaluada con base en los datos. La pregunta central puede no abordarse directamente, pero es respondida.
Puede que la hipótesis no se evalúe con base en los datos. La pregunta central no se responde correctamente.
Habilidades conceptuales (epistémicas)
Conexiones con conceptos aprendidos
El estudiante incorpora correctamente al menos 3 conceptos del capítulo al describir el experimento.
El estudiante incorpora correctamente 1–2 conceptos al describir el experimento.
El estudiante no menciona conceptos, los usa incorrectamente o no los incorpora en la descripción.
Hipótesis
Justificada y respaldada por conceptos científicos mencionados anteriormente.
No es comprobable, no responde a la pregunta.
No está respaldada por evidencia del conocimiento científico aprendido en el capítulo.
Conclusión

Se hacen conexiones claras entre las afirmaciones conclusivas y los conceptos científicos de la introducción. El valor del experimento está relacionado con conceptos, eventos o problemas fuera del aula.
Hay algunas conexiones entre conceptos científicos. Se intenta conectar el valor del experimento con conceptos externos, aunque dicha conexión puede no estar clara.
Las conexiones entre conclusiones y conceptos son escasas o incorrectas. El valor del experimento no se aborda o no está claro.

Como se ha revisado existen diversas rúbricas que permiten caracterizar y medir las habilidades científicas en indagación del alumnado. Entre otras, para esta tesis están las propuestas de: Ferrés (2017), Crujeiras-Pérez y Cambeiro (2017), Crujeiras-Pérez y Cambeiro (2018) y Otero y Crujeiras-Pérez (2016). Se considera valioso el trabajo de Crujeiras-Pérez (2014) pero no será usada su rúbrica, pues esta tesis evaluará un producto final y no un constructo de progreso, pero su inclusión permite matizar la evaluación de la indagación.

Todas las estrategias anteriores para medir el desempeño en indagación del alumnado se pueden complementar en su alcance con enfoques que exploren no solo el desempeño en tareas o criterios específicos como la identificación de variables, sino también en marcos más transversales que detallen tanto la construcción del conocimiento y la forma en que los estudiantes operacionalizan conceptos científicos en sus investigaciones. El trabajo de los autores Gott y Duggan (1995), Gott y Duggan (2002), Roberts y Gott (2003), Gott y Roberts (2008) y Gott et al. (2020), ofrece una perspectiva en este sentido, ya que aborda la relación entre conceptos procedimentales y su evaluación dentro del proceso de indagación y creación de pruebas mediante indagación. Su enfoque permite entender de manera más global y detallada cómo los estudiantes utilizan conocimientos científicos para estructurar investigaciones y diseñar experimentos, lo que sugiere la posibilidad de integrar sus aportes en la construcción de rúbricas más robustas y contextualizadas para evaluar desempeños en indagación.

2.2.2.1 Concepts to Evidence (CoE)

El estudio de Roberts y Gott (2003) y Gott et al. (2020)[1] señala que existe un conjunto de conocimientos que subyacen a la comprensión de las pruebas científicas. Según estos autores, los Concepts to Evidence (CoE) se refieren a un conjunto de ideas que sustentan la

[1] Los trabajos de Gott y Roberts son de principios de los 90, la cita es del 2020 porque corresponde a la página web https://cofev.webspace.durham.ac.uk/, que menciona tiene su última actualización el 11 de diciembre del 2020. Según la norma APA7 se debe citar considerando la última actualización.

recopilación, el análisis y la interpretación de los datos, y que deben entenderse por parte del alumnado antes de poder manejar las pruebas, esto implica por un lado que la evaluación de la indagación podría realizarse evaluando estos COE y por otro lado la idea contenida en el corpus mismo, de que los datos deben ser sometidos a algún tipo de proceso de fiabilidad y validez para que sea posible asignarles un "peso" como prueba, y que luego puedan ser usados como pruebas para la construcción de modelos y/o explicaciones.

El trabajo de Gott y Roberts (2008, p.6) nace en el año 1987, cuando el "Assessment of Performance Unit in Science" de Inglaterra comenzó a evaluar la capacidad de los alumnos para llevar a cabo investigaciones abiertas. Siguiendo la noción de ciencia de Duschl et al. (2007, p.6): "[…] la ciencia se trata fundamentalmente de establecer líneas de pruebas y usar las pruebas para desarrollar y refinar explicaciones usando teorías, modelos, hipótesis, mediciones y observaciones.", los autores señalan que las investigaciones abiertas les permitirán hacerlo: "[…] se espera que también los estudiantes se comprometan con el papel central de las pruebas en la ciencia" (Gott y Roberts, 2008). Estas investigaciones abiertas las definen como aquellas donde el alumnado desconocen la respuesta "correcta", donde hay muchos caminos diferentes para encontrar una solución válida y donde diferentes fuentes de incertidumbre conducen a variaciones en los datos repetidos. La idea es que los estudiantes reflexionen y modifiquen su práctica a la luz de las pruebas que han recogido. El equipo de Gott et al. (2020) han desarrollado todo un trabajo alrededor de la idea de que existe un conjunto de conocimientos que subyace a la comprensión de las pruebas científicas. Este trabajo que se encuentra disponible en el sitio web https://cofev.webspace.durham.ac.uk/ (Gott et al., 2020) y tiene el esquema de la tabla 2.2.F.

Tabla 2.2.F - Resumen de los COE de Gott et al. (2020)

COE	Conceptos y/o ideas
1. Ideas fundamentales: Da los conceptos e ideas básicas sobre investigaciones prácticas	**1.1 Opinión y datos:** Es necesario distinguir entre opiniones basadas en evidencia científica e ideas, por un lado, y opiniones basadas en ideas no científicas (prejuicio, capricho, rumores, etc.) por otro. **1.2 Relaciones o vínculos:** Una investigación científica busca establecer vínculos (y la forma de esos vínculos) entre dos o más variables. **1.3 Asociación y causalidad:** Los vínculos pueden ser causales (un cambio en el valor de una variable causa un cambio en otra) o asociativos (los cambios en una variable y los cambios en otra están vinculados a una tercera, posiblemente no reconocida, variable adicional). **1.4 Tipos de medición:** Los datos de intervalo (mediciones de una variable continua) son más poderosos que los datos ordinales (ordenamiento por rangos), que a su vez son más poderosos que los datos categóricos (una etiqueta).

	1.5 Tareas extendidas: Algunas mediciones, por ejemplo, pueden ser muy complicadas y constituir una tarea por sí mismas, pero solo tienen sentido cuando se enmarcan en investigaciones más amplias de las que formarán parte.
2. Observación: La observación conecta el mundo real con ideas científicas abstractas, generando descripciones y preguntas, sin incluir la medición.	**2.1 Observando objetos:** Los objetos pueden ser "vistos" de manera diferente dependiendo de la ventana conceptual utilizada para observarlos. **2.2 Observando eventos:** Los eventos pueden, de manera similar, ser observados a través de diferentes ventanas conceptuales. **2.3 Uso de una clave:** La forma en que un objeto puede ser "visto" puede estar determinada por el uso de una clave, lo que implica que el observador sigue un conjunto de criterios predefinidos para interpretar lo que está viendo. **2.4 Taxonomías:** Las taxonomías son un medio de usar observaciones impulsadas conceptualmente para establecer clases de objetos u organismos que exhiben características o propiedades similares/diferentes, con el fin de usar la clasificación para resolver un problema. **2.5 Observación y experimento:** La observación puede ser el inicio de una investigación, experimento o estudio. **2.6 Observación y dibujo de mapas:** Técnica utilizada en trabajos de campo biológicos y geológicos para mapear un sitio basado en observaciones impulsadas conceptualmente que ilustran características de interés científico.
3. Medición: contempla la variación inherente por variables no controladas y las cualidades de los instrumentos usados.	**3.1 Variación inherente:** El valor medido de cualquier variable nunca se repetirá a menos que todas las variables posibles estén controladas entre las mediciones, circunstancias que son muy difíciles de crear. **3.2 Error humano:** Como es evidente, el valor medido de cualquier variable puede estar sujeto a error humano, que puede ser aleatorio o sistemático.
4. Relaciones subyacentes: Todos los instrumentos dependen de una relación subyacente que convierte la variable que se está midiendo en otra que es fácil de leer.	**4.1 Relaciones lineales:** La mayoría de los instrumentos se basan en una relación subyacente, y preferiblemente lineal, entre dos variables. **4.2 Relaciones no lineales:** Algunos "instrumentos", por necesidad, se basan en relaciones no lineales. **4.3 Relaciones complejas:** La relación puede no ser directa y puede estar confundida por otros factores. **4.4 Relaciones múltiples:** A veces varias relaciones están vinculadas entre sí para que la medición de una variable sea indirecta.
5. Calibración y error: Los instrumentos deben calibrarse para minimizar las incertidumbres en las lecturas, para que la relación subyacente se refleje con precisión en la escala, si la relación es no lineal, la escala debe calibrarse con mayor frecuencia para representar esa no linealidad. Todos	**5.1 Puntos finales:** el instrumento debe calibrarse en los puntos finales de la escala. **5.2 Puntos intermedios:** el instrumento debe calibrarse en los puntos intermedios para verificar la linealidad de la relación subyacente. **5.3 Errores de cero:** puede haber un desplazamiento sistemático en la escala, y los instrumentos deben verificarse regularmente. **5.4 Sobrecarga, sensibilidad límite / límite de detección:** existe un valor máximo (desviación completa de la escala) y una cantidad mínima que puede medirse de manera confiable con un instrumento y técnica dados. **5.5 Sensibilidad:** la sensibilidad de un instrumento es una medida de la cantidad de error inherente al propio instrumento. **5.6 Resolución y error:** la resolución es la división más pequeña que puede leerse fácilmente. La resolución puede expresarse como un porcentaje. **5.7 Especificidad:** un instrumento debe medir solo aquello que pretende medir.

los instrumentos están sujetos a errores y cada instrumento tiene límites finitos.	**5.8 Uso del instrumento:** existe un procedimiento prescrito para usar un instrumento que, si no se sigue, provocará errores sistemáticos y/o aleatorios. **5.9 Error humano:** incluso cuando se elige y utiliza un instrumento de manera adecuada, puede ocurrir un error humano.
6. Fiabilidad y validez de una única medición	**6.1 Fiabilidad:** una medición fiable requiere un promedio de una serie de lecturas repetidas; el número necesario depende de la precisión requerida en cada circunstancia particular. **6.2 Fiabilidad:** los instrumentos pueden estar sujetos a inexactitudes inherentes, por lo que usar diferentes instrumentos puede aumentar la fiabilidad. **6.3 Fiabilidad:** los errores humanos en el uso de un instrumento pueden superarse mediante verificaciones independientes y aleatorias. **6.4 Validez:** las mediciones que dependen de relaciones complejas o múltiples deben asegurar que están midiendo lo que pretenden medir.
7. La elección de un instrumento para medir un dato: Las mediciones nunca son completamente precisas por diversas razones. Es de suma importancia elegir el instrumento que proporcione la exactitud y precisión requeridas; una elección proactiva en lugar de darse cuenta, después de haber tomado mediciones o realizado el experimento, de que el instrumento elegido no era el adecuado, lo que puede invalidar los datos obtenidos.	**7.1 Veracidad o exactitud:** la veracidad es una medida del grado en que las lecturas repetidas de la misma cantidad dan un promedio que es igual al promedio "verdadero". **7.2 No repetibilidad:** las lecturas repetidas de la misma cantidad con el mismo instrumento nunca dan exactamente la misma respuesta. **7.3 Precisión:** la precisión se refiere a las variaciones observadas en las mediciones repetidas realizadas con el mismo instrumento. En otras palabras, la precisión es una indicación de la dispersión de las mediciones repetidas alrededor del promedio. Una medición precisa es aquella en la que las lecturas se agrupan estrechamente. Cuanto menor sea la precisión del instrumento, mayor será su incertidumbre. Una medición precisa no necesariamente será una medición exacta o verdadera (y viceversa). El concepto de precisión también se denomina "fiabilidad" en algunos campos. Un descriptor o evaluación más formal de la precisión podría ser el rango de las lecturas observadas, la desviación estándar de esas lecturas o el error estándar del propio instrumento. **7.4 Reproducibilidad:** mientras que la repetibilidad (precisión) se relaciona con la capacidad del método para dar el mismo resultado en pruebas repetidas de la misma muestra en el mismo equipo (en el mismo laboratorio), la reproducibilidad se relaciona con la capacidad del método para dar el mismo resultado en pruebas repetidas de la misma muestra en equipos de diferentes laboratorios. **7.5 Valores atípicos en las relaciones:** los valores atípicos, aberrantes o anómalos en los conjuntos de datos deben examinarse para descubrir posibles causas. Si una medición o dato aberrante puede explicarse por procedimientos de medición deficientes (cualquiera que sea la fuente del error), entonces puede eliminarse.
8. Muestreo de un dato: Una serie de mediciones del mismo dato puede utilizarse para determinar la fiabilidad de la medición.	**8.1 Muestreo:** una o más mediciones constituyen una muestra de todas las mediciones que podrían realizarse. **8.2 Tamaño de la muestra:** el número de mediciones tomadas. Cuanto mayor sea el número de lecturas realizadas, más probable es que sean representativas de la población. **8.3 Reducción del sesgo en la muestra / muestreo representativo:** las mediciones deben tomarse utilizando una estrategia de muestreo adecuada, como el muestreo aleatorio, estratificado o sistemático, de manera que la muestra sea lo más representativa posible. **8.4 Un dato anómalo:** un dato inesperado podría ser indicativo de variación inherente en los datos o la consecuencia de una variable no controlada reconocida.

	Gott et al. (2003) usa el término "muestreo" para referirse a cualquier subconjunto de una "población". La "población" podría ser la población de una especie de animal o planta. También consideraremos la población como el número infinito de lecturas repetidas que podrían tomarse de cualquier medición en particular.
9. Tratamiento estadístico de mediciones de un único dato: Un grupo de mediciones del mismo dato puede describirse de diversas maneras matemáticas. El tratamiento estadístico de un dato se ocupa de la probabilidad de que una medición esté dentro de ciertos límites de la lectura verdadera.	**9.1 Rango:** el rango es una descripción simple de la distribución y define los valores máximo y mínimo medidos. **9.2 Moda:** la moda es el valor que ocurre con mayor frecuencia. **9.3 Mediana:** la mediana es el valor por debajo y por encima del cual se encuentran la mitad de las mediciones. **9.4 Media:** la media (promedio) es la suma de todas las mediciones dividida por el número de mediciones. **9.5 Distribuciones de frecuencia:** una serie de lecturas del mismo dato puede representarse como una distribución de frecuencia agrupando mediciones repetidas que caen dentro de un rango dado y trazando las frecuencias de las mediciones agrupadas. **9.6 Desviación estándar:** la desviación estándar (SD) es una forma de describir la dispersión de datos distribuidos normalmente. Indica qué tan estrechamente las mediciones se agrupan alrededor de su media. En otras palabras, es una medida del grado en que las mediciones se desvían de su media. Cuanto más se agrupen alrededor de la media, menor será la desviación estándar. La desviación estándar depende del instrumento y técnica de medición: cuanto más precisos sean, menor será la desviación estándar de la muestra o de las mediciones repetidas. **9.7 Desviación estándar de la media (error estándar):** la desviación estándar de la media describe la distribución de frecuencia de las medias de una serie de lecturas repetidas muchas veces. Depende del instrumento de medición, la técnica y el número de repeticiones. El error estándar de una medición es una estimación del rango probable dentro del cual cae la "media verdadera"; es decir, una estimación de la incertidumbre asociada al dato. **9.8 Coeficiente de variación:** el coeficiente de variación es la desviación estándar expresada como un porcentaje de la media (CV = SD*100/media). **9.9 Límites de confianza:** los límites de confianza indican el grado de confianza que puede asignarse a un dato. Por ejemplo, "límite de confianza del 95%" significa que el "dato verdadero" se encuentra dentro de 2 errores estándar de la media calculada el 95% del tiempo. De manera similar, "límite de confianza del 68%" significa que el "dato verdadero" se encuentra dentro de 2 errores estándar de la media calculada el 68% del tiempo.
10. Fiabilidad y validez de un dato	**10.1 Fiabilidad:** un dato solo puede ser considerado como evidencia una vez que se haya determinado la incertidumbre asociada con el instrumento y los procedimientos de medición. **10.2 Validez:** una medición debe ser del dato adecuado o permitir el cálculo de este.
11. Estructura de las variables: Identificar y comprender la estructura básica de una investigación en términos de las variables y sus tipos	**11.1 La variable independiente:** la variable independiente es aquella cuyos valores son cambiados o seleccionados por el investigador. **11.2 La variable dependiente:** la variable dependiente es aquella cuyo valor se mide para cada cambio en la variable independiente. **11.3 Variables correlacionadas:** en algunas circunstancias buscamos únicamente una correlación, en lugar de una causalidad implícita.

ayuda a evaluar la validez de los datos.	**11.4 Variables categóricas:** una variable categórica tiene valores que se describen mediante etiquetas. Las variables categóricas también se conocen como datos nominales. **11.5 Variables ordenadas:** una variable ordenada tiene valores que también son descripciones, etiquetas o categorías, pero estas categorías pueden ser ordenadas o clasificadas. La medición de variables ordenadas resulta en datos ordinales. **11.6 Variables continuas:** una variable continua es aquella que puede tener cualquier valor numérico y cuya medición resulta en datos de intervalo. **11.7 variables discretas:** una variable discreta es un caso especial en el que los valores de la variable están restringidos a múltiplos enteros. **11.8 Diseños multivariados:** una investigación multivariada es aquella en la que hay más de una variable independiente.
12. Validez, 'pruebas justas' y controles: La variación no controlada puede reducirse mediante una variedad de técnicas.	**12.1 Prueba justa:** es aquella en la que solo se ha permitido que la variable independiente afecte a la variable dependiente. **12.2 Variables de control en el laboratorio:** otras variables pueden afectar los resultados de una investigación a menos que sus efectos se controlen manteniéndolas constantes. **12.3 Variables de control en estudios de campo:** algunas variables no pueden mantenerse constantes y lo único que se puede hacer es asegurarse de que cambien de la misma manera. **12.4 Variables de control en encuestas:** el efecto potencial sobre la validez de las variables no controladas puede reducirse seleccionando datos de condiciones que sean similares con respecto a otras variables. **12.5 Experimentos con grupos de control:** los grupos de control se utilizan para garantizar que los efectos observados se deban a las variables independientes y no a alguna otra variable no identificada. No son más que el valor predeterminado de la variable independiente.
13. Elegir valores	**13.1 Prueba preliminar:** una prueba preliminar puede usarse para establecer los parámetros generales requeridos para el experimento (escala, rango, número) y ayudar en la elección de instrumentación y otros equipos. **13.2 La muestra:** las cuestiones de tamaño de muestra y representatividad se aplican de la misma manera que en el muestreo de un dato (ver Medición de un dato). **13.3 Escala relativa:** la elección de valores sensatos para las cantidades es necesaria si las mediciones de la variable dependiente han de ser significativas. **13.4 Rango:** el rango sobre el cual se eligen los valores de la variable independiente es importante para garantizar que se detecte algún patrón. **13.5 Intervalo:** la elección del intervalo entre valores determina si el patrón en los datos puede identificarse o no. **13.6 Número:** es necesario un número suficiente de lecturas para determinar el patrón.
14. Exactitud y precisión	**14.1 Determinando diferencias:** existe un nivel de precisión suficiente para proporcionar datos que permitan discriminar entre dos o más mediadas. **14.2 Determinando patrones:** existe un nivel de precisión requerido para determinar la tendencia en un patrón.
15. Tablas	**15.1 Tablas:** las tablas pueden usarse como organizadores para el diseño de un experimento al preparar la tabla antes de todo el experimento. Una tabla tiene un formato convencional. Las tablas pueden usarse para diseñar un experimento antes de la recopilación de datos y, de esta manera, contribuir a su validez. De este

	modo, las tablas pueden ser mucho más que un simple medio para presentar datos después de haber sido recopilados.
16. Fiabilidad y validez del diseño	**16.1 Fiabilidad del diseño:** la fiabilidad del diseño incluye una consideración de todas las ideas asociadas con la medición de cada uno de los datos. **16.2 Validez del diseño:** la validez del diseño incluye una consideración de la fiabilidad (como se mencionó anteriormente) y la validez de cada uno de los datos.
17. Presentación de datos: Existe un vínculo estrecho entre las representaciones gráficas y el tipo de variable que representan.	**17.1 Tablas:** una tabla es un medio para reportar y mostrar datos. Pero una tabla por sí sola presenta información limitada sobre el diseño de una investigación, por ejemplo, las variables de control o las técnicas de medición no siempre se describen de manera explícita. **17.2 Gráficos de barras:** los gráficos de barras pueden usarse para mostrar datos en los que la variable independiente es categórica y las variables dependientes son continuas. **17.3 Gráficos de líneas:** los gráficos de líneas pueden usarse para mostrar datos en los que tanto la variable independiente como la dependiente son continuas. Permiten la interpolación y extrapolación. **17.4 Gráficos de dispersión (o diagramas de dispersión):** también pueden usarse para mostrar datos en los que tanto la variable independiente como la dependiente son continuas. Los gráficos de dispersión se utilizan a menudo donde hay mucha fluctuación en los datos, ya que permiten detectar una asociación. Los puntos dispersos ampliamente pueden mostrar una correlación débil, mientras que los puntos agrupados alrededor, por ejemplo, de una línea pueden indicar una relación. **17.5 Histogramas:** los histogramas pueden usarse para mostrar datos en los que una variable independiente continua ha sido agrupada en rangos y en los que la variable dependiente es continua. **17.6 Diagramas de caja y bigotes**: la caja, en los diagramas de caja y bigotes, representa el 50% de los datos delimitados por el percentil 25 y el percentil 75. La línea central es la mediana. Los límites de los "bigotes" pueden mostrar los extremos del rango o los valores del 2,5% y 97,5%. **17.7 Datos multivariados:** los gráficos de barras 3D y los gráficos de líneas (superficies) son adecuados para algunas formas de datos multivariados. **17.8 Otras formas de representación:** los datos pueden transformarse, por ejemplo, a escalas logarítmicas para que cumplan con los criterios de normalidad, lo que permite el uso de estadísticas paramétricas.
18. Tratamiento estadístico de mediciones de datos: Existen numerosas técnicas estadísticas para analizar datos que abordan tres preguntas principales: ¿Los dos grupos de datos difieren entre sí (solo por probabilidad)?	**18.1 Diferencias entre medias:** se puede usar una prueba t para estimar la probabilidad de que dos medias de poblaciones normalmente distribuidas, derivadas de una investigación que involucra una variable independiente categórica, sean diferentes. Es decir, ¿cuál es la probabilidad de que las dos medias hayan ocurrido por casualidad? Si las mediciones se repiten con las mismas muestras o pares emparejados, se puede usar una prueba t pareada. **18.2 Análisis de varianza:** el análisis de varianza es una técnica que puede usarse para estimar los efectos de un número de variables en un problema multivariado que involucra variables independientes categóricas. **18.3 Regresión lineal y no lineal:** la regresión puede usarse para derivar la "línea de mejor ajuste" para los datos resultantes de una investigación que involucra una variable independiente continua. **18.4 Medidas no paramétricas:** cuando las mediciones no están normalmente distribuidas, se pueden usar pruebas no paramétricas, como la prueba U de Mann-Whitney, para estimar la probabilidad de cualquier

¿Los datos cambian cuando se toman mediciones repetidas en una segunda ocasión separada? ¿Existe una asociación entre dos conjuntos de datos?	diferencia. **18.5 Datos categóricos:** cuando los datos resultan de una investigación en la que tanto las variables independientes como dependientes son categóricas, el análisis de los datos debe usar, por ejemplo, una prueba de chi-cuadrado.
19. Patrones y relaciones en los datos Los datos deben ser inspeccionados en busca de patrones subyacentes. Los patrones representan el comportamiento de las variables, por lo que no pueden tratarse de forma aislada del sistema físico que representan. Los patrones pueden observarse en tablas o gráficos, o pueden informarse utilizando los resultados de un análisis estadístico adecuado. La interpretación de patrones y relaciones debe respetar las limitaciones de los datos: generalización excesiva o de implicar causalidad cuando puede haber un tipo de asociación diferente y menos directa.	**19.1 Tipos de patrones:** existen diferentes tipos de asociaciones como causales, consecuentes, indirectas o asociaciones por azar. "Asociación por azar" significa que las diferencias observadas en los conjuntos de datos, o los cambios en los datos a lo largo del tiempo, ocurren simplemente por azar. Debemos estar escépticamente abiertos a la posibilidad de que un patrón haya surgido por azar. Las pruebas estadísticas nos dan una forma racional de estimar esta probabilidad. **19.2 Relaciones lineales:** las relaciones de línea recta (pendientes positivas, negativas, verticales y horizontales como casos especiales) pueden estar presentes en los datos en tablas y gráficos de líneas, y tales relaciones tienen un importante poder predictivo ($y = mx + c$). **19.3 Relaciones proporcionales:** la proporcionalidad directa es un caso particular de relaciones de línea recta con características predictivas consecuentes. La relación a menudo se expresa en la forma ($y = mx$). **19.4 Curvas 'predecibles':** los patrones pueden seguir curvas predecibles ($y = x^2$, por ejemplo), y tales patrones probablemente representan regularidades significativas en el comportamiento del sistema. **19.5 Curvas complejas:** algunos patrones pueden modelarse matemáticamente para proporcionar aproximaciones a diferentes partes de la curva. **19.6 Relaciones empíricas:** los patrones pueden ser puramente empíricos y no representarse fácilmente mediante una relación matemática simple. **19.7 Datos anómalos:** los patrones en tablas o gráficos pueden mostrar puntos de datos anómalos que requieren mayor consideración antes de excluirlos de análisis futuros. **19.8 Línea de mejor ajuste:** para gráficos de líneas (y gráficos de dispersión en algunos casos) se puede usar una "línea de mejor ajuste" para ilustrar la relación subyacente, suavizando parte de la variación inherente (no controlada) y el error humano.
20. De datos a pruebas – fiabilidad y validez de toda la investigación	**20.1 Una serie de experimentos:** una serie de experimentos puede aumentar la fiabilidad y validez de la evidencia incluso si, individualmente, su precisión no permite dar mucho peso a los resultados de un solo experimento. **20.2 Datos secundarios:** los datos recopilados por otros son una fuente valiosa de evidencia adicional, siempre que su valor como evidencia pueda ser juzgado. Por ejemplo, metaanálisis. **20.3 Triangulación:** la triangulación con otros métodos puede fortalecer la validez de la evidencia.
21. Problemas sociales relevantes: Las pruebas deben considerarse a la luz	**21.1 Credibilidad de la evidencia:** la credibilidad tiene mucho que ver con la validez aparente: consistencia de la evidencia con ideas convencionales, con el sentido común y con la experiencia personal. La credibilidad aumenta con el grado de consenso científico sobre la evidencia

de la experiencia personal y social, así como del estatus de los investigadores. Si nos enfrentamos a las pruebas y queremos llegar a un juicio, entonces otros factores también entrarán en juego.	o sobre teorías que apoyan la evidencia. La credibilidad también puede depender del tipo de evidencia presentada, por ejemplo, evidencia estadística versus evidencia anecdótica. **21.2 Practicidad de las consecuencias:** las implicaciones de la evidencia pueden ser prácticas y rentables, o pueden no serlo. Cuanto más imprácticas o costosas sean las implicaciones, mayor será la demanda de estándares más altos de validez y fiabilidad de la evidencia. **21.3 Sesgo del experimentador:** la evidencia debe ser examinada para detectar sesgos inherentes de los experimentadores. El posible sesgo puede deberse a fuentes de financiamiento, rigidez intelectual o una lealtad a una ideología como el cientificismo, el fundamentalismo religioso, el socialismo o el capitalismo, por nombrar algunos. El sesgo también está directamente relacionado con el interés: ¿quién se beneficia? ¿quién se ve perjudicado? **21.4 Estructuras de poder:** la evidencia puede recibir un peso indebido o descartarse demasiado a la ligera, simplemente por su importancia política o debido a cuerpos influyentes. La confianza a menudo puede ser un factor aquí. A veces, las personas se ven influenciadas por eventos pasados de pérdida de confianza por parte de agencias gubernamentales, voceros de la industria o grupos de interés especial. **21.5 Paradigmas de práctica:** diferentes investigadores pueden trabajar dentro de diferentes paradigmas de investigación. Por ejemplo, los ingenieros operan desde una perspectiva diferente a la de los científicos. Por lo tanto, la evidencia obtenida dentro de un paradigma puede adquirir un estatus diferente cuando se ve desde otro paradigma de práctica. **21.6 Aceptabilidad de las consecuencias:** la evidencia puede ser negada o descartada por razones que pueden parecer ilógicas, como el miedo público y político a sus consecuencias. Aquí juegan un papel los prejuicios y las preconcepciones. **21.7 Estatus de los experimentadores:** el estatus académico o profesional, la experiencia y la autoridad de los experimentadores pueden influir en el peso que se otorga a la evidencia. **21.8 Validez de las conclusiones:** las conclusiones deben limitarse a los datos disponibles y no ir más allá mediante generalizaciones, interpolaciones o extrapolaciones inapropiadas.

Para Gott et al. (2020) un dato es la unidad básica de información y representa el resultado de una medición. Puede tratarse de un único valor o del promedio de múltiples lecturas realizadas para aumentar la precisión, como, por ejemplo, el volumen de un gas o la masa de un objeto. La medición es el proceso de determinar un valor cuantitativo o cualitativo de un parámetro específico mediante un instrumento o técnica, y puede implicar múltiples lecturas del instrumento para mejorar la confiabilidad. Mientras que prueba es un estatus superior que logra adquirir un dato. Los datos se convierten en pruebas cuando han sido sometidos a un proceso de fiabilidad, validación y análisis crítico. Esto incluye evaluar la calidad de los datos, las condiciones del experimento y su reproducibilidad en diferentes contextos.

El marco conceptual propuesto por Gott et al. (2020) ofrece una estructura para entender cómo los datos se transforman en pruebas dentro del contexto de la ciencia y la ingeniería. Este

enfoque aborda la definición de los conceptos clave (tabla 2.2.F) y los pasos involucrados en este proceso jerárquico y estructurado, que se ilustra en la figura 2.2.B.

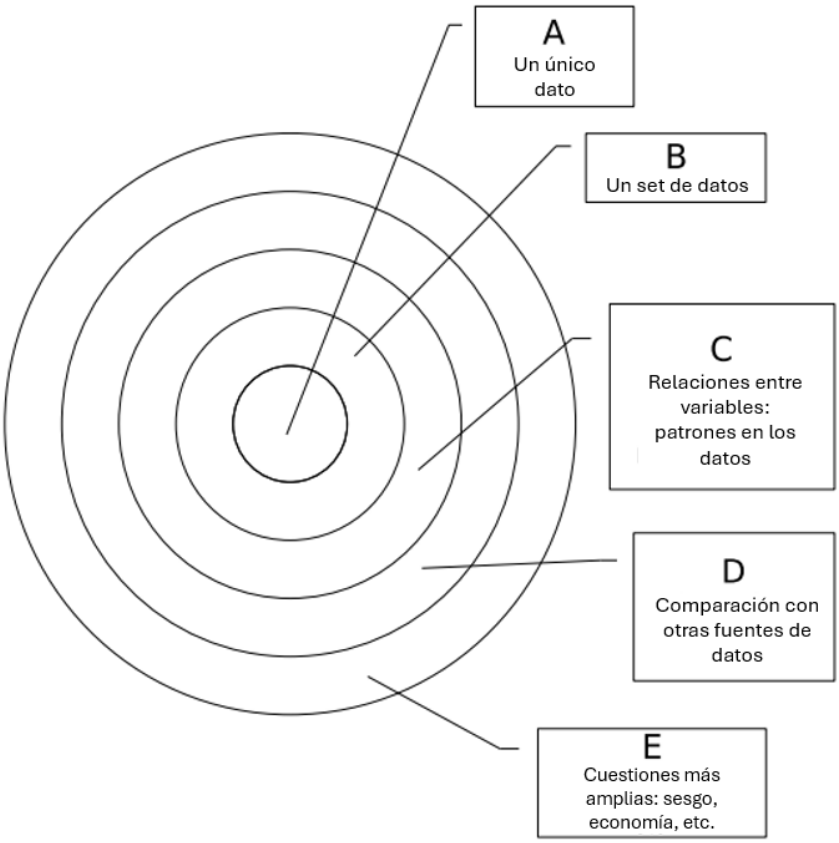

Figura 2.2.B - Representación de las capas de análisis en la recopilación y evaluación de datos en investigaciones científicas (Gott y Roberts, 2008, p.14).

Esta representa el proceso en círculos concéntricos, donde cada nivel (denotados por A, B, C, D, E) se construye sobre el anterior:

1. Una medición individual (A): El punto de partida es una sola medición que proporciona un valor inicial para un parámetro específico.
2. Conjunto de mediciones (A): Un dato puede incluir múltiples mediciones o lecturas, cuyo promedio mejora la precisión y confiabilidad.
3. Conjunto de datos (B): La reunión y organización de varios datos constituye el conjunto sobre el cual se realizarán análisis más amplios.
4. Relaciones entre datos (C): Los datos individuales se analizan para identificar patrones o relaciones significativas, lo que permite un entendimiento más profundo del sistema estudiado.
5. Transformación en pruebas (D): Los datos validados, sometidos a evaluación crítica y reproducibilidad, se convierten en pruebas que respaldan conclusiones específicas.

6. Comparación con otros datos (D): Las pruebas generadas se contrastan con datos de otras investigaciones para contextualizar los hallazgos y reforzar su validez.
7. Implicaciones sociales (E): En el nivel más externo, las pruebas se interpretan considerando su impacto social, ético y cultural. Esto incluye evaluar su relevancia en un contexto amplio y las posibles consecuencias de su aplicación.

Este modelo jerárquico destaca cómo un dato inicial, generado a partir de una medición, puede evolucionar a través de un proceso sistemático para convertirse en una prueba significativa, relevante tanto para la ciencia como para la sociedad.

2.2.2.2 Fiabilidad y validez en una investigación científica

Según Couso y Jiménez-Liso (2020) en una investigación científica se deben tener en cuenta criterios de fiabilidad y validez al momento de realizar la construcción del conocimiento. Diversos autores en didáctica han expresado definiciones para estos dos criterios, pero en general semejantes entre ellas. Gott y Roberts (2008) afirman que la perspectiva de la comprensión de ideas y conceptos que subyacen a las pruebas representa una forma diferente de conceptualizar el componente procedimental del currículo. Se requiere que el alumnado construya significado, específicamente en torno a la validez y la fiabilidad, a partir de ideas específicas sobre las pruebas, en lugar de considerarlas simplemente como habilidades que se desarrollan implícitamente mediante la práctica. Estas ideas pueden aplicarse y sintetizarse en investigaciones abiertas, junto con las ideas sustantivas tradicionales de la ciencia.

Así, los CoE buscan especificar la base de conocimiento procedimental que sostiene la recopilación y evaluación de pruebas, es decir, lo que el alumnado necesita saber para juzgar la validez y fiabilidad de una investigación, ya sea propia o reportada por otros. Estas ideas (tabla 2.2.F) constituyen un set de herramientas conceptuales esenciales para planificar y realizar investigaciones prácticas con comprensión, en lugar de hacerlo como un procedimiento mecanizado. No se debe dejar de lado, que Gott y Roberts (2008) consideran necesarios los CoE, pero no suficientes, y que para una indagación científica son también necesarios las habilidades manuales, el conocimiento tácito descrito por Polanyi (1966) y las ideas sustantivas tradicionales de la ciencia (de contenido).

La misma figura 2.2.B sirve para explicar como la fiabilidad y validez se van ejecutando en cada capa del proceso. Un investigador debe asegurar la fiabilidad y validez en parte del proceso de indagación si quiere formular y defender una afirmación que luego estará basada en

esa prueba empírica. Dado que los datos están en el centro de las investigaciones científicas, primero se comienza de ese punto como partida (Gott y Roberts, 2008):

- Para cada dato individual, se debe considerar la fiabilidad y validez de cualquier observación o medición realizada. Si un dato es inválido o poco fiable, toda la investigación queda comprometida.
- Para un set de datos, es necesario evaluar si se han realizado suficientes repeticiones para capturar la variabilidad y garantizar la confiabilidad en los datos, así como si estos han sido resumidos de manera válida.
- Al examinar relaciones entre variables, debe considerarse la validez y fiabilidad del diseño de la investigación y la interpretación de los datos, teniendo en cuenta cualquier incertidumbre derivada del proceso de recolección de datos.
- Al comparar con otras fuentes de datos, se debe juzgar la validez y fiabilidad de los estudios previos que hayan influido en el diseño o desarrollo de la investigación.

Debido a lo anterior, corresponde profundizar algo en como los conceptos de fiabilidad y validez se van trabajando en el enfoque de Gott y Roberts (2008) y Gott et al. (2020), primero describen la validez y fiabilidad de las mediciones que se realizan en un experimento. Por fiabilidad de la medida explican que:

- Una medición fiable requiere un promedio de varias lecturas repetidas; el número necesario depende de la precisión requerida en las circunstancias particulares. (donde, precisión se refiere a las variaciones observadas en mediciones repetidas del mismo instrumento, cuantificada como la dispersión alrededor de la media). Hay que cuestionarse siempre si un dato tiene suficiente precisión, lo que implica que cuanto mayor sea la incertidumbre, menos fiable será el dato. Una medición precisa es aquella en la que las lecturas se agrupan muy juntas. Debido a que fiabilidad implica muchas mediciones con una cierta precisión, esto implica considerar que los instrumentos pueden estar sujetos a inexactitudes inherentes, por lo que el uso de diferentes instrumentos puede aumentar la fiabilidad; y que el error humano en el uso de un instrumento puede superarse mediante comprobaciones aleatorias e independientes de una medición.

Respecto a la validez se señala que:

- La validez garantiza que se está midiendo lo que pretenden medir. Por lo anterior señalan hay que cuestionarse: ¿se ha medido el valor de la variable apropiada? ¿Se ha muestreado el parámetro de modo que el dato represente a la población?

Lo anterior implica que para los autores fiabilidad está asociada netamente al valor que se le asigna a una variable, por ello importa ejecutar acciones orientadas a ello: varias lecturas repetidas y aumentar la precisión. Esto último lleva a minimizar el error humano, a ocuparse en seleccionar y usar el instrumento adecuadamente y la comprobación aleatoria e independiente de la misma medida. En cambio, validez implica garantizar que se está midiendo lo que se pretende medir, lo clave es que es válido si se ha muestreado el parámetro de modo que el dato represente a la población.

Luego, extienden estos dos conceptos para hablar de la fiabilidad y validez del diseño experimental. Al hacerlo señalan que hay dos preguntas generales: ¿Las mediciones darán como resultado datos suficientemente fiables para responder a la pregunta? ¿El diseño dará como resultado datos suficientemente válidos para responder la pregunta?

Las preguntas se contestan si sal seguir las mismas orientaciones, las mediciones darán como resultado datos suficientemente fiables si en cada una de las mediciones se ha medido reiteradas veces, tomado la media como representante, definido una tolerancia para la precisión, minimizado el error humano, analizado el rango y el intervalo de las mediciones, usado el instrumento apropiado y realizado comprobaciones aleatorias e independientes. Por otro lado, el diseño experimental dará resultados suficientemente válidos si este permite muestrear el parámetro de modo que el dato represente a la población.

Finalmente, señalan que se debe garantizar la fiabilidad y validez en toda la investigación, no solo a nivel de mediciones individuales, sino en el conjunto de evidencias recopiladas a lo largo del proceso. Para ello, destacan tres estrategias fundamentales: serie de experimentos, uso de datos secundarios y triangulación, definidas en la tabla 2.2.F con la numeración 20.1, 20.2 y 20.3, respectivamente.

El enfoque de Gott et al. (2020) es fundamental para la investigación empírica, pero también existen otras perspectivas sobre la fiabilidad y validez en la literatura científica. En la siguiente sección, exploraremos otros tipos de validez y fiabilidad, como la validez de constructo, la validez de contenido y otras formas de evaluar la calidad de una investigación en distintos enfoques metodológicos.

2.2.2.3 Otras conceptualizaciones de fiabilidad y validez

Si bien el enfoque de Gott et al. (2020) sobre fiabilidad y validez fundamenta la investigación empírica y experimental, en la literatura científica se han desarrollado otras perspectivas para evaluar la calidad de una investigación. Estas conceptualizaciones han surgido en respuesta a la diversidad de enfoques metodológicos utilizados en distintas disciplinas y a la necesidad de establecer criterios de evaluación más amplios, especialmente en estudios donde la medición directa de variables es compleja.

Las investigaciones empíricas, como las que estudia Gott et al. (2020), se caracterizan por la recolección de datos a partir de experimentos o mediciones observables, donde la fiabilidad y validez dependen en gran medida de la precisión instrumental, el control de variables y la replicabilidad de los resultados. Sin embargo, en otros campos de la investigación los fenómenos estudiados no siempre pueden medirse con la misma exactitud, ya que involucran conceptos abstractos y constructos teóricos. En estos casos, la validez y fiabilidad se amplían para abarcar dimensiones que permitan evaluar la calidad de los instrumentos de medición y la coherencia interna de los modelos teóricos empleados.

Para Martínez (2006) la validez es "es el grado en que un instrumento de medida mide lo que realmente pretende o quiere medir; es decir, lo que en ocasiones se denomina exactitud". Por otro lado, validez es "el criterio para valorar si el resultado obtenido en un estudio es el adecuado". Algo diferente a Gott et al. (2020) es que plantean tipos de validez definidas por Martínez (2006) como:

- *Validez de contenido*: grado en el cual la medición empírica refleja un dominio específico del contenido.
- *Validez de criterio*: comparación entre la medida de la investigación y otra medida estándar que se denomina criterio y de la cual se conoce su validez. En tal caso, existen varios tipos de criterios: 1) concurrente: instrumento y estándar medidos a la vez, y 2) predictiva: instrumento y estándar no son medidos a la vez.
- *Validez de constructo*: medida en que una variable es abstracta y latente, más que concreta y observable, se denomina constructo, porque no existe una dimensión (variable) observable. Por lo tanto, la medida de un constructo se obtiene al combinar los resultados de diversas medidas. De este modo, existen dos tipos de validez de constructo: 1) validez convergente: es el grado en que dos o más intentos de medir el mismo concepto están de acuerdo entre sí y se determina con la aplicación del análisis

factorial confirmatorio, y 2) validez discriminante: grado en el que un concepto difiere de otros y se determina con el coeficiente Phi del análisis factorial confirmatorio.

Para Hernández-Sampieri y Fernández-Collado (2016) la validez de contenido hace referencia al: "Grado en que un instrumento refleja un dominio específico de contenido de lo que se mide". La validez de criterio es aquella que se establece al correlacionar las puntuaciones resultantes de aplicar el instrumento con las puntuaciones obtenidas de otro criterio externo que pretende medir lo mismo. Si el criterio se fija en el presente de manera paralela, se habla de validez concurrente (los resultados del instrumento se correlacionan con el criterio en el mismo momento o punto de tiempo). Si el criterio se fija en el futuro, se habla de validez predictiva.

Para Martínez (2006) la fiabilidad se refiere a: "la consistencia interna de la medida; es decir que la fiabilidad de una medida analiza si ésta se halla libre de errores aleatorios y, en consecuencia, proporciona resultados estables y consistentes". Además, Martínez (2006) categoriza métodos para obtener tipologías de fiabilidad que son:

- Método de Aplicaciones Repetidas: consiste en la medición repetitiva de las variables, con el fin de determinar hasta qué punto un conjunto de medidas es reproducible en el tiempo. En tal sentido, fiabilidad sería sinónimo de estabilidad; es decir, el grado en que las puntuaciones son estables, sería el grado de fiabilidad.
- Método de Formas Paralelas: se emplea para medir el grado de acuerdo entre los observadores o más evaluadores que valoran a los mismos sujetos con el mismo instrumento; es decir, la coherencia que existe entre palabras, órdenes o respuestas diferentes.
- Método de División en Mitades: mide la coherencia interna de una escala cuando se divide la muestra en dos mitades. Se estudia la correlación entre cada uno de los grupos.
- Método de Coherencia Interna: mide la coherencia entre todos los ítems de una misma escala y no se puede usar en aquellas medidas que usan pocos ítems.

2.2.2.4 La validez y la fiabilidad de un instrumento de medida

La disciplina denominada metrología, aborda cuestiones como la calibración o la verificación de un instrumento de medida que son importantes para determinar su validez y fiabilidad. La calibración se define como una operación que, bajo condiciones especificadas, establece una relación entre los valores y sus incertidumbres, obtenidos a partir de los patrones de medida, y utiliza esta información para establecer una relación que permita obtener un

resultado de medida a partir de una indicación (Centro español de Metrología, 2012, p.37). En contraste, la verificación es un proceso posterior a la calibración, que implica confirmar o comprobar si un instrumento o equipo de medición cumple con ciertos requisitos especificados. Durante esta etapa, se compara el rendimiento del instrumento con estándares o criterios predefinidos para determinar si está funcionando dentro de los límites aceptables. Generalmente, estas inspecciones son llevadas a cabo por laboratorios sujetos a estándares estatales regulados, los cuales dan cuenta de la validez y fiabilidad del instrumento. En metrología un instrumento tiene asociado los conceptos de exactitud referido a validez, y precisión referida a fiabilidad (Centro español de Metrología, 2012, p.31):

- La exactitud es un concepto cualitativo que se refiere a la proximidad entre un valor medido y un valor verdadero de un mensurando, y suele cuantificarse con el error de medida o absoluto.
- La precisión es la proximidad entre las indicaciones o los valores medidos obtenidos en mediciones repetidas de un mismo objeto, o de objetos similares, realizadas en condiciones similares o a las mediciones de este mensurando hechas por diferentes personas. Esta dispersión se expresa de manera más adecuada en términos de una función de distribución de probabilidad, o en términos de parámetros que miden la dispersión de una distribución de probabilidad: (a) Para variables numéricas, se suele calcular la varianza para su determinación, y (b) para variables de tipo dicotómica, se usa el porcentaje de acierto, matriz de confusión o medidas más sofisticadas como el coeficiente alfa de Cronbach.

2.3 La práctica científica de la argumentación

La argumentación en didáctica de las ciencias es una línea muy extensa y con mucho desarrollo, esto se correlaciona con el hecho que en el área pura del estudio de la argumentación:

> Aún no existe una teoría de la argumentación en el sentido de teoría como cuerpo establecido y sistemático de conocimientos al respecto; la denominación teoría de la argumentación más bien designa un campo de estudios, por más señas interdisciplinarios (Vega, 2013, p.97).

Lo que implica que existen distintas perspectivas para abordarla. Según Marruad (2014, p.5) la argumentación puede verse como:

- Una función del lenguaje que estudiaría la lingüística,

- Un proceso cuyo estudio corresponde a la retórica,
- Un procedimiento cuyo estudio compete a la dialéctica,
- Como un producto cuyo estudio corresponde a la lógica.

Cuando se examinan los propósitos aparecen nuevamente tres perspectivas clásicas sobre la argumentación. Según Marruad (2017, p.2) argumentar es:

- Tratar de persuadir o convencer a alguien de algo por medio de razones (retórica).
- Tratar de lograr el asentimiento o el compromiso de alguien con algo por medio de razones (dialéctica).
- Justificar ante alguien una pretensión de validez por medio de razones (lógica).

Podríamos por lo tanto encontrar definiciones puras basadas en la retórica, dialéctica o lógica, y otras que las combinan.

En didáctica de las ciencias, según expresa Aduriz-Bravo en la conferencia "La Argumentación basada en modelos. Perspectivas teóricas" (Canal Asociación Docentes de Ciencias Biológicas, 2016, 21m40) existen al menos cinco líneas de argumentación en didáctica que lista como:

- Los estudios de carácter cognitivo realizados por la Dra. Deanna Kuhn como una forma de dialogo personal, del departamento de Psicología y Educación del Colegio de Profesores de la Universidad de Columbia.
- La línea de hablar y escribir ciencias representada por el grupo LIEC de la Universidad Autónoma de Barcelona.
- La línea de competencias argumentativas y uso de pruebas como formadora de ciudadanía y pensamiento crítico, principalmente por la Dra. Marilar Jiménez-Aleixandre de la Universidad de Santiago de Compostela.
- Los estudios de etnografía del aula sobre la interacción discursiva en entornos multiculturales como los de la Dra. Antonia Candela en México.
- El reconocimiento de una gran cantidad de estudios metateóricos sobre el lenguaje de la ciencia, por ejemplo, los trabajos del Dr. Maurice Finocchiaro.

Podemos señalar, que la argumentación aporta a la educación científica, en ciertas dimensiones que Erduran y Jiménez-Aleixandre (2008, p.5) y que Revel (2012, p.5) resumen en:

- Hacer públicos los procesos cognitivos (Revel, 2012, p.5; Erduran y Jiménez-Aleixandre, 2008, p.6)

- Desarrollar el pensamiento crítico (Revel, 2012, p.6; Erduran y Jiménez-Aleixandre, 2008, p.7)
- Las habilidades para hablar y escribir ciencias (Revel, 2012, p.7; Erduran y Jiménez-Aleixandre, 2008, p.8)
- La inmersión en la cultura científica y el desarrollo de prácticas epistémicas (Revel, 2012, p.8; Erduran y Jiménez-Aleixandre, 2008, p.9)

Que se relacionan con la lista anterior dada por Aduriz-Bravo (Canal Asociación Docentes de Ciencias Biológicas, 2016, 21m40).

En la línea de investigación didáctica hablar y escribir ciencias (Márquez, 2005) señala diferencias entre acciones relacionadas con la argumentación. Definen que:

- Describir consiste en producir enunciados que enumeren cualidades, propiedades o características de un objeto, organismo o fenómeno.
- Explicar implica producir razones o argumentos de manera ordenada, estableciendo una relación causa-efecto entre los elementos.
- Justificar se refiere a la producción de razones o argumentos en relación con un corpus de conocimiento o teoría que sustente una afirmación.
- Argumentar es el proceso de generar razones o argumentos con la finalidad de convencer a otros sobre la validez de una postura o idea.

Entre ellas existen las siguientes diferencias, describir se enfoca en enumerar o identificar características de algo observable, sin hacer inferencias ni establecer relaciones causales. Explicar va más allá de la descripción al buscar relacionar hechos y establecer conexiones causales entre ellos. Justificar implica conectar hechos o fenómenos con teorías o modelos ya establecidos, es decir, relaciona la explicación con un marco teórico. Argumentar, por su parte, es un proceso más amplio, ya que no solo justifica o explica, sino que también busca convencer a un interlocutor sobre la validez de una opción o interpretación frente a otras.

Se desprende de las definiciones que tanto la explicación, justificación y argumentación se basan en razones o evidencia. Todas requieren algún tipo de razonamiento lógico. Justificar y argumentar implican referirse a una base teórica o de conocimiento; sin embargo, la argumentación introduce la idea de la competencia entre puntos de vista. Todas estas acciones son actividades cognitivas que implican la organización de la información para lograr una comprensión más profunda o convencer a otros.

Si se ordenan jerárquicamente, describir es el nivel más básico ya que es el primer paso, se limita a enumerar características observables de un fenómeno sin realizar interpretaciones

ni establecer causas. Es un nivel más básico de análisis, ya que no involucra razonamiento complejo, solo observación. Explicar ocupa un nivel superior porque implica ir más allá de la simple descripción al relacionar los hechos observados y establecer una conexión entre causas y efectos. Aquí se busca entender el "por qué" detrás de lo observado. Justificar sigue a explicar, porque además de ofrecer una explicación, requiere que se conecte esa explicación con una teoría o marco de referencia. No solo se busca entender el "por qué", sino también demostrar cómo ese "por qué" encaja dentro de un conjunto de conocimientos previos. Argumentar está en el nivel más alto, ya que involucra no solo la justificación o explicación, sino también la capacidad de defender un punto de vista frente a otros, con el fin de convencer. La argumentación requiere un razonamiento crítico, que implique considerar y refutar puntos de vista alternativos, lo que la hace más compleja que justificar o explicar.

Así, la didáctica de las ciencias reconoce que construir argumentos es un proceso complejo que requiere orientación y oportunidades para generar explicaciones sustentadas en pruebas. Esta perspectiva promueve un enfoque cíclico, donde el alumnado no solo expone sus ideas, sino que aprende a justificarlas, confrontarlas y comunicarlas adecuadamente. En la figura 2.3.A se presenta un ciclo de argumentación articulando fases específicas tanto para el profesorado como para el alumnado.

El ciclo se representa dividido en seis fases. En el centro, se destaca la meta común: construir argumentos. El ciclo se apoya en una interacción entre las acciones del alumnado y la guía del profesorado. Mientras el profesorado diseña situaciones auténticas, ofrece criterios y guía la interpretación de los datos, el alumnado identifica explicaciones, genera datos, evalúa alternativas y comunica sus ideas de forma clara. Se incluyen fases instruccionales, ejecutadas por el profesorado, y fases de práctica, protagonizadas por el alumnado. Las fases de la instrucción consisten en diseñar proyectos significativos, guiar el diseño experimental, orientar la interpretación de datos, apoyar la conexión entre datos y teorías, promover el uso de criterios de calidad argumentativa y enseñar cómo comunicar de forma clara. En cambio, el alumnado desarrolla la argumentación recorriendo seis pasos interrelacionados: (1) identificar explicaciones u opciones alternativas; (2) generar datos a través de observaciones o experimentos; (3) evaluar las alternativas en función de los datos; (4) identificar teorías relevantes y elaborar justificaciones; (5) evaluar argumentos contrarios al propio; y (6) comunicar de forma clara el argumento construido.

Figura 2.3.A - Ciclo de argumentación. Tomado de Jiménez-Aleixandre (2020, p.78).

Desde la instrucción, en la fase 1 el profesorado propone tareas auténticas que invitan a explorar múltiples explicaciones posibles; el alumnado, en respuesta, identifica explicaciones alternativas. En la fase 2, el docente guía el diseño de los experimentos u observaciones, mientras que los estudiantes recogen datos para respaldar o refutar sus ideas. En la fase 3, se orienta la interpretación de los datos, lo que permite que el alumnado valore críticamente las alternativas posibles. En la fase 4, el profesorado ayuda a conectar los datos con teorías científicas, y los estudiantes justifican sus ideas a partir de esos marcos teóricos. En la fase 5, se promueve el uso de criterios argumentativos, lo que habilita al alumnado a comparar su postura con argumentos contrarios. Finalmente, en la fase 6, el profesorado explicita cómo comunicar con claridad, y el estudiante comunica su argumento de forma comprensible, estructurada y fundamentada.

2.3.1 La argumentación en el aula de ciencias

Existen varios marcos teóricos que trabajan la argumentación y cada una de las líneas didácticas mencionadas anteriormente podría usarlos, estos son:

- Nueva retórica de Chaim Parelman
- Pragma-dialéctica de Franz van Eemeren y Rob Grootendorst
- Lógica sustantiva de Stephen Toulmin
- Lógica natural de Jean-Blaise Grize
- Lógica informal de Douglas Walton

Sin embargo, entre los marcos revisados, solo se han encontrado investigaciones en educación científica escolar que se basan en la lógica sustantiva e informal. En los artículos analizados, los enfoques de la "nueva retórica" y la "pragma-dialéctica" no se abordan, posiblemente debido a la tendencia predominante de estudiar los productos argumentativos finales del alumnado, en lugar de centrarse en el proceso que conduce a dichos productos. En cuanto al marco de la lógica natural, se considera que aún no ha surgido un investigador pionero que haya incorporado su aplicación en este contexto.

De las líneas de investigación revisadas anteriormente, hay tres que trabajan con el Modelo de Toulmin, “Estudios de carácter cognitivo”, “Hablar y escribir ciencias” y “Competencias argumentativas y uso de pruebas”. Desde la revisión de artículos en revistas de didáctica, se ha encontrado sólo un artículo que no usa un modelo distinto al de Toulmin, ese artículo usa los esquemas de Walton para estudios dialécticos de argumentación (Erduran y Jiménez, 2009, p.168). Por lo anterior, en esta tesis se utilizará también el Modelo de Toulmin, dado que la mayoría de los estudios revisados coinciden en emplearlo para analizar el producto argumentativo final del alumnado en lugar de los procesos intermedios que conducen a dicho argumento. Además, existen diversos estudios que emplean el Modelo de Toulmin para desarrollar rúbricas o progresiones de aprendizaje en argumentación. Estos últimos trabajos proporcionan guías para evaluar el nivel de desempeño argumentativo, basado en las componentes del modelo de Toulmin.

El modelo de Toulmin es autoría del filósofo y epistemólogo inglés Stephen Toulmin (1958), tiene seis componentes que forman su estructura y cumplen determinadas funciones dentro de la argumentación. El diagrama del modelo y sus componentes se muestra en la Figura 2.3.B.

La función que cumplen cada componente del modelo en la argumentación se puede ilustrar con la Figura 2.3.C, allí se muestra la relación de las componentes con las preguntas que realiza un “retador” hipotético que desafía a quien presenta una argumentación. Es interesante ver cómo mediante la interrogación se va construyendo la estructura y función de

cada componente, como se profundiza de la justificación al conocimiento básico y como aparecen las componentes pruebas, calificadores y refutación en función del desafío. El ejemplo, ¿es Harry o no un ciudadano británico? corresponde al presentado por el mismo Toulmin (2007, p.142) para la deducción de su modelo.

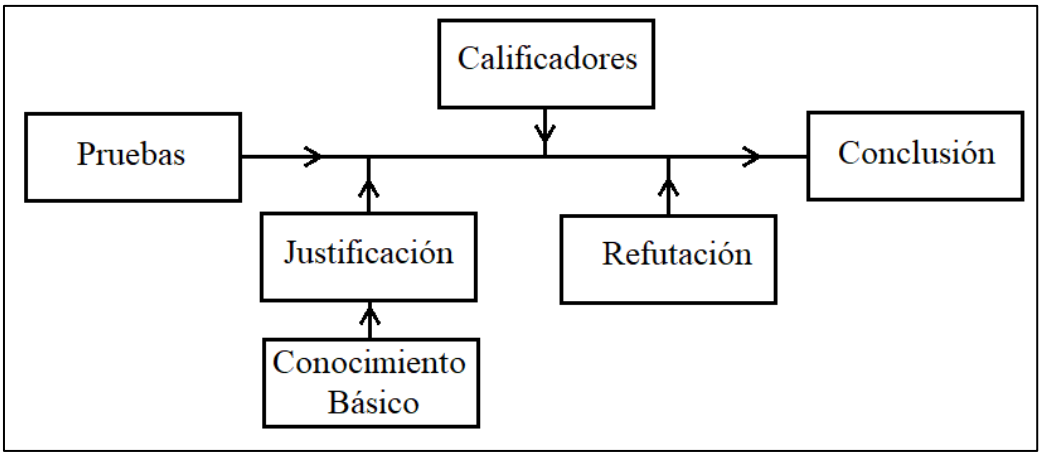

Figura 2.3.B - Modelo argumentativo de Toulmin. Adaptado de Los usos de la argumentación de Toulmin (2007, p.141). Para cada componente se ha usado la traducción de Jiménez-Aleixandre (2010).

Conclusión C	Harry es Británico	¿En qué se apoyas?
Prueba P	Harry nació en Bermudas	¿Cómo se llega a esa conclusión?
Justificación J	Un hombre nacido en Bermudas será un súbdito británico.	¿Es necesariamente así?
Calificadores M	Presumiblemente es así	¿Cuándo no se aplica la regla general?
Refutación R	Si sus padres son extranjeros o si se ha convertido en un estadounidense naturalizado (o ha adquirido otro tipo de nacionalidad que excluye ser súbdito británico).	¿Qué te autoriza entonces a concluir de P (el hecho de que alguien haya nacido en Bermudas) a C (que se puede presumir que es súbdito británico)?
Conocimiento Básico B	Está incorporado en la siguiente legislación: …	

Figura 2.3.C - Componentes argumentativas en función de un "retador" hipotético.

Algunos investigadores educativos en sus artículos nombran ciertas componentes del modelo de manera distinta entre sí y al original (Toulmin, 1958). Por ejemplo, la traducción de

cada componente usada en la línea de investigación de competencias argumentativas y uso de pruebas (Jiménez-Aleixandre et al., 2010) es: Conclusión (Claim en inglés), Pruebas (Evidence en inglés), Justificación (Warrant en inglés), Conocimiento básico (Backing en inglés), Calificadores modales (Modals en inglés) y Refutación (Rebuttals en inglés).

Las definiciones que da Jiménez-Aleixandre (2010) a cada componente son las siguientes. La componente Conclusión la define como el enunciado de conocimiento que se pretende probar o refutar (Jiménez-Aleixandre, 2010, p.70). Añade que hay distintas formas en que se podría referir el estatus de estos enunciados, siendo este término el más general. Cuando el enunciado no está explicado entonces si las ideas sobre sus posibles causas son expresadas, se habla de hipótesis.

A lo anterior se puede agregar que no siempre se establece en argumentación una distinción entre conclusión e hipótesis (Jiménez-Aleixandre, 2010). Así muchas veces cuando se hace explicita una hipótesis, la argumentación completa gira en torno a probar si esta es verdadera o falsa, lo cual constituye una buena manera de identificarla en una argumentación.

Para la componente Prueba señala que: "hablamos de datos para referirnos a informaciones, magnitudes, cantidades, relaciones o testimonios con el fin de llegar a la solución de un problema o la comprobación de un enunciado" (Jiménez-Aleixandre, 2010, p.72). Afirma que por lo general se piensa en datos en términos de cifras, pero a veces también son informaciones cualitativas no reducibles a números. Además, distingue entre datos hipotéticos, que son los que se entregan al alumno en una actividad o que no son medidos y datos empíricos, que son los que el propio alumnado obtiene (Jiménez-Aleixandre, 2010, p.73). La diferencia entre prueba y dato señala, es bastante sutil, y radica en el contexto de uso: "Lo que hace que nos refiramos a algo como prueba es su función o papel en la evaluación del enunciado" (Jiménez-Aleixandre, 2010, p.74), razón por la cual se entenderá por Prueba: observación, hecho o experimento al que se apela para evaluar un enunciado (Jiménez-Aleixandre, 2010, p.72).

Por la componente Justificación, esta se define como el elemento del argumento que relaciona la conclusión o explicación con las pruebas (Jiménez-Aleixandre, 2010, p.75). Agrega que hay que tener en cuenta que para el profesorado puede ser obvio que las pruebas tratan de confirmar un enunciado y que no necesita explicarse, pero que la experiencia muestra que a menudo para el alumnado no es así. Por lo anterior, para el alumnado la justificación puede ser implícita, y a veces hay que buscarla en el discurso pues no está de forma explícita.

Esto suele suceder cuando se trata de algo conocido por todos los interlocutores, un conocimiento compartido que se da por supuesto (Jiménez-Aleixandre, 2010, p.75).

Para Jiménez-Aleixandre (2010, p.77), el Conocimiento Básico es la apelación a conocimientos teóricos o empíricos, a modelos, leyes o teorías que respaldan la justificación, dándole mayor solidez al argumento. Por ejemplo, si una Conclusión es “Este metal es conductor" y la Prueba usada es "Este metal es cobre", la Justificación, que puede ser implícita o explicita sería "Los metales, en general, son conductores". Aquí, la Justificación es una regla general que conecta el hecho de que el metal es cobre con la conclusión de que es conductor. Si esta es criticada, se debería apelar a un conocimiento básico (backing), en este caso: "La teoría de conducción eléctrica establece que los metales permiten el paso de electricidad debido a la movilidad de sus electrones.".

Las últimas dos componentes son: Calificadores (también llamados calificadores modales) y Refutación. Los Calificadores y Refutación son definidos respectivamente como; “los calificadores expresan el grado de certeza o incertidumbre del argumento” (Jiménez-Aleixandre, 2010, p.77) y “las refutaciones son el reconocimiento de las restricciones o excepciones que se aplican a la conclusión” (Jiménez-Aleixandre, 2010, p.79).

Revisaremos las definiciones dadas por el mismo Toulmin (2007, p.132). Para Toulmin una aseveración, “Conclusión” (Jiménez-Aleixandre, 2010), es la sentencia que se ha hecho. Si está es puesta en duda, hay que apoyarla, probarla y demostrar que estaba justificada. Luego menciona:

> Por consiguiente, como punto de partida contamos ya con una distinción establecida: entre la aseveración o afirmación o conclusión cuyo valor estamos tratando de establecer y los elementos justificatorios que alegamos como base de la aseveración realizada, a los que me referiré como las pruebas (Toulmin, 2007, p.133).

Toulmin (2007) señala que es posible que no se nos solicite aportar nueva información factual adicional a la ya presentada, sino que se nos demande señalar cómo se relacionan los datos proporcionados con la conclusión alcanzada. Así una Justificación, que él llama garantía, es la componente que permite conectar las pruebas con la aseveración (conclusión o afirmación). Específicamente, afirma que la justificación es una proposición general o regla que legitima el paso de las pruebas a la aseveración. Señala que no se trata de agregar más información o pruebas a un argumento, sino de ofrecer un principio que permita inferir la aseveración a partir de las pruebas: “lo que se necesita son enunciados hipotéticos, de carácter

general, que actúen como puente entre unos y otras, legitimando el tipo de paso que el argumento en particular que hemos enunciado nos obliga a dar" (Toulmin, 2007, p.134). Es decir, la justificación es el puente lógico que da el por qué la conclusión es válida a partir de las pruebas proporcionadas. Entonces (Toulmin, 2007):

- La justificación no aporta nuevos datos: "El objetivo no consiste ya en reforzar la base sobre la que hemos elaborado nuestro argumento, sino en mostrar cómo a partir de esos datos hemos pasado a la afirmación original o conclusión" (Toulmin, 2007). Esto indica que la función de la justificación no es proporcionar información adicional, sino mostrar que el razonamiento entre las pruebas y la conclusión es válido.
- Las justificaciones son reglas generales: Toulmin (2007) explica que las garantías son enunciados generales, como "Si P, entonces C o "Pruebas tales como, permiten extraer conclusiones como C". Este tipo de estructura permite inferir la conclusión a partir de los datos de manera lógica y consistente.
- La relación entre pruebas y justificación puede ser cuestionada: Toulmin (2007) plantea que puede surgir la duda sobre si lo que se está cuestionando es la prueba o la justificación: "¿se puede trazar una distinción clara entre los datos y la justificación?". Esto implica que la legitimidad de una justificación puede ser objeto de debate, lo que lleva a la necesidad de ofrecer respaldo adicional. En este punto aparece la componente "Conocimiento básico".

En resumen, la justificación en el Modelo de Toulmin son proposiciones generales o reglas que justifican el paso de las pruebas a la aseveración, permitiendo hacer inferencias sin agregar nueva información.

Luego, Toulmin (2007) señala que, si la justificación es criticada, aparece la componente "Conocimiento Básico" que el de forma más general llama Respaldo (backing). Esta corresponde al soporte teórico que valida la justificación. Según Toulmin (2007) son enunciados categóricos sobre hechos. Toulmin (2007) establece que los respaldos suelen ser enunciados de tipo categórico porque para defender la garantía, se recurre a principios, reglas o conocimientos aceptados como válidos en una comunidad. Estos son formulados de manera categórica, es decir, como afirmaciones generales que se consideran verdaderas, sin condición o limitación inmediata. Un enunciado categórico es simplemente un enunciado que afirma algo de manera general, como "El calor mata bacterias" o "Los metales se expanden al calentarse".

No es hipotético ("si pasa tal cosa, entonces tal otra") ni condicional. Es una afirmación rotunda que sirve como base para sostener que la justificación es razonable.

Así la diferencia clave para Toulmin (2007) es que la justificación es el principio general que justifica por qué, dadas ciertas pruebas, se puede llegar a la conclusión. No necesariamente tiene que ser un conocimiento científico detallado, sino una afirmación que explique la lógica subyacente. En cambio, el respaldo (conocimiento básico) es el soporte adicional que fortalece la justificación y le da más validez.

Finalmente, Toulmin (2007) señala que los calificadores y refutación son por su propia naturaleza distintos de la justificación, y que suponen un comentario implícito a la importancia de la justificación e: "indican la fuerza conferida por la garantía en el paso adoptado, mientras que las condiciones de refutación apuntan las circunstancias en que la autoridad general de la garantía ha de dejarse a un lado" (Toulmin, 2007, p.137).

Es importante que en los debates que enfrentan dos explicaciones opuestas, se entiende por refutación la crítica a las pruebas del adversario. Así calificadores modales y refutación no caen directamente dentro del proceso de justificación en el sentido estricto que se ha discutido. Los calificadores modales no son parte central del proceso de justificación en sí, ya que no se enfocan en proporcionar razones o conexiones teóricas entre los datos y la afirmación, sino en el nivel de seguridad con el que se realiza esa conclusión. Sin embargo, sí ayudan a matizar hasta qué punto se justifica la conclusión. Por otro lado, si bien la refutación se relaciona con la validez del argumento, no es parte esencial de la justificación. La refutación se utiliza para fortalecer el argumento anticipando objeciones o mostrando situaciones en las que la justificación no se aplica. Por lo tanto, aunque la refutación puede enriquecer la argumentación, no forma parte del proceso de justificar una afirmación en un sentido simple.

Para complementar las definiciones propuestas por Jiménez-Aleixandre (2010) y por el mismo Toulmin (2007), se presenta el trabajo de Sardà y Sanmartí (2000, p. 480), quienes diseñaron una actividad en la que se pedía a los estudiantes argumentar sobre distintos métodos de conservación de alimentos, con el propósito de analizar cómo construían sus argumentos y qué dificultades enfrentaban durante el proceso. El alumnado participante, partía de datos suministrados previamente (hipotético) y elaboraban textos argumentativos en los que justificaban su postura, comparaban su método con otros y respondían a las críticas de sus compañeros. Luego, Sardà y Sanmartí (2000) analizaron los textos producidos por los estudiantes para evaluar cómo estructuran sus argumentos. Entre otras cosas, encontraron que

la conclusión generalmente aparece después de conectores como “por lo tanto”, “en conclusión”, “así”, “en consecuencia” y “luego”. El conector “porque” fue el más común para determinar las componentes justificación y conocimiento básico, y en segundo lugar, “ya que”. Señalan que ejemplos de calificadores son: quizá, seguramente, totalmente, algunas veces, probablemente, la mayoría de las veces y algunas veces. Además de lo anterior, el trabajo de Sardà y Sanmartí (2000, p.412) permite entender cómo podríamos evaluar la calidad argumentativa de un texto o discurso argumentativo mediante las nociones de validez formal, secuencia textual, análisis de conectores, concordancia entre las pruebas y la conclusión, aceptabilidad de la justificación principal y relevancia de los argumentos, nociones distintas a la más común y usada en muchos artículos para evaluar el desempeño del alumnado en argumentación, que consiste en estudiar las componentes del modelo de Toulmin y que se revisará a continuación.

2.3.2 Estrategias para evaluar el desempeño en argumentación del alumnado

Existen varias investigaciones que tratan sobre el nivel argumentativo del alumnado y en ellas hay varias propuestas de Rúbricas para evaluar este desempeño. Según lo revisado se ha encontrado dos tipos principales de rúbricas holísticas. El primer tipo corresponde a aquellas donde las componentes del modelo de Toulmin se gradúan en niveles de desempeño de manera nuclear, una a una (Cho y Jonassen, 2002, p.12; Cebrián-Robles et al., 2018; Yeh, K., y She, H., 2010; Lin, 2019). En algunos casos se asignan puntajes y se suma el global para cuantificar en nivel alcanzado. Lo interesante es que van por separado y dan la idea de ser autónomas de las demás, un ejemplo es la rúbrica mostrada en la tabla 2.3.A.

El segundo tipo corresponde a aquellas que van apilando las componentes del modelo de Toulmin, esto implica que un mayor número de ellas es equivalente a un mejor desempeño argumentativo, como el ejemplo de la Tabla 2.3.B Aquí no se pierde aparentemente la interrelación entre ellas (Lin, Hong y Lawrenz, 2012; Torun, 2019, Sadler y Fowler, 2006). Este último tipo esta más en acuerdo con los desarrollos de progresiones de aprendizaje trabajados por Osborne et al. (2015) y Berland y McNeaill (2010).

Tabla 2.3.A - Un ejemplo de rúbrica para determinar el desempeño en argumentación, tomada de Cho y Jonassen (2002, p.12). Traducción propia.

Conclusión	
Calidad	Criterio
6	El escritor establece generalizaciones que están relacionadas con la proposición y son claras y completas.
4	El escritor establece generalizaciones que están relacionadas con la proposición, pero las afirmaciones no son completas. La información disponible permite entender la intención del escritor, pero deja mucho al lector para interpretar.
2	El escritor realiza generalizaciones que están relacionadas con la proposición, pero las afirmaciones carecen de especificidad o presentan referentes poco claros. El escritor deja mucho al lector para inferir con el fin de determinar el impacto de la afirmación.
0	No hay ninguna afirmación relacionada con la proposición o las aseveraciones son poco claras.
Justificación	
Calidad	Criterio
6	El escritor explica los datos de manera que queda claro cómo respaldan la afirmación.
4	El escritor explica los datos de alguna manera, pero la explicación no está específicamente vinculada a la afirmación.
2	El escritor reconoce la necesidad de conectar los datos con la afirmación y ofrece cierta elaboración de los datos, pero no logra establecer la conexión. O bien, la mayoría de las reglas y principios no son válidos ni relevantes.
0	No se ofrecen reglas ni principios.
Conocimiento básico	
Calidad	Criterio
6	El escritor presenta fuentes de respaldo correctas, relevantes y específicas.
4	El escritor presenta fuentes de respaldo correctas y relevantes, pero las fuentes son muy generales y no específicas.
2	El escritor presenta fuentes de respaldo incorrectas o irrelevantes.
0	No se presentan fuentes de respaldo.

Tabla 2.3.B - Ejemplo de una rúbrica para argumentación diseñada mediante agregación de componentes. Tomada de Lin et al. (2012).

Nivel	**Descripción**	**Ejemplos de argumento estudiantil**
1	Afirmación simple o proposición	Estoy de acuerdo con el argumento a favor de los OMG (Organismos Modificados Genéticamente) porque es más persuasivo y los OMG han sido ampliamente aceptados por el público. *(proposición simple)*
2	Afirmaciones con datos válidos, garantías o respaldos	Una planta de energía nuclear necesita una gran cantidad de agua para el sistema de enfriamiento. La descarga de agua caliente al mar podría causar un efecto perjudicial en la ecología oceánica de las áreas circundantes. *(garantía)* Por lo tanto, me opongo a la construcción de cualquier planta de energía nuclear. *(afirmación)*
3	Afirmaciones con más de una evidencia válida de respaldo (por ejemplo, datos, garantías o respaldos) o una evidencia	No apoyo a los OMG porque su riesgo potencial no se comprende completamente (datos) y nadie puede garantizar su seguridad para la salud humana. Además, su cultivo masivo podría dañar la biodiversidad (garantía). Resolver la escasez de alimentos con OMG no es realista, ya que el problema principal es la distribución, algo que puede abordarse mediante operaciones de la ONU (refutación débil).

	válida y una refutación débil	
4	Afirmación, evidencia válida y una refutación identificable	El mundo apoya los OMG: en 2008 se cultivaron más de 125 millones de hectáreas (datos), y el arroz dorado con betacaroteno puede combatir la deficiencia de vitamina A (datos). Además, no es cierto que beneficien solo a empresarios, ya que 13.3 millones de agricultores en 25 países dependen de ellos para su sustento (refutación).
5	Argumento extenso con afirmaciones respaldadas por datos y garantías con más de dos refutaciones identificables	No apoyo los OMG porque muchas plantas derivan de "enemigos naturales", lo que podría afectar la tolerancia a antibióticos (datos). Además, los OMG pueden causar alergias, como se evidenció cuando el gen de la nuez brasileña transmitido a la soja provocó contaminación por metionina y un compuesto alergénico (refutación). También cuestiono que sus riesgos potenciales sean solo excusas, ya que, al igual que el DDT en la década de 1950, los efectos perjudiciales de nuevas tecnologías no siempre se comprenden completamente, como el daño a la vida acuática y a las aves (datos y refutación).

2.4 La práctica científica de la modelización

La modelización constituye una de las prácticas más significativas dentro de la actividad científica escolar y su enseñanza. Comprender en qué consiste esta práctica y cómo se diferencia de otros conceptos afines resulta esencial para promover un aprendizaje más auténtico y participativo en ciencias.

Según Moraga-Toledo y Espinet-Blanch (2023), hay dos conceptos clave a diferenciar: modelo y modelización. Por cada uno se entiende:

- **Modelo**: Un modelo se entiende como una construcción didáctica diseñada con propósitos educativos específicos, considerando el currículo, las capacidades y dificultades de los estudiantes. Según Izquierdo y Adúriz-Bravo (2003) este concepto está orientado a facilitar el aprendizaje y la comprensión de fenómenos científicos de manera estructurada.
- **Modelización**: La modelización se refiere al proceso de construir el significado de fenómenos científicos, ya sea individualmente o en grupo. Es un proceso clave en la enseñanza de las ciencias, ya que permite a los estudiantes generar conocimiento científico que les ayuda a explicar los hechos o fenómenos que observan.

Couso (2020) explica que la modelización es una actividad fundamental tanto en la ciencia erudita como en la enseñanza de las ciencias. Esta práctica implica "expresar, usar, evaluar y revisar modelos" (Couso, 2020, p. 66), lo que permite a los estudiantes y docentes

explorar fenómenos mediante representaciones simplificadas que sirven para describir, predecir y explicar aspectos relevantes de la naturaleza. Lejos de ser un mero recurso pedagógico, la modelización representa un enfoque activo de construcción del conocimiento, en el que las ideas iniciales se revisan y transforman progresivamente hasta alcanzar explicaciones más sólidas y coherentes con los fenómenos estudiados.

Según Oliva (2019) la modelización se entiende como un proceso que favorece la evolución de los modelos mentales de los estudiantes acerca del mundo y propicia su inmersión en prácticas científicas auténticas. Un modelo mental es una representación interna que los sujetos construyen para interpretar fenómenos o procesos del mundo real, mientras que un modelo científico formaliza esta representación en un lenguaje compartido, utilizando signos, códigos y formatos visuales, verbales, simbólicos, matemáticos, analógicos o digitales.

Modelizar implica plantear problemas, formular predicciones, recoger y analizar información, elaborar explicaciones y validar modelos mediante contrastaciones experimentales, mentales o simuladas. El proceso de modelización es iterativo: se construyen modelos iniciales que se ponen a prueba, se refuerzan o se modifican según su capacidad predictiva y explicativa.

La construcción de un modelo requiere justificar su propósito sobre un fenómeno conocido, seleccionar un sistema de signos adecuado, ensamblar un lenguaje de razonamiento causal, y representar el modelo en uno o varios formatos. Los modelos deben ser utilizados para describir, explicar o predecir fenómenos, ser evaluados y revisados para mejorar su potencial explicativo (Oliva, 2019, p.10).

Oliva (2019, p.10) menciona cinco formas de interacción con modelos: aprender modelos existentes, usar modelos, revisar modelos, reconstruir modelos y crear modelos nuevos, en una progresión creciente de complejidad. La modelización exige la movilización de capacidades epistémicas y competencias científicas, asociadas a la formulación, uso y evaluación crítica de modelos (Couso, 2014; Oliva, 2019). Para Oliva (2014, p.10) un modelo es considerado adecuado mientras cumpla su propósito explicativo; cuando las predicciones no se cumplen, el modelo debe ser modificado o reemplazado. Así, modelizar no implica solo reproducir información científica, sino construir y validar representaciones del mundo basadas en prácticas de razonamiento y evidencia.

Entre las ventajas de esta metodología, se destaca que "la modelización aporta autenticidad a la enseñanza de las ciencias, al permitir que el alumnado participe en prácticas

que muestran la ciencia como una actividad centrada en la generación de nuevo conocimiento en lugar de como un cuerpo de conocimiento acabado" (Couso, 2020, p. 66). Este enfoque no solo fomenta la participación de los estudiantes, sino que también los introduce a la naturaleza dinámica de la construcción científica. Sin embargo, Couso (2020) señala desafíos en su implementación, como la necesidad de "tiempos de enseñanza y aprendizaje mucho más largos y revertir el orden habitual de las secuencias tradicionales y los libros de texto" (Couso, 2020, p. 70), lo que podría ser complejo en contextos educativos con currículos rígidos o limitaciones de tiempo.

La modelización sigue un ciclo que organiza y estructura esta práctica Couso (2020, p. 70). Este ciclo, que se muestra en la figura 2.4.A, está compuesto por seis fases que se desarrollan tanto desde las prácticas de modelización del alumnado como desde las fases de la secuencia de enseñanza y aprendizaje a cargo del profesorado.

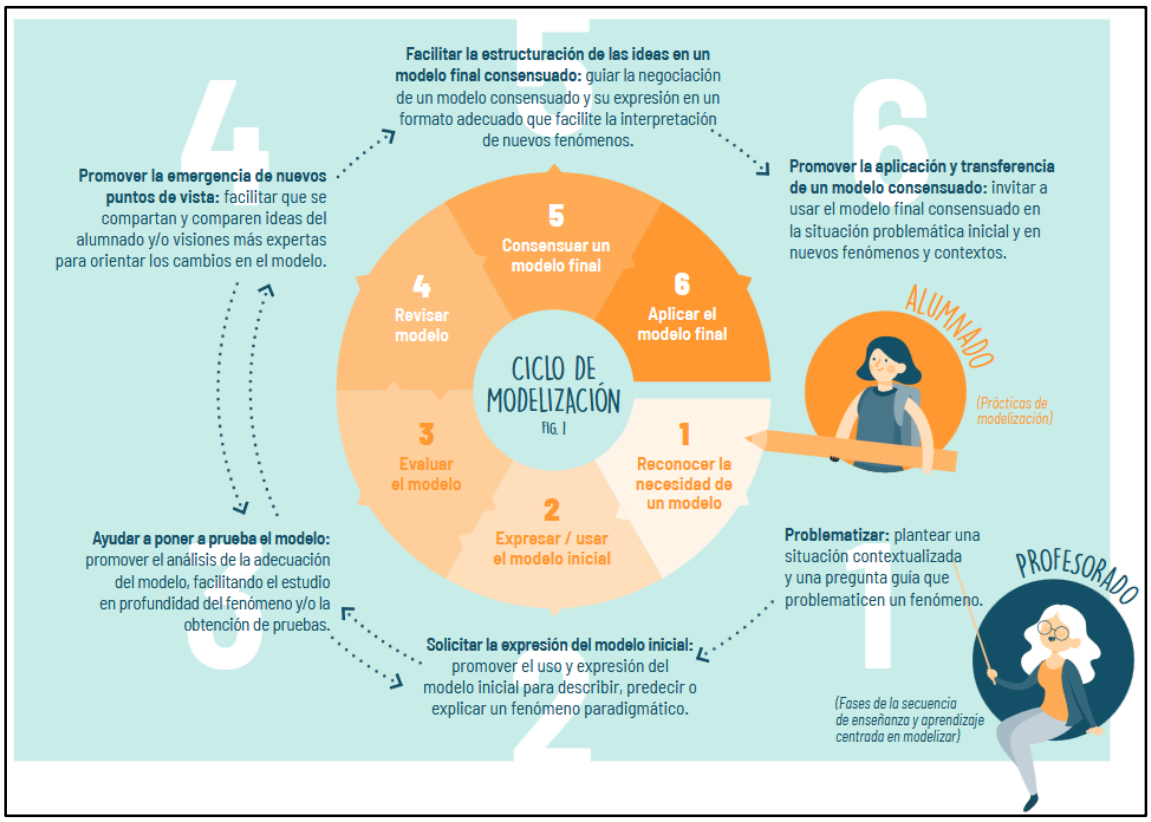

Figura 2.4.A - Ciclo de modelización. Tomado de Couso (2020, p.68).

El proceso comienza con el reconocimiento de la necesidad de un modelo, donde el profesorado plantea una situación contextualizada y una pregunta guía que problematice un fenómeno y motive la búsqueda de una representación explicativa. En esta fase, el alumnado reconoce la necesidad de generar un modelo inicial que dé cuenta del fenómeno observado. A

continuación, el alumnado expresa o utiliza un modelo inicial para describir, predecir o explicar el fenómeno, ya sea de manera gráfica, verbal o simbólica. El profesorado promueve la expresión explícita de este modelo inicial, facilitando un espacio de comunicación de las ideas del alumnado. Posteriormente, el estudiante pone a prueba el modelo mediante el análisis de su adecuación al fenómeno, observando datos o realizando experimentos, mientras que el docente guía este proceso fomentando el estudio en profundidad del fenómeno o la obtención de nuevas pruebas.

En la cuarta fase, el alumnado revisa el modelo, contrastándolo con los nuevos datos o puntos de vista emergentes. El profesorado, en paralelo, promueve la comparación de ideas entre los estudiantes y orienta los cambios necesarios en el modelo. Luego, en la quinta fase, los estudiantes consensuan un modelo final mediante la negociación y estructuración colectiva de las mejores explicaciones posibles. El rol del profesorado consiste en facilitar la expresión clara y adecuada de este modelo consensuado para permitir su aplicación a nuevos fenómenos.

Finalmente, en la sexta fase, el alumnado aplica el modelo final a situaciones problemáticas nuevas o al fenómeno inicial, transfiriendo lo aprendido a distintos contextos. El profesorado apoya esta transferencia, motivando el uso reflexivo del modelo y su adaptación a otros escenarios científicos.

2.4.1 Modelos mentales en el aula de ciencias

Los modelos no son representaciones exactas de la realidad, sino aproximaciones que permiten comprender y explicar fenómenos específicos. (Couso, 2020, p. 70) afirma que "los modelos científicos son representaciones simplificadas e intencionadas de la realidad, pero no la realidad". En el contexto escolar, estos modelos no buscan replicar los modelos científicos profesionales, sino que son "reconstrucciones didácticas del conocimiento científico consensuado realizadas especialmente para favorecer su enseñanza y aprendizaje" (Couso, 2020, p. 70). Esta diferencia es clave para comprender su función en el aula.

En términos de características, un modelo debe ser evaluable, revisable y expresable en diversos formatos, como verbal, gráfico o algebraico (Couso, 2020, p. 66). Estas propiedades aseguran que los modelos no solo sean útiles como herramientas explicativas, sino que también puedan adaptarse y perfeccionarse conforme se obtienen nuevos datos o se profundiza en el fenómeno estudiado. Además, la autora ofrece ejemplos concretos para aclarar qué constituye un modelo. Por ejemplo, señala que "el modelo Sol-Tierra no se refiere a una maqueta

tridimensional o animación digital, sino a las ideas que permiten explicar cómo varían las horas de luz solar en distintas localidades" (Couso, 2020, p. 70). Esto subraya que los modelos científicos trascienden las representaciones físicas y se centran en las conceptualizaciones que explican fenómenos. Según Couso (2020), la evaluación o puesta a prueba de los modelos se debe hacer en base a la observación, análisis e investigación del fenómeno (Couso, 2020, p. 66).

En su estudio, López-Simó y Simarro (2023) explican que en el contexto de la educación primaria y secundaria no se trabaja habitualmente con modelos científicos en su sentido estricto, sino con adaptaciones denominadas modelos científicos escolares o MCE, los cuales "tienen la función de interpretar el mundo natural al nivel adecuado para cada etapa del aprendizaje" y pueden concebirse "en forma de ramificación, es decir, en base a grandes familias de modelos entorno de las grandes ideas de la ciencia, que a su vez contienen modelos y submodelos más concretos y específicos" (López-Simó y Simarro, 2024, p. 99). Esta concepción implica reconocer la complejidad inherente a la enseñanza de las ciencias, donde no se busca replicar exactamente los modelos científicos profesionales, sino construir versiones adaptadas que sirvan al desarrollo cognitivo del alumnado.

Sobre esta base, los autores añaden que los procesos de aprendizaje de las ciencias "no solamente dependen de estos modelos científicos escolares, sino también de los modelos mentales de los que dispone el alumnado en cada momento, que a menudo son alternativos o alejados de los que se pretende que aprendan" (López-Simó y Simarro, 2024, p. 99). Desde esta perspectiva, trabajar con modelos no se reduce a transmitir ideas científicas ya elaboradas, sino que requiere involucrar activamente al alumnado en la expresión y revisión de sus propios modelos mentales. Según los autores, es fundamental que los estudiantes participen en "el proceso de expresión y revisión sus propios modelos mentales para acercarlos a dichos modelos escolares".

Este enfoque de modelos mentales reconoce que los estudiantes construyen interpretaciones propias del mundo natural, las cuales deben ser explicitadas, confrontadas con la evidencia y refinadas progresivamente. La modelización, entendida de esta forma, implica un proceso dinámico en el que "la expresión de los modelos mentales propios precede a su evaluación y revisión a la luz de las evidencias de que se dispone" (López-Simó y Simarro, 2024, p. 99), constituyendo un ciclo continuo de construcción, contraste y modificación de representaciones científicas personales hacia formas más consensuadas y acordes al conocimiento científico escolar.

Gutiérrez (2004) describe que un modelo mental está compuesto por tres elementos esenciales. El primer elemento "constituye una ontología del sistema físico representado, reflejo del sistema de creencias del usuario del modelo" (p.12). Este primer componente consiste en identificar las entidades principales que configuran el sistema de interés, seleccionadas en función de lo que el usuario desea modelizar. El segundo elemento "hace posible la explicación del comportamiento del sistema y la predicción de futuros comportamientos" (p.11), es decir, consiste en establecer reglas de funcionamiento que definan cómo se comportan esas entidades y sus propiedades. Finalmente, el tercer elemento corresponde a "la ejecución del modelo mental", entendida como una simulación que permite la evaluación de los modelos mentales construidos, de manera que, si no hay correspondencia entre el comportamiento del sistema simulado en la mente y el comportamiento del sistema físico representado, el modelo mental no sería válido. Así, los modelos mentales no solo representan entidades y reglas, sino que son puestos en funcionamiento mentalmente para comprobar su coherencia y validez respecto al sistema físico real.

En contraposición a otras representaciones, el modelo mental se caracteriza por permitir su ejecución de manera interna. Tal como describe Gutiérrez (2004, p.12):

> La segunda representación se puede ejecutar mentalmente, es decir, permite poner en funcionamiento mentalmente el modelo mental construido, de manera que se puede comparar el comportamiento de este con el comportamiento real del sistema físico modelizado.

Así el modelo mental incorpora una dinámica de simulación y comparación activa con la realidad. Su validez depende de la correspondencia entre el comportamiento del sistema simulado mentalmente y el comportamiento del sistema físico real, pues si no hay correspondencia entre el comportamiento del sistema simulado en la mente y el comportamiento del sistema físico representado, el modelo mental no sería válido.

Un trabajo que ha sido usado para evaluar modelos mentales del alumnado (López-Simó y Simarro, 2024) es el de Louca et. al. (2011a), que revisaremos a continuación.

2.4.2 Estrategias para evaluar modelos mentales del alumnado

Louca et al. (2011a, p. 174) define los modelos como “representaciones sistemáticamente construidas de sistemas físicos, utilizadas para describir, representar y explicar los mecanismos subyacentes a los fenómenos físicos”. Según esta perspectiva, los modelos tienen como

objetivo principal proporcionar un marco para interpretar hallazgos experimentales y predecir el comportamiento de sistemas físicos. Los autores citan a Gilbert et al. (1998), destacando que "los atributos más importantes de un modelo son su poder explicativo, su potencial para realizar predicciones, su capacidad para hacer visibles entidades abstractas y su utilidad como base para interpretar hallazgos experimentales" (Gilbert et al., 1998). Según Louca et al. (2011a) hay una ausencia de un marco analítico para evaluar los modelos construidos por estudiantes, se necesita una herramienta que pueda ayudar a evaluar el progreso de los estudiantes al construir modelos. Para los autores una forma de abordar la evaluación de un modelo es identificar sus componentes estructurales y evaluar cada uno de ellos por separado, y luego considerar si el modelo resultante representa con precisión el fenómeno representado.

Por ello Louca et al. (2011a, p.177) identifica cinco características que deben cumplir los modelos:

- **Objetos físicos**: es una entidad tangible dentro del sistema modelado. Podría tratarse de cuerpos o estructuras que interactúan dentro del fenómeno físico. Son los principales "actores" en el sistema físico y están directamente relacionados con la estructura sistémica del sistema físico representado. Las representaciones de sistemas o fenómenos físicos se basan en la representación de los objetos físicos, que constituyen la base del resto del modelo.
- **Entidades físicas**: Son utilizadas para representar las características de los objetos físicos. Comprenden conceptos abstractos inventados por los humanos para interpretar fenómenos físicos, como velocidad, fuerza, energía, etc.
- **Comportamientos de los objetos físicos**: Reflejan el mecanismo que subyace al comportamiento de cada objeto por separado. El comportamiento de un objeto a menudo se basa en relaciones causales entre entidades físicas relacionadas con él.
- **Interacciones entre objetos físicos, entidades físicas y comportamientos de los objetos**: Estas interacciones son necesarias para un mecanismo unificado que subyace al comportamiento del fenómeno en su conjunto.
- **Precisión científica**: Refleja la precisión científica de cómo el modelo representa el fenómeno representado. Se agrega que, "si un modelo representaba adecuadamente la estructura superficial del fenómeno, independientemente de la presencia o ausencia del mecanismo causal subyacente al fenómeno". (Louca et al., 2011a, p.182).

Para el primer componente los autores profundizan mencionando que los objetos físicos involucrados en un fenómeno pueden clasificarse en categorías según su relación con el

mecanismo subyacente al fenómeno. Esta subclasificación permite distinguir entre aquellos que tienen un papel funcional directo en el modelo y aquellos que son meramente contextuales. La subclasificación incluye (Louca et al., 2011a, p.182):

1. **Objetos internos**: Son los que desempeñan un papel funcional en el mecanismo subyacente al fenómeno en estudio. Estos objetos son esenciales para comprender y representar cómo se desarrolla el fenómeno, ya que contribuyen directamente a su dinámica y comportamiento.
2. **Objetos externos**: Son aquellos que no participan en el mecanismo subyacente al fenómeno, pero pueden estar presentes en el modelo como elementos adicionales, decorativos o contextuales, sin afectar el desarrollo del fenómeno representado.

Para identificar estos elementos desde los modelos de los estudiantes, se pueden usar los ejemplos que aparecen en la tabla 2.4.A que están dados en el trabajo de Louca et al. (2011a).

Tabla 2.4.A - Ejemplos de identificación de los elementos en modelos estudiantiles según Louca et al. (2011a).

Elemento	Ejemplo
Objeto físico	"En un modelo de movimiento acelerado, una pelota que cae es un objeto interno, porque participa en el mecanismo subyacente al fenómeno representado", Louca et al. (2011a, p.183). Además, es un objeto interno porque participa en el mecanismo que describe el fenómeno de caída libre. En contraste, cualquier elemento adicional del entorno, sería un objeto externo, ya que no influye en el funcionamiento del fenómeno representado.
Entidades físicas	La velocidad de caída es una entidad física que designa el estado cinético de un objeto, Louca et al. (2011a, p.184). La velocidad como la aceleración se representan como variables que interactúan entre sí para determinar el comportamiento del objeto (la posición de la pelota). Esto ilustra cómo las entidades físicas pueden ser utilizadas para interpretar fenómenos físicos, ya sea en términos descriptivos (valores numéricos fijos) o dinámicos (variables que cambian con el tiempo)
Comportamiento de los objetos físicos	Según Louca et al. (2011a), el comportamiento de un objeto refleja el mecanismo que subyace en el comportamiento de cada objeto por separado. El comportamiento de un objeto a menudo se basa en relaciones causales entre entidades físicas relacionadas con él. En otras palabras, describen lo que hace un objeto en el modelo. Por ejemplo: El cambio de posición de un objeto o la variación en su velocidad o aceleración. En este ejemplo el aumento de velocidad en caída libre, no se representan de manera causal en un modelo inicial. En lugar de incluir una relación explícita entre la aceleración y la velocidad, un modelo inicial solo podría mostrar el movimiento de un objeto sin explicar el mecanismo subyacente que lo causa.
Interacciones	La interacción entre entidades físicas se representa cuando una entidad física tiene un efecto sobre otra. Por ejemplo, en un modelo que la aceleración afecte directamente a la velocidad, lo que a su vez define el comportamiento del objeto en caída libre. Esta relación es representada mediante reglas, que describen cómo

	la velocidad del objeto aumenta progresivamente debido a la influencia de la aceleración.
Precisión científica	El alumnado puede representar varios modelos, pero la sofisticación y la precisión científica de la representación del mecanismo subyacente al fenómeno varía entre los diferentes modelos (Louca et al., 2011a, p. 187). La precisión científica depende de cómo los mecanismos subyacentes, como la interacción entre aceleración y velocidad, son representados en el modelo. Por ejemplo, en algunos modelos, la caída libre podría mostrarse correctamente en términos de movimiento visual, pero carecer de una representación causal de la relación entre aceleración y velocidad, lo que limita su precisión científica.

La tabla 2.4.A ofrecen un esquema de codificación para cada uno de los elementos del modelo. Estas codificaciones permiten analizar cómo los estudiantes representan los distintos componentes de los modelos y diferenciar el nivel de sofisticación. Este último es trabajado con la rúbrica de la tabla 2.4.B diseñada por Louca et al. (2011a), que permite según los autores estudiar cómo los modelos progresan o evolucionan en cada componente por separado.

Tabla 2.4.B - Esquema de codificación de los elementos del modelo según Louca et al. (2011a).

Elemento	Representaciones e interacciones
Objetos físicos	**Objetos internos:** Participan funcionalmente en el mecanismo subyacente al fenómeno. **Objetos externos:** No participan en el mecanismo subyacente al fenómeno.
Entidades físicas	**Sin representación:** No hay entidades físicas representadas. **Representadas con valores numéricos no variables:** Las entidades físicas están representadas mediante valores fijos. **Representadas con una combinación de valores numéricos no variables y variables:** Parte de las entidades se representan con valores fijos y parte con variables. **Representadas exclusivamente con variables:** Todas las entidades están representadas mediante variables.
Comportamientos de los objetos físicos	**Representación no causal:** No se representan relaciones causales en los comportamientos. **Representación semi-causal:** Combinación de relaciones causales y no causales. **Representación causal:** Las relaciones causales están completamente representadas.
Interacciones	**Interacciones entre objetos físicos**: Dos objetos tienen un efecto directo entre sí. **Interacciones entre entidades físicas**: Una entidad física influye sobre otra (por ejemplo, aceleración afecta a velocidad). **Interacciones entre comportamientos**: Las reglas que representan los comportamientos están relacionadas entre sí.
Precisión científica	**Representación válida:** El modelo representa el fenómeno de manera científica y precisa. **Representación no válida:** La representación del fenómeno carece de precisión científica.

Al estudiar la rúbrica (tabla 2.4.B) se puede determinar que no proporciona directamente un índice global o una puntuación acumulativa que determine cuál modelo es "mejor" de manera digamos absoluta. La rúbrica no parece imponer restricciones explícitas sobre qué combinaciones de atributos son permitidas o prohibidas. Por ejemplo, un modelo podría tener objetos físicos internos bien representados, pero carecer de precisión científica o tener únicamente representaciones no causales de los comportamientos, en ese sentido es del tipo de rúbrica de la tabla 2.3.B para argumentación. Por ello la rúbrica no dicta que un modelo "mejor" necesariamente tiene que incluir más componentes o que estos deban estar representados en niveles más sofisticados. Más bien, permite una evaluación cualitativa y cuantitativa de cada componente, lo que puede generar una diversidad de combinaciones posibles en los resultados, aunque existe una valoración final de precisión que parece ser disonante con las demás componentes.

Louca et al. (2011a) llevaron a cabo un estudio centrado en analizar cómo los estudiantes construyen modelos de fenómenos físicos utilizando dos tipos distintos de entornos de programación basados en computadoras. El diseño del estudio incluyó la participación de dos clases de sexto grado, cada una conformada por 20 estudiantes, en una escuela primaria de Chipre. A lo largo de cinco meses, los estudiantes se dedicaron a la construcción de modelos de fenómenos físicos. Los estudiantes trabajaron en cuatro temas diferentes, relacionados con fenómenos como el movimiento relativo, la caída libre, el movimiento de proyectiles y una combinación de ambos. Para cada fenómeno, los estudiantes pasaron por un proceso iterativo que comenzaba con discusiones iniciales en las que generaban ideas sobre el fenómeno. Posteriormente, trabajaban en pequeños grupos para desarrollar modelos iniciales en el entorno de programación asignado. Estos modelos eran evaluados colaborativamente en clase, donde se ofrecía retroalimentación que los estudiantes incorporaban en rondas de refinamiento hasta alcanzar una versión final del modelo. Cada grupo completó al menos dos versiones del modelo para cada fenómeno: una inicial y otra final.

Para evaluar los modelos construidos, Louca et al. (2011a) utilizaron la rúbrica de la tabla 2.4.B. Se destaca el hecho que en al evaluar los comportamientos de los objetos físicos, estos son clasificados como no causales, semi-causales o causales, dependiendo de cómo estuvieran representadas las relaciones en el modelo. Esto podría parecer algo contradictorio con lo que se expresa en la tabla 2.4.A, donde estos son definidos como: “El comportamiento de un objeto refleja el mecanismo que subyace en el comportamiento de cada objeto por separado”, sin embargo, el propósito de la rúbrica no es contradecir alguna parte de la descripción inicial, sino

evaluar si los estudiantes han avanzado desde representaciones iniciales (generalmente no causales) hacia representaciones más sofisticadas, que incluyan relaciones causales. Este es un indicador de mayor sofisticación en el modelo y esta alineado con el propósito metodológico de evaluar si los andamios trabajados (discusiones grupales, compartir y criticar modelos iniciales) logran un modelo final más refinado.

Finalmente, en la dimensión de "Interacciones" la rúbrica asegura que, aunque un comportamiento pueda incluir relaciones causales internas, las interacciones evalúan explícitamente las conexiones externas entre elementos del modelo. Por ejemplo:

- Comportamiento: La velocidad de un objeto aumenta debido a la aceleración.
- Interacción: La aceleración de un objeto afecta la velocidad de otro objeto.

Louca et al. (2011b) usan las ideas anteriores para que un grupo de estudiantes estudien fenómenos físicos y creen modelos de estos utilizando herramientas computacionales, específicamente el software Stagecast Creator, que permite la creación de micromundos basados en un lenguaje de programación gráfico. Así los estudiantes planifican e implementan sus propias investigaciones con el propósito de construir modelos que simulen el fenómeno físico estudiado y representen el mecanismo que explica cómo se genera el fenómeno. Para estudiar los modelos de los estudiantes desarrollan una rúbrica que surge del análisis de los modelos construidos por los estudiantes y las conversaciones de la clase. Utilizan un enfoque inductivo basado en codificación abierta (Strauss & Corbin, 1998) para identificar patrones en los datos discursivos y en los artefactos (modelos computacionales). El desarrollo de la rúbrica se basa en el trabajo desarrollado anteriormente por Louca et al. (2011a). La rúbrica en cuestión es la de la tabla 2.4.C.

Tabla 2.4.C - Esquema de codificación de los elementos del modelo según Louca et al. (2011b).

Códigos	Descripción de los códigos/Ejemplos
1. Representación de objetos físicos	
1.1. Objetos físicos internos al sistema físico	Representación de objetos que desempeñan un papel funcional en un fenómeno físico (por ejemplo, una pelota en movimiento acelerado).
1.2. Objetos físicos externos al sistema físico	Representación de objetos que no desempeñan un papel funcional en un fenómeno físico (por ejemplo, una nube en el cielo en el movimiento acelerado de una pelota en caída libre).
2. Representación de procesos físicos	
2.1. No causal	No hay representación de una entidad física (por ejemplo, velocidad) que cause cambios en un proceso físico (por ejemplo, movimiento).

2.2. Semi-causal	Representación parcial de cómo una entidad física (por ejemplo, velocidad) causa cambios en un proceso físico (por ejemplo, movimiento).
2.3. Causal	Representación de cómo una entidad física (por ejemplo, velocidad) causa cambios en un proceso físico (por ejemplo, movimiento).
2.4. Causal (científicamente correcta)	Representación científicamente correcta de cómo una entidad física (velocidad) causa cambios en un proceso físico (movimiento).
3. Representación de entidades físicas	
3.1. Sin representación de entidades físicas	No hay representación de las entidades físicas involucradas en el fenómeno.
3.2. Representadas con un valor numérico	Entidades físicas representadas con un valor numérico no variable (por ejemplo, la velocidad de un objeto se determina por un número).
3.3. Representadas con una variable y un valor numérico	Entidades físicas representadas con una combinación de un valor numérico no variable y una variable de programa.
3.4. Representadas con una variable	Entidades físicas representadas mediante una variable de programa (por ejemplo, la velocidad de un objeto se representa con una variable que define la distancia recorrida por ciclo de máquina).
4. Representación de interacciones entre:	
4.1. Objetos físicos	Dos objetos interactúan entre sí (por ejemplo, dos pelotas que chocan entre sí provocarán cambios en sus velocidades).
4.2. Entidades físicas	Dos variables que representan entidades físicas interactúan entre sí (por ejemplo, una variable nombrada como aceleración provoca cambios en la variable nombrada como velocidad).
4.3. Procesos físicos	Representación de la interacción entre dos procesos físicos (por ejemplo, la interacción entre el cambio de velocidad y el cambio de distancia recorrida en movimiento acelerado).

En esta rúbrica (tabla 2.4.C) Louca et al. (2011b) trabajan solo en cuatro dimensiones. La primera dimensión de objetos físicos es similar en esta como en la rúbrica de la tabla 2.4.B. Además, a diferencia de esta, la rúbrica de la tabla 2.4.C no incluye un criterio explícito de precisión científica. En su lugar, se enfoca en la corrección científica dentro del nivel de causalidad. Esto se evidencia en la categorización de los procesos físicos, que incluyen un nivel de "causal científicamente correcta", es decir la precisión pasa a esta dimensión y criterio.

En ambas rúbricas se incluyen valores numéricos y variables como cualidades de los objetos físicos, aunque el segundo estudio (Louca et al., 2011b) se desarrolla el análisis con más detalle y especificidad. Aquí, las entidades físicas pueden representarse de tres formas: Con valores numéricos fijos, con una combinación de valores numéricos y variables o exclusivamente con variables. También, en Louca et al. (2011a), las interacciones se dividen en interacciones entre objetos físicos, entre entidades físicas y entre comportamientos. Sin

embargo, en Louca et al. (2011b), se reformulan como interacciones entre objetos físicos, entre entidades físicas y entre procesos físicos, con un enfoque más amplio en los modelos de procesos físicos

Basados en las definiciones de Louca et. al (2011a) otros autores también las han adaptado para generar sus propias rúbricas y valorar modelos. Por ejemplo, López-Simó y Simarro (2024), las usan para valorar los modelos planteados por futuros maestros de primaria. Les asignan la tarea de producir animaciones, simulaciones o videojuegos educativos utilizando el entorno de programación Scratch. La secuencia didáctica incluye un total de seis horas distribuidas en diferentes sesiones: dos horas dedicadas a enseñar el funcionamiento de la plataforma Scratch y su lenguaje de programación, dos horas enfocadas en abordar las ideas de modelo científico y modelo computacional, y dos horas de trabajo dirigido para orientar el diseño de las producciones digitales. Los estudiantes trabajan en pequeños grupos de tres o cuatro integrantes y desarrollan sus proyectos de manera autónoma, con la intención de que estas producciones se conviertan en recursos educativos para la enseñanza de contenidos curriculares de primaria, abarcando diversas disciplinas científicas como biología, geología, física o química.

Desde este enfoque, López-Simó y Simarro (2024) plantean dos preguntas fundamentales de investigación: ¿Qué tipo de producciones digitales hechas con Scratch priorizan los futuros docentes de primaria según el tipo de producto y según el contenido científico? y ¿Hasta qué punto estas producciones digitales incorporan un modelo computacional subyacente y cuál es su robustez? Para responder a la segunda pregunta, los autores diseñaron una rúbrica basada en componentes estructurales de los modelos computacionales, fundamentada en el marco teórico de Louca et al. (2011a). Esta rúbrica permite evaluar la existencia y "nivel de elaboración" de los modelos subyacentes en las producciones digitales de los estudiantes a partir de cuatro componentes esenciales: objetos materiales, entidades abstractas, procesos y comportamientos, e interacciones y relaciones. La rúbrica se muestra en la tabla 2.4.D.

Tabla 2.4.D - Esquema de codificación de los elementos del modelo según López-Simó y Simarro (2024).

Componente	Nivel 0	Nivel 1	Nivel 2	Nivel 3
Objetos	No hay objetos representados	1 solo objeto representado	Varios objetos representados, pero no integrados en un sistema	Varios objetos representados, que configuran un sistema

Entidades	No hay entidades representadas	1 entidad implícita	1 o más entidad definida	Varias variables definidas relacionadas entre ellas.
Procesos y comportamientos	No hay procesos representados (no sucede nada)	1 proceso lineal (antes – después).	Más de un proceso que puede ocurrir en paralelo o en cadena.	Variedad de procesos posibles que generan un diagrama de árbol de actividades.
Interacciones y relaciones	No hay interacciones representadas	1 interacción simple	1 o más interacción compleja	Variedad de interacciones entre varios objetos o entidades

La rúbrica asigna a cada componente un nivel de desarrollo que varía de 0 a 3, dependiendo de su complejidad y del grado en que aparecen representados en las simulaciones o animaciones creadas en Scratch.

La dimensión de "procesos y comportamientos" se diferencia de la de "interacciones y relaciones" en función del tipo de dinámica que representan dentro del modelo computacional subyacente. Según los autores, "los procesos y comportamientos de los objetos, que reflejan el mecanismo que subyace el comportamiento de cada objeto por separado" (López-Simó y Simarro, 2024, p.104). En este sentido, esta dimensión evalúa cambios internos o transformaciones que afectan a un solo elemento dentro del sistema modelado, lo que incluye variaciones en magnitudes físicas, desplazamientos, transformaciones químicas o incluso eventos biológicos como la muerte de un individuo.

En contraste, la dimensión de "interacciones y relaciones" se refiere a los vínculos entre diferentes objetos o entidades dentro del modelo computacional. Los autores afirman que "las interacciones y relaciones entre objetos, entidades y comportamientos, permiten unificar los mecanismos subyacentes del modelo" (López-Simó y Simarro, 2024, p.104). Esto implica que, mientras que los procesos y comportamientos están centrados en los cambios individuales dentro de un objeto o entidad, las interacciones y relaciones se ocupan de las conexiones y dependencias entre distintos elementos del sistema, como la atracción y repulsión de fuerzas en la física, los intercambios de materia y energía en la química o las relaciones ecológicas en biología.

Las diferencias entre estas dos dimensiones pueden entenderse con el ejemplo del caso de una simulación del ciclo del agua, el cambio de estado del agua, de líquido a gas debido a la evaporación, se consideraría un "proceso y comportamiento", ya que involucra una transformación interna en la sustancia sin necesidad de otra entidad. Sin embargo, la relación entre la evaporación y la formación de nubes, donde el vapor de agua asciende y luego se

condensa, implicaría una "interacción y relación" porque conecta diferentes procesos dentro del sistema atmosférico.

Además, en la rúbrica de evaluación, los niveles de elaboración de estas dimensiones también presentan diferencias fundamentales. En "procesos y comportamientos", el nivel más bajo (0) corresponde a la ausencia de procesos representados, mientras que el nivel más alto (3) implica "una variedad de procesos posibles que generan un diagrama de árbol de actividades" (López-Simó y Simarro, 2024, p.107), lo que significa que existen múltiples transformaciones encadenadas o paralelas en el modelo computacional. Por otro lado, en la dimensión de "interacciones y relaciones", el nivel 0 indica que "no hay interacciones representadas", mientras que el nivel 3 describe "una variedad de interacciones entre varios objetos o entidades" (López-Simó y Simarro, 2024, p.107).

En términos generales, mientras que los procesos y comportamientos hacen referencia a transformaciones internas dentro de un objeto o entidad, las interacciones y relaciones enfatizan la forma en que estos objetos y entidades se conectan y afectan mutuamente dentro del sistema. Estas distinciones permiten a los autores evaluar con mayor precisión la complejidad de los modelos computacionales creados por los estudiantes y su capacidad para representar fenómenos científicos con una estructura coherente y dinámica.

Por otro lado, existen trabajos que usan las categorías de Louca et al. (2011a) objetos, entidades, comportamientos, interacciones y precisión, pero ampliando y adaptando estos conceptos y sus definiciones para usarlas en contextos distintos. Un ejemplo es el trabajo de Lane y Headley (2021) en que adapta las categorías de Louca et al. (2011a) para describir la estructura y dinámica dentro de una comunidad de práctica en física. En lugar de centrarse en la representación de sistemas físicos en modelos computacionales, los redefine en términos de agentes sociales, sus cualidades, sus acciones y las relaciones que establecen dentro de la comunidad (Lane y Headley, 2021, p.6):

- **Objetos**: individuos e instituciones, como profesionales y departamentos. Este conjunto de objetos incluye al aprendiz, otros miembros de la comunidad académica local (como compañeros de clase y profesores), y miembros de la comunidad profesional global (como presentadores en conferencias o autores publicados).
- **Entidades**: cualidades de esos individuos e instituciones que describen su pertenencia dentro de la comunidad. Estas cualidades incluyen las expectativas del aprendiz sobre su participación en la comunidad de práctica, su confianza, posición y trayectoria.

- **Comportamientos**: acciones y prácticas en las que los individuos e instituciones participan. Este conjunto de comportamientos incluye la participación periférica legítima del aprendiz y los ejemplos que observa llevados a cabo por miembros centrales de la comunidad.

- **Interacciones**: el sentido de empresa conjunta que guía a los individuos e instituciones, establece estándares para las cualidades y mediatiza las acciones y prácticas.

- **Precisión**: la alineación entre el modelo de comunidad de práctica del aprendiz y la comunidad de práctica en la realidad.

Estas definiciones reformulan las categorías de Louca et al. (2011a) en función del contexto educativo, trasladan estas categorías al análisis de la participación de los estudiantes en una comunidad académica y profesional. Una de las principales diferencias radica en el concepto de objetos. Louca et al. (2011a) los definen como elementos físicos internos o externos al sistema modelado, mientras que en la adaptación del nuevo estudio, los objetos incluyen individuos e instituciones dentro de una comunidad académica. Aquí, el objeto deja de ser algo tangible dentro de un modelo físico y se convierte en un agente activo dentro de una estructura social. En cuanto a las entidades, Louca et al. (2011a) las conciben como variables medibles dentro del sistema modelado, como la velocidad o la aceleración en el caso de un modelo de movimiento. En cambio, en la reformulación del nuevo estudio, las entidades describen cualidades subjetivas y sociales, como la confianza del estudiante o su sentido de pertenencia dentro de la comunidad. Además, en este estudio, los comportamientos no se limitan a cambios medibles dentro de un sistema físico, sino que incluyen acciones y prácticas dentro de una comunidad, como la participación de los estudiantes en actividades académicas o la observación de prácticas realizadas por expertos. Respecto a las interacciones, representan el sentido de "empresa conjunta" que orienta a los participantes en la comunidad académica. En lugar de ser relaciones de causa y efecto en un sistema físico, se convierten en normas y dinámicas que regulan la integración y evolución de los estudiantes dentro de su contexto profesional.

2.5 El pensamiento crítico en la enseñanza de las ciencias

En los últimos años ha existido un interés creciente en el estudio del pensamiento crítico ya que su desarrollo es una parte fundamental de varias competencias a lo largo de su escolaridad, y ha sido, durante décadas, un tema de gran relevancia dentro del ámbito de investigación e innovación en educación. También es un objeto de estudio fundamental en la didáctica y en la enseñanza de las ciencias (Jiménez-Aleixandre, 2010; Oliveras y Sanmartí, 2009; Blanco-López et al., 2017), ya que permite a las personas evaluar de manera crítica la información que reciben, identificar la desinformación, tomar decisiones basadas en pruebas (Puig y Uskola, 2021) y adoptar una postura crítica ante la desinformación, fomentando así una ciudadanía informada y responsable (Covitt y Anderson, 2022). Sin esta habilidad, los individuos están más vulnerables a aceptar afirmaciones infundadas y a ser influenciados por emociones y sesgos ideológicos o de información, en lugar de hechos objetivos. A medida que nos adentramos en paradigmas como la “post-verdad”, la necesaria capacitación de los estudiantes para discernir de forma crítica y reflexiva la calidad de la información que recibe ha tomado todavía más relevancia (Couso y Puig, 2021; Osborne et al., 2022). Algunas de las tradiciones didácticas en ciencia, como la argumentación o el uso educativo de controversias socio-científicas han estado estrechamente ligadas a su desarrollo proponiéndose diferentes formas de definirlo, promoverlo y evaluarlo (Blanco-López et al., 2017; Cebrián-Robles et al., 2018; Vila et al., 2023).

Si bien existen diversas definiciones y aproximaciones al pensamiento crítico y su relación con la educación científica, en general lo entenderemos como un conjunto de disposiciones afectivas y habilidades cognitivas y metacognitivas que configuran la capacidad de los individuos para emitir juicios en un contexto y de forma razonada (Facione, 1990).

Según McPeck (1981, p.8), el pensamiento crítico, se define como "una disposición y habilidad para participar en una actividad con escepticismo reflexivo", e incluye dos componentes esenciales: 1) la evaluación o juicio, es decir, el desarrollo y evaluación de argumentos; y 2) la inclusión tanto de habilidades como de hábitos o disposiciones en su conceptualización (Facione, 1990).

Según Van Gelder (2005) el pensamiento crítico no es una capacidad innata, sino que debe desarrollarse a través de la "instrucción, práctica y tiempo". Específicamente, en la enseñanza de la ciencia, se destaca la necesidad de superar una tendencia a la enseñanza "instrumentalista, descontextualizada, formulista y transmisora de verdades absolutas", que es

opuesta a la naturaleza del pensamiento crítico (Solbes y Torres, 2013). El pensamiento crítico, por tanto, implica habilidades como el análisis y la argumentación, que ayudan a los estudiantes a "distinguir opiniones de hechos" y detectar informaciones sesgadas o falsas.

Una propuesta para el diseño de actividades que promuevan el desarrollo del pensamiento crítico en el aula de ciencias es realizada por Vila et al. (2023), partiendo de dos premisas clave. La primera premisa es que el desarrollo del pensamiento crítico requiere instrucción, práctica y tiempo, por lo que debe ser abordado de manera continua y transversal, en lugar de mediante actividades puntuales. La segunda premisa señala que el aula de ciencias es un entorno ideal para este desarrollo, especialmente cuando se utilizan contextos socialmente relevantes, como controversias socio-científicas, que motivan al alumnado a participar en prácticas científicas como la indagación, la argumentación y la modelización.

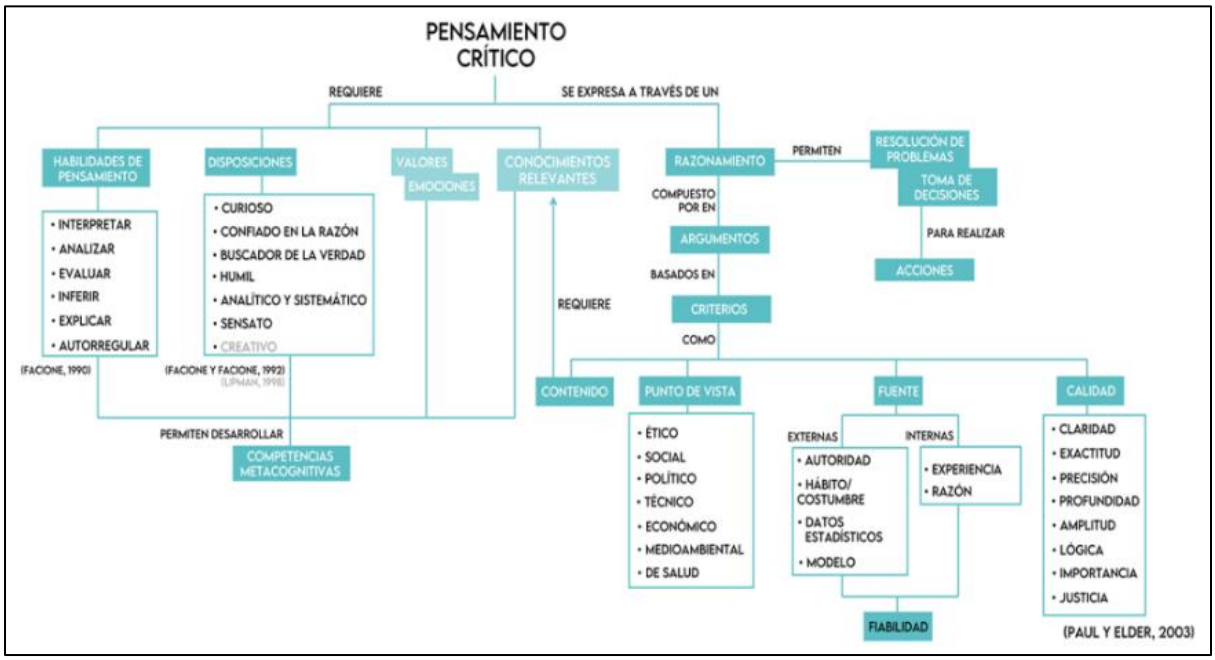

Figura 2.5.A - Mapa Operativo del Pensamiento Crítico publicado por Vila et al. (2023).

A partir de estas premisas, destacan una herramienta diseñada para ayudar al profesorado en la creación de actividades que fomenten el pensamiento crítico, que nombran Mapa Operativo del Pensamiento Crítico (MOPC), y que define los elementos del pensamiento crítico desde un enfoque práctico (Figura 2.5.A). Este MOPC muestra cuales son las características del pensamiento crítico encontradas en su investigación (Vila et al., 2023). Tiene dos ramas principales, aquellas dimensiones necesarias a la hora de activar el pensamiento crítico y el resultado o expresiones de un pensamiento crítico. Para la primera rama referida con activar el pensamiento crítico se mencionan 4 dimensiones: (1) Habilidades de Pensamiento, tomada de

Facione (1990), donde incluyen: Interpretar, Analizar, Evaluar, Inferir, Explicar y Autorregular; (2) Disposiciones, tomada de Facione y Facione (1992) y Lipman (2016), donde se incluyen las características: Curioso, confiado en la razón, buscador de la verdad, humilde, analítico y sistemático, sensato y creativo; (3) Valores y emociones y (4) Conocimientos relevantes. Estas 4 dimensiones del pensamiento crítico necesarias para su activación permiten que un individuo exprese "pensamiento crítico" mediante el "razonamiento" el cual este compuesto por "argumentos" basados en "criterios". Estos criterios, se desprende de Vila et al. (2023), necesitan para su construcción:

- Contenido que se relaciona con los conocimientos relevantes del individuo.
- Puntos de vista: Ético, Social, Político, Técnico, Económico, Medioambiental, de Salud
- Fuentes Externas: Autoridad, Hábito/Costumbre, Datos Estadísticos, Modelo
- Fuentes Internas: Experiencia, Razón.
- Calidad: Claridad, Exactitud, Precisión, Profundidad, Amplitud, Lógica, Importancia, Justicia.

Todos ellos, para las autoras, en un mismo nivel jerárquico. Finalmente, se afirma que el pensamiento crítico, manifestado mediante el razonamiento, capacita al individuo para la resolución de problemas, la toma de decisiones y la ejecución de acciones.

En la figura 2.5.A los conceptos de validez y fiabilidad se integran al marco del pensamiento crítico desde un enfoque que destaca la importancia de los criterios y las fuentes en la evaluación de los argumentos. La fiabilidad aparece de manera explícita como una cualidad que se analiza dentro del proceso de razonamiento crítico, en particular en lo que respecta al tratamiento de las fuentes utilizadas en la construcción de argumentos. Dichas fuentes se organizan en dos categorías: externas, como la autoridad, los hábitos o costumbres, los datos estadísticos y los modelos; e internas, como la experiencia y la razón. En este marco, la fiabilidad se presenta como el resultado del escrutinio riguroso de estas fuentes, lo que implica que un argumento será considerado más fiable en la medida en que se base en fuentes sólidas, verificables y pertinentes.

En cuanto a la validez, aunque no se menciona de forma explícita en la figura, puede inferirse a partir de los criterios de calidad que se indican en el análisis del razonamiento. Estos criterios comprenden dimensiones como la claridad, la exactitud, la precisión, la profundidad, la amplitud, la lógica, la importancia y la justicia. La validez de un argumento, por tanto, puede interpretarse como su adecuación a estos estándares de calidad, entendiendo que un argumento

válido no solo debe ser coherente y estructurado, sino también estar profundamente fundamentado y responder a una lógica interna rigurosa.

Ambos conceptos, fiabilidad y validez, se entienden en este modelo como resultados del ejercicio riguroso del pensamiento crítico. No basta con disponer de habilidades cognitivas o de ciertas disposiciones personales, es necesario también aplicar criterios precisos que permitan evaluar la calidad del razonamiento, así como la confiabilidad de las fuentes empleadas. En este sentido, la figura 2.5.A sugiere que la construcción de argumentos sólidos y significativos requiere tanto de la evaluación crítica de la procedencia y el contenido de la información como del cumplimiento de estándares formales de calidad argumentativa.

2.6 El pensamiento crítico en la confiablidad de la Apps

Un contexto que puede favorecer el trabajo del pensamiento crítico en el aula de ciencias es alguno que desafié las creencias previas del alumnado, combinando aspectos tecnológicos cotidianos con un análisis fundamentado en prácticas científicas. Una de estas situaciones es la evaluación de la confiabilidad de las aplicaciones móviles (Apps) que aseguran medir fenómenos científicos o sociales. ¿Podemos confiar en lo que mide una App que detecta metales, resuelva problemas matemáticos, analice la voz o prediga la compatibilidad amorosa? ¿Qué nos dicen realmente estas herramientas sobre la validez y fiabilidad de los datos que ofrecen?

Este tipo de experiencias pueden tener valor didáctico, ya que introducen un conflicto cognitivo: el alumnado suele conocer o incluso haber utilizado este tipo de Apps, pero rara vez se ha cuestionado su funcionamiento interno o su legitimidad como instrumentos de medición. Esto crea una oportunidad didáctica para activar el pensamiento crítico poniendo en juego habilidades como el análisis, la argumentación, la indagación y los modelos mentales.

Como se muestra en la Figura 2.6.A, algunas de estas Apps, disponibles en tiendas digitales, prometen funcionalidades que van desde medir la temperatura corporal hasta analizar patrones de voz o detectar mentiras. Este tipo de recursos, con apariencia científica o tecnológica, generan un entorno ideal para que el alumnado explore nociones como fiabilidad, validez, modelos de funcionamiento y justificaciones basadas en pruebas, cuestionando la apariencia de objetividad que ofrecen muchas soluciones digitales.

Existen autores que han publicado artículos relacionados con el uso de Apps y celulares en contextos educativos, resaltando sus potencialidades como herramientas didácticas en la

enseñanza de las ciencias. Estas investigaciones coinciden en señalar que el empleo de dispositivos móviles permite ampliar las oportunidades de indagación, representación y análisis de fenómenos físicos mediante sensores accesibles, entornos familiares y actividades de bajo costo.

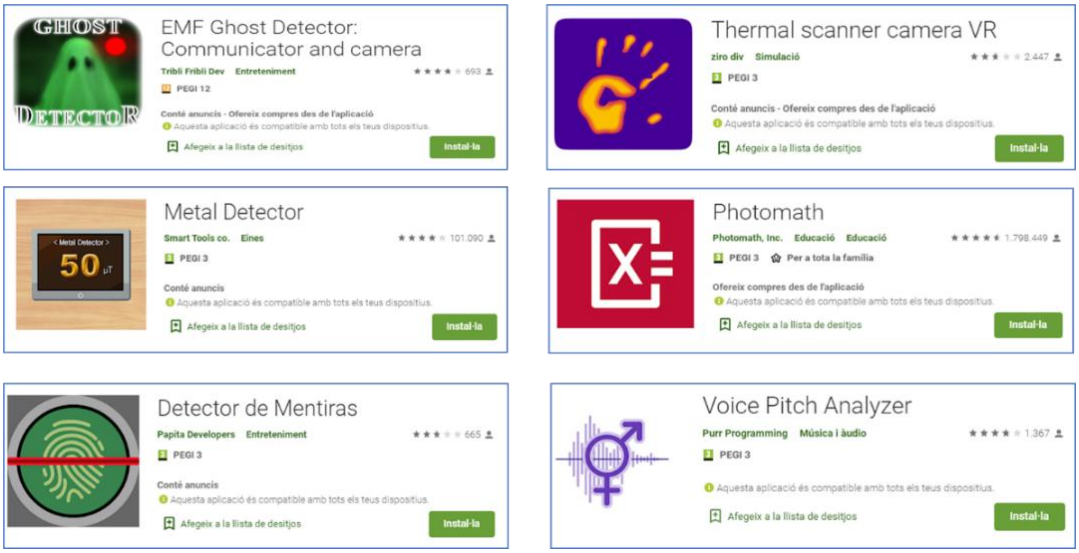

Figura 2.6.A - Ejemplos de Apps que declaran realizar mediciones a través de sensores del teléfono o procesamiento de datos, utilizadas en el proyecto App Checkers.

El artículo de Torres Climent, Bañón García y López-Simó (2017) ofrece una visión general sobre el potencial educativo de las Apps en la enseñanza de la Física y la Química. Su propuesta consistió en seleccionar y clasificar Apps gratuitas, agrupándolas según su función (enciclopedia, social, sensores, correctores, simuladores y realidad aumentada), y ensayar metodologías de uso en distintos espacios y niveles educativos. Se trabajó con estudiantes de ESO y Bachillerato, dentro y fuera del aula, identificando ventajas como la motivación, la posibilidad de trabajar de forma autónoma, la evaluación inmediata y el uso interdisciplinar. Aunque también se reportaron dificultades como la necesidad de conexión a internet en algunas Apps, la disponibilidad en idiomas no siempre accesibles para todo el alumnado, y la limitación de memoria o compatibilidad de ciertos dispositivos móviles. El trabajo concluye que el uso de aplicaciones móviles puede enriquecer la enseñanza de ciencias cuando se acompaña de una planificación metodológica adecuada y guía docente con criterios de aprendizaje claros.

Además, el artículo de Torres Climent et al. (2017) destaca que muchas de las experiencias didácticas revisadas incorporan prácticas científicas como la toma de datos experimentales mediante sensores, la formulación y validación de hipótesis, y el análisis crítico

de resultados, todas ellas vinculadas a propuestas de indagación escolar. En particular, se menciona que las Apps que operan como sensores (como Science Journal) o simuladores permiten al alumnado construir evidencia empírica en torno a fenómenos físicos, lo cual favorece la aproximación al trabajo científico en el aula. Así, el artículo no solo clasifica Apps, sino que también vincula su uso con el desarrollo del pensamiento crítico y la alfabetización científica mediante indagación.

En básica y secundaria, hay estudios que han documentado experiencias concretas con estudiantes que emplean Apps y smartphones como parte de actividades de indagación. Rodríguez-Arteche et al. (2024) describen una experiencia con alumnado de 5º y 6º de primaria, quienes participaron en una propuesta STEAM orientada a diseñar y construir instrumentos musicales, valiéndose de aplicaciones como afinadores o sonómetros. El trabajo se organizó en cinco fases que incluyeron la formulación de preguntas investigables, la emisión de hipótesis, la planificación de un diseño experimental, la recolección y expresión de datos, y la obtención de conclusiones. El alumnado midió, por ejemplo, cómo afecta el volumen de aire en un instrumento de viento a su tono, o cómo varía la intensidad de un instrumento de cuerda según el tamaño de su caja de resonancia. Para ello, emplearon Apps que permitían medir con precisión frecuencia e intensidad sonora, como afinadores digitales y medidores de decibelios. Se trabajaron explícitamente prácticas científicas como la identificación de variables independientes y dependientes, el control de condiciones experimentales, la representación gráfica de resultados, y la validación de hipótesis. Esta aproximación permitió que el estudiantado explorara fenómenos acústicos de forma activa, creativa y crítica, integrando herramientas tecnológicas de uso cotidiano en el desarrollo de competencias científicas, técnicas y artísticas.

En secundaria, González y González (2016) desarrollaron un conjunto de actividades prácticas, en las cuales estudiantes realizaron experiencias con sus propios teléfonos móviles, tanto dentro como fuera del laboratorio. Se propusieron trabajos como el estudio del movimiento de un péndulo, frecuencias de vibración de varillas de diferentes materiales y secciones o estudiar el coeficiente de roce entre diferentes superficies. En el caso del péndulo, se construyó un sistema oscilante atando el smartphone dentro de una bolsa a un hilo, de modo que el propio teléfono actuaba como masa pendular. Se registraron las aceleraciones durante el movimiento oscilatorio. Esto permitió analizar parámetros como el período y medir la aceleración de gravedad. En general, González y González (2016) muestran que con el uso de smartphone se pueden trabajar habilidades como la observación, recolección y análisis de

datos, interpretación crítica de errores y contraste con bibliografía. Una de las fortalezas del estudio fue permitir el diseño libre de experiencias por parte del alumnado, estimulando la creatividad, el análisis crítico y la conexión entre saberes científicos y la vida cotidiana. De manera que el uso del dispositivo no se limite al registro y comprobación, sino que promueva el razonamiento, la comparación y la argumentación espontánea sobre diseño y resultados obtenidos por el alumnado.

Monteiro, Stari y Martí (2022) emplearon teléfonos inteligentes como laboratorios portátiles para la enseñanza de la física en niveles secundarios y universitarios. Los dispositivos móviles, gracias a sus sensores integrados como acelerómetros, giroscopios y magnetómetros, permitieron realizar experimentos que antes requerían equipamiento costoso o inaccesible. El objetivo principal fue promover un aprendizaje activo, desarrollando habilidades en medición, análisis de datos y comprensión de fenómenos físicos en contextos cotidianos. Los resultados mostraron que los estudiantes lograron una mayor implicación y comprensión conceptual al utilizar sus propios dispositivos, facilitando además la toma masiva de datos y el trabajo colaborativo. Esta metodología demostró ser una estrategia eficaz para integrar tecnología accesible en la enseñanza experimental (Monteiro et al., 2022).

El libro Experimentación en Física con dispositivos móviles 2ª Edición de Lorenzo Ramírez (2022) propone muchas experiencias didácticas para secundaria y bachillerato que integran teléfonos móviles y tabletas como herramientas de medición experimental. A través de sensores integrados o externos, conectados por USB o bluetooth, los dispositivos permiten recopilar datos en áreas como mecánica, ondas, electromagnetismo o física del cuerpo humano. Cada experimento incluye el objetivo didáctico, el protocolo y las apps necesarias, como Phyphox o Physics Toolbox. Aunque no se presentan resultados cuantitativos detallados, el autor destaca el potencial pedagógico de estas experiencias para promover el trabajo autónomo y extender el laboratorio más allá del aula. El libro es de libre acceso y puede descargarse en formato PDF desde el sitio web https://experimentaciolliure.com (Ramírez, 2022), donde también se alojan materiales complementarios y enlaces a aplicaciones.

López Simó (2018), utilizó un teléfono móvil para registrar en video la caída de un imán a través de una bobina, con el fin de analizar experimentalmente la ley de Faraday-Lenz. mediante el software Tracker, obtuvo la posición del imán en función del tiempo, permitiendo estimar indirectamente la variación del flujo magnético. Paralelamente, un sensor de voltaje medía la FEM inducida en la bobina. El objetivo era que el alumnado comprendiera la relación entre electricidad y magnetismo mediante datos reales. Al comparar la FEM calculada con la

medida experimental mediante los análisis gráficos el alumnado pudo determinar la validez del modelado numérico lo que facilitó el aprendizaje del concepto de inducción electromagnética (López Simó, 2018).

Un enfoque complementario aparece en Sans et al. (2013), donde se utiliza el sensor de luz ambiental del teléfono para estudiar oscilaciones acopladas de resortes en laboratorio universitario. Aunque se trata de un entorno más avanzado, los datos muestran que los estudiantes universitarios pueden trabajar con modelos matemáticos, estimación de parámetros físicos, y análisis de ajuste, promoviendo prácticas científicas como la modelización y la validación experimental. Las ventajas observadas incluyen la alta precisión de los resultados, el bajo costo del equipamiento y la motivación generada al usar dispositivos propios. Sin embargo, se requiere una guía didáctica del profesorado rigurosa para interpretar correctamente los datos.

En niveles superiores, también se han reportado usos exitosos de Apps y smartphone como herramienta experimental y formativa. González de los Reyes (2024) desarrolló una propuesta en la Universidad Tecnológica de La Habana para trabajar conceptos de Física Moderna como el efecto fotoeléctrico y las curvas volt-ampéricas de diodos, usando Apps como: EveryCircuit y Efecto Fotoeléctrico. Participaron estudiantes universitarios de la carrera de Telecomunicaciones, organizados en tres subgrupos con actividades diferenciadas. Se implementó un diseño cuasi-experimental con pre-test y post-test, mostrando mejoras en la comprensión conceptual tras el uso de las Apps. Además, el alumnado expresó altos niveles de satisfacción por el aprendizaje autónomo y contextualizado. La investigación destacó como ventajas el acceso libre a simuladores interactivos y la posibilidad de vincular teoría y práctica sin requerir equipamiento costoso. No obstante, se identificó la necesidad de acompañamiento docente y diseño metodológico para evitar usos superficiales.

Otro estudio en nivel universitario fue realizado por Salzedas (2024) en la Universidad de Oporto, donde se aprovechó el magnetómetro del smartphone para medir campos magnéticos locales. En este caso, el objetivo fue aplicar la física a entornos reales, como las catenarias del metro de Porto, y enseñar a calibrar sensores usando la App Phyphox. Esta actividad permitió trabajar procedimientos experimentales rigurosos y fomentar la interpretación crítica de datos. Aunque la implementación requiere conocimientos técnicos, se constató que el alumnado desarrolla competencias en el uso de tecnología científica real y conectada al entorno urbano.

Finalmente, el artículo de Pereira (2024) documenta una experiencia en formación docente universitaria, donde se usó el smartphone para estudiar el momento de inercia mediante un experimento de rotación con sensores de campo magnético. Participaron futuros profesores de física de la Universidad Federal do Sul e Sudeste do Pará (Brasil). Se promovió el desarrollo de habilidades como el diseño experimental, la interpretación de datos y la validación de modelos. Las ventajas incluyen el uso de materiales de bajo costo, la posibilidad de repetir experiencias en distintos contextos, y la apropiación crítica del dispositivo. El análisis de datos, guiado por regresión y ecuaciones físicas, permitió trabajar prácticas científicas avanzadas de forma accesible.

En síntesis, estos artículos muestran que el uso educativo de Apps y del uso del celular se ha extendido a múltiples niveles del sistema educativo. Estas herramientas pueden ser altamente efectivas para promover el pensamiento crítico, siempre que se integren con prácticas científicas auténticas y se acompañen de un diseño didáctico intencionado. Las evidencias empíricas disponibles muestran que es posible articular lo cotidiano, lo digital y lo científico en experiencias que desafían las creencias previas del alumnado y potencian el aprendizaje significativo en ciencias.

2.6.1 El funcionamiento subyacente de las Apps

Existe un modelo que podemos llamar estándar en computación y teorías de sistemas: entrada-proceso-salida, que sirve para representar el funcionamiento de dispositivos ingenieriles y/o sistemas (Ibañez et al., 2010, p.11): “Los diagramas entrada-proceso-salida son una representación de las entradas (condiciones) y las salidas (requerimientos) de un problema”, este modelo describe el flujo del procesamiento de datos. Señala Ibañez et al. (2010) que el proceso corresponde a un algoritmo (en el caso de programación) o reglas predefinidas que se ejecutan en el procesador (CPU), sensores, memoria RAM, tarjeta gráfica (GPU), software o tecnología en la nube (Ingeniería Cloud), o distintas combinaciones de ellas.

El modelo de entrada-proceso-salida es fundamental para el funcionamiento de las Apps, ya que describe cómo los datos ingresan a la Apps, cómo son procesados y qué tipo de información se obtiene como resultado.

Cada Apps recibe datos a través de distintos sensores, cámaras o la interacción del usuario con la pantalla táctil. Por ejemplo, en el caso de App FotoMath (https://photomath.es/), la cámara del dispositivo captura imágenes de ecuaciones matemáticas, mientras que la App, Detector de metales (https://play.google.com/store/apps/details?id=kr.sira.metal&hl=es)

obtiene información del magnetómetro del teléfono para detectar variaciones en los campos magnéticos cercanos.

Una vez ingresados los datos, cada aplicación los procesa mediante algoritmos específicos que transforman la información en resultados útiles. La App FotoMath emplea reconocimiento óptico de caracteres para identificar números y símbolos matemáticos, resolviendo ecuaciones con algoritmos de álgebra. Detector de metales analiza los datos del magnetómetro para determinar la presencia de objetos metálicos.

Finalmente, los datos procesados se presentan de diferentes maneras, dependiendo de la App. FotoMath muestra el resultado de la ecuación resuelta junto con el procedimiento paso a paso. Detector de metales indica la intensidad del campo magnético en la pantalla y emite alertas cuando se detectan niveles altos.

Este modelo de entrada-proceso-salida sirve para comprender el funcionamiento de las Apps, y ayuda a que el diseño garantice que los datos ingresados sean interpretados correctamente y se presenten de manera útil al usuario.

2.6.1.1 Tipos de Apps

Según Montiel (2017) existe diferentes tipos de Apps, las principales son las Apps nativas, web, hibridas y en la nube. Las Apps nativas (Montiel, 2017, p.80) se desarrollan específicamente para un sistema operativo móvil en particular, como Android o iOS, utilizando los lenguajes de programación propios de cada plataforma. Este tipo de Apps están diseñadas para integrarse de manera eficiente con el sistema operativo. Además, permiten el acceso total a las funcionalidades del hardware del dispositivo, como GPS, cámara y sensores en general. Sin embargo, su desarrollo implica un costo mayor, ya que requiere la creación de versiones independientes para cada sistema operativo, además de un proceso de actualización más complejo, dado que cualquier modificación debe pasar por una revisión en las tiendas de Apps.

Las Apps web (Montiel, 2017, p.80), en cambio, se ejecutan directamente en el navegador de un dispositivo móvil sin necesidad de ser descargadas ni instaladas. Desarrolladas con tecnologías como HTML, CSS y JavaScript, permiten el acceso desde cualquier dispositivo con conexión a internet. Su mayor ventaja radica en la facilidad de actualización, ya que los cambios se aplican de manera instantánea sin requerir intervención del usuario. También presentan un costo de desarrollo menor, ya que utilizan un único código base para todas las plataformas. No obstante, su rendimiento suele ser inferior al de las Apps nativas, dependen

completamente de la conexión a internet y no pueden acceder de manera completa a las funciones del hardware del dispositivo.

Las Apps híbridas (Montiel, 2017, p.80) combinan características de las Apps nativas y web, ya que se desarrollan con tecnologías web, pero se ejecutan dentro de un contenedor nativo, lo que permite su distribución en las tiendas de Apps. Esta estrategia permite reducir costos y tiempos de desarrollo, ya que se puede utilizar un solo código para múltiples plataformas. A pesar de su versatilidad, este tipo de Apps presenta algunas limitaciones en cuanto a rendimiento y acceso al hardware del dispositivo, ya que no aprovechan por completo las ventajas de las Apps nativas. Además, la experiencia de usuario puede variar dependiendo del dispositivo en el que se ejecute la aplicación.

Por último, las Apps en la nube (Montiel, 2017, p.81) funcionan a través de servidores remotos y no requieren almacenamiento significativo en el dispositivo. Se accede a ellas mediante una conexión a internet y permiten la sincronización de datos entre distintos dispositivos. Su principal ventaja es que no ocupan espacio en el almacenamiento del usuario y las actualizaciones se realizan automáticamente sin necesidad de intervención. Sin embargo, dependen completamente de la conexión a internet para su funcionamiento y pueden presentar problemas de rendimiento si los servidores no cuentan con la capacidad suficiente para soportar múltiples usuarios simultáneamente. Además, requieren medidas de seguridad para proteger los datos almacenados en la nube.

2.6.1.2 Elementos de una App

Una aplicación móvil (App) es un programa informático creado para realizar una tarea específica dentro de un sistema operativo móvil. Según Almanza (2014), las Apps surgen para responder a necesidades concretas de los usuarios, facilitando o posibilitando la ejecución de tareas puntuales. Gracias a su diversidad, pueden estar dirigidas a múltiples sectores, tales como el educativo, empresarial, recreativo, de comunicaciones o bancario, entre otros.

Uno de los elementos fundamentales en toda App es la interfaz de usuario, la cual permite la interacción entre el usuario y el dispositivo. Esta interfaz comprende todos los puntos de contacto, desde botones hasta menús y elementos visuales, y desempeña un papel crucial en la experiencia del usuario. Para construir una interfaz efectiva, resulta esencial seleccionar una paleta de colores adecuada, elegir tipografías legibles, utilizar formatos compatibles y adoptar un diseño responsivo. Este último garantiza una correcta adaptación a distintos dispositivos y tamaños de pantalla. En el ámbito educativo, un diseño de interfaz bien estructurado no solo

aporta valor estético, sino que mejora la funcionalidad y accesibilidad del entorno digital. Además, facilita la navegación, optimiza el acceso a contenidos y refuerza la comunicación entre estudiantes y docentes. Al incorporar elementos interactivos como chats en vivo o formularios personalizados, se promueve una mayor participación y motivación por parte del alumnado.

En relación con el funcionamiento técnico de una App, por ejemplo, en el caso de la App Cámara Térmica (figura 2.6.B) permite ilustrar algunos aspectos del hardware, el desarrollador indica que necesita un sistema operativo Android 6.0 o versiones posteriores y que tiene un tamaño de descarga aproximado de 25 MB. Además, se menciona su compatibilidad con diversos modelos de teléfonos móviles, sin detallar los componentes internos de cada uno. De este modo, se deduce que cualquier dispositivo de gama baja o media con soporte para Android 6.0 debería ser capaz de ejecutar esta App sin dificultades.

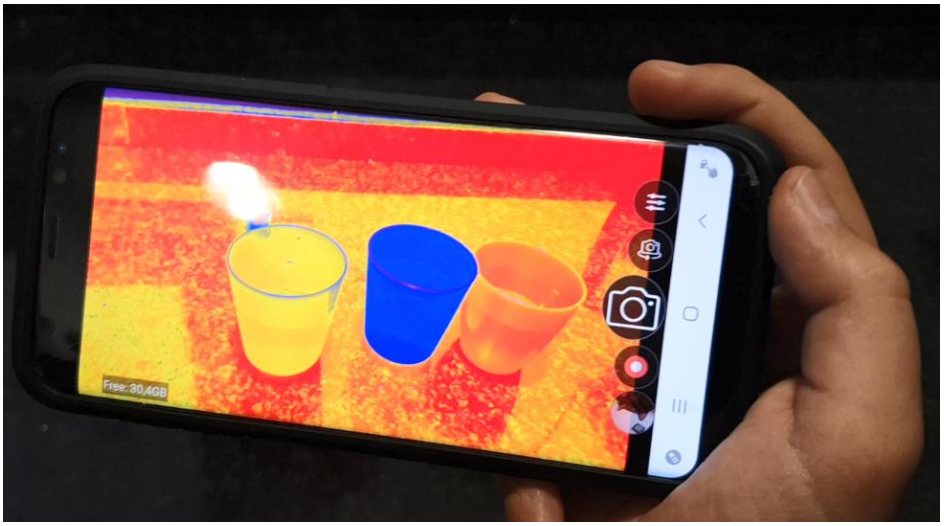

Figura 2.6.B Estudiante trabajando con la App Cámara Térmica.

Una cuestión interesante que suele surgir es si una App es un software o si posee un software de funcionamiento propio. La respuesta es afirmativa en ambos casos. Por un lado, toda App es un tipo de software, ya que el software se define como un conjunto de instrucciones, datos y programas que permiten ejecutar tareas en un dispositivo. Este concepto abarca desde sistemas operativos hasta Apps, controladores y herramientas de desarrollo. Así, Apps como WhatsApp, Instagram o la Cámara Térmica son ejemplos concretos de software en forma de App, mientras que Android, Windows o Linux son sistemas operativos, también considerados software, pero no catalogados como Apps.

Por otro lado, cada App tiene un software de funcionamiento específico, que consiste en el conjunto de códigos, algoritmos y procesos internos que permiten su operatividad. Una App no es únicamente una interfaz gráfica, sino que incluye una arquitectura funcional que interpreta y ejecuta las acciones del usuario. En el caso de la App Cámara Térmica, su software interno toma imágenes desde la cámara del dispositivo, aplica filtros y genera una imagen simulada con efecto térmico. Para lograrlo, utiliza algoritmos que procesan la información visual en tiempo real, lo que constituye su núcleo de funcionamiento.

CAPÍTULO III

PRESENTACIÓN DE LOS ESTUDIOS Y OBJETIVOS DE INVESTIGACIÓN

El Capítulo III detalla las tres variantes del proyecto implementadas en distintos contextos educativos, abarcando secundaria y bachillerato en Cataluña, y escuelas en Chile en enseñanza secundaria. Cada variante se asocia a un producto final diferente (videos, pósteres o plantilla) y a un eje de análisis específico (indagación y argumentación, estrategias o modelos mentales del alumnado). Además, se presentan los objetivos de investigación para cada estudio relacionados con cada una de las variantes del proyecto App Checkers.

3.1 Variantes del proyecto App Checkers

El proyecto App Checkers que se ha definido en la introducción en 1.2.2 ha sido implementado en diferentes contextos. En cada contexto se ha implementado de una forma distinta, con variaciones tanto en el diseño como en la metodología aplicada, según las necesidades y características del alumnado y los recursos disponibles. Contextos que implican diferentes edades, diferentes lugares geográficos, diferentes demandas y aproximaciones metodológicas.

En la primera variante, el proyecto se aplicó en España, específicamente en Cataluña, con estudiantes de 2º de ESO, trabajando desde la indagación y argumentación científica escolar con un énfasis en la verificación de Apps y la construcción de una escala de confiabilidad que sirvió como andamiaje al alumnado. En la segunda variante, se implementó en primer año de bachillerato también en Cataluña, con un nivel de exigencia más alto, centrándose en el proceso de indagación y las estrategias usadas para determinar la confiabilidad de las Apps mediante triangulación y elaboración de pósteres científicos. Finalmente, en la tercera variante, realizada en Santiago de Chile, se trabajó con estudiantes de entre 13 y 17 años de diferentes niveles escolares, utilizando una plantilla estructurada como herramienta de andamiaje, incluyendo la construcción de modelos de funcionamiento de Apps como parte del proyecto.

Cada una de estas implementaciones presenta diferencias en cuanto a edad del alumnado, ubicación geográfica, nivel de autonomía otorgado, tipo de producto final exigido y nivel de estructuración de las tareas, permitiendo estudiar distintas facetas del pensamiento crítico en ciencia escolar.

A continuación, en la Tabla 3.A, se presentan de forma comparativa las principales características de cada una de las tres implementaciones del proyecto App Checkers.

Tabla 3.1.A - Características principales de las tres implementaciones del proyecto App Checkers.

Aplicación	Lugar	Edades	Demanda pedida al estudiante	Análisis hecho en el estudio	Producto final del alumnado	Principales diferencias
Primera variante	Instituto Pau Vila de Sabadell, Cataluña, España	13-14 años (2º de ESO)	Diseñar y ejecutar un experimento de verificación, construir una escala de confiabilidad (sólo en 2019-20), y argumentar resultados.	Estudio 1: Análisis del desempeño en prácticas de indagación y argumentación.	Video argumentativo (2-4 minutos)	Implementación en secundaria baja, uso opcional de escala de fiabilidad como andamiaje, libertad en selección de Apps. Foco en indagación y argumentación.

Segunda variante	Instituto Francesc Ferrer i Guàrdia, Cataluña, España	16-17 años (1º bachillerato)	Diseñar investigación para evaluar validez y fiabilidad de una App, empleando estrategias de triangulación.	Estudio 2: Análisis cualitativo de estrategias de validez y niveles de fiabilidad en pósteres científicos.	Póster científico	Nivel de exigencia superior, foco explícito en estrategias de verificación, foco en indagación, uso de póster como formato de divulgación científica.
Tercera variante	Colegios en Santiago de Chile	13 a 17 años (8° básico a 4° medio)	Construir un modelo de funcionamiento de la App, diseñar y ejecutar experimentos guiados mediante plantilla estructurada.	Estudio 3: Análisis de modelos mentales del alumnado de funcionamiento de las Apps y evaluación del nivel de desempeño en la construcción de modelos de funcionamiento de las Apps.	Plantilla estructurada con actividades	Implementación en cuatro colegios, uso de plantilla como andamiaje, enfoque explícito en modelos de funcionamiento de las Apps, selección cerrada de Apps.

La Tabla 3.A permite observar que, aunque todas las implementaciones comparten un objetivo general relacionado con el desarrollo de habilidades de investigación y pensamiento crítico alrededor del proyecto App Checkers, cada una de ellas ha adaptado un enfoque metodológico específico. Las principales diferencias residen en el nivel de autonomía del alumnado (más alta en las primeras variantes, más guiada en la tercera), el producto final esperado (video, póster o plantilla) y el grado de estructuración de las tareas (menos en el primer estudio, intermedio en el segundo, y mucho mayor en el tercero). Además, el rango de edades y niveles educativos involucrados varía, abarcando desde los 13 hasta los 17 años, y extendiéndose geográficamente de Cataluña a Santiago de Chile.

Cada una de las variantes descritas en la Tabla 3.A se convirtió en un estudio distinto, con objetivos específicos, marcos analíticos diferenciados y productos finales particulares. Tal como se muestra en la columna "Análisis hecho en el estudio", cada implementación dio lugar a un abordaje metodológico adaptado al contexto educativo correspondiente: el primer estudio se centró en el análisis del desempeño en prácticas de indagación y argumentación; el segundo en la exploración cualitativa de estrategias epistémicas de validez y niveles de fiabilidad; y el tercero en la evaluación de modelos mentales construidos por el alumnado sobre el funcionamiento de las Apps. Estas diferencias metodológicas y de enfoque permiten abordar el pensamiento crítico en ciencia escolar desde múltiples dimensiones complementarias, que a continuación se explicarán y desarrollarán en el cuerpo de esta tesis.

3.2 Objetivos de la investigación

En primer lugar, como objetivo central al que responde esta tesis es analizar la relación entre un proyecto de verificación de Apps y el desarrollo de las prácticas científicas de

argumentación, indagación y modelización. Posteriormente, se establecieron objetivos particulares correspondientes a cada uno de los estudios. Estos son:

Primer estudio:

- Caracterizar el nivel de desempeño en las prácticas científicas de indagación y argumentación de los estudiantes participantes en el proyecto App Checkers.
- Analizar el nivel de desempeño caracterizado en indagación y argumentación según dos variables didácticas: tipo de app y andamiaje.

Segundo estudio:

- Identificar las estrategias que emplean los y las estudiantes para evaluar la validez de los resultados obtenidos por una App.
- Analizar hasta qué punto los y las estudiantes incorporan procedimientos de fiabilidad para evaluar la estabilidad de los resultados que ofrece cada App.

Tercer estudio:

- Evaluar los modelos mentales que usan los estudiantes entre 13 y 17 años en Chile para explicar el funcionamiento de las Apps cuando tratan de verificar su confiabilidad.
- Analizar el nivel de elaboración de los modelos expresados en función de dos variables didácticas: tipo de app y edad.

CAPÍTULO IV

ESTUDIO 1: ANÁLISIS DEL DESEMPEÑO EN INDAGACIÓN Y ARGUMENTACIÓN DEL ALUMNADO DE 2° DE ESO EN UN VIDEO DE APP CHECKERS

El capítulo IV presenta el Estudio 1, que analiza cómo estudiantes de 2º de ESO desarrollan prácticas de indagación y argumentación al evaluar la confiabilidad de Apps. Se aplicaron rúbricas basadas principalmente en los Conceptos de Evidencia (CoE) y el modelo de Toulmin a 76 videos, permitiendo caracterizar diferentes niveles de desempeño.

Se identificaron patrones y habilidades relacionadas con calibración, repetitividad y límites operativos, útiles para definir perfiles competenciales. Este estudio dio lugar a una publicación académica en la ***Revista Electrónica de Enseñanza de las Ciencias*** (Aguilera y López-Simó, 2025), disponible en http://reec.uvigo.es/

Aguilera, M., y López-Simó, V. (2025). Estudio del desempeño en prácticas científicas de indagación y argumentación de alumnado de enseñanza secundaria mediante una aplicación de móvil. *Revista Electrónica de Enseñanza de las Ciencias*, *24*(1), 52–73.

4.1 Introducción al estudio 1

Un grupo de alumnos del Instituto Pau Vila de Sabadell en Cataluña participó en el proyecto App Checkers, como se explica en la tabla 3.A de esta tesis. El objetivo del profesorado fue introducirlos en el diseño de investigaciones a partir de situaciones cotidianas y así promover sus habilidades de indagación y argumentación. Este consistió en trabajar con sus celulares, seleccionando alguna App disponible de manera gratuita, que mida algo, o que supuestamente lo haga, y que pueda tener utilidad tanto en contextos sociales como profesionales. El alumnado llevó a cabo un estudio científico sobre la confiabilidad de la App, independientemente de si la fiabilidad y validez se presumen de antemano. En este proyecto, la pregunta de investigación o problema de investigación es proporcionada al alumnado: ¿Qué tan confiable es la aplicación seleccionada?, donde por confiabilidad se entenderá la validez y fiabilidad de la App.

Como la elección de la App fue libre, esto se tradujo en una diversidad de enfoques que permitió explorar diferentes áreas de medición y poner a prueba las habilidades de indagación científica, argumentación y pensamiento crítico del alumnado. El proceso incluyó el diseño y la ejecución de experimentos, así como la evaluación de la confiablidad de las Apps seleccionadas, fomentando la reflexión sobre la importancia de la verificación en la ciencia y la tecnología.

4.2 Contexto y objetivos de investigación

El proyecto fue realizado en el Instituto Pau Vila de Sabadell en Cataluña, un instituto público con aproximadamente 150 estudiantes por curso organizado en 5 grupos-clases heterogéneos. Se trabajó con estudiantes de segundo año de ESO durante dos años académicos distintos, con tres profesores distintos como se muestra en la tabla 4.2.A.

Tabla 4.2.A - Profesores que implementaron App Checkers.

Profesor	Año académico	N° videos
Profesor 1	2019-20	29
Profesor 2	2020-21	35
Profesor 3	2020-21	12

El proyecto fue concretado en 4 etapas definidas como:

1a Etapa - Discusión inicial sobre concepto de verificación: Trabajo de dos horas en torno a la pregunta: ¿Qué significa que una App que tiene por objetivo la medición, sea falsa o

engañosa? Se realizó una introducción donde los estudiantes visualizaron vídeos de Fact Check para reflexionar sobre el significado de "verificación". Posteriormente se trabajó alrededor de sus ideas iniciales sobre "medir" y "confiabilidad", para luego explorar diferentes Apps, gratuitas y de contenido apto para menores de edad, indagaron sobre su funcionamiento y analizaron los comentarios y sus valoraciones para finalmente elaborar de un listado de Apps posibles de ser verificadas.

2ª Etapa - Selección de una App y diseño de una investigación: Trabajo de dos horas en torno a la pregunta: ¿Qué significa que una App que tiene por objetivo la medición, sea falsa o engañosa? Se realizó una introducción donde los estudiantes visualizaron vídeos de Fact Check para reflexionar sobre el significado de "verificación". Posteriormente se trabajó alrededor de sus ideas iniciales sobre "medir" y "confiabilidad", para luego explorar diferentes Apps, gratuitas y de contenido apto para menores de edad, indagaron sobre su funcionamiento y analizaron los comentarios y sus valoraciones para finalmente elaborar de un listado de Apps posibles de ser verificadas.

3ª Etapa - Definición de una escala de confiabilidad: Los estudiantes, durante 2 horas de clase, diseñaron y ejecutaron un experimento para poner a prueba una App. Las acciones que los estudiantes llevaron a cabo fueron la elección de la App, seguidamente de la planificación de uno o varios experimentos para poner a prueba la App y recoger resultados.

4ª Etapa - Elaboración de un video: Se dedicaron 3 horas de trabajo y consistió en la elaboración de un guion por parte de los estudiantes que posteriormente facilitó la elaboración y edición de una video-presentación con una duración de entre 2 y 4 minutos.

Además, el hecho de implementar el proyecto durante 3 cursos distintos generó algunas variaciones internas. En el año académico 19-20 se implementó el proyecto con el profesor 1 que había participado en el diseñado, y en el curso 2020-21 se volvió a implementar con otros dos profesores distintos (profesor 2 y profesor 3), que, por limitaciones de tiempo, decidieron prescindir de la actividad 3 del proyecto, la que correspondió a definir de una escala de confiabilidad con el alumnado, que se muestra en la figura 4.2.A. Esta escala sirvió como andamiaje para que los estudiantes del curso 2019-20 tuviera criterios para valorar la confiabilidad de la App en su argumentación

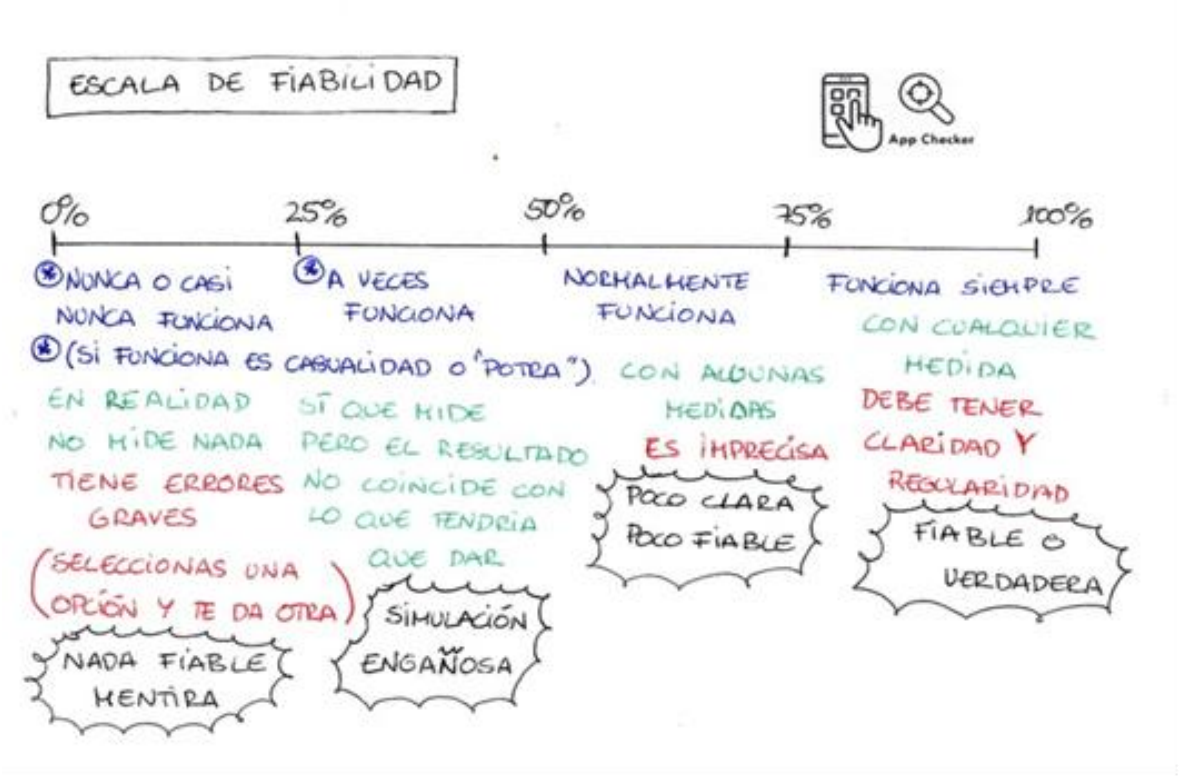

Figura 4.2.A - Fotografía del dosier de un estudiante donde construye una escala de confiabilidad entre 0% y 100%. Esta actividad se usó en el curso 2019-20, pero no en el 2020-21.

Durante el proyecto se observó que la variedad de Apps elegidas por los estudiantes era muy grande, por lo que surgió la pregunta de si el tipo de App había influido en el desempeño en indagación y argumentación que mostraban los estudiantes.

Dada la variedad de videos elaborados por los estudiantes y el posible efecto tanto del andamiaje usado como del tipo de App elegida, se plantearon como objetivos de investigación:

- Caracterizar el nivel de desempeño en las prácticas científicas de indagación y argumentación de los estudiantes participantes en el proyecto App Checkers.
- Analizar el nivel de desempeño caracterizado en indagación y argumentación según dos variables didácticas: tipo de app y andamiaje.

4.3 Metodología

A continuación, se presenta la metodología seguida en el estudio 1, incluyendo el proceso de recolección de datos, el tratamiento de la información y la categorización de los desempeños observados en indagación y argumentación. Se describe el proceso de recolección y generación de categorías de análisis de datos desarrollado para el proyecto App Checkers. Para el análisis

de indagación y argumentación se utilizaron marcos teóricos 2.2 y 2.3, definiendo cinco niveles de desempeño para ambas prácticas científicas.

4.3.1 Recogida de datos

En las implementaciones del proyecto App Checkers 2019-20 y 2020-21 el producto final ha sido un video, que fue realizado por el alumnado en sus casas, flexibilidad otorgada inicialmente a causa del confinamiento por la pandemia del COVID-19. Del curso 19-20 se recibieron un total de 29 vídeos, de los cuales 22 eran individuales, 5 en parejas, 1 en trío y 1 de cuatro estudiantes. De las cohortes 20-21 se recibieron 47 videos de los cuales 20 fueron individuales, 14 en parejas, 9 en tríos, 3 de a cuatro y 1 de a cinco estudiantes. En la tabla 4.3.A se muestra la mayoría de las Apps con las cuales trabajaron los estudiantes.

Tabla 4.3.A - Número de videos y Tipo de App para cada implementación. La categoría otros corresponden a Apps elegidas que sólo aparecen una vez.

Tipo de App	**Número de vídeos**			
	Curso 2019-20 Profesor 1	**Curso 2020-21 Profesor 2**	**Curso 2020-21 Profesor 3**	**Total**
Detector de Mentiras	11	3	0	14
Detector de Metales	3	5	0	8
Cámara de Rayos X	4	1	2	7
Medidor de longitudes	0	7	0	7
Sonómetro	2	5	0	7
Detector de fantasmas	2	2	2	6
Medidor de temperatura	0	2	1	3
Detector de edad	1	2	0	3
Detector Infrarrojo	1	0	1	2
Buscador de celular	0	1	1	2
Medidor de luz	0	2	0	2
Otros	5	5	5	15

El procedimiento de tratamiento de datos correspondió primero a una codificación con la finalidad de sistematizar los videos en una base de datos y luego una transcripción de cada video con la finalidad de estudiarlos en profundidad tanto en indagación y argumentación.

4.3.2 Construcción de categorías para el análisis del desempeño en indagación

Para determinar criterios y niveles de desempeño que permitan evaluar las habilidades procedimentales del alumnado, definidas como Ferrés (2017), es que se planteó un diseño que consistió en los siguientes pasos:

1. **Revisión de la literatura:** Se realizó una búsqueda en Google Scholar de rúbricas para medir habilidades procedimentales en indagación. Se formularon términos de búsqueda con operadores booleanos sencillos y luego como criterios de inclusión se tomaron en cuenta aquellos artículos académicos que explícitamente mostraran criterios de indagación (o dimensiones si se prefiere) graduados en niveles de desempeño, abordando la evaluación de habilidades procedimentales en indagación científica.
2. **Análisis Comparativo:** Consistió en contrastar de forma sistemática los criterios de evaluación presentes en las rúbricas de los artículos seleccionados, manteniendo la integridad de los instrumentos originales. Se diseñó una tabla que permitió establecer equivalencias semánticas reales entre los criterios de diferentes autores encontrados en el punto anterior, considerando tanto su formulación como los niveles de desempeño asociados. Asimismo, se incorporó una columna adicional construida a partir del marco de los Conceptos de Evidencia (CoE), lo que permitió extender el análisis hacia dimensiones conceptuales y transversales no siempre presentes en los instrumentos revisados.
3. **Selección de criterios y creación de niveles de desempeño en indagación**. Tras el análisis comparativo de rúbricas y marcos conceptuales, se seleccionaron aquellos criterios que resultaban relevantes para evaluar el desempeño de los estudiantes en el proyecto. Estos criterios fueron extraídos principalmente del marco de los Conceptos de Evidencia (CoE), pero reformulados en términos contextuales que reflejaran las acciones reales observadas en los videos generados por el alumnado. Este proceso de reformulación consistió en traducir cada criterio a un lenguaje funcional y adaptado a la tarea concreta de los estudiantes. Por ejemplo, en lugar de evaluar genéricamente si el alumno “calibra el instrumento”, se adoptó el descriptor “calibra la App si esta lo requiere”. Cada descriptor fue construido como una unidad de observación contextualizada y vinculado explícitamente al CoE correspondiente.

4. **Codificación iterativa y refinamiento de criterios mediante análisis de la muestra**. Una vez definido el conjunto de criterios contextualizados, se procedió a usar el análisis iterativo de los 76 videos del alumnado. En el análisis iterativo se aplicaron los criterios a cada video, y se fueron ajustando tanto los criterios como sus niveles de desempeño conforme se detectaban matices y patrones en las acciones del alumnado. Esta iteración permitió refinarlos y determinar criterios cuya variabilidad no era significativa. En algunos casos los niveles de desempeño fueron obtenidos de las rúbricas en los artículos derivados de la revisión de la literatura, a los cuales se le realizaron ajustes menores en función de las acciones observadas en los videos.

5. **Determinación de la frecuencia de desempeños observados.** Con los criterios y sus niveles de desempeño se construyó una tabla de frecuencias que recoge cuántos videos se ubican en cada nivel de desempeño para cada criterio. Este procedimiento permitió mapear las acciones más comunes, identificar tendencias generales y determinar los criterios con mayor capacidad discriminativa. Las frecuencias obtenidas constituyen una base empírica para clasificar los tipos de indagación desarrollados por el alumnado.

6. **Derivación de patrones y determinación de niveles de indagación**. A partir del análisis de los niveles de desempeño observados se elaboró esquema de trabajo que sintetiza los patrones típicos de actuación del alumnado. Luego, con el propósito de clasificar los niveles de desempeño competencial del alumnado se realizará un análisis sistemático de combinaciones de desempeños en criterios procedimentales observados en 76 videos evaluados. Esta estrategia permite agrupar dichas combinaciones en niveles progresivos de competencia y da lugar a una rúbrica sustentada en patrones empíricos consistentes y las decisiones metodológicas anteriores.

7. **Codificación de los niveles de indagación por parte del director de tesis**. Para garantizar la fiabilidad intercodificadores del proceso de análisis, el director de tesis codificó de manera independiente el 30% de los datos. Esto permitió evaluar la consistencia en la aplicación de la rúbrica y detectar posibles discrepancias en la categorización.
8. **Comparación y consenso en la categorización**. Los resultados obtenidos en la fase de codificación independiente fueron comparados, discutiendo las diferencias hasta alcanzar un consenso en la categorización de los niveles de indagación. Este proceso permitió afinar la rúbrica y asegurar su validez, ajustando las dimensiones y criterios de evaluación en función de la concordancia interevaluador.

4.3.2.1 Revisión de la literatura

La revisión de la literatura condujo a la identificación de varios artículos de interés, centrando la atención en aquellos que incluían de forma explícita una rúbrica con criterios desarrollados en niveles de desempeño. Los artículos seleccionados se presentan resumidos en la tabla 4.3.B.

Tabla 4.3.B - Artículos seleccionados de la búsqueda.

Fórmula de búsqueda	Artículos seleccionados
"rúbrica" AND "indagación"	Ferrés, C.; Marbà, A. y Sanmartí, N. (2015) Otero, S., y Crujeiras Pérez, B. I. (2016) Crujeiras-Pérez, B. I. y Cambeiro, F. (2017) Crujeiras-Pérez, B. I., y Cambeiro, F. (2018)
("rubric" OR "analytical rubric") AND ("laboratory skills" OR "inquiry skills" OR "inquiry")	Tamir, P., Friedler, Y., y Nussionwitz, R. (1982) Knaggs, C. M., y Schneider, R. M. (2012) Arnold, J. C., Kremer, K., y Mayer, J. (2014)

Su selección recae también en una metodológica más explícita, diversidad de contextos (educación secundaria y bachillerato, asignaturas como física, química o biología) y, especialmente, por proporcionar rúbricas útiles para mapear el desarrollo de habilidades de indagación del alumnado.

4.3.2.2 Análisis Comparativo

El análisis comparativo consistió en contrastar los criterios (o dimensiones) de cada rúbrica presente en los artículos seleccionados (tabla 4.3.B). Se diseñó la tabla 4.3.C, donde se consideró mantener la integridad de cada uno de los criterios de las rúbricas no alterando ni desechando ninguno. No se modificó la redacción original, de modo que las celdas vacías de color naranjo del análisis indican únicamente la ausencia explícita de un criterio en determinada rúbrica, y no una eliminación por parte de los autores de este estudio. En la tabla 4.3.C se comparan horizontalmente las rúbricas, esta estrategia permite identificar correspondencias, ausencias y grados de especificidad entre rúbricas, respetando sus marcos teóricos de origen. Además, busca una lógica de equivalencia semántica real entre los criterios utilizados por los distintos autores de las rúbricas, es decir en algunos casos se determinó que los criterios tienen un significado o función evaluativa muy similar, aunque estén redactados de forma distinta o usen etiquetas diferentes, y considerando no solo los nombres de los criterios, sino especialmente la definición operativa y los niveles de desempeño que tienen asociados. Algunos de ellos se visibilizan en las tablas 2.2.B hasta la 2.2.E

Cuando un criterio general de una rúbrica es abordado de forma más específica por algún otro autor, se registró en la medida de los posible, esta especialización como subcriterios dentro de una misma fila, en lugar de separarlos. Si esta profundización está contenida en otra rúbrica con un criterio más genérico, ya que lo relevante es la equivalencia conceptual. Esto se hace combinando las celdas.

Por ejemplo, Tamir et al. (1982) incluyen criterios como “Formulación de problemas”, “Identificación de la variable dependiente” e “Identificación de la variable independiente”. Knaggs y Schneider (2012), por su parte, evalúan acciones como “Pregunta científica” e “Hipótesis”. Arnold et al. (2014) desagregan aspectos como “Medición de la variable dependiente”, “Variación de la variable independiente” y “Consideración de variables de confusión”. Ferrés et al. (2015), en cambio, trabaja criterios como “Identificación de problemas investigables”, “Formulación de hipótesis” e “Identificación de variables”. Estas diferencias y similitudes se representan en la tabla mediante la exclusión o combinación de celdas cuando es necesario.

Para ejemplificar el desarrollo de la tabla 4.3.C, en la primera fila y columna de la tabla, está el criterio “Formulación de problemas”, de Tamir et al. (1982) que corresponde a:

> Cualquier ítem que requiera que el estudiante formule un problema a investigar. Por ejemplo, se presenta un determinado sistema experimental y se le pide al examinado: “Formula un problema que puedas investigar con este sistema”. Otro ejemplo: después de haber realizado un experimento en particular, se le pregunta al estudiante: “¿Cuál fue el problema que intentaste investigar?”.

De manera análoga, Ferrés et al. (2015) plantean el criterio “identificación de problemas investigables” como la capacidad de plantear problemas adecuados para la investigación escolar, evaluando si estos son abordables, específicos y bien formulados. A pesar del cambio terminológico, el foco está en valorar la capacidad de traducir una situación en una pregunta científica viable. En la misma línea, Crujeiras-Pérez y Cambeiro (2017) incluyen entre sus dimensiones evaluativas la “identificación del problema”, entendida como la habilidad del alumnado para delimitar una cuestión precisa y coherente que oriente el diseño de una investigación científica. En los tres casos, se trata de evaluar el reconocer, enunciar y precisar el objeto del estudio. Por tanto, aunque cada rúbrica emplea su propia formulación y contexto, los tres criterios comparten una misma función evaluativa centrada en la articulación de un problema investigable, lo que permite establecer entre ellos una equivalencia semántica real.

La equivalencia semántica entre los criterios “Identificación de la variable dependiente” e “Identificación de la variable independiente” de la rúbrica de Tamir et al. (1982), y los ítems “Medición de la variable dependiente”, “Medición de la variable dependiente” y “Variación de la variable independiente” de Arnold et al. (2014), se justifica al considerar que ambas rúbricas evalúan el reconocimiento y uso adecuado de variables experimentales como parte central del diseño de investigaciones. Tamir et al. (1982) pone énfasis en la identificación de los tipos de variables, lo cual es un paso necesario en el planteamiento de un experimento válido. Arnold, en cambio, desagrega el proceso de manera más específica y operacional: mide si el estudiante no solo reconoce las variables, sino si efectivamente las utiliza en un contexto de medición o manipulación, evaluando así la aplicación del conocimiento de variables en el proceso experimental. Por su parte, Ferrés et al. (2015) agrupan ambos aspectos bajo el criterio “Identificación de variables”, evaluando la capacidad del estudiante para reconocer cuáles son las variables implicadas en la investigación, tanto independientes como dependientes, y para relacionarlas con la hipótesis. Esta misma lógica es adoptada en la rúbrica de Crujeiras-Pérez y Cambeiro (2018), quienes integran la identificación y operacionalización de las variables dentro del criterio general “Preparación”, que incluye la planificación de la investigación y el reconocimiento de factores relevantes. En todos los casos, las rúbricas apuntan a evaluar si el estudiante logra discernir y manejar correctamente las variables que intervienen en una investigación, por lo que, pese a las diferencias de formulación, cumplen una función evaluativa esencialmente equivalente. Este mismo criterio de equivalencia semántica fue aplicado de manera sistemática en el análisis del resto de los criterios de cada una de las rúbricas consideradas para construir la tabla 4.3.C.

Como ya se mencionó las celdas de color naranjo que no contienen texto indican que el artículo específico en esa columna no incluye un criterio explícito para evaluarla un criterio de otra u otras rúbricas. Esta observación indica la necesidad de contar con marcos de referencia más integral si se desea completar genéricamente estas celdas. Un enfoque tal lo ofrecen los Conceptos de Evidencia (CoE) de Gott et al. (2020), que permiten identificar de manera sistemática un espectro más amplio de habilidades científicas, tanto procedimentales como epistémicas. Su incorporación permite no solo completar partes sin criterios evaluativos, sino también diseñar instrumentos transversales a todas las fases de la indagación o a diseños educativos particulares como App Checkers tomando una porción de los descriptores.

Tabla 4.3.C - Análisis comparativo entre las Rúbricas de los artículos de la Tabla 4.3.B junto a los conceptos de evidencia de Gott et al. (2024).

Tamir et al. (1982)	Knaggs y Schneider (2012), habilidades proceso	Arnold et al. (2014)	Ferrés et al. (2015)	Otero y Crujeiras-Pérez (2016)	Crujeiras-Pérez y Cambeiro (2017)	Crujeiras-Pérez y Cambeiro (2018)	COE Habilidades procedimentales
Formulación de problemas	Pregunta científica		Identificación de problemas investigables		Identificación del problema		- Analiza relaciones entre variables propuestas en una investigación (COE 1) - Selecciona el tipo de medición más adecuado según el fenómeno a estudiar (COE 1) - Realiza observaciones detalladas y pertinentes sobre objetos o eventos científicos (COE 2) - Aplica criterios predefinidos para guiar sus observaciones (uso de claves o categorías) (COE 2) - Registra observaciones mediante representaciones visuales adecuadas (como mapas o esquemas) (COE 2)
Formulación de hipótesis	Hipótesis		Formulación de hipótesis	Hipótesis	Formulación de hipótesis		- La hipótesis esta formulada en términos de variables dependientes e independientes
Identificación de la variable dependiente		Medición de la variable dependiente	Identificación de variables		Identificación de variables		- Clasifica correctamente el tipo de variable según su naturaleza (categórica, ordinal, continua, discreta) (COE 11) - Identifica la variable independiente (COE 11) - Identifica la variable dependiente (COE 11)
Identificación de la variable independiente		Variación de la variable independiente					
Comprensión del papel del control en el experimento	Procedimiento	Consideración de variables de confusión					- Identifica variables de control (confusión) en estudios de laboratorio (COE 12). - Identifica variables de control (confusión) en estudios de campo (COE 12). - Identifica variables de control en encuestas (COE 12). - Identifica el grupo de control en un experimento (COE 12).
Adecuación del experimento al problema o hipótesis			Planificación de investigación	Procedimiento	Propuesta de procedimiento	Preparación	- Ajusta la técnica de medición según la situación experimental para asegurar la calidad de los datos (COE 3). - Interpreta correctamente los datos obtenidos considerando que el instrumento mide una variable relacionada matemáticamente —y no necesariamente de forma directa— con la variable de interés. - Realiza comprobaciones cruzadas con otros instrumentos o métodos (COE 6). - Justifica el uso de grupos de control o condiciones equivalentes en distintos diseños experimentales (COE 12).

	Procedimiento	Tiempo y periodo de prueba			Selección del criterio de medida		- Diseña investigaciones considerando múltiples variables independientes o relaciones de correlación (COE 11). - Selecciona valores adecuados de la variable independiente (rango, intervalo y número de valores) para facilitar la detección de patrones (COE 13). - Ajusta el tamaño de muestra según criterios de representatividad y análisis previsto (COE 13). - Evalúa el diseño experimental identificando posibles fuentes de error y falta de control (COE 16). - Revisa el diseño considerando cómo cada dato contribuye a la validez general de la investigación (COE 16).
					Selección de materiales y equipamiento		- Realiza pruebas preliminares para ajustar escala, rango y equipos antes de la recolección definitiva de datos (COE 13). - Utiliza correctamente instrumentos que operan con diferentes tipos de relaciones (lineales o no) (COE 4). - Ajusta o verifica el instrumento antes de su uso mediante calibración (COE 5). - Evalúa los límites operativos del instrumento antes de su uso (COE 5). - Utiliza el instrumento según procedimientos técnicos correctos (COE 5). - Elige el instrumento adecuado en función de la precisión y exactitud requeridas (COE 7).
Diseño del control				Control de variables	Control de variables		- Aplica estrategias de control de variables para asegurar una prueba justa, según el tipo de investigación (laboratorio, campo o encuesta) (COE 12). - Establece una prueba justa al diseñar un experimento (COE 12). - Diseña Controles de variables en un experimento de laboratorio (COE 12). - Diseña condiciones de control en estudios de campo (COE 12). - Selecciona participantes bajo condiciones similares en encuestas (COE 12). - Identifica correctamente el grupo de control en un diseño experimental (COE 12).
Completitud del diseño experimental		Repetición			Repetitividad		- Repite mediciones y aplica procedimientos para minimizar errores durante la medición (COE 3). - Repite mediciones y calcula un promedio para aumentar la fiabilidad (COE 6). - Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato (COE 8).
Elaboración de gráficos	Datos y análisis						- Selecciona representaciones gráficas adecuadas al tipo de variable (COE 17). - Elabora representaciones gráficas multivariadas para mostrar relaciones complejas (COE 17).

Elaboración de tablas			Recogida y procesamiento de datos				- Usa tablas como herramienta de planificación experimental antes de la recolección de datos (COE 15). - Registra datos en tablas utilizando un formato convencional (columnas, encabezados, unidades, etc.) (COE 15). - Adapta el formato de una tabla a las necesidades específicas de cada investigación (COE 15). - Utiliza la tabla para evaluar la completitud y consistencia del diseño experimental (COE 15). - Organiza y presenta los datos en tablas según convenciones científicas estándar (COE 17).
Toma y registro de mediciones						Experimentación y toma de datos	- Reconoce y justifica la exclusión de valores atípicos basándose en fallos del procedimiento (COE 7). - Identifica y justifica la exclusión o consideración especial de datos anómalos (COE 8). - Aplica estrategias de muestreo para reducir el sesgo y garantizar representatividad (COE 8).
Interpretación de datos observados			Análisis de datos y obtención de conclusiones	Interpretación de resultados		Análisis y establecimiento de conclusiones	- Aplica medidas estadísticas básicas para describir la dispersión y tendencia central de los datos (COE 9). - Utiliza parámetros estadísticos avanzados para estimar incertidumbre y confiabilidad en las mediciones (COE 9). - Estima la incertidumbre de una medición considerando el instrumento y el procedimiento (COE 10). - Verifica si la medición corresponde efectivamente al dato que se desea conocer (COE 10). - Ajusta el diseño experimental para lograr un nivel de precisión adecuado que permita distinguir entre resultados distintos (COE 14). - Establece el grado de precisión necesario para detectar patrones o tendencias en los datos (COE 14). - Aplica técnicas estadísticas básicas para comparar conjuntos de datos (COE 18). - Utiliza herramientas estadísticas para ajustar relaciones funcionales entre variables (regresión) (COE 18). - Identifica patrones en representaciones gráficas (tablas, gráficos de líneas, dispersión, etc.) (COE 19). - Detecta y justifica la exclusión de datos anómalos en una interpretación (COE 19). - Ajusta modelos (como líneas de tendencia) a los datos experimentales (COE 19). - Evalúa la validez de las conclusiones a partir de múltiples experimentos (COE 20).
Formulación de conclusiones	Conclusiones			Conclusiones			
Explicación de hallazgos de la investigación							
Examen crítico de los resultados							
Comprensión e interpretación de datos presentados en un gráfico							

							- **Compara los resultados obtenidos con hallazgos previos o datos esperados para evaluar su consistencia interna y externa (Análisis de datos/ conclusiones) (COE 20).** - **Aplica estrategias de triangulación metodológica para aumentar la validez de una investigación. (Diseño experimental /Análisis de datos/ Conclusiones) (COE 20).** - **Evalúa el peso probatorio del conjunto de datos considerando su coherencia, replicabilidad y diversidad metodológica (triangulación) (Diseño experimental /Análisis de datos/ Conclusiones) (COE 20).**
Aplicación de conocimientos							- **Evalúa la credibilidad de una evidencia científica considerando el tipo de evidencia y el consenso disciplinar (Conclusiones / Metarreflexión) (COE 21).** - **Identifica posibles sesgos del experimentador al analizar o reportar evidencia (Conclusiones / Metarreflexión) (COE 21).** - **Discute la aceptabilidad de las consecuencias prácticas de una conclusión científica (Aplicación del conocimiento / Metarreflexión) (COE 21).** - **Reconoce el peso social o político que puede tener una evidencia al ser comunicada (Comunicación de resultados / Metarreflexión) (COE 21).**
			Metarreflexión				
						Comunicación de resultados	
Propuesta de ideas y formas de continuar la investigación							
						Trabajo Cooperativo	

La dualidad considerada existente en los COE se puede justificar con trabajos como el de Knaggs y Schneider (2012) que muestra un enfoque dual, evaluar el desempeño de los estudiantes en términos de sus acciones concretas en el diseño de experimentos, la manipulación de variables o la recolección de datos, como también considerar el nivel de comprensión conceptual (epistémica) que subyace a dichas acciones, como el uso de conocimientos previos, el establecimiento de relaciones causales o la integración de conceptos científicos al interpretar resultados.

Lo anterior implica que, si bien los COE son compresiones acerca de la indagación, es posible desde este enfoque obtener criterios de evaluación procedimentales que sean más "finos o nucleares" en cuanto a detallar el nivel de sofisticación de la indagación del alumnado.

El conjunto de CoE propuesto por Gott et al (2008, 2020) constituye una referencia más completa y transversal para analizar la indagación, es un corpus organizado de conocimientos procedimentales y epistemológicos que subyacen a cualquier proceso de generación de pruebas mediante indagación. Estos conceptos no se limitan a describir lo que se debe hacer, sino que explicitan el porqué, cuándo y cómo de cada decisión experimental. Por ejemplo, conceptos como validez, fiabilidad, relaciones entre variables, diseño de instrumentos, selección de escalas, control de errores o interpretación estadística constituyen un enfoque que permite interpretar y fundamentar las decisiones tomadas por el alumnado en contextos investigativos de diversos tipos.

Desde el análisis detallado del marco de los Conceptos de Evidencia (COE) se generaron criterios evaluativos análogos a los que proponen diversos autores (tabla 4.3.B) en sus rúbricas. Para ello, se revisaron uno por uno los 21 bloques temáticos del marco de Gott et al. (2020), analizando sus definiciones, características y acciones. Este procedimiento permitió identificar cuáles podían traducirse en desempeños procedimentales observables en tareas de indagación, o bien en criterios evaluativos comparables con los presentes en las rúbricas seleccionadas.

Como resultado, se derivaron criterios procedimentales (Ferrés, 2017), lo que posibilitó extender la tabla 4.3.C con una octava columna que los organiza, los cuales resultaron ser transversales a muchos de los criterios de las demás rúbricas. Aquellos criterios de rúbricas que no presentan correspondencia con los CoE, como "Propuesta de ideas y formas de continuar la investigación" y "Trabajo cooperativo", están vacíos por las siguientes razones: el primero se encuentra fuera del foco central de los CoE, ya que no forma parte del proceso de transformar datos en pruebas, que constituye su objetivo principal; mientras que el segundo se refiere a la

evaluación de habilidades sociales dentro de estrategias de enseñanza-aprendizaje cooperativas, no al desarrollo de competencias propias o "puras" del proceso de indagación científica.

4.3.2.3 Selección de criterios y determinación de niveles de desempeño

Se analizaron los 76 videos de la tabla 4.3.A para determinar qué criterios son acordes con las acciones que en los videos realizó el alumnado para evaluar la confiabilidad de sus Apps. Al mismo tiempo se fueron determinando niveles de desempeño para esos indicadores, que fueron refinados durante el proceso. En algunos casos estos eran dicotómicos, y solo interesaba si estaban o no presentes, en otros casos lo niveles de desempeño debían ser más de dos pues las acciones de los estudiantes tenían niveles diferenciados. Además, cuando existían, los niveles de desempeño propuestos de los criterios seleccionados fueron contrastados con los presentes en las rúbricas de los artículos de la tabla 4.3.C para establecerlos.

La tabla 4.3.D presenta la selección final de criterios, adaptados al contexto específico del proyecto App Checkers en su primera implementación. Su propósito es operacionalizar estos conceptos en indicadores evaluables que permitan analizar con mayor precisión el desempeño del alumnado en tareas de indagación científica escolar en el contexto del análisis de la confiabilidad de Apps.

La tabla está organizada en cuatro columnas. La primera columna, titulada "Criterio", identifica el indicador específico basado en los CoE, incluyendo su código correspondiente. La segunda columna, "App Checkers", contiene una reformulación contextualizada del criterio, ajustada al tipo de tareas desarrolladas por los estudiantes en los 76 videos analizados. La tercera columna, "Justificación", explica la pertinencia del criterio en función de las evidencias observadas en la muestra, destacando patrones, excepciones o implicancias metodológicas. Finalmente, la cuarta columna, "Niveles de desempeño", establece una gradación cualitativa que permite valorar el nivel alcanzado por cada estudiante o grupo en relación con el criterio específico, utilizando escalas ordinales que van desde el nivel más bajo de desempeño (0) hasta niveles avanzados (1, 2 o 3), dependiendo del caso, en general construidos usando como ejemplos las correspondencias con los criterios análogos de las rúbricas de la tabla 4.3.C (columnas 1 a 7).

Esta tabla permite describir como categorizar de forma sistemática las habilidades procedimentales del alumnado, integrando los marcos teóricos abordados con las observaciones empíricas del proyecto. Su construcción implicó un proceso de revisión teórica,

análisis iterativo de los videos del alumnado y ajuste de los descriptores según la diversidad de Apps observadas en el proyecto.

Tabla 4.3.D - Criterios procedimentales derivados de los COE aplicados al contexto del proyecto App Checkers, su formulación contextualizada, justificación y niveles de desempeño asociados.

Criterio	App Checkers	Justificación	Niveles de desempeño
Identifica la variable dependiente (COE 11)	Identifica la variable dependiente o magnitud de entrada de la App.	La mayoría de los videos de los estudiantes identifican correctamente las magnitudes de entrada de las Apps, pues la misma App actúa como andamiaje. Sólo en un caso un video no lo hace, el estudiante en cuestión sólo explica como esta funciona.	0. No identifica la variable dependiente o magnitud de entrada de la App. 1. Identifica correctamente la variable dependiente o magnitud de entrada de la App.
Identifica la variable independiente (COE 11)	Identifica la variable independiente o magnitud de entrada de la App.	La mayoría de los videos de los estudiantes identifican correctamente las magnitudes de entrada de las Apps, pues la misma App actúa como andamiaje. Sólo en un caso un video no lo hace, el estudiante en cuestión sólo explica como esta funciona.	0. No identifica la variable independiente o magnitud de salida de la App. 1. Identifica correctamente la variable independiente o magnitud de salida de la App.
Ajusta o verifica el instrumento antes de su uso mediante calibración (COE 5)	Calibra la App si está lo requiere.	Algunas Apps usadas por el alumnado necesitan calibración, en algunos casos está no fue realizada obteniendo resultados incorrectos.	0. La App no requiere calibración. 1. La App requiere calibración, pero está mal calibrada. 2. La App requiere calibración y la calibración está bien hecha.
Utiliza el instrumento según procedimientos técnicos correctos (COE 5)	Las lecturas realizadas con las Apps e instrumentos están bien realizadas.	En algunos casos las lecturas realizadas por el alumnado no has sido realizadas bien, por ello es necesario un criterio para determinar el desempeño de ellas.	0. Todas las lecturas están mal medidas. 1. Sólo una parte de las lecturas están bien realizadas. 2. Todas las medidas han sido realizadas correctamente.
Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato (COE 8).	Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato.	Se ha mantenido pues es lo estándar para asegurar la fiabilidad de un dato. Para los niveles de desempeño se han usado las rúbricas de Arnold et al. (2014) y Crujeiras-Pérez y Cambeiro (2017) para plantear niveles iniciales y luego iterar con los videos de los estudiantes.	0. No especifica un número de mediciones. 1. Realiza sólo una medición 2. Realiza de dos a cinco mediciones por dato. 3. Realiza más de 5 mediciones por dato.
Evalúa los límites operativos del instrumento antes de su uso (COE 5).	Evalúa los límites operativos de la App.	Se ha reescrito ya que en el contexto evaluar los limites operativos es durante todo el desarrollo del proyecto. Además, lo consideramos aplicable para todo tipo de App, los límites operativos no siempre implican una escala explícita como en instrumentos tradicionales, incluso en Apps de clasificación algunos estudiantes observan variaciones en la respuesta de la App en función de variables como la distancia o el entorno. Esto indica que, aunque la App entregue un resultado binario, su comportamiento puede depender de condiciones operativas. Por tanto, este criterio se ha usado si el alumnado investiga el comportamiento de la App frente a variaciones sistemáticas que permitan identificar un mínimo, máximo o sensibilidad de respuesta, rangos operativos y condiciones del entorno.	0. No hace referencia a ningún límite operativo del instrumento ni intenta caracterizarlo. 1. Menciona de forma general que la App funciona mejor o peor bajo ciertas condiciones, pero sin explorar sistemáticamente sus límites. 2. Investiga los límites de la App realizando mediciones en una condición. 3. Diseña mediciones sistemáticas en condiciones distintas: rangos de escala, condiciones operativas o de entorno.
Registra datos en tablas utilizando un formato convencional (COE 15) Organiza y presenta los datos en tablas según convenciones	Registra y organiza datos en tablas según convenciones científicas estándar.	En algunos videos los estudiantes resumen datos en tablas y los presentan de forma ordenada con la intención de mostrar variaciones o patrones.	0. No hay registro de datos en tablas. 1. Registra y organiza datos en tablas según convenciones científicas estándar.

científicas estándar (COE 17)			
Verifica si la medición corresponde efectivamente al dato que se desea conocer (COE 10) Compara los resultados obtenidos con hallazgos previos o datos esperados para evaluar su consistencia interna y externa (COE 20)	Compara los resultados obtenidos de un experimento con valores de referencia fiables y válidos.	En varios videos los estudiantes comparan sus resultados con valores de referencia. En algunos casos los valores de referencia usados pueden ser mejorados en validez y fiabilidad.	0. No hay valores de referencia el diseño. 1. Compara sus resultados con valores de referencia que son parcialmente fiables y válidos. 2. Compara sus resultados con valores de referencia que son fiables y válidos.
Compara los resultados obtenidos con hallazgos previos o datos esperados para evaluar su consistencia interna y externa (COE 20)	Compara los resultados obtenidos de una App medidora con los de un instrumento similar.	En algunos casos las Apps seleccionadas tienen como objetivo replicar instrumentos de medición concretos, como las Apps Sonómetros, Medidor de longitudes o Medidor de temperatura. En muchos videos los estudiantes usan como patrón el instrumento al que la App pretende emular para determinar la confiabilidad.	0. No compara sus resultados con un instrumento similar cuando debería hacerlo. 1. Compara sus resultados con un instrumento similar, pero no está calibrado o es parcialmente fiable y válido. 2. Compara sus resultados con un instrumento similar, y está calibrado, es fiable y válido.
Aplica estrategias de triangulación metodológica para aumentar la validez de una investigación. (COE 20). Evalúa el peso probatorio del conjunto de datos considerando su coherencia, replicabilidad y diversidad metodológica (triangulación) (COE 20)	Compara los resultados obtenidos de una App con una triangulación entre distintos celulares.	Consideramos la triangulación entre celulares se considera una variante de triangulación metodológica dentro de un mismo entorno experimental. Además, utilizar múltiples celulares aumenta la replicabilidad y coherencia metodológica del diseño, aunque el criterio está más enfocado en la evaluación de resultados que en el diseño experimental en sí. En los 76 videos los grupos que han realizado una triangulación entre distintos celulares fueron 6 grupos.	0. No hay una comparación de los resultados obtenidos de una App con una triangulación entre distintos celulares. 1. Compara los resultados obtenidos de una App con una triangulación entre distintos celulares.

Cabe mencionar que el criterio relacionado con el diseño de control de variables se ha dejado fuera de la tabla 4.3.D. Las rubricas de la tabla 4.3.C muestran que el control de variables es un aspecto importante de los procesos de indagación y por tanto tiende a ser un criterio de evaluación. Por ello se plantearon forma de generar niveles de desempeño para este criterio. Inicialmente se propuso un criterio dicotómico basado en la pregunta ¿Hay un diseño de control de variables? En una segunda oportunidad se buscó determinar desde los videos ¿Cuántas variables (si hay alguna) son controladas de forma explícita o implícita por el alumnado cuando hay un diseño de control? En ambos enfoques se constató que la App es un andamio para este criterio.

Un ejemplo es el de un grupo que trabajó con la App PlantNet, diseñada para identificar plantas mediante reconocimiento de imágenes. En su procedimiento, las estudiantes seleccionaban una planta, le tomaban una fotografía, la subían a la App y verificaban si la respuesta era coherente con la especie ingresada. Es decir, el alumnado siempre eligió “planta”, lo cual puede interpretarse como una forma implícita de control de una categoría de entrada.

Sin embargo, *esto se basa más en el flujo natural de la App que en un diseño experimental consciente*, es decir la App actúa como un andamio, lo que se consideró ensucia su intención, parece responder más al flujo natural de uso de la App que a un diseño de control deliberado.

A diferencia de tareas donde los estudiantes diseñan un experimento (con manipulación sistemática de variables y diseño de condiciones controladas), en App Checkers se verifica un instrumento que ya está dado: la App. En este sentido, el criterio derivado de COE 5 —"Evalúa los límites operativos de la App"— se consideró absorbe parte de las funciones evaluativas del control de variables, pero adaptadas al contexto específico de la demanda en App Checkers, donde no siempre se pueden aislar o controlar las variables de forma tradicional. Se puede entonces, argumentar que la categoría "control de variables" está reconceptualizada en términos operativos y contextuales, integrándose en el análisis de los límites del instrumento (la App). Esto permite una evaluación más ajustada a la tarea dada.

También sucede que en la tarea de App Checkers se parte de una lógica de diseño abierta, exploratoria y bastante diversa. A los estudiantes no se les solicitó seguir protocolos experimentales con control de condiciones, sino que interactuaran con Apps en diferentes contextos, esta diversidad generó dificultad metodológica en poder determinar variables de control. El principio de diseño seguido en la rúbrica es la discriminación empírica entre niveles. Durante la codificación iterativa no emergieron patrones claros o diferenciables en cuanto al control de variables por los motivos dados, se consideró por tanto sería metodológicamente incorrecto forzar su inclusión. Así, se buscó coherencia metodológica y la validez empírica del instrumento incluyendo criterios donde existía suficiente variabilidad en los desempeños observados y una discriminación e interpretación menos ambigua de las habilidades de indagación que se lograron derivar desde el discurso del alumnado cuando realizó la tarea de App Checkers.

4.3.2.4 Determinación de la frecuencia de desempeños observados

Desde la iteración de los criterios se determinó la frecuencia de cada desempeño en la tabla 4.3.E, donde se presenta la distribución de niveles de desempeño observados en los 76 videos analizados del proyecto App Checkers en esta primera implementación, según los diez criterios seleccionados (tabla 4.3.D). Las columnas indican el número de videos que se ubicaron en cada nivel de desempeño, del 0 al 3, de acuerdo con la escala específica establecida para cada indicador.

Tabla 4.3.E - Distribución de niveles de desempeño del alumnado en 76 videos del proyecto App Checkers, primera implementación, según diez criterios procedimentales.

Criterio	0	1	2	3
Identifica la variable dependiente o magnitud de entrada de la app.	1	75	-	-
Identifica la variable independiente o magnitud de entrada de la app.	1	75	-	-
Calibra la App si está lo requiere.	70	3	3	-
Las lecturas realizadas con las Apps e instrumentos están bien realizadas.	4	0	72	-
Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato.	8	3	32	33
Evalúa los límites operativos de la App.	37	26	6	7
Registra y organiza datos en tablas según convenciones científicas estándar.	71	4	-	-
Compara los resultados obtenidos de un experimento con valores de referencia fiables y válidos.	23	2	51	-
Compara los resultados obtenidos de una App medidora con los de un instrumento similar.	62	10	4	-
Compara los resultados obtenidos de una App con una triangulación entre distintos celulares.	70	6	-	-

Se determinó que la mayoría del alumnado logra identificar adecuadamente las variables independiente y dependiente de las Apps utilizadas. Esta tendencia se atribuye a que las propias Apps funcionan como elementos de andamiaje: sus nombres y funcionalidades explícitas orientan al estudiante sobre qué magnitud se mide y qué condiciones o acciones modifican su comportamiento. Por ejemplo, Apps como *Detector de metales*, *Medidor de longitud*, *Detector de edad* o *Detector de mentiras* anticipan desde su nombre especificaciones como qué tipo de fenómeno está siendo observado y qué entrada modifica el resultado. Esto reduce la ambigüedad conceptual en la etapa inicial de la indagación y del diseño experimental y permite al alumnado identificar con mayor facilidad las variables implicadas, quizá incluso sin una formación experimental avanzada. El único caso donde no se identifican variables corresponde a un estudiante que se remitió a describir la interfaz de usuario de la App y como se usa, lo que representa a un estudiante que no comprendió la demanda solicitada por el profesorado.

Se observó que muchas de las Apps seleccionadas por los estudiantes no requerían una calibración previa por parte del usuario, ya que es realizada por los desarrolladores de la App. Sin embargo, existen Apps, especialmente aquellas que utilizan tecnología de realidad aumentada (RA) para medir distancias, que sí requieren una calibración inicial, como las de la figura 4.3.A.

En cuanto al criterio “Las lecturas realizadas con las Apps e instrumentos están bien realizadas”. Cuando los estudiantes no realizan una lectura correcta en un laboratorio, la

experiencia indica que las causas más comunes son una comprensión inadecuada de la precisión de la escala del instrumento, errores al leer la escala o problemas de paralaje, especialmente si esta es analógica. Sin embargo, en la muestra de 76 videos no se identificaron Apps con escalas analógicas. Esto sugiere que, en este caso, el error humano relacionado con la lectura directa de valores no es una fuente relevante de error. Sólo se observaron mediciones incorrectas en situaciones que la App no fue calibrada y el estudiante no cuestionó la validez de la medición obtenida.

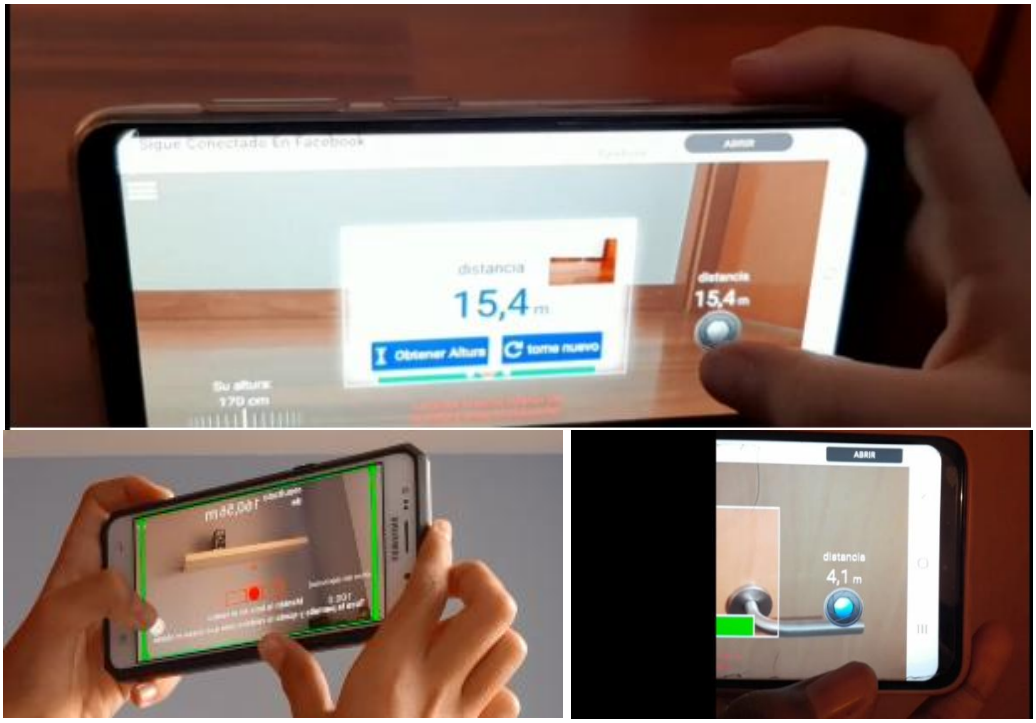

Figura 4.3.A - Apps que miden distancia y necesitan ser calibradas por el usuario antes de su uso.

En cuanto al criterio "Evalúa los límites operativos de la app", se observan diversos escenarios importantes, n=7 para el nivel de desempeño superior, n =6 para el nivel 2, n=26 para el nivel 1 y n=37 para el nivel 0. Hay casos en que los estudiantes intentan realizar mediciones, pero se enfrentan a que los valores están fuera del rango operativo de la App; otros que, de manera intencionada, exploran todo el rango posible —desde el mínimo hasta el máximo— con el objetivo de evaluar la confiabilidad del instrumento. Incluso en Apps sin escala explícita, como aquellas de tipo clasificatorio (detector de metales, detector de edad), algunos estudiantes intentan identificar empíricamente los límites operativos mediante pruebas sistemáticas. Estos casos revelan que la exploración de los límites de funcionamiento de la App constituye un descriptor clave para evaluar su confiabilidad en el contexto del proyecto.

Para el criterio "Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato", este indicador resulta clave al momento de evaluar la calidad de las conclusiones experimentales. La conclusión que el estudiante pueda establecer sobre la confiabilidad de una App será mejor en la medida en que realice múltiples lecturas, estas sean estables, presenten una baja dispersión (lo que se traduce en una desviación estándar baja) y la media de los valores obtenidos sea comparable con un valor de referencia fiable u otro método de validez. Estas condiciones permiten garantizar tanto la reproducibilidad como la veracidad de los resultados.

Del análisis de los 76 videos evaluados, se observa que 8 grupos no realizaron mediciones repetidas suficientes, ubicándose en el nivel más bajo de desempeño. Solo 3 grupos realizaron algunas mediciones, pero sin alcanzar una cantidad adecuada. En contraste, 32 grupos lograron una cantidad aceptable de mediciones repetidas para representar la población del dato, mientras que 33 grupos alcanzaron el nivel más alto de desempeño, realizando un número amplio y adecuado de mediciones repetidas.

Esto sugiere que, aunque algunos estudiantes lograron alcanzar un nivel avanzado, la mayoría aún no incorpora sistemáticamente estrategias de repetición y muestreo representativo. Esta situación plantea una oportunidad didáctica de reforzar la importancia de la repetición como estrategia para validar resultados y construir evidencia científica confiable.

Para el criterio "Registra y organiza datos en tablas según convenciones científicas estándar", se considera que el uso adecuado de tablas no solo es una herramienta de registro, sino también un medio para visualizar variaciones, identificar patrones y organizar la evidencia recogida durante la experimentación. Para presentar los datos existen diferentes herramientas (tablas, gráficos, esquemas, etc.). Cuando el número de mediciones es adecuado, las tablas se vuelven especialmente útiles para analizar tendencias, establecer comparaciones y facilitar la interpretación sistemática de los resultados.

Se pudo observar que, según los datos recolectados, 71 de los 76 videos no presentan ningún tipo de tabla, lo cual indica una ausencia casi generalizada de esta práctica en la muestra. Solo 4 videos muestran registros organizados en tablas que cumplen con criterios mínimos de formato científico, tales como encabezados, unidades o disposición coherente (figura 4.3.B).

Este resultado puede mostrar una debilidad transversal en las habilidades de representación y organización de datos por parte del alumnado. Pese a que en algunos casos se observa una preocupación por medir y repetir lecturas, la falta de sistematización mediante tablas puede limitar la posibilidad de transformar los datos para un análisis posterior.

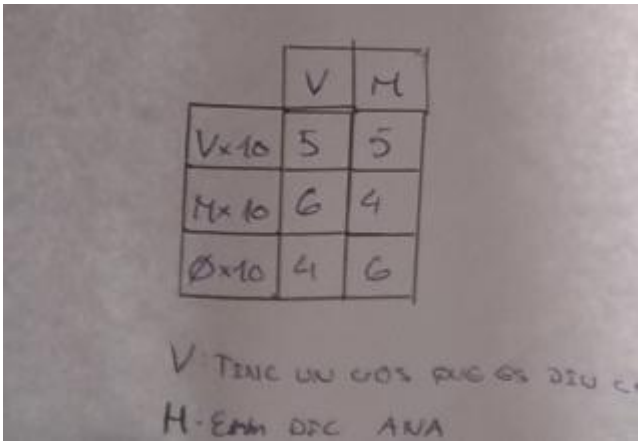

Figura 4.3.B - Grupo de estudiantes que presenta sus resultados en una tabla

Se identificaron tres estrategias recurrentes entre los estudiantes para evaluar la confiabilidad de las Apps analizadas, las cuales fueron operacionalizadas en tres criterios distintos: comparación con valores de referencia, comparación con un instrumento similar y triangulación entre distintos dispositivos. Estos criterios fueron marcados de color negro en la tabla 4.3.C pues son más transversales y asociarlos a una determinada fase de la indagación es más complejo.

El criterio: Compara los resultados obtenidos de un experimento con valores de referencia fiables y válidos, contempla situaciones en las que el alumnado utiliza referencias previas conocidas, teóricas o empíricas, para contrastar la información entregada por la App. En la tabla 4.3.A, 51 de los 76 videos (67%) muestran comparaciones con valores de referencia que son considerados fiables y válidos; 2 videos (3%) contienen referencias parcialmente válidas y fiables; y 23 videos (30%) no incluyen ningún valor de referencia en su diseño experimental.

Los videos categorizados como “Hay valores de referencia y son fiables y válidos” corresponden a diversos casos. Por ejemplo:

- En Apps clasificadoras binarias como detectoras de mentiras, los estudiantes formulan preguntas con respuestas conocidas; en Apps detectoras de metales, prueban con materiales conocidos; y en Apps que clasifican voces por género, se conocen de antemano las características esperadas.
- En Apps clasificadoras múltiples, los estudiantes utilizan elementos previamente etiquetados, como especies de plantas o partes del cuerpo (App que simulan rayos X), para validar la clasificación realizada por la App.

- En Apps que miden magnitudes, se observan ejemplos donde se utilizan tablas teóricas para verificar lecturas de intensidad lumínica (Figura 4.3.C) o diseños experimentales para detectar temperaturas en distintas condiciones térmicas. Esto último está en la figura 4.3.D donde los alumnos diseñan varios experimentos con agua caliente y fría, vasos de distintos colores, vasos transparentes y un radiador, para probar la fiabilidad en detectar el calor y concluir que la App solo cambia los colores.

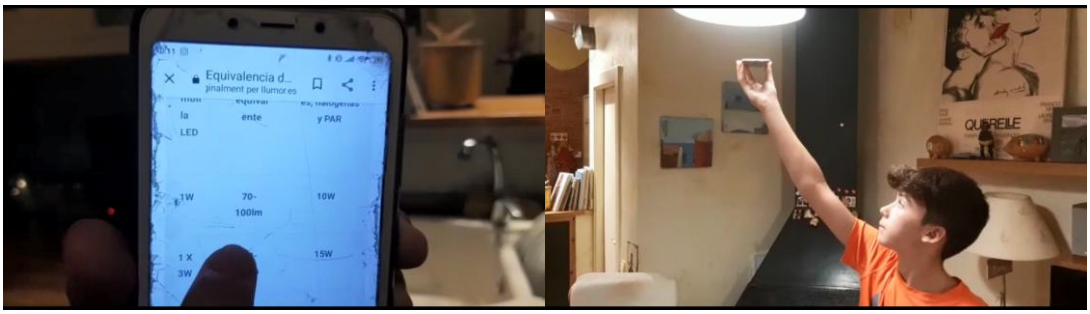

Figura 4.3.C - Alumnos usando una tabla de equivalencia watts versus lúmenes para determinar la confiabilidad de la App.

Figura 4.3.D - Estudiantes diseñan un experimento con vasos de distintos colores y temperaturas para evaluar si la App responde al calor real o solo altera los colores.

- En Apps que resuelven operaciones matemáticas, los estudiantes conocen la solución con anterioridad, figura 4.3.E(a), mientras que en Apps lúdicas como las que calculan compatibilidad amorosa, se usan combinaciones de nombres para analizar la coherencia del resultado, figura 4.3.E(b).

Los casos con referencias parcialmente fiables corresponden a situaciones como:

- Verificar el estado del clima observando si está nublado para contrastarlo con lo que indica la App.
- Usar una cámara fotográfica para probar una App detectora de metales, sin tener certeza total de la composición interna de la cámara, señalando que puede que la App no

funcione porque es plástica y “algo marca”, sin considerar los componentes internos del cuerpo de referencia.

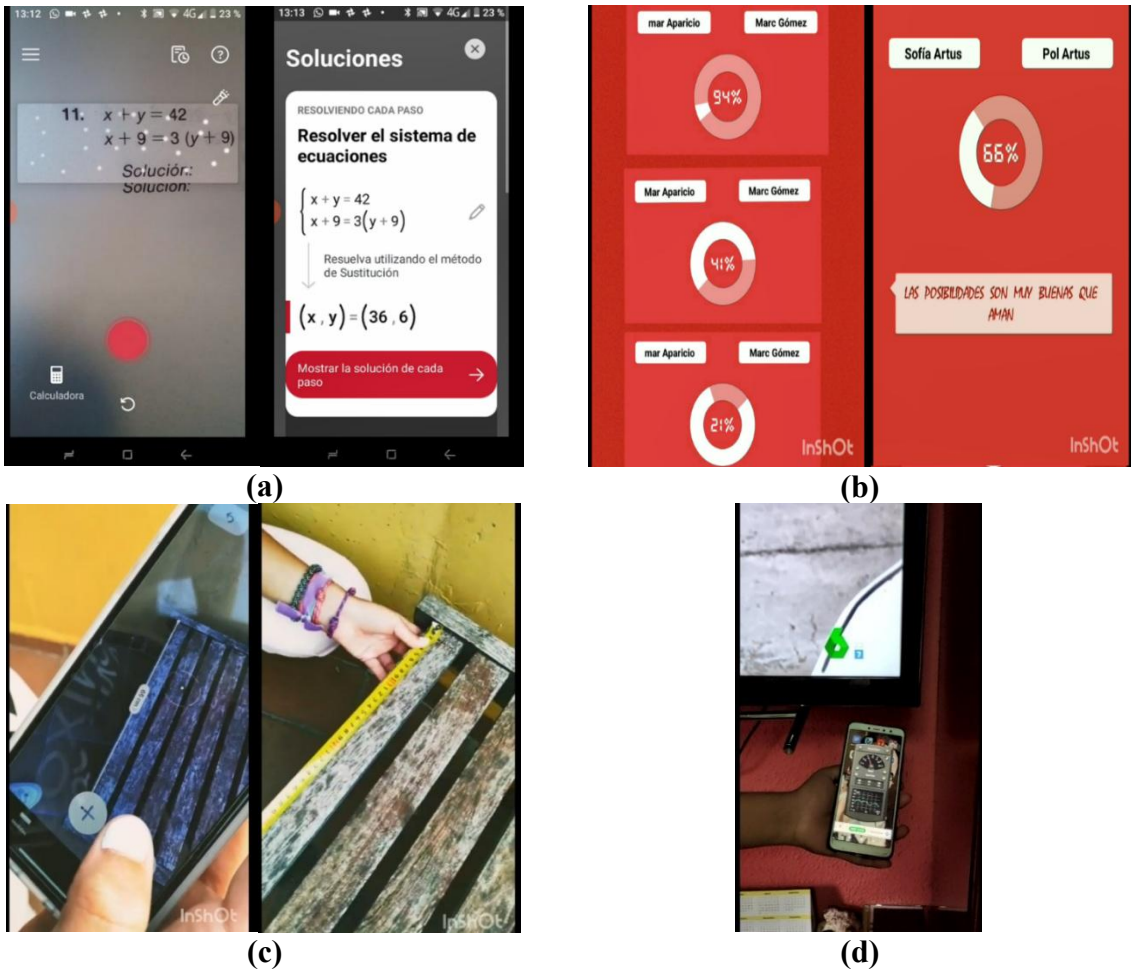

Figura 4.3.E - Alumnado evaluando la confiabilidad de una App que: (a) resuelve problemas matemáticos, (b) calcula el “porcentaje de amor” entre dos personas, (c) mide la distancia, comparándola con una regla de medir y (d) mide el volumen y frecuencias del sonido.

Por otro lado, los videos en los que no hay valores de referencia incluyen comparaciones con instrumentos o fenómenos, como reglas de medir o termómetros no calibrados, y pruebas entre celulares con la misma App.

El criterio: Compara los resultados obtenidos de una App con una triangulación entre distintos celulares, fue observado en solo 6 de los 76 videos (8%). En estos casos, los estudiantes instalaron la App en distintos celulares y verificaron la consistencia de las lecturas obtenidas. Esta estrategia refuerza la confiabilidad de los datos al probar la independencia del resultado respecto del dispositivo. Sin embargo, los 70 videos restantes (92%) no incorporaron

este tipo de diseño, lo que indica que la triangulación metodológica es aún una estrategia poco frecuente entre los estudiantes.

El criterio: Compara los resultados obtenidos de una App medidora con los de un instrumento similar, evalúa el uso de un instrumento físico, ya sea profesional o improvisado, que funcione como patrón para validar la App. Según los datos, solo 4 videos (5%) utilizaron un instrumento apropiado y calibrado; 10 (13%) usaron un patrón parcialmente calibrado, que permitió correlaciones razonables; y 62 videos (82%) no realizaron ninguna comparación con instrumentos similares.

En los niveles más altos de desempeño, se encuentran:

- Comparaciones entre Apps de medición de distancias por realidad aumentada y reglas físicas, figura 4.3.E(c).
- En Apps que miden sonido, se utilizaron televisores con control de volumen o la propia voz en distintos niveles, figura 4.3.E(d).
- En Apps que miden intensidad de luz o temperatura, se emplearon fuentes conocidas, como hornos o refrigeradores con medidores integrados (Figura 4.3.F).

En contraste, la mayoría de los estudiantes no recurrió a ningún instrumento patrón, lo que refleja una omisión metodológica importante en el proceso de verificación empírica.

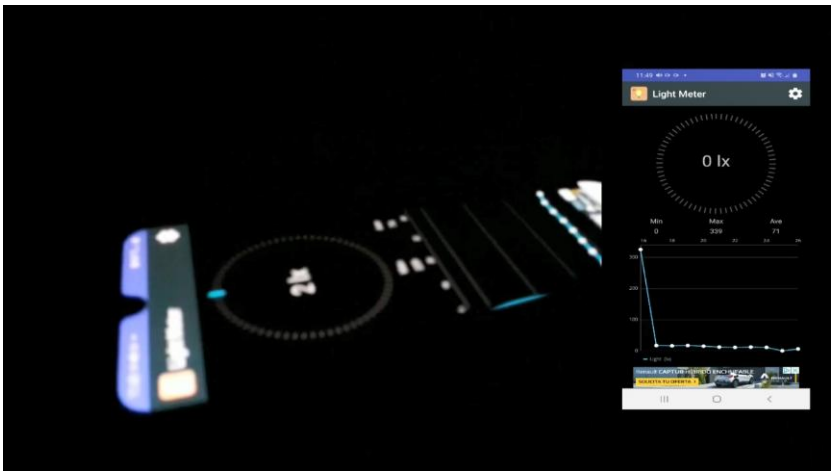

Figura 4.3.F - Alumnos comparando y correlacionando Apps medidoras de intensidad de luz en más de 2 puntos.

En conclusión, los videos muestran que, si bien más de la mitad del alumnado realiza comparaciones con valores de referencia fiables, aún es bajo el uso de instrumentos patrón y muy escasa la triangulación entre dispositivos. Este patrón sugiere una tendencia hacia evaluaciones basadas en el conocimiento previo o la intuición, pero con escaso fortalecimiento

del diseño experimental mediante estrategias más rigurosas. Reforzar estas dimensiones en el aula permitiría mejorar la calidad de los juicios emitidos sobre problemas relacionados con confiabilidad (validez y fiabilidad).

Además de los criterios analizados en las secciones anteriores, se identificaron algunos que, si bien podrían parecer relevantes para evaluar el desempeño en tareas de indagación, no fueron incluidos en la rúbrica final por razones metodológicas. Específicamente, se trata de criterios cuya presencia fue uniforme en toda la muestra —es decir, todos los estudiantes los cumplieron o ninguno lo hizo—, lo que impide establecer gradualidad en los desempeños. En una rúbrica, lo relevante no es únicamente constatar la presencia de una acción, sino que esa acción permita discriminar niveles de logro entre los participantes.

Un caso ilustrativo es la formulación de hipótesis. En los 76 videos analizados no todos los estudiantes formularon explícitamente una hipótesis. Sin embargo, incluso en los casos donde esta no aparece de forma explícita, se infiere una postura implícita frente a la pregunta de investigación. Por ejemplo, frases como "vamos a ver si esta App funciona" o "queremos saber si esta App es verdadera o falsa" suponen una hipótesis implícita del tipo "la App es confiable" o "la App no lo es". En este sentido, la formulación de hipótesis no varió de manera suficiente ni significativa como para establecer niveles diferenciados de desempeño evaluables.

Otro conjunto de criterios se refiere a operaciones estadísticas básicas y su uso en la validación experimental. El criterio "Calcula el promedio de un conjunto de datos para aumentar la fiabilidad", relacionado con los CoE 3 y 6, fue inicialmente considerado como indicador relevante. No obstante, los datos muestran que ningún grupo de estudiantes realizó este cálculo, lo que impide establecer niveles intermedios o avanzados: los 76 grupos obtuvieron nivel 0 en este aspecto. Por este motivo, y considerando que el promedio es una estrategia estándar para mejorar la precisión de una medida, el criterio fue tratado como dicotómico y posteriormente excluido de la rúbrica final, al no aportar información discriminativa en este contexto.

La misma situación se presenta para el criterio "Calculan el error típico, cuando realizan varias lecturas" (asociado a los COE 9 y 10). Al igual que en el caso anterior, ninguno de los 76 grupos de estudiantes utilizó esta medida estadística, lo que sugiere una brecha significativa en el uso de herramientas básicas para estimar la dispersión de los datos. El nivel de enseñanza secundaria en el que se ubican estos estudiantes podría explicar esta ausencia, ya que conceptos como "error típico" o "desviación estándar" no siempre forman parte del currículo práctico de laboratorio a esta edad.

Finalmente, se consideraron dos criterios estrechamente vinculados entre sí: “Comparan el error típico de las lecturas obtenidas con la App contra el error típico o precisión del instrumento patrón”, repetido con una redacción equivalente para distinguir su presencia en distintos contextos instrumentales. Sin embargo, en ambos casos los resultados fueron idénticos: 76 de 76 grupos no realizaron este tipo de comparación, lo que refuerza la conclusión anterior sobre la falta de uso de herramientas estadísticas más avanzadas.

En suma, si bien estos criterios reflejan buenas prácticas de investigación científica, en este estudio específico no contribuyeron a establecer diferencias entre desempeños estudiantiles observables. Su exclusión de la rúbrica no responde a una falta de relevancia conceptual, sino a la ausencia de variabilidad en los datos, lo que imposibilita su uso evaluativo en un instrumento orientado a discriminar niveles de desempeño.

4.3.2.5 Determinación de patrones y de niveles de desempeño en indagación

Con el objetivo de identificar patrones recurrentes en el trabajo del alumnado durante la implementación 1 del proyecto App Checkers, se procedió a derivar un esquema de trabajo a partir del análisis sistemático de los criterios evaluativos previamente definidos, este se muestra en la figura 4.3.G. Este esquema no pretende representar un modelo prescriptivo, sino más bien una síntesis empírica del comportamiento observado, destacando las rutas más comunes seguidas por los estudiantes para llegar (o no llegar) a la formulación de patrones o relaciones a partir de la evidencia recogida. En particular, se ha enfatizado el rol de ciertos criterios clave como la identificación de variables, la existencia de un diseño válido, la calibración del instrumento, la repetición de mediciones, la consideración de límites operativos del instrumento y el uso de estrategias de triangulación, como elementos que configuran la calidad del proceso de indagación y de sus conclusiones.

Cabe señalar que no todos los criterios presentes en la Tabla 4.3.E han sido incorporados en el esquema. En primer lugar, el criterio "Las lecturas realizadas con las Apps e instrumentos están bien realizadas" se excluyó del esquema ya que, aunque importante, su frecuencia de aparición (solo 4 casos) depende previamente de la calibración. Sólo los grupos que no calibraron la App tienen mediciones incorrectas, pero debido a la calibración, no a paralaje, mala lectura desde la escala u otro. En segundo lugar, los criterios "Compara los resultados obtenidos de un experimento con valores de referencia fiables y válidos" y "Compara los resultados obtenidos de una App medidora con los de un instrumento similar" fueron considerados dentro del bloque genérico de *Triangulación*, que agrupa distintas formas de

validación cruzada de resultados, sin detallar explícitamente cada una de ellas en el esquema para mantener claridad visual. Por último, "Compara los resultados obtenidos de una App con una triangulación entre distintos celulares" fue igualmente absorbido dentro de la categoría de triangulación, siendo este uno de los tres tipos de diseño considerados en el flujo lógico.

El uso de los nuevos nombres de los criterios en el esquema, de criterios como "Identifica la variable dependiente o magnitud de entrada de la App", "Evalúa los límites operativos de la App", “Registra y organiza datos en tablas según convenciones científicas estándar” o "Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato", responde a la necesidad de vincular directamente cada indicador con las acciones concretas observables, mejorando así la interpretabilidad y claridad del esquema resultante.

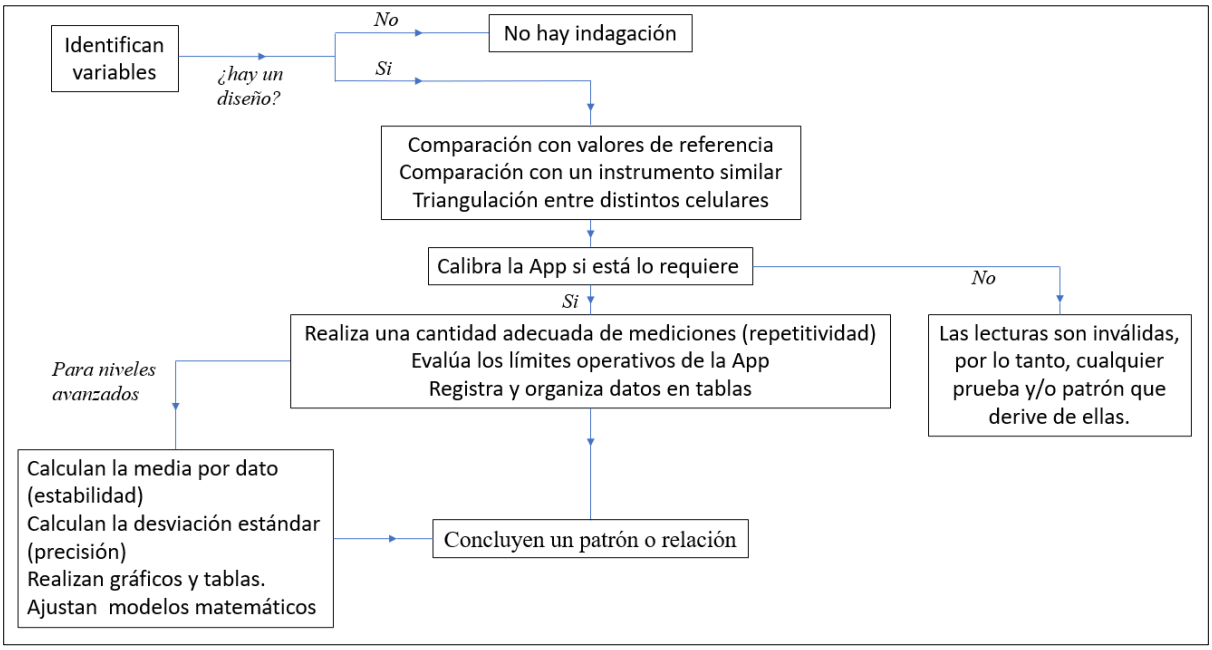

Figura 4.3.G - Patrón de indagación seguido por el alumnado para evaluar la confiabilidad de Apps.

El esquema de la figura 4.3.G representa de forma estructurada el patrón de trabajo seguido por el alumnado en el proyecto App Checkers, y tiene varias fortalezas didácticas y analíticas:

- Hay una lógica en el “flujo” de decisiones. El esquema inicia con la identificación de variables, lo que se ajusta a los fundamentos del diseño experimental según los CoE y a partir de ahí, se distingue entre quienes diseñan una investigación y quienes no lo hacen, permitiendo clasificar el trabajo como indagación o no indagación desde el inicio.

- Hay una integración coherente de tres estrategias de confiabilidad experimental. Se incorporan explícitamente las tres formas más frecuentes de validación detectadas en los videos: comparación con valores de referencia, uso de un instrumento patrón y triangulación con distintos celulares. Esta integración evidencia un vínculo directo con el CoE 20 sobre la evaluación del peso de una prueba.
- Si existe, se destaca la importancia metodológica de la calibración. La calibración aparece como un punto crítico que condiciona la validez de todas las mediciones posteriores. Esta ubicación jerárquica refuerza el papel del CoE 5 en la fiabilidad experimental.
- Se visibilizan prácticas avanzadas de análisis. El bloque inferior izquierdo del esquema destaca acciones avanzadas como el cálculo de media, desviación estándar y la elaboración de modelos. Esto permite identificar con precisión a los estudiantes que alcanzan niveles superiores en la evaluación de confiabilidad, en línea con los CoE 9, 14, 17 y 18 y propone un camino a considerar en posibles futuras implementaciones.
- Hay un énfasis en las consecuencias de una medición inválida: Se subraya que las lecturas mal realizadas —por falta de calibración o diseño defectuoso— invalidan cualquier conclusión posterior. Esto es importante para que el alumnado comprenda que no toda experimentación genera automáticamente conocimiento confiable.

En conjunto, este esquema permite comprender los procesos y decisiones que el alumnado toma durante la indagación, evidenciando tanto aciertos como omisiones clave en la construcción de evidencia científica escolar que pueden ayudar tanto del lado del profesorado para usar en el diseño de actividades de indagación, como para la evaluación de estas.

Por otro lado, con el objetivo de clasificar los niveles de desempeño competenciales del alumnado se optó por una estrategia metodológica basada en el análisis sistemático de combinaciones de desempeños en criterios procedimentales observados en los 76 videos evaluados. Esta aproximación permitió agrupar dichas combinaciones en niveles progresivos de competencia, dando origen a una rúbrica fundamentada en patrones empíricos consistentes y las decisiones metodológicas ya expuestas.

Se definieron usaron los cuatro criterios con más variabilidad (tabla 4.3.E): calibración del instrumento (criterio A), evaluación de los límites operativos de la App (criterio B), repetición de mediciones (criterio C) y la generación de tablas (criterio D). Cada uno de estos criterios fue codificado en niveles de desempeño, como se muestra en la tabla 4.3.F.

Tabla 4.3.F - Criterios usados para determinar niveles de desempeño competenciales en indagación.

Criterio	Niveles de desempeño
A. Calibra la App si está lo requiere.	A0. La App no requiere calibración. A1. La App requiere calibración, pero está mal calibrada. A2. La App requiere calibración y la calibración está bien hecha.
B. Evalúa los límites operativos de la App.	B0. No hace referencia a ningún límite operativo del instrumento ni intenta caracterizarlo. B1. Menciona de forma general que la App funciona mejor o peor bajo ciertas condiciones, pero sin explorar sistemáticamente sus límites. B2. Investiga los límites de la App realizando mediciones en una condición. B3. Diseña mediciones sistemáticas en condiciones distintas: rangos de escala, condiciones operativas o de entorno.
C. Realiza una cantidad adecuada de mediciones repetidas para representar la población del dato.	C0. No especifica un número de mediciones. C1. Realiza sólo una medición C2. Realiza de dos a cinco mediciones por dato. C3. Realiza más de 5 mediciones por dato.
D. Registra y organiza datos en tablas según convenciones científicas estándar.	D1. No hay registro de datos en tablas. D2. Registra y organiza datos en tablas según convenciones científicas estándar.

Una vez codificados los desempeños individuales, se construyeron las combinaciones posibles de los niveles A, B, C y D para cada video, esto se encuentra en la tabla 4.3.G. Estas combinaciones reflejan trayectorias metodológicas concretas seguidas por los estudiantes durante la resolución del problema planteado.

Tabla 4.3.G - Combinaciones por video según criterios A, B, C y D.

Código video	Calibrar App (A)	Límites App (B)	Repetitividad (C)	Tablas (D)	Combinaciones
2C2PlantNet	A0	B0	C2	D0	A0B0C2D0
3C2Sonometro	A0	B2	C1	D0	A0B2C1D0
4C2Clima	A0	B1	C2	D0	A0B1C2D0
5C2GhostDetector	A0	B1	C3	D0	A0B1C3D0
6C2GoldMetalDetector	A0	B0	C2	D0	A0B0C2D0
7C2LieDetector	A0	B0	C2	D0	A0B0C2D0
8C2Decibeles	A0	B1	C2	D0	A0B1C2D0
9C2PhysicsTools	A0	B1	C2	D0	A0B1C2D0
10C2LightMeter	A0	B2	C3	D0	A0B2C3D0
11C2GhostDetector	A0	B0	C2	D0	A0B0C2D0
12C2MedidorDistancia	A2	B1	C2	D0	A2B1C2D0
13C2SoundMeter	A0	B3	C3	D0	A0B3C3D0
14C2SoundMeter	A0	B3	C3	D0	A0B3C3D0
15C2MedidorDistancia	A1	B1	C1	D0	A1B1C1D0

Posteriormente, se procedió a agrupar dichas combinaciones según patrones comunes, generando clústeres que reflejan distintos niveles de desempeño competencial procedimental en indagación científica escolar. Esta clasificación se presenta en la tabla 4.3.H.

Tabla 4.3.H - Niveles de desempeño competenciales en indagación para la implementación del proyecto App Checkers.

Nivel	Combinaciones	Competencia	Nivel de desempeño
0	A0B0C0D0 A1B0C0D0	**I0** Sin indagación	No hay diseño experimental. No se realiza calibración adecuada, no se repiten mediciones y no se consideran los límites operativos.
1	A0B1C1D0 A1B1C1D0	**I1** Indagación deficiente	Se realizan una sola medición y la exploración de límites es superficial o la calibración es incorrecta.
2	A0B0C2D0 A0B1C2D0 A0B2C1D0 A0B2C2D0 A2B1C2D0	**I2** Indagación simple	Hay un diseño experimental, con calibración correcta si es necesaria, mediciones a lo más cinco o límites explorados parcialmente, pero no se integran todos los elementos procedimentales esperados.
3	A0B2C3D0 A0B1C3D0 A0B0C3D1 A0B0C3D0	**I3** Indagación intermedia	Las mediciones se repiten adecuadamente, con calibración correcta si es necesaria, se exploran límites de escala o condiciones operativas, pero aún esta exploración deja lugar a mejoras.
4	A0B3C3D0 A0B2C3D0 A0B3C3D1	**I4** Indagación compleja	Diseño experimental con calibración correcta si es necesaria, mediciones sistemáticas y en alto número, análisis avanzado de los límites de la App.

Con el fin de garantizar la fiabilidad del proceso de codificación y asegurar la validez de los niveles definidos en la tabla 4.3.H, se incorporó una fase de codificación independiente realizada por el director de tesis. Este codificó de forma autónoma el 30% de los datos, utilizando la rúbrica previamente construida, lo que permitió evaluar la consistencia en la aplicación de los criterios y detectar posibles discrepancias en la categorización. Posteriormente, se compararon los resultados de ambas codificaciones, llevando a cabo un proceso de discusión y consenso sobre los casos divergentes. Esta etapa no solo fortaleció la consistencia interevaluador, sino que también permitió afinar la rúbrica inicial, ajustando sus dimensiones y descriptores a partir de la concordancia alcanzada. Como resultado de este procedimiento, se validaron las combinaciones presentadas en la tabla 4.3.J y se establecieron los cinco niveles de desempeño competencial en indagación.

La tabla 4.3.H muestra las combinaciones identificadas y la competencia a la que fueron asignadas. Los cinco niveles de competencia definidos —desde I0 (Sin indagación) hasta I4 (Indagación compleja)— reflejan una progresión desde desempeños con baja articulación

metodológica hasta diseños experimentales más rigurosos, con análisis sistemático de variables y condiciones de medición.

Si bien en la columna dos de la tabla 4.3.H aparece el criterio D incluido en las combinaciones, este se terminó por excluir. Se consideró que su poder para discriminar era limitado en el contexto de este estudio. La decisión se fundamenta en que, como se observa en los datos, la mayoría de las combinaciones relevantes presentan el mismo nivel (D0), indicando ausencia de tablas, lo que impide que dicho criterio contribuya de manera efectiva a diferenciar niveles de desempeño. Su inclusión habría introducido una distorsión en la progresión lógica de los niveles competenciales, al generar subdivisiones basadas en un criterio que no guarda correspondencia directa con la calidad de la indagación del alumnado. Además, desde una perspectiva competencial, el uso de tablas, si bien deseable, podemos argumentar que constituye una habilidad que permite organizar y descubrir patrones en caso más complejos, más que contribuir transversalmente al procedimiento de indagación demandado al estudiante para determinar la confiablidad de una App. Por estas razones, se trabajó con los criterios A, B y C —calibración, evaluación de límites operativos y repetitividad—, para desarrollar los niveles de desempeño de las columnas tres y cuatro de la tabla 4.3.H, los cuales sí permiten observar diferencias más sustantivas en las habilidades procedimentales del alumnado.

4.3.3 Construcción de categorías para el análisis del desempeño en argumentación

Dado el marco teórico del apartado 2.3, se determinó que la mayoría de los autores que investigan el desempeño en argumentación científica escolar utilizan el modelo de Toulmin como herramienta principal de análisis. Este predominio se debe a que dicho modelo permite descomponer el producto argumentativo del alumnado en componentes bien definidas (Jiménez-Aleixandre, 2010): conclusión, pruebas, justificación, conocimiento básico, calificadores y refutación, facilitando así una evaluación estructurada y coherente del nivel de desempeño argumentativo.

Además, existen numerosas rúbricas públicas, alguna de ellas revisadas en la sección 2.3.2, que emplean explícitamente el modelo de Toulmin como base para la evaluación en argumentación de productos finales del alumnado. Estas rúbricas no solo han sido validadas empíricamente, sino que también se articulan con propuestas de progresión de aprendizaje,

como las desarrolladas por Osborne et al. (2016) y Berland y McNeill (2010), lo que refuerza su pertinencia didáctica y evaluativa.

A diferencia de otros marcos teóricos en argumentación, el modelo de Toulmin ha demostrado una mayor aplicabilidad en contextos educativos que priorizan el análisis de productos argumentativos finales. Esta característica resulta especialmente útil un estudio como el presente, donde el objetivo es caracterizar el nivel de desempeño en argumentación a partir de producciones del alumnado en formato de videos, lo cual justifica su uso.

El modelo de Toulmin, además, se alinea de forma natural con la distinción jerárquica entre describir, explicar, justificar y argumentar, propuesta por la línea de investigación "Hablar y escribir ciencias" (Márquez, 2005). En este sentido, argumentar se considera el nivel más alto del razonamiento escolar, ya que implica no solo establecer una conclusión a partir de pruebas, sino también conectar dicha conclusión con principios justificativos, contextualizarla en marcos teóricos relevantes y anticipar o responder a posibles objeciones.

Se decidió adoptar el segundo tipo de rúbricas de argumentación (tabla 2.3.B), aquellas basadas en la agregación progresiva de componentes del modelo de Toulmin. Esta decisión responde tanto a razones metodológicas como a os objetivos del estudio 1.

Primero, este tipo de rúbricas permite representar de forma más clara y funcional el desarrollo progresivo de la competencia argumentativa, al asumir que la calidad del argumento mejora en la medida en que se incorporan más componentes relevantes. A diferencia de las rúbricas analíticas que evalúan cada componente por separado, el enfoque por agregación reconoce que un argumento no se construye como una suma aislada de partes, sino como una estructura integrada cuyo valor depende de la presencia, articulación y función conjunta de esas partes.

Segundo, la naturaleza del proyecto App Checkers en la implementación 1, exige un tipo de rúbrica que privilegie trayectorias globales de desempeño. En el caso anterior, al analizar habilidades de indagación, se diseñó una rúbrica específica basada en criterios más finos, que requerían evaluar acciones metodológicas concretas observadas en los videos, como por ejemplo, la calibración de una App. Esa complejidad exigía una rúbrica más técnica y detallada, pues las habilidades a evaluar no estaban presentes de forma explícita en las rúbricas estándar disponibles, como el NPTAI, sino contenidas en categorías generales. En cambio, en argumentación, la situación es diferente: existen múltiples investigaciones previas, marcos teóricos bien establecidos y propuestas de rúbricas que ya han categorizado niveles de

desempeño en función del número y tipo de componentes argumentativos utilizados por el alumnado. Por lo tanto, no resulta necesario construir desde cero una rúbrica basada en acciones nucleares, sino que se puede adoptar un enfoque validado y coherente con la literatura revisada en el marco teórico.

Además, las rúbricas por agregación de componentes permiten establecer niveles competenciales de forma más sencilla y transparente para su posterior codificación y análisis. Al basarse en el número y tipo de componentes presentes en el argumento (por ejemplo, evidencia, justificación, refutación), estas rúbricas facilitan el trabajo de codificadores, reducen la ambigüedad en la interpretación y hacen posible realizar análisis comparativos más eficientes.

Finalmente, este enfoque puede resultar más didácticamente accesible para el profesorado que busque usar la rúbrica como herramienta formativa o sumativa. Una rúbrica en progresión coincide con los enfoques de enseñanza de la literatura, permitiendo a los profesores acompañar el desarrollo del alumnado en argumentación mediante retroalimentaciones específicas centradas en qué componente agregar o fortalecer. De este modo, la rúbrica no solo cumple puede cumplir una función de evaluar, sino también puede orientar el aprendizaje.

Con el objetivo de construir una rúbrica para evaluar el nivel de desempeño en argumentación del alumnado en el proyecto App Checkers, se diseñó un proceso metodológico compuesto por una serie de etapas sistemáticas. A continuación, se presenta la secuencia de pasos desarrollados, que permitieron definir tanto los criterios de evaluación como los niveles de desempeño argumentativo observados:

1. **Adopción del modelo de Toulmin como base conceptual y selección del tipo de rúbrica**. Esto lo consideramos como un paso previo que fue justificado al inicio de este apartado de metodología. Lo mencionamos porque consideramos que se debe referir al modelo usado.
2. **Definición de las componentes del modelo de Toulmin**: A partir del modelo de Toulmin y su adaptación al contexto del proyecto, se definirá como se trabajarán las seis componentes: Prueba, Conclusión, Justificación, Conocimiento básico, Calificadores y Refutación.
3. **Revisión de los videos del alumnado e identificación de componentes.** Se examinaron los 76 videos generados por los estudiantes, junto con sus transcripciones.

Se identificaron las componentes argumentativas presentes en cada producto del alumnado y luego se representaron mediante diagramas de caja y flecha.

4. **Clasificación argumentativa mediante agregación de componentes.** Con el objetivo de analizar la relación entre el desempeño argumentativo y el desempeño en indagación científica, se optó por una metodología basada en la agregación de componentes del modelo de Toulmin. Esta decisión se justifica por el tipo de análisis que se desea realizar: observar cómo varía la argumentación como producto en función del nivel de indagación alcanzado, manteniendo las componentes como una variable dependiente y progresiva. A partir de las componentes identificadas en paso anterior, se establecerá un procedimiento de clasificación basado en la presencia acumulativa de las componentes del modelo de Toulmin.
5. **Codificación de los niveles de desempeño en indagación por parte de la dirección de tesis**. Al igual que la rúbrica de indagación, para garantizar la fiabilidad intercodificadores del proceso de análisis, la dirección de tesis codificó de manera independiente el 30% de los datos. Esto permitió evaluar la consistencia en la aplicación de la rúbrica y detectar posibles discrepancias en la categorización.
6. **Comparación y consenso en la categorización**. Los resultados obtenidos en la fase de codificación independiente fueron comparados, discutiendo las diferencias hasta alcanzar un consenso en la categorización. Este proceso permitió afinar la rúbrica y asegurar su validez, ajustando las dimensiones y criterios de evaluación en función de la concordancia interevaluador.

4.3.3.1 Definición de las componentes del modelo de Toulmin

En base al marco teórico, se tomaron varias decisiones y consideraciones al momento de definir las componentes.

Como traducción de “claim” se usará desde ahora “aserción” en lugar de conclusión. Conclusión en castellano sugiere algo más elaborado, derivado, o que requiere un razonamiento previo completo, como sucede en un ensayo o una investigación formal. La palabra "conclusión" implica muchas veces el cierre de un razonamiento. Sin embargo, el “claim” en este estudio 1 es considerado como la afirmación que el alumnado está defendiendo o proponiendo, sin que necesariamente este correcta.

La componente se denominará pruebas en lugar de evidencia, con el objetivo de evitar confusiones derivadas de la traducción directa de anglicismo evidence. Tal como advierte

Jiménez-Aleixandre (2010), en castellano el término “evidencia” tiende a asociarse con aquello que es evidente o incuestionable, lo cual no se ajusta con precisión al uso técnico del modelo de Toulmin. Por ello, se opta por el término pruebas, que refleja de manera más clara y rigurosa su función dentro del modelo de argumentación.

No hablaremos de refutaciones porque en los videos no hay debates, aquí se trata de restricciones de la App que presenta el alumnado en sus videos para el profesor. Como explica Jiménez-Aleixandre (2010, p.78) “[…] Toulmin llama condiciones de refutación es el reconocimiento de las restricciones o excepciones que se aplican a la conclusión, circunstancias en que la conclusión no sería válida […]”, en contraste con la situación en que dos posiciones contrarias se enfrentan, es decir las críticas a las pruebas del contrario.

Una limitación metodológica al usar el modelo de Toulmin es que se han reportado problemas respecto al "tamaño de grano" de la información que se busca y utiliza, es decir que parte de la respuesta del alumnado corresponde a cada componente, sucede a veces que las categorías construidas para cada componente se solapan (Duschl, 2007; Díaz de Bustamante y Jiménez-Aleixandre, 1999). Así, una de las dificultades que se ha establecido, está en esclarecer lo que cuenta como prueba, justificación o conocimiento teórico (respaldo). En este sentido seguiremos la línea de algunos autores que han integrado estas dos últimas en una sola categoría, con el fin de eludir algunos de estos problemas metodológicos de solapamiento (Erduran et al., 2004; McNeill y Krajcik, 2007; Berland y McNeill, 2010; Sampson et al., 2013; Osborne et al., 2016). Usaremos una sola categoría que llamaremos razonamiento ya que esto permite abordar la diversidad de Apps con que trabajan los estudiantes.

Del trabajo de Sardà y Sanmartí (2000, p.414), se puede usar su idea que hay dos tipos de validez formal, la de argumentos “completos”, definidos como aquellos que contienen la triada prueba-justificación -conclusión, que para nuestro análisis será prueba-razonamiento-aserción, y los incompletos, en los que la triada no está presente, es decir la tríada es lo mínimo presente para que algo califique como argumento. Además, existen antecedentes sólidos que indican que la presencia progresiva de componentes del modelo de Toulmin puede utilizarse como indicador válido de calidad argumentativa (Berland y McNeill, 2010; Lin et al., 2012; Osborne et al., 2016). Así, hay que considerar que un argumento simple se compone de la tríada mínima prueba-razonamiento-aserción, mientras que los niveles más avanzados incluyen modulaciones (calificadores) y restricciones (refutaciones), tal como sugiere la progresión de Osborne et al. (2016). La progresividad se entiende aquí como profundización estructural del discurso argumentativo.

En concreto, se considerarán cinco componentes: pruebas, razonamiento (fusión de justificación y conocimiento básico), aserción (conclusión), calificadores modales y refutación.

La tabla 4.3.I presenta una sistematización de las definiciones de las componentes del modelo, indicadores para determinarlas en el modelo y ejemplos contextualizados al proyecto App Checkers.

Tabla 4.3.I - Definición de las componentes del modelo de Toulmin para usar en el proyecto.

Definición	Identificadores para la componente
Pruebas: Son observaciones, hechos, experimentos a los que se apela para evaluar la confiabilidad de la App.	-Pueden ser informaciones, observaciones, experimentos, magnitudes, cantidades, relaciones o testimonios. Es todo lo apelado para mostrar si el enunciado es cierto o falso, -Se ha considerado su presencia explícita por parte del alumnado. -Pueden ser hipotéticas (dadas al alumno), pero en el proyecto serán empíricas (buscadas por el alumno).
Aserción: Es un enunciado que pone en relación la explicación con las pruebas.	-Se identifica exclusivamente al ser de todos los enunciados presentes en el relato, aquel que esta sostenido por las pruebas presentadas. -Se ha considerado su presencia explícita de parte del alumnado. -Debe que ser entendida dentro del contexto de la demanda solicitada al alumnado, -No necesariamente tiene que estar al final del relato. La argumentación no es un proceso lineal. -Puede aparecer luego de conectores como “por lo tanto”, “entonces”, “en conclusión”, “así”, “en consecuencia” y “luego”.
Razonamiento: Es un enunciado que pone en relación la explicación con las pruebas.	-Puede ser expresada de forma explícita o implícita. -Si es implícita la explicitaremos. -Se suele construir implícitamente con enunciados de la forma: • Pruebas P entonces conclusiones C. • Pruebas tales como P permiten extraer o realizar tales conclusiones. • Dadas las pruebas P, puede asegurarse que C. - Se puede buscar en el discurso respondiendo ¿cómo el relator relaciona pruebas con conclusión y/o como autoriza pasar de las pruebas a la conclusión? -Puede aparecer luego de conectores como “porque” o “ya que” o “puede asegurarse”. -El razonamiento puede apelar a conocimientos teóricos o empíricos, a modelos, leyes o teorías e incluso a pruebas informales. Todas buscan fundamentar y/o justificar la conexión entre aserción y prueba. -Se puede buscar en el discurso respondiendo
Restricciones: Son las restricciones o excepciones a la conclusión.	-Pueden verse como una oposición al razonamiento, conclusión o pruebas empíricas (o teóricas). -Se puede buscar en el discurso respondiendo ¿Cuándo no se aplica la regla general (o razonamiento o pruebas)?
Calificadores modales (modulación):	-Son expresadas explícitamente. -Aparecen como una palabra u enunciado que concede algún tipo de probabilidad a la argumentación.

Expresan el grado de certeza o incertidumbre del argumento.	-Se puede buscar en el discurso respondiendo ¿Es todo esto (conclusión, justificación o pruebas) necesariamente así? ¿es siempre así o probable? ¿qué tan probable? -Suelen ser adverbios que modifican un verbo o adjetivos de sustantivos clave. Son ejemplos: “quizá”, “seguramente”, “totalmente”, “algunas”, “probablemente”, “algunas veces”, “la mayoría de las veces”.

4.3.3.2 Revisión e identificación de componentes desde los videos

Se analizaron las transcripciones de los 76 videos del alumnado usando la tabla 4.3.I para identificar las partes de cada video en las cuales el alumnado usa las componentes del modelo de Toulmin. Se construyó un diagrama de caja y flecha como el de la figura 4.3.H, que permitió observar qué componentes estaban presentes y cómo se articulaban en el discurso.

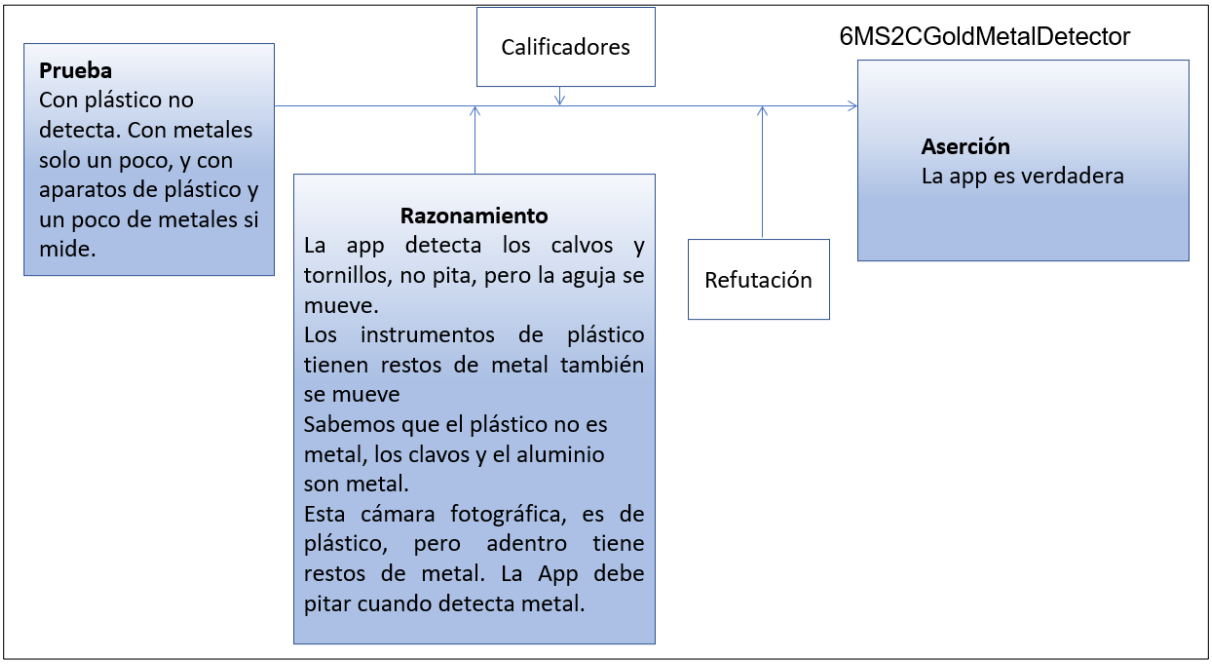

Figura 4.3.H - Ejemplo de un diagrama de caja y flecha construido para cada video.

4.3.3.3 Clasificación por agregación de componentes.

Desde los diagramas de flecha y caja, las pruebas fueron clasificadas en dos tipos: pruebas válidas, cuando provenían de observaciones o experimentaciones realizadas en condiciones confiables y fundamentadas empíricamente, y pruebas inválidas, cuando carecían de validez, por ejemplo, por errores de calibración. Las pruebas de tipo informal como comentarios o clasificación de Apps por usuario, por ejemplo, en Google Play también fueron

consideradas válidas. Esta distinción es importante pues se espera que el nivel de indagación influya directamente en la validez de la prueba utilizada.

Luego, se procedió a construir una base de datos que sistematizara la presencia o ausencia de cada componente. Para ello, se adoptó una codificación alfabética específica. Si una componente estaba presente en el video, se representaba con las siguientes letras:

- **PI**: prueba inválida
- **P**: prueba válida
- **R**: razonamiento (justificación y/o conocimiento básico)
- **A**: aserción
- **M**: modulación (calificadores modales)
- **F**: refutación

En caso de que la componente no estuviera presente, se dejaba la celda correspondiente en blanco. Estas codificaciones permitieron generar combinaciones únicas por video, que fueron luego agrupadas para analizar su frecuencia de aparición.

A partir de la información contenida en esta base de datos, se diseñó una rúbrica específica para clasificar los niveles de competencia argumentativa. Dicha rúbrica no evalúa la calidad interna de cada componente, sino la acumulación de componentes dentro del discurso argumentativo del alumnado. Esta elección metodológica fue realizada para mantener la coherencia con el objetivo principal del estudio, que es explorar la relación entre el desempeño en indagación y la presencia estructural de elementos argumentativos, sin introducir subcriterios que podrían sesgar la comparación entre dominios. Cada combinación observada en los productos del alumnado fue analizada en términos de cantidad de componentes presentes y tipo de prueba utilizada (válida o inválida). A partir de esta combinación, se construyó una rúbrica con cinco niveles de desempeño argumentativo.

Desde el punto de vista metodológico, se quiere representar la relación entre el desempeño en indagación y argumentación mediante un gráfico de bidimensional, donde se cruzarán niveles de indagación (eje X) y niveles de argumentación (eje Y), resulta crítico que la rúbrica de argumentación no contenga subcriterios internos que alteren la progresión natural de las componentes. De este modo, si se desagregaran las componentes en niveles internos de calidad —por ejemplo, diferencias entre justificaciones implícitas o explícitas, o tipos de respaldo teórico en el conocimiento básico— se estaría introduciendo ruido conceptual en el análisis de su relación con la indagación, transformando la variable dependiente en una mezcla

de componentes y subcomponentes. Se perdería así la capacidad de analizar dominios de influencia de forma limpia.

Se hace, no obstante, una única distinción en la progresión entre los niveles A1 y A2, relativa a la validez de las pruebas presentadas. Esto se debe a que la confiabilidad de la prueba está condicionada en gran parte por el diseño experimental, es decir, por la calidad de la indagación subyacente. Así, mantener esta distinción permite observar cómo el desempeño argumentativo responde a la calidad del proceso de indagación, manteniendo la coherencia con el modelo teórico y con los objetivos empíricos del análisis. La rúbrica resultante se presenta en la Tabla 4.3.J.

Como parte del proceso de validación de la rúbrica de niveles de desempeño argumentativo presentada en la Tabla 4.3.J, el director de tesis llevó a cabo una codificación independiente del 30% de los productos del alumnado, aplicando los criterios definidos en la rúbrica basada en el modelo de Toulmin. Este procedimiento permitió contrastar la aplicación de los descriptores en distintos evaluadores, determinando la coherencia interpretativa en la identificación y acumulación de componentes argumentativas. Las discrepancias detectadas fueron discutidas, lo que dio lugar a una revisión consensuada de los casos y a ajustes en los criterios. Esto contribuyó a consolidar la estructura progresiva de los niveles argumentativos, alineando su formulación con los objetivos del estudio y los datos analizados.

Posteriormente, cada video fue clasificado de manera única según el nivel alcanzado en la rúbrica. Esta clasificación final permitió generar una nueva columna en la base de datos, correspondiente al nivel de argumentación (A0–A4), que se utilizará Posteriormente en el análisis conjunto con los niveles de indagación. Este procedimiento permitió construir la gráfica bidimensional entre ambas dimensiones competenciales del gráfico de la figura 4.4.A, donde se estudia la asociación entre los dos dominios del estudio: indagación y argumentación.

Tabla 4.3.J - Rúbrica de niveles de argumentación construida a partir de la agregación de componentes del modelo de Toulmin.

Nivel	Descripción del desempeño argumentativo
A0 Sin argumentación	Los estudiantes no presentan un argumento. Aunque muestren datos, no los utilizan para plantear ninguna argumentación.
A1 Argumentación deficiente	La argumentación consta de prueba, razonamiento y aserción, pero la prueba presentada es inválida.

A2 Argumentación simple	La argumentación consta de prueba, razonamiento y aserción, y la prueba es válida y fiable. Sin embargo, no se incluye modulación ni refutación del argumento.
A3 Argumentación intermedia	La argumentación, además de prueba, razonamiento y aserción, también incluye una modulación sobre la confiabilidad de la App elegida, pero no especifica las restricciones (Refutación) que lo lleva a dicha modulación.
A4 Argumentación compleja	Además de conectar adecuadamente la prueba, razonamiento y aserción, modula su propia conclusión añadiendo las restricciones del contexto.

4.4 Resultados del estudio 1

4.4.1 Relación entre niveles de desempeño en indagación y argumentación

Para construir la tabla de frecuencias con las combinaciones entre niveles de indagación y argumentación, en primer lugar, se sistematizó la información de los 76 videos en una base de datos, registrando para cada uno el nivel de desempeño en indagación y el nivel de desempeño en argumentación, según las respectivas rúbricas desarrolladas previamente, tabla 4.3.H y tabla 4.3.J. A continuación, se clasificó cada video asignándole una categoría específica de desempeño en indagación (I0 a I4) y otra en argumentación (A0 a A4), de acuerdo con los criterios definidos. Una vez completada esta codificación, se generó una nueva columna que combinaba ambas clasificaciones en un solo código (por ejemplo, I2-A2 o I4-A3), representando así cada cruce posible entre ambas dimensiones. Finalmente, se utilizó una tabla dinámica para resumir las frecuencias de aparición de cada combinación, permitiendo visualizar cuántos videos se ubicaban en cada cruce específico entre desempeño en indagación y en argumentación. La tabla 4.4.A tiene los resultados.

Tabla 4.4.A - Frecuencia de combinaciones entre niveles de desempeño en indagación y argumentación en los 76 videos analizados.

Combinaciones indagación argumentación	Frecuencia
I0-A0	7
I0-A3	1
I1-A1	2
I2-A2	20
I2-A3	10
I2-A4	3

I3-A2	2
I3-A3	15
I3-A4	3
I4-A2	3
I4-A3	6
I4-A4	4
Total general	**76**

A partir de la categorización para los niveles de desempeño en argumentación e indagación de la tabla 4.4.A, se construyó el gráfico de la figura 4.4.A, en el cual se presenta la cantidad de videos situados en cada nivel, mostrando los niveles de indagación en el eje horizontal y los niveles de argumentación en el eje vertical. Estos niveles mantienen la notación I0, I1, I2, I3 e I4 para indagación y A0, A1, A2, A3 y A4 para argumentación. El recuento de videos que caen en una intersección de niveles se representa con un círculo cuyo diámetro es proporcional con el número de videos en cada coordenada.

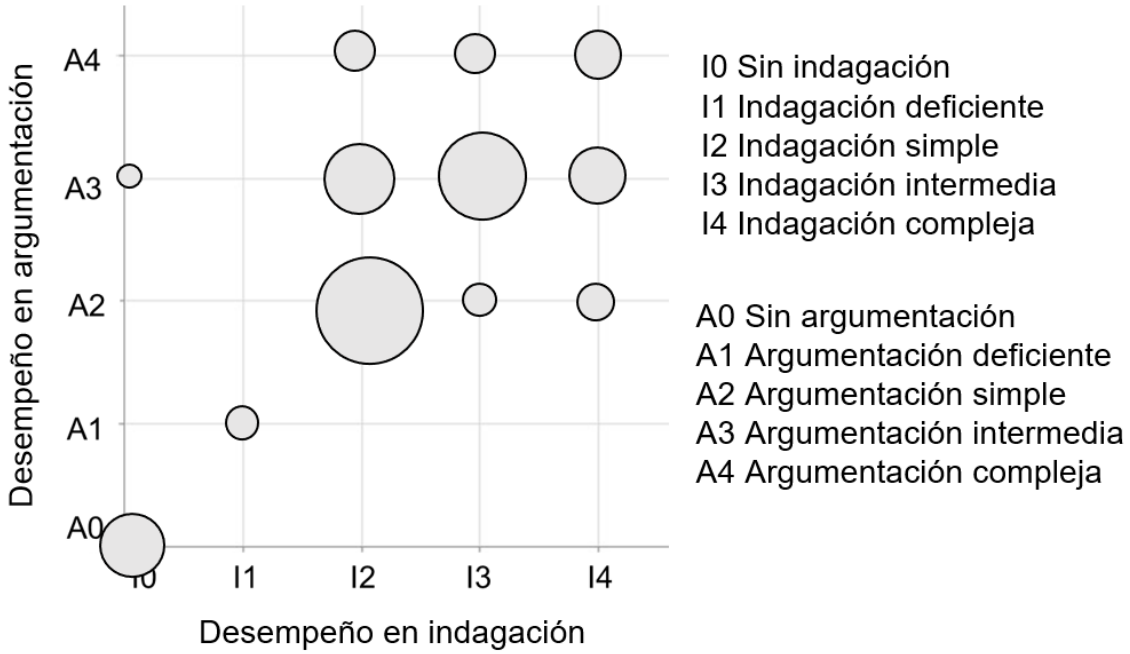

Figura 4.4.A - Distribución de los niveles de desempeño para los n = 76 videos.

En la figura 4.4.A se aprecia que frente a la misma pregunta investigativa encontramos una amplia variedad de desempeño en indagación y argumentación, pero con algunas regularidades. Podemos observar que la mayoría de los videos se sitúan en el cuadrante superior - derecho del gráfico, en los rangos entre I2 e I4 y, A2 y A4, aunque con menor presencia en los niveles I4 y A4 que en los niveles inferiores, coincidiendo así con lo expuesto en investigaciones previas (Osborne et al. 2016; Berland y McNeill, 2010; Yeah y She, 2010; Hsu

et al., 2015). Además, del mismo gráfico se puede observar que las combinaciones más concurridas son aquellas que se sitúan en el eje diagonal del gráfico en las coordenadas I0-A0, I2-A2 y I3-A3, por lo que parecería que una mejor indagación coincide con una mejor argumentación.

Así, en primer lugar, los videos categorizados en el nivel de coordenadas I0-A0 (n=7) son aquellos donde no se ha planteado una investigación que permita obtener pruebas, por lo que tampoco existe argumentación alguna, tal como se ejemplifica en la siguiente transcripción:

> He elegido la App medidor de distancias. Puedes medir distancia corta o larga [*Muestra las opciones por pantalla, enfoca una estantería de su habitación y usa la App para medir su longitud*]. El resultado es 160,56 m [*El resultado no tiene sentido, ya que la estantería se encuentra a unos 2 metros (Figura 4.4.B)*]. Lo bueno que tiene la App es que el resultado se queda guardado. Esta es mi App, espero que os haya gustado.

Figura 4.4.B - Capturas para un video situado en la coordenada I0-A0.

En segundo lugar, los videos de nivel I2-A2, que son los más numerosos (n=20), muestran un nivel de indagación y argumentación simples, ya que los estudiantes sólo toman una o dos lecturas válidas en cuanto al uso instrumental, y además no incluyen ningún tipo de refutación ni modulación. Un ejemplo es la App "PlantNet" de la figura 4.4.C que, dada una imagen, determina qué tipo de planta contiene esa imagen. En este ejemplo, un grupo de estudiantes trabajó la App tomando dos mediciones, y comprobaron que las imágenes de salida y entrada eran similares, dando así por válida la App sin aportar mayor matiz. La transcripción es la siguiente:

> Verificaremos la App PlantNet, detecta plantas mediante la cámara haciendo una foto [*Explican las opciones y como operar la App*]. Comenzaremos con un cactus [*toman una foto*] (Figura 4.4.C

izquierda), y aquí pone fotos de cactus, o sea que sí reconoce que planta es. Ahora tomaremos una foto a esta planta que es un Aloe Vera [*toman la foto*] (Figura 4.4.C derecha), y aparece un Aloe Vera, si la reconoce. Hemos verificado que esta aplicación es cierta, porque salen fotos de la misma planta que hemos hecho la foto, ha funcionado con todas las plantas mostradas.

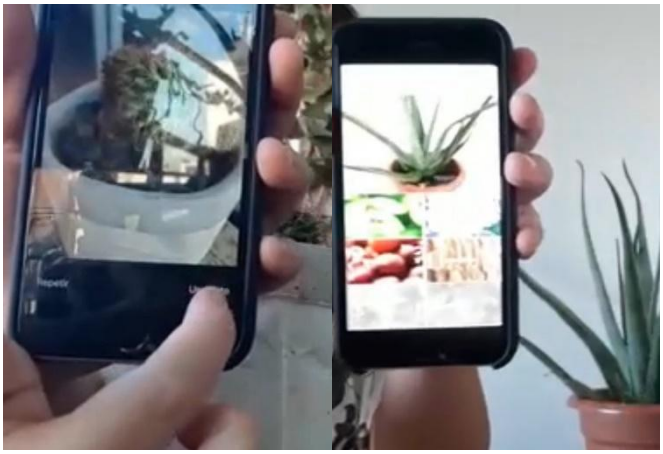

Figura 4.4.C - Transcripción y capturas de un video situado en la coordenada I2-A2.

La segunda combinación de niveles más frecuente es I3-A3 (n=15), que corresponde a una indagación y argumentación intermedia. Un ejemplo de videos en este grupo corresponde a la figura 4.4.D, donde un estudiante usó la App "Thermal Scanner" que simula un escáner térmico, llevó a cabo tres experimentos diferentes recogiendo varias mediciones y concluyendo que la App cambiaba los colores entre imágenes de forma aleatoria. La transcripción es la siguiente:

Utilizaremos la App Thermal Scanner. Es una aplicación de cámara térmica [*luego, explica cómo funciona la App y sus opciones*]. Para este experimento utilizaremos agua fría en vasos de colores y vasos transparentes y también utilizaremos una tostadora. Para empezar, en los dos vasos transparentes he puesto agua fría y agua caliente, y pone que están a la misma temperatura. Ahora en estos tres vasos de colores me he puesto agua fría a igual temperatura (Figura 4.4.D izquierda), y sale que uno está muy frío y los demás están calientes, además se ve cómo cambia los colores. Por tanto, esta aplicación sólo cambia los colores. Y para terminar utilizaremos la tostadora. Ahora está fría y después la encenderemos, para ver si cambia o no. Vamos a ver [*la enciende y muestra la imagen de la App*], pone que las partes que están calientes están frías (Figura 4.4.D derecha). Las líneas estas azules indican frío, pero están calientes. Después de realizar este experimento llegamos a la conclusión que la App es falsa y sólo cambia los colores de los objetos por otros.

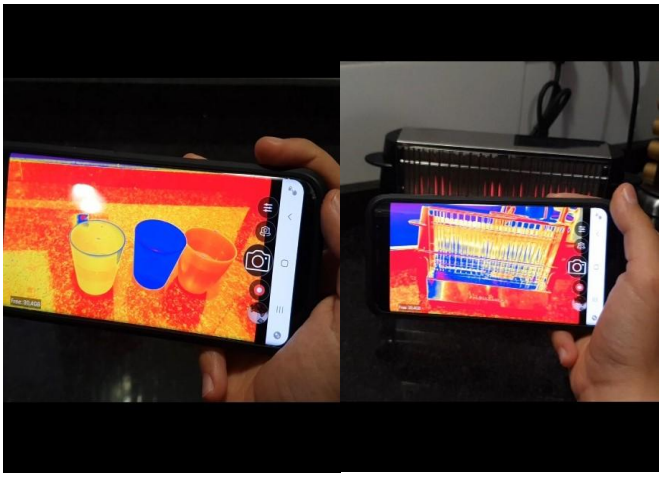

Figura 4.4.D - Transcripción y capturas de un video situado en la coordenada I3-A3.

Finalmente, dentro de la diagonal del gráfico de la figura 4.4.A cabe destacar los videos situados en la coordenada I4-A4 (n=4), donde encontramos simultáneamente una indagación y argumentación complejas. En la figura 4.4.E se muestra un ejemplo que corresponde al trabajo de un estudiante que indagó sobre la App "Photomath", puso a prueba la App de distintas maneras, usando tablas para obtener gráficos, y explorando sus alcances en cuanto a reconocimiento de imágenes entregándole diferentes tipos, y finalmente asignando una confiabilidad del 75%, presentando así una modulación en su argumentación y explicando estos matices en términos de refutaciones sobre la capacidad de la App en reconocer imágenes. La transcripción es la siguiente:

> Usamos la App "Photomath", que resuelve cuestiones matemáticas. Primero plantearemos una ecuación escrita a mano para ver si sabe detectarla y resolver (Figura 4.4.E izquierda), pasa la prueba. Ahora le pediremos que resuelva un sistema de ecuaciones. También enfocaremos el número del ejercicio, y la palabra solución. Lo hace perfectamente y, además, nos deja elegir el método de resolución y muestra todos los pasos hasta la solución. Quizás uno de los mayores problemas está en el enfoque de la cámara. Aquí se puede ver cómo el escaneo de la imagen falla. Ahora intentaremos que nos represente en forma de gráfica las siguientes tablas [usan dos]. No las detecta. La siguiente prueba consistirá dar las fórmulas de las tablas. En ese caso sí lo resuelve en forma de gráfica. Por último, intentaremos que resuelva un problema de enunciado y algo sin sentido, y no sabe detectarlo (Figura 4.4.E derecha). Teniendo en cuenta que le hemos puesto retos muy difíciles, creo que sería justo darle una confiabilidad del 75%.

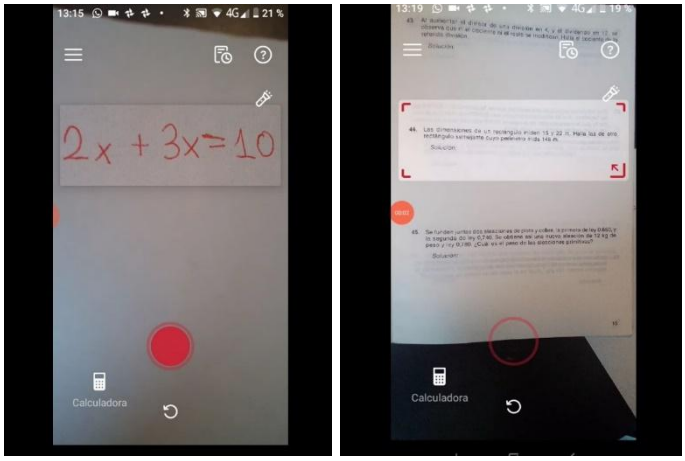

Figura 4.4.E - Transcripción y capturas de un video situado en la coordenada I4-A4.

A pesar de que el grueso de los videos se encuentre en la diagonal del gráfico, esto no implica que un mayor desempeño en indagación siempre vaya de la mano de un mayor desempeño en argumentación. Por un lado, encontramos algunos pocos videos en las coordenadas I3-A2 (n=2) e I4-A2 (n=3), que corresponde a desempeños de indagación altos que aportan pruebas robustas que luego no usan para argumentar adecuadamente, y también en la coordenada I4-A3 (n=6), donde a pesar de presentar una indagación compleja solamente desarrollan una argumentación intermedia.

Por otro lado, los videos que se sitúan por encima de la diagonal del gráfico se corresponden con indagaciones más simples y argumentaciones más elaboradas, incluyendo las coordenadas I2-A3 (n=10), I2-A4 (n=3) e I3-A4 (n=3). En todos estos casos observados, los datos que obtienen los estudiantes en su investigación son simples, pero los usan adecuadamente en su argumentación e incluyen otros datos informales que le permite modular su argumento o refutarlo parcialmente. Por ejemplo, en uno de los videos de la coordenada I2-A4 vemos un estudiante usando la App "Sonómetro" para tomar solamente dos medidas de sonido del televisor (con el volumen alto y bajo respectivamente), y mostrando cómo la aguja que medía los dB de la App subía y bajaba. A pesar de la simplicidad del experimento (indagación de nivel I2), donde no toma un número de mediciones suficientemente amplio que permita asegurar el patrón presentado como prueba, posteriormente señala que en ausencia de sonido la App no marca 0 dB tal como él había previsto, por lo que añade matices al grado de confiabilidad que le ha dado a la App, en forma de modulación y refutación. Otro ejemplo, quizás el que presenta mayor discrepancia entre el nivel de indagación y argumentación, es el situado en la coordenada I0-A3 (n=1), donde un estudiante no investiga científicamente la App, pero argumenta basándose en una prueba informal como son los comentarios y la valoración

de la App dados por otros usuarios, y presentando estas valoraciones para construir su argumento.

En paralelo, si estos resultados (n=76) se desagregan según las tres implementaciones hechas, una en el 2019-20 (n=29) y dos en el 2020-21 (n=35 y n=12), se obtienen los tres gráficos de la figura 4.4.F, donde se observan pequeñas diferencias en la distribución de niveles de desempeño.

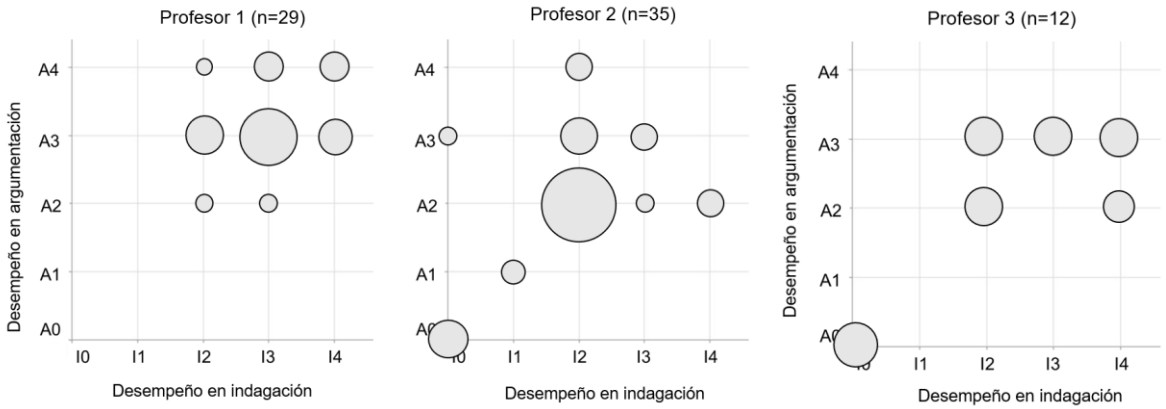

Figura 4.4.F - Distribución de los niveles de desempeño para indagación y argumentación según implementación.

El gráfico de la izquierda corresponde a los 29 videos de los estudiantes de la primera implementación, donde el profesor 1 usó una escala de confiabilidad (Figura 4.2.A) para orientar a los estudiantes en su veredicto respecto a su App, las fases 1 a 3 fueron realizadas en el aula y los videos en sus casas. En este primer gráfico se observa que 11 de 29 videos (38%) se sitúan en la coordenada I3-A3 y que en 29 de 29 (100%) se puede identificar la triada Prueba – Razonamiento – Aserción (nivel de indagación y argumentación ≥ 2). En cambio, el gráfico del centro corresponde a 35 videos con el profesor 2, que no usó en su proyecto la escala de confiabilidad y además abordó el proyecto de una forma más superficial, por lo que muchos estudiantes se limitaron a demostrar si la App funcionaba o no funcionaba. Se observa que 17 de 35 videos (49%) se sitúan en la coordenada I2-A2. Finalmente, en el gráfico de la derecha se muestra el análisis de los 12 vídeos recogidos por el profesor 3 que tampoco usó la escala de confiabilidad y planteó el proyecto como un contexto para trabajar el método científico. Se observa que tanto con el profesor 2 como con el profesor 3 algunos videos se sitúan en la coordenada I0-A0 mostrada en la figura 4.4.B, y que corresponde a los estudiantes que no llegaron a entender la finalidad del proyecto. Todas estas diferencias en el nivel de desempeño del alumnado según la implementación realizada por cada profesor (poniendo más o menos

énfasis en la escala de confiabilidad, en el diseño experimental, etc.) sugieren una cierta influencia didáctica en el producto final, pero que debido a las limitaciones metodológicas no permiten establecer una única relación causal, sino simplemente una tendencia a explorar en futuras implementaciones del proyecto.

Finalmente, al desagregar los datos correspondientes a los 76 videos según el tipo de App elegida, se obtienen los gráficos de la figura 4.4.G, donde se presentan las 6 Apps más usadas por los estudiantes de acuerdo con la tabla 4.3.A. En estos gráficos se muestra que una misma App lleva a los estudiantes a diferentes niveles de desempeño tanto en indagación como en argumentación. Si comparamos los videos producidos por diferentes estudiantes que analizan una misma App, todos se distribuyen como mínimo en 3 niveles distintos de indagación y/o de argumentación, de manera que la variedad en el nivel de desempeño dentro de cada conjunto de videos según la App elegida es mayor que las diferencias que pudiera haber entre diferentes Apps.

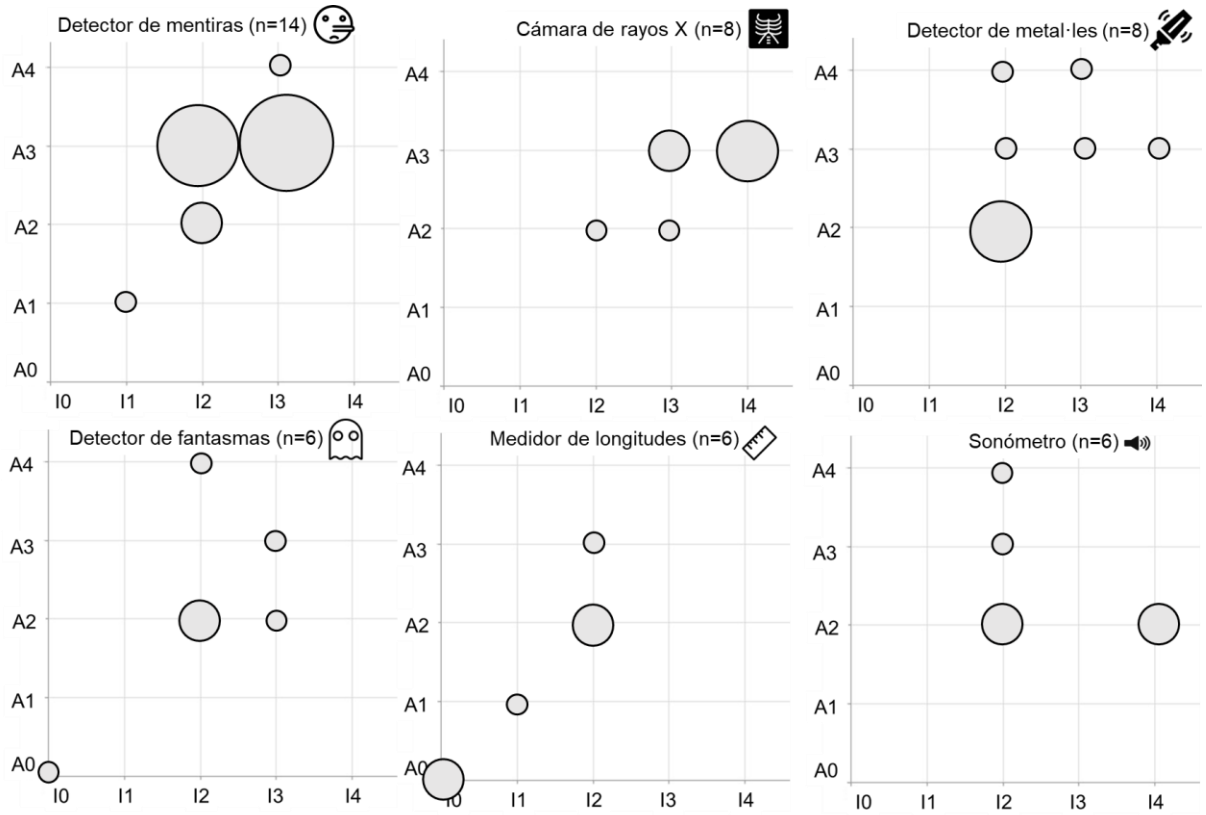

Figura 4.4.G - Distribución de los niveles de desempeño para indagación y argumentación según Tipo de App.

Las 3 Apps con una apariencia más científica (Rayos X, Detector de Metales y Sonómetro) son las únicas donde algunos estudiantes logran el nivel 4 de indagación, mientras

que en las otras tres (Mentiras, Fantasmas y Longitudes) ningún grupo de estudiantes logra plantear indagaciones complejas. Del mismo modo, parecería que las Apps con una confiabilidad a priori más clara son en las que encontramos argumentaciones más sofisticadas, mientras que las Apps más confusas ningún estudiante logra un nivel de argumentación compleja. En el caso de Mentiras y Fantasmas todos los estudiantes creen que serán falsas y Metales y Sonómetro suelen recibir la confianza de los estudiantes desde el principio. Esto podría sugerir que los estudiantes argumentan mejor (añadiendo modulación y refutación) si se enfrenta a Apps cuya confiabilidad conocen de antemano, cosa que coincide con lo descrito por Osborne et al. (2016), donde el desempeño en argumentación es mayor en contextos donde los estudiantes se sienten más cómodos y con menor complejidad conceptual (Osborne et al., 2016). No obstante, debido al limitado tamaño de la muestra analizada, no se puede generalizar esta tendencia, y sería necesario un análisis más detallado y sistemático.

4.5 Conclusiones del estudio 1

El estudio 1 basado en la primera implementación del proyecto App Checkers (Tabla 3.A) logró cumplir su primer objetivo, que consistía en caracterizar los niveles de desempeño en indagación científica del alumnado mediante una rúbrica construida a partir de combinaciones de criterios procedimentales. A partir del análisis sistemático de 76 videos producidos por el alumnado participante, se codificaron distintas dimensiones del diseño experimental, tales como el uso de calibración o la repetición de mediciones. Esta codificación permitió derivar una escala de cinco niveles (I0 a I4), que fue aplicada a toda la muestra. La construcción del instrumento y su uso en el análisis permitieron distinguir entre diferentes niveles de elaboración en la práctica de indagación, desde niveles en los que no se planteaba investigación alguna, hasta niveles donde se identificaban múltiples controles, uso de tablas, y toma sistemática de datos.

El segundo objetivo, centrado en caracterizar los niveles de argumentación científica en los productos audiovisuales del alumnado mediante el modelo de Toulmin, también fue cumplido. Se diseñó una rúbrica específica que permitió evaluar la presencia de seis componentes: pruebas válidas, pruebas inválidas, aserción, razonamiento, moduladores y refutaciones. Esta rúbrica permitió identificar niveles de desempeño argumentativo desde A0 a A4, que fueron asignados a cada video en función de su estructura y grado de sofisticación del argumento. La presencia o ausencia de cada componente fue registrada de forma sistemática, y se utilizó para categorizar a los estudiantes según la complejidad de su

argumentación. Gracias a esta herramienta fue posible reconocer desde productos sin ningún tipo de argumentación, hasta producciones que incluían razonamientos complejos, matices y contraargumentaciones explícitas.

El análisis cruzado de ambas dimensiones, representado en la Figura 4.4.A, revela una amplia variedad de combinaciones, lo cual confirma que las habilidades científicas del alumnado no se desarrollan de forma homogénea. Tal como señalan Osborne et al. (2016), es posible observar estudiantes que alcanzan niveles complejos de argumentación a partir de procesos de indagación simples, así como casos inversos, donde una buena indagación no se traduce necesariamente en una argumentación sólida, hay estudiantes que a pesar de plantear una indagación alta no usan las pruebas obtenidas.

Uno de los hallazgos del estudio 1 es que existe una cierta tendencia según la cual los niveles altos de indagación están asociados con mayores niveles argumentativos, pero esta correlación no es absoluta (Castellanos et al., 2002; Koerber y Osterhaus, 2019). Por ejemplo, se identificaron casos donde un diseño experimental complejo (I4) solo da lugar a una argumentación intermedia (A2), lo cual sugiere la necesidad de una intervención docente que potencie explícitamente la conexión entre pruebas empíricas y construcción de argumentos.

El cruce de ambos niveles permitió observar con nitidez las combinaciones de desempeño más frecuentes, como I2-A2 e I3-A3, así como algunas discordancias interesantes, como casos con buena argumentación basada en datos informales pese a una pobre indagación. Este análisis evidencia que los objetivos planteados inicialmente se cumplieron de manera sólida, permitiendo no solo identificar niveles diferenciados de competencia, sino también comprender cómo se relacionan entre sí la indagación y la argumentación en un contexto auténtico de investigación escolar.

Una de las relaciones teóricas exploradas en este estudio fue la correspondencia entre la presencia de restricciones o refutaciones en los argumentos (componente “F” de Toulmin) y el tratamiento explícito del criterio “límites operativos de las Apps” durante la indagación (criterio C de la rúbrica de indagación). Desde un punto de vista epistemológico, ambos elementos cumplen una función similar: establecer condiciones de validez o invalidez en las afirmaciones científicas.

Sin embargo, el análisis de las figuras revela que esta relación no es directa ni sistemática entre el alumnado. Aunque algunos de los estudiantes que alcanzan el nivel A4 de argumentación (argumentación compleja con restricciones) también se sitúan en niveles altos

de indagación (I3 o I4), también se observan casos donde esta refutación argumentativa emerge a partir de indagaciones más simples (I2). Esto sugiere que la capacidad de incorporar restricciones en los argumentos no depende exclusivamente del diseño experimental, sino también de habilidades cognitivas y discursivas que deben ser promovidas intencionalmente por el profesorado (Jiménez-Aleixandre y Crujeiras-Pérez, 2017).

El análisis desagregado por docente (Figura 4.4.F) permitió identificar patrones diferenciados en la distribución del desempeño entre indagación y argumentación. En particular, se observan tres aspectos relevantes: (1) Variabilidad de trayectorias formativas: El Profesor 1 que implementó la escala de fiabilidad, presenta una mayor diversidad de combinaciones, lo que puede indicar una implementación más abierta o flexible del proyecto. En contraste, el Profesor 3 muestra una agrupación más uniforme en los niveles A2 e I2, lo cual sugiere una tendencia a resultados medios en ambas competencias. (2) Predominio de patrones simples en ciertas aulas: El Profesor 2, con una muestra numerosa (n=35), muestra una fuerte concentración en la combinación I2–A2. Este patrón puede deberse a una guía metodológica más centrada en el "método experimental". (3) Pocos casos extremos de desempeño: En los tres profesores, la cantidad de estudiantes que alcanzan simultáneamente los niveles máximos (I4–A4) o mínimos (I0–A0) es baja. Esto puede reflejar que el diseño de la actividad propicia niveles intermedios de desempeño, pero requiere ajustes para favorecer trayectorias más ambiciosas de aprendizaje.

El cumplimiento de ambos objetivos puede considerarse alto, tanto por la consistencia metodológica del análisis como por la riqueza de los datos obtenidos. A nivel didáctico, este estudio ofrece implicaciones. En primer lugar, pone de manifiesto la importancia de trabajar la indagación y la argumentación como competencias interrelacionadas, pero no necesariamente paralelas. El hecho de que algunos estudiantes argumenten bien con escasa indagación o que, por el contrario, indaguen rigurosamente sin lograr articular argumentos sólidos, muestra que ambas dimensiones requieren de un trabajo pedagógico intencionado y posiblemente diferenciado.

También la variedad de desempeños en indagación y argumentación identificadas señala algunas oportunidades de enseñanza y aprendizaje, como son las ideas de verificación, de fiabilidad y validez, del concepto de medición y de error humano en la medición, de réplica, de refutación, etc. Sin embargo, dado que un grueso importante de estudiantes se queda en un nivel de argumentación e indagación simple o inferior, se pone de manifiesto la necesidad de promover las habilidades de indagación de los estudiantes para trabajar pruebas científicas y

evaluar su confiabilidad, especialmente en un momento donde la enseñanza de las ciencias se enfrenta a retos propios del siglo XXI (Osborne et al., 2022). Para ello, será necesario seguir investigando sobre qué estrategias educativas son más efectivas para dicho propósito, centrándose no solamente en los productos finales que presentan los estudiantes al finalizar su proyecto de verificación, sino también adentrarse en los procesos y las prácticas en que participa el alumnado.

Además, los resultados sugieren que el uso de herramientas como escalas de confiabilidad, el andamiaje explícito del diseño experimental y la discusión de criterios de validez pueden elevar significativamente el nivel de desempeño del alumnado. Finalmente, el análisis muestra que las decisiones didácticas del profesorado, como el uso o no de instrumentos guía, inciden de forma directa en la calidad del producto final, lo que refuerza la necesidad de diseñar propuestas de enseñanza que integren estrategias de modelado, exploración guiada y reflexión crítica sobre los datos.

Además de las implicancias didácticas ya discutidas, resulta pertinente señalar algunas consideraciones metodológicas derivadas de las decisiones adoptadas en el diseño del análisis. La forma en que se construyó y aplicó la rúbrica de argumentación, basada en la agregación progresiva de componentes del modelo de Toulmin (A0 a A4), posee fortalezas que justifican su uso en este estudio 1. Entre ellas, se destaca su coherencia con modelos de progresión ampliamente reconocidos en la literatura especializada (Osborne et al., 2016; Berland y McNeill, 2010), así como su capacidad para conservar el rol de la argumentación como variable dependiente en relación con la calidad de la indagación, un aspecto clave para responder a la pregunta de investigación planteada. La elección de este enfoque también permitió construir un instrumento evitando la sobrecarga analítica derivada de considerar demasiadas dimensiones internas que dificulten la comparación entre competencias.

No obstante, esta opción también conlleva limitaciones metodológicas que deben ser reconocidas. La agrupación por número de componentes puede resultar algo laxa si no se controla la calidad interna de cada uno, la cual se consideró lograda e intencionada al especificar como se identificaría cada componente (tabla 4.3.1) y por las codificación y consensos entre evaluadores. A su vez, no se han explorado con profundidad matices cualitativos vinculados al tipo de conocimiento básico activado o a la pertinencia de las pruebas utilizadas, elementos que estudios como el que Sardà y Sanmartí (2000) sugieren. Por otra parte, este enfoque requiere ajustes si se pretende aplicar la rúbrica en contextos distintos al actual, particularmente en estudios centrados exclusivamente en la argumentación científica.

En conjunto, el enfoque adoptado, basado en la codificación de componentes argumentativos, su sistematización mediante diagramas, la agrupación por progresión y su cruce con la calidad de la indagación, ha demostrado ser eficaz para analizar el doble producto generado por el alumnado en este estudio 1. La elección de una rúbrica agregativa, centrada en la acumulación de componentes, no solo se alinea con el marco teórico de Toulmin y sus aplicaciones educativas, sino que además permite establecer relaciones interdimensionales sin comprometer la claridad del análisis. Esta estrategia ofrece, por tanto, una base para futuras investigaciones sobre la articulación entre prácticas escolares y competencias científicas que se consideran esenciales.

De forma complementaria, la estrategia metodológica basada en el análisis y agrupación de combinaciones como vía para construir una rúbrica de evaluación competencial responde a un enfoque holístico, donde la competencia se concibe como la integración de habilidades múltiples. Desde esta perspectiva, el desempeño no puede evaluarse de manera fragmentada, sino como una manifestación integrada de saberes, habilidades y actitudes en contextos auténticos, tal como lo plantea Ferrés (2017). Agrupar combinaciones de criterios permite observar cómo el alumnado articula distintos aspectos del trabajo experimental, brindando una visión más completa y coherente del nivel competencial alcanzado.

Este tipo de análisis permite también una evaluación más auténtica, ya que se centra en las competencias científicas puestas en juego durante la resolución de problemas, no en suposiciones abstractas. Interpretar las decisiones metodológicas del alumnado, como calibrar antes de medir o repetir ensayos, dentro del contexto en el que se desarrolló la actividad fortalece la validez de la evaluación.

Asimismo, estructurar estas combinaciones en niveles progresivos permite construir categorías con sentido didáctico, útiles no solo para discriminar distintos niveles de desempeño, sino también para orientar la intervención docente. Cada nivel permite identificar fortalezas y debilidades específicas, ofreciendo así una base para la retroalimentación formativa y el diseño de estrategias de mejora ajustadas a las necesidades detectadas. Además, los niveles emergen de la evidencia empírica observada en los videos, no de suposiciones teóricas previas, lo que garantiza una evaluación fundamentada en datos reales. En conjunto, este enfoque metodológico —centrado en patrones observables y decisiones epistemológicas— proporciona una herramienta válida, fiable y útil para valorar aprendizajes en contextos de indagación científica, con potencial para ser transferida a tareas similares donde el foco esté puesto no solo en los resultados, sino también en los procesos que los generan.

CAPÍTULO V

ESTUDIO 2: ESTRATEGIAS PARA EVALUAR LA CONFIABILIDAD DE APLICACIONES MÓVILES EN ESTUDIANTES DE BACHILLERATO

El Capítulo V presenta el segundo estudio centrado en el análisis de estrategias para evaluar la confiabilidad de Apps mediante pósteres científicos. Se identifican cuatro estrategias de validez (comparación con valor esperado, comparación instrumental, triangulación y poner a prueba la App) y cinco niveles de fiabilidad.

Este trabajo ha dado lugar a un artículo aceptado para publicación en la revista ***Didáctica de las Ciencias Experimentales y Sociales***, con el título Estrategias para evaluar la confiabilidad de aplicaciones móviles en estudiantes de bachillerato (Aguilera, López-Simó y Domènech Amador, en prensa).

5.1 Introducción al estudio 2

En el estudio 1 correspondiente a la primera implementación del proyecto App Checkers se identificaron patrones metodológicos en la indagación del alumnado, desde la identificación de variables hasta la elaboración de pruebas sustentadas en la práctica de indagación para su uso en argumentación, lo que permitió establecer niveles progresivos de desempeño. No obstante, se detectaron también ciertas estrategias de validación no previstas inicialmente, como la comparación con valores esperados, el uso de instrumentos externos y la instalación de la App en distintos dispositivos para realizar un tipo de triangulación. Estas estrategias, aplicadas espontáneamente por el estudiantado, parecían no pertenecer a una única fase del trabajo científico, sino actuar como mecanismos epistémicos transversales. Esta hipótesis se vio reforzada al contrastar los hallazgos con los Conceptos de Evidencia (CoE), los cuales muestran que acciones como la triangulación o el uso de datos secundarios no se limitan necesariamente a un momento específico del proceso, sino que aparecen en distintas fases de la indagación. Estos criterios para analizar el desempeño en indagación se marcaron de color negro en la tabla 4.3.C y se resumen en la tabla 5.1.A Por ejemplo, el criterio "Utiliza datos secundarios como fuente de evidencia en un informe de investigación", implica una acción aplicada donde el estudiante integra efectivamente datos ajenos, esto puede ocurrir durante la fase de diseño de la indagación como parte de la planificación donde debe buscarlos y justificar su uso. Luego, ocurre en el análisis, cuando compara, triangula o apoya sus datos con estos datos secundarios, pero también en las aserciones realizadas cuando esos datos se invocan como parte de la justificación.

A partir de las observaciones anteriores, se realizó un segundo estudio o estudio 2 para caracterizar y clasificar con mayor precisión las estrategias que el alumnado emplea para evaluar la confiabilidad de las Apps trabajadas. Este nuevo análisis se centra en los pósteres científicos elaborados por estudiantes de bachillerato, examinando de forma cualitativa tanto las estrategias de validez como los niveles de fiabilidad presentes en sus diseños de investigación. La intención es profundizar en cómo estas estrategias se manifiestan en contextos reales de investigación escolar, y de qué manera podrían ser enseñadas, fortalecidas y sistematizadas en futuros diseños didácticos.

En síntesis, el segundo estudio se justifica por tres razones principales:

1. La aparición de estrategias de triangulación en los trabajos estudiantiles que no habían sido objeto de evaluación sistemática en el primer estudio.
2. Ciertos criterios CoE, que en el proyecto App Checkers no operan como indicadores de una fase, sino como estrategias distribuidas a lo largo de toda la investigación.
3. La necesidad de caracterizar el pensamiento epistémico del alumnado, especialmente en lo relativo al uso consciente de procedimientos que refuerzan la validez y la fiabilidad de sus conclusiones, más allá de las prácticas procedimentales de indagación científica.

Este nuevo enfoque busca, por tanto, ampliar la comprensión sobre cómo el estudiantado evalúa la confiabilidad en la producción de conocimiento, contribuyendo a un marco más robusto para el análisis de prácticas auténticas de investigación escolar.

Tabla 5.1.A - Criterios CoE transversales que operan como estrategias epistémicas en diversas fases de indagación.

Criterio	Fase de la indagación
Compara los resultados obtenidos con hallazgos previos o datos esperados para evaluar su consistencia interna y externa (CoE 20).	Diseño experimental /Análisis de datos/ Conclusiones
Aplica estrategias de triangulación metodológica para aumentar la validez de una investigación. (CoE 20).	Diseño experimental /Análisis de datos/ Conclusiones
Evalúa el peso probatorio del conjunto de datos considerando su coherencia, replicabilidad y diversidad metodológica (triangulación) (CoE 20).	Diseño experimental /Análisis de datos/ Conclusiones
Evalúa la credibilidad de una evidencia científica considerando el tipo de evidencia y el consenso disciplinar (CoE 21).	Conclusiones / Metarreflexión
Identifica posibles sesgos del experimentador al analizar o reportar evidencia (CoE 21).	Conclusiones / Metarreflexión
Discute la aceptabilidad de las consecuencias prácticas de una conclusión científica (CoE 21).	Aplicación del conocimiento / Metarreflexión
Reconoce el peso social o político que puede tener una evidencia al ser comunicada (CoE 21).	Comunicación de resultados / Metarreflexión

5.2 Contexto y objetivos de investigación

El proyecto App Checkers volvió a ser implementado. En esta segunda ocasión fue realizado en un curso de primer año de bachillerato en el Instituto Francesc Ferrer i Guàrdia, del municipio Sant Joan Despí de la región de Cataluña. El número de participantes en el

estudio fueron 38 durante el curso 2020-21, en la asignatura de “Ciencias para el mundo contemporáneo”. A diferencia de la implementación del primer estudio (Capitulo IV), el producto final ha sido un póster científico no un video. Se recibió un total de 19 pósteres científicos. La App sobre la cual trataba cada póster se muestran en la tabla 5.2.A.

Tabla 5.2.A - Número de pósteres y Tipos de Apps.

Tipo de App	Frecuencia
Calculadora de amor	3
Detector de edad	2
Detector de mentiras	4
Detector de metales	1
Detector de ronquidos	1
Examen óptico	1
Infrarrojo	1
Medidor de frecuencia cardíaca	3
Medidor de inclinación	1
Podómetro	1
Termómetro	1
Total	**19**

El alumnado tuvo la libertad de trabajar de forma individual o grupal. Se contabilizó lo siguiente: 1 grupo de cuatro participantes, 4 grupos de tres, 8 grupos de dos y 6 participantes que trabajaron de forma individual.

La secuencia aplicada del proyecto App Checkers en esta segunda implementación consistió en:

1a Etapa - Discusión inicial sobre concepto de verificación: Trabajo de dos horas en torno a la pregunta: ¿Qué significa que una App que tiene por objetivo la medición, sea falsa o engañosa? Se realizó una introducción para reflexionar sobre el significado de "verificación". Posteriormente se trabajó alrededor de sus ideas iniciales sobre "medir" y "confiabilidad", para luego explorar diferentes Apps, gratuitas y de contenido apto para menores de edad, indagaron sobre su funcionamiento y analizaron los comentarios y sus valoraciones para finalmente elaborar de un listado de Apps posibles de ser verificadas.

2a Etapa - Selección de una App y diseño de una investigación: Los participantes durante 6 horas de clase, seleccionaron una App, diseñaron y ejecutaron uno o varios experimentos para poner a prueba una App.

3a Etapa - Elaboración de un póster científico: Se dedicó 3 horas de trabajo y consistió en la elaboración de un póster científico para ser presentados en el instituto a modo de divulgación científica (Figura 5.2.A). Se les solicitó que su póster tuviera la siguiente estructura: Introducción, Hipótesis y Objetivos, Diseño experimental, Resultados y Conclusiones. Un ejemplo de póster se muestra en la Figura 5.2.B.

Figura 5.2.A - Los pósteres fueron presentados a la comunidad estudiantil del Instituto.

LOS RESULTADOS FALSOS DEL CARDIÓGRAFO

Institut Francesc Ferrer i Guàrdia, Sant Joan Despí (IES FFG)

Introducción

Actualmente existen muchas FakeApps que dicen medir una magnitud o controlar algo cuando en realidad no funcionan, lo buscan en Internet o simplemente lo hacen de manera aleatoria.
He escogido la aplicación Cardiograph (Cardiógrafo), que consiste en medir la frecuencia cardíaca. Esta app tiene más de 10 M de descargas y tiene una puntuación de 3,9. En general, las opiniones son buenas, aunque hay gente que dice que no les deja medir y que no sirve.

Hipótesis de funcionamiento y objetivos

HIPÓTESIS
Mi hipótesis es que la aplicación basa sus resultados en la manera en que detecta la yema del dedo.

OBJETIVOS
- Comprobar si la app funciona como debe funcionar o como se dice que funciona.
- Investigar los factores que influyen en la toma de medidas.

Diseño experimental

He hecho un experimento sobre si la aplicación Cardiograph es fiable o no.
Materiales - Papel y lápiz/bolígrafo
- Tensiómetro
- Móvil con la app descargada
- Cinta de correr

Para comprobarlo, he medido mis pulsaciones mientras estaba descansando, caminando y corriendo. Después de anotar los resultados, los he comparado para ver si eran los mismos o se acercaban entre sí.

Resultados

* *Calentamiento: 3 km/h. Corriendo: 6 km/h.* ppm: pulsaciones por minuto

	ACTIVIDAD		
INSTRUMENTO	DESCANSO	CALENTAMIENTO*	CORRIENDO*
CARDIOGRAPH	76ppm*	56ppm*	66ppm*
CINTA DE CORRER	81ppm*	115ppm*	132ppm*
TENSIÓMETRO	79ppm*	110ppm*	124ppm*

Conclusiones

Los resultados que he obtenido con el experimento que he realizado me han demostrado que esta app no funciona, ya que los primeros coinciden más o menos, pero los otros varían mucho y no es lógico que mi frecuencia cardíaca en reposo sea superior a la que tengo cuando estoy calentando o corriendo.
Mi hipótesis es falsa, porque los resultados no dependen de la manera en que la aplicación detecta la yema del dedo.
Esta app no tiene confiabilidad, es **falsa**.

Figura 5.2.B - Póster realizado por una estudiante participante del proyecto App Checkers en la segunda implementación.

El propósito de esta investigación es comprender cómo el grupo de alumnos diseña sus investigaciones para evaluar la confiabilidad de las aplicaciones seleccionadas, determinando cómo aplican los conceptos tanto de validez como de fiabilidad. A través del análisis de los pósteres científicos, se busca ofrecer una visión detallada del enfoque metodológico utilizado por cada estudiante, respondiendo a los siguientes objetivos:

- Identificar las estrategias que emplean los y las estudiantes para evaluar la validez de los resultados obtenidos por una App.
- Analizar hasta qué punto los y las estudiantes incorporan procedimientos de fiabilidad para evaluar la estabilidad de los resultados que ofrece cada App.

5.3 Metodología

La metodología utilizada se basa en un enfoque cualitativo de análisis interpretativo del contenido de los pósteres elaborados por los y las estudiantes al término del proyecto. El análisis se centra en las estrategias empleadas por cada estudiante en sus investigaciones sobre la confiabilidad de una App, con el fin de clasificar sus métodos de evaluación en cuanto a la validez y fiabilidad de la medida que ofrece cada App. Como instrumento de recopilación de datos, se han utilizado los pósteres elaborados por el alumnado.

5.3.1 Análisis de las estrategias para determinar la validez

Respecto a la primera pregunta de investigación, para cada póster se identificaron los experimentos realizados, ya que en un mismo póster podían existir varios. Así, el póster 1 se dividió en experimento 1.1 y 1.2, siendo cada uno de ellos, una unidad de análisis independiente, esto se muestra en la tabla 5.3.A. Aquí se incluyen todos los experimentos analizados, sus variables, número de celulares usados por el alumnado, si hay algún instrumento externo o si se considerado en el experimento algún valor esperado (valor de referencia o patrón teórico esperado).

Tabla 5.3.A - Descripción de los experimentos identificados en los pósteres científicos: unidades de análisis, variables consideradas y elementos de validez.

Póster Experimento	Tipo de Variable	Variables usadas	Número de celulares Instrumento externo Valor esperado
1.1 Medidor de frecuencia cardíaca	Numérica continua	VI: Tipo de celular o instrumento VD: frecuencia cardíaca VC: misma persona (no explícito)	3 1 (tensiómetro) No
1.2 Medidor de frecuencia cardíaca	Numérica continua	VI: Color de la superficie VD: frecuencia cardíaca VC: Ritmo de movimiento	1 No No
2.1 Medidor de frecuencia cardíaca	Numérica continua	VI: situación de actividad física VD: frecuencia cardíaca VC: misma persona (no explícito)	1 2 (tensiómetro y cinta de correr) No
3.1 Medidor de frecuencia cardíaca	Numérica continua	VI: situación de actividad física VD: frecuencia cardíaca VC: misma persona (no explícito)	2 2 (tensiómetro y pulsera inteligente) No
4.1 Termómetro	Numérica continua	VI: Instrumento (App v/s termómetro digital) VD: temperatura corporal VC: no se menciona	1 1 (termómetro digital) No
4.2 Termómetro	Numérica continua	VI: frecuencia cardiaca VD: temperatura corporal VC: no se menciona	1 1 (medidor de pulsaciones) No
5.1 Medidor de inclinación	Numérica continua	VI: Tipo de celular o instrumento VD: ángulo de inclinación VC: ángulo de inclinación de la superficie	3 1 (nivel de burbuja tradicional) No
6.1 Detector de mentiras	Dicotómica	VI: Veracidad del enunciado VD: Veredicto de la App (verdadero/falso) VC: La posición de los dedos	1 No Si
6.2 Detector de mentiras	Dicotómica	VI: Usar dos dedos distintos a los que pide la App (índice y anular) VD: Veredicto de la App (verdadero/falso); VC: la veracidad de la pregunta.	1 No Si
7.1 Detector de mentiras	Dicotómica	VI: Veracidad del enunciado VD: Veredicto de la App (verdadero/falso) VC: Afirmaciones distintas o la misma afirmación	1 No Si
8.1 Detector de mentiras	Dicotómica	VI: Veracidad del enunciado VD: Veredicto de la App (verdadero/falso) VC: Afirmaciones distintas o la misma afirmación y usar el mismo dedo índice	1 No Si
8.2 Detector de mentiras	Dicotómica	VI: Tipo de dedo (índice y meñique) VD: Veredicto de la App (verdadero/falso) VC: La misma sentencia o falsa o verdadera	1 No Si
9.1 Detector de mentiras	Dicotómica	VI: Veracidad del enunciado VD: Veredicto de la App (verdadero/falso) VC: Afirmaciones distintas o la misma afirmación	1 No Si

9.2 Detector de mentiras	Dicotómica	VI: Las varias repeticiones usando distintos dedos y codos VD: Veredicto de la App (verdadero/falso) VC: Afirmaciones distintas o la misma afirmación	1 No Si
9.3 Detector de mentiras	Dicotómica	VI: Las varias repeticiones usando distintos tonos de voz VD: Veredicto de la App (verdadero/falso) VC: Afirmaciones distintas o la misma afirmación	1 No Si
10.1 Detector de metales	Dicotómica y numérica	VI: El tipo de objeto colocado cerca del celular VD: El valor de la intensidad de campo magnético VC: no se menciona	1 No Si
10.2 Detector de metales	Dicotómica y numérica	VI: Estado del celular (wifi, bluetooth, sin conexión) VD: El valor de la intensidad de campo magnético VC: El objeto siempre el mismo	1 No No
10.3 Detector de metales	Dicotómica y numérica	VI: Distancia del celular al objeto VD: El valor de la intensidad de campo magnético VC: El objeto siempre el mismo	1 No No
11.1 Detector de ronquidos	Dicotómica y numérica	VI: las distintas personas VD: La respuesta proporcionada por la App VC: No se indica	1 No No
12.1 Detector de edad	Numérica discreta	VI: Las diferentes fotografías seleccionadas VD: la predicción de la App de la edad VC: No se indica	1 No Si
13.1 Detector de edad	Numérica discreta	VI: Las diferentes fotografías seleccionadas VD: la predicción de la App de la edad VC: No se indica	1 No Si
14.1 Podómetro	Numérica discreta	VI: La cantidad de pasos realizados por la persona (10, 20 y 100 pasos) VD: la predicción de la App de los pasos dados VC: No se indica	1 No Si
15.1 Calculadora de amor	Numérica continua	VI: La situación sentimental de las parejas, amigos o famosos VD: El porcentaje de compatibilidad o amor generado por la aplicación VC: No se indica	2 No Si
16.1 Calculadora de amor	Numérica continua	VI: La situación sentimental de parejas de ficción casadas o con relación romántica VD: El porcentaje de compatibilidad o amor generado por la aplicación VC: No se indica	1 No Si
17.1 Calculadora de amor	Numérica continua	VI: La variación en la posición de los nombres de las personas ingresadas en la aplicación (hombre-mujer, mujer-hombre) VD: El porcentaje de compatibilidad o amor generado por la aplicación VC: No se indica	1 No No
17.2 Calculadora de amor	Numérica continua	VI: La introducción de símbolos y emoticonos como nombres alternativos en la aplicación	1 No

		VD: El porcentaje de compatibilidad o amor generado por la aplicación VC: No se indica	No
18.1 Detector de radiación	Numérica continua y categórica ordinal	VI: La ubicación del dispositivo con la aplicación, colocándolo junto al router wifi y el televisor VD: La medición de la cantidad de radiación en la zona, representada en la escala analógica del 0 al 100 proporcionada por la App en μT VC: No se indica	1 1 (router wifi) No
18.2 Detector de radiación	Numérica continua y categórica ordinal	VI: alimentos a distintas temperaturas VD: La variación en las tonalidades de la imagen de la App VC: No se indica	1 No Si
19.1 Examen óptico	Numérica continua	VI: El número de veces que se realiza el examen una persona VD: Los resultados del examen de agudeza visual proporcionados por la App VC: No se indica	1 No Si

De los 19 pósteres se obtuvo un total de 29 unidades de análisis. Seguidamente, cada una de estas 29 unidades de análisis se caracterizó en un conjunto de ítems, cuyo propósito final fue determinar las estrategias de validez usadas por el alumnado. Los ítems se presentan a continuación:

- Tipo de variable: se refiere a que "magnitud" mide (o dice medir) la App (categórica ordinal o nominal, dicotómica, numérica discreta, numérica continua o algún tipo de combinación de ellas).
- Experimento: Se refiere al experimento realizado por el alumnado. Se identificaron las variables independientes, dependientes y controladas.
- Número de celulares: se refiere a los usados en el experimento, ya que algunos experimentos se hacían instalando la App en un solo dispositivo, y otros experimentos hechos por el alumnado incluían resultados de la misma App en distintos celulares con el objetivo de compararlos posteriormente.
- Instrumento externo: La existencia de comparación de los datos obtenidos con otros datos externos obtenidos de otro dispositivo experimental (por ejemplo, comparar una App termómetro con un termómetro digital).
- Valor esperado: La existencia de comparación de los datos obtenidos con otros datos externos conocidos, por ejemplo, comparar una App de detección de edad con la edad real conocida de la persona en la imagen.

Usando la tabla 5.3. A se identificaron 4 tipos de estrategias. La primera estrategia, la llamamos "Comparación con un valor esperado", consiste en comparar el resultado de la App

con valores esperados (conocidos con antelación) en un contexto específico. En términos teóricos se consideró equivalente a determinar la validez en un experimento usando "datos secundarios" (Gott y Duggan, 2003). Las Apps que promovieron esta estrategia tienen en común que ofrecen mediciones, predicciones o evaluaciones cuantificables que pueden ser comparadas con valores conocidos o esperados. Los tipos de variables que corresponden a esta estrategia son de todo tipo, numéricas continuas, numéricas discretas y dicotómicas.

La segunda estrategia, "Comparación instrumental", consiste en comparar el resultado de la App con instrumentos de medición independientes. En términos teóricos es también similar a determinar la validez en un experimento usando "datos secundarios" (Gott y Duggan, 2003), pero provenientes de un instrumento de medida. Es también similar al proceso de "verificación" que define la metrología. Es importante considerar que los resultados obtenidos desde esta estrategia pueden variar según el tipo de fenómeno medido y la calidad de los instrumentos de medición utilizados como referencia.

La tercera estrategia, que hemos llamado "triangulación", consiste en comparar el resultado de la App instalada en diferentes dispositivos, la cual se ha considerado equivalente a determinar la validez en un experimento usando una "triangulación" (Gott y Duggan, 2003). Al usar distintos celulares, el estudiantado establece la validez en función de qué tan similares sean las mediciones de cada fenómeno cuando comparan las lecturas entre los dispositivos. En la tabla 1, si hay sólo un celular, no se ha realizado una triangulación, si hay dos o más se ha hecho una triangulación. Esta estrategia permite evaluar la precisión de las mediciones realizadas por la App, identificando posibles variaciones debido a diferencias en los sensores u otras características del celular. Si los resultados obtenidos por la App varían entre dispositivos para la medición del mismo fenómeno los y las estudiantes establecen que la confiabilidad de la App es menor.

La cuarta estrategia, "Poner a prueba la App", consiste precisamente en poner a prueba la App con algún montaje experimental que busca fallas o determinar la consistencia en el funcionamiento de la App. Así, quienes usaron esta estrategia evaluaron el desempeño de la App en situaciones específicas o bajo diferentes condiciones. En términos teóricos es equivalente a "serie de experimentos" (Gott y Duggan, 2003), donde se realizan distintos controles de variables para poner a prueba la App.

5.3.2 Análisis de los niveles de fiabilidad

En cada uno de los experimentos de la tabla 5.3.A, se tuvieron en cuenta diferentes ítems relacionados con los procedimientos de medida que permiten determinar la fiabilidad de un instrumento. Estos se presentan en la tabla 5.3.B y se describen a continuación:

- Número de medidas: Se refiere al número de fenómenos o situaciones en que se realiza la medición con la App. Por ejemplo, en la tabla 5.3.A, fila "1.1 Medidor de frecuencia cardíaca", tenemos 1 fenómeno medido por el estudiantado que trabajó con la App medidor de frecuencia cardiaca. Además, se aprecia que instalaron la App y midieron con ella en 3 celulares diferentes, junto con un tensiómetro. Luego, al determinar las repeticiones desde el póster correspondiente los estudiantes tomaron 1 medida que fue repetida 3 veces con cada celular y tensiómetro, como se ve en la figura 5.3.A.
- Número de repeticiones en cada medición: El número de lecturas (repeticiones) en cada fenómeno. Consideramos lecturas en las mismas condiciones y una medición correspondiente a un valor (el mismo fenómeno).
- Medida de centralización: Corresponde a determinar el uso de alguna medida de tendencia central. En la figura 5.3.A izquierda, el grupo ha calculado la media de los tres fenómenos medidos.
- Medida de dispersión: Corresponde al uso de alguna medida de dispersión. Por ejemplo, en la figura 5.3.A derecha, el grupo cuantifica la dispersión promediando diferencias positivas (en valor absoluto) acercándose a un concepto cuantitativo de varianza o desviación estándar.

El desarrollo del segundo objetivo permitió identificó 5 categorías basadas en el procedimiento estándar de la metrología para determinar la fiabilidad de un medidor. Los límites de cada categoría se generaron mediante la adicción de alguna acción en el procedimiento de fiabilidad que fue contrastado con la tabla 5.3.B. Por ejemplo, para cuantificar la estabilidad, primero se debe calcular una medida de centralización y luego determinar cuánto se aleja, en promedio, cada lectura respecto de ella, equivalente a una dispersión alrededor de la medida central. Los niveles alcanzados están definidos en la tabla 5.3.C.

Tabla 5.3.B - Ítems relacionados con los procedimientos de fiabilidad para cada unidad de análisis de la tabla 5.3.A.

Póster	Número de medidas	Número de repeticiones	Medida central Medida de dispersión
1.1	1	3	Promedio No
1.2	No se indica	No se indica	Ninguna
2.1	3 (niveles de intensidad de actividad física)	1	Ninguna
3.1	6 (para la misma persona, 2 medidas, un descanso y otra con actividad)	1	Ninguna
4.1	6	1	Ninguna
4.2	6	1	Ninguna
5.1	9	1	Ninguna
6.1	4	10	Tasa de acierto No
6.2	3 (para tres preguntas personales distintas)	1	Ninguna
7.1	3 (afirmaciones distintas)	15	Tasa de acierto No
8.1	4 (para una sentencia cierta y otra falsa)	20	Tasa de acierto No
8.2	2 (una con sentencia cierta otra con una verdadera)	20	Ninguna
9.1	No se indica	No se indica	Tasa de acierto No
9.2	No se indica	No se indica	Tasa de acierto No
9.3	No se indica	No se indica	Tasa de acierto No
10.1	10 mediciones	1	Ninguna
10.2	3 mediciones	1	Ninguna
10.3	6 (3 para un objeto y 3 para otro)	1	Ninguna
11.1	3 (para tres personas distintas, en distintos lugares físicos mientras duermen)	1	Ninguna
12.1	7	1	Tasa de acierto No
13.1	18	1	Tasa de acierto Promedio de diferencias absolutas
14.1	3 (persona y cantidad de pasos dados)	1	Ninguna
15.1	4 (1 por personas famosas) 3 (1 por personas no famosas)	10 10	Ninguna
16.1	6	1	Ninguna
17.1	7	1	Ninguna
17.2	4	1	Ninguna
18.1	2	1	Ninguna
18.2	2	1	Ninguna
19.1	3	1	Ninguna

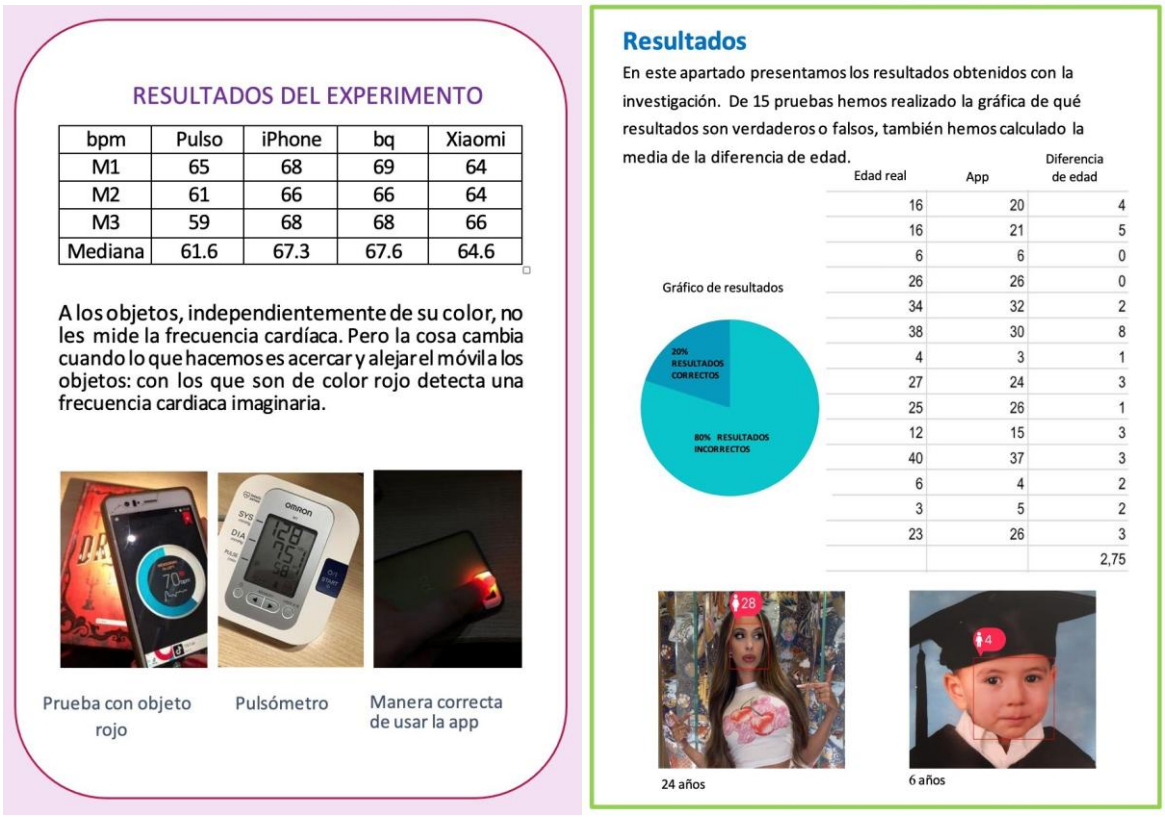

Figura 5.3.A - Resultados de los pósteres de grupos que trabajaron con la App: "Medidor de frecuencia cardíaca" (Imagen izquierda) y "Detector de edad" (Imagen derecha).

Tabla 5.3.C - Niveles de fiabilidad.

Nivel	Definición
Nivel 0	**Ausencia de análisis de fiabilidad:** No se realiza ningún estudio de fiabilidad. Puede que no se realicen mediciones repetidas o que solo se analice la validez de la aplicación sin considerar su estabilidad.
Nivel 1	**Repetición de varias mediciones:** Este es un nivel mecánico, el o la estudiante mide muchas veces porque sabe que debe hacerlo, pero no menciona nada respecto a la estabilidad del conjunto de mediciones y no realiza ningún tipo de cálculo para estudiar la dispersión del conjunto de datos.
Nivel 2	**Análisis cualitativo de la fiabilidad:** Se realizan mediciones repetidas y un análisis cualitativo de la estabilidad del conjunto de datos medido.
Nivel 3	**Medida de centralización sin cuantificación de la dispersión:** El o la estudiante mide muchas veces y realiza un análisis cualitativo de la estabilidad del conjunto de datos medido. Además, del cálculo de una medida de centralización del conjunto de datos, pero no la usa para cuantificar algún parámetro de dispersión alrededor de esa medida de centralización.
Nivel 4	**Cuantificación de la dispersión y uso para evaluar la fiabilidad:** Se realizan mediciones repetidas, se calcula una medida de centralización del conjunto de datos y se cuantifica algún parámetro de dispersión alrededor

de esta medida. Se utiliza este parámetro de dispersión para evaluar la estabilidad de las mediciones de la aplicación.

5.4 Resultados del estudio 2

En la tabla 5.4.A, desde las columnas dos a la cuatro, se presentan los resultados obtenidos para las estrategias de validez usadas y el nivel de fiabilidad que han logrado. Al usar las categorías de la tabla 3, se obtuvieron los resultados de la columna "Nivel de fiabilidad" de la tabla 5.6.

Tabla 5.4.A - Resultados estrategias de validez y niveles de fiabilidad.

Póster - App	Comparación valor esperado	Comparación instrumental	Triangulación	Poner a prueba la App	Nivel de fiabilidad
1. Medidor de frecuencia cardíaca		x	x	x	3
2. Medidor de frecuencia cardíaca		x		x	1
3. Medidor de frecuencia cardíaca		x	x		1
4. Termómetro		x		x	1
5. Medidor de inclinación		x	x		0
6. Detector de mentiras	x			x	2
7. Detector de mentiras	x				2
8. Detector de mentiras	x			x	2
9. Detector de mentiras	x			x	1
10. Detector de metales	x			x	0
11. Detector de ronquidos				x	0
12. Detector de edad	x				2
13. Detector de edad	x				4
14. Podómetro	x				2
15. Calculadora de amor	x		x		1
16. Calculadora de amor	x				0
17. Calculadora de amor				x	0
18. Detector de radiación	x	x			0
19. Examen óptico	x				0

5.4.1 Estrategias de validez

En el caso de las estrategias de validez usadas, cada App tendrá más o menos estrategias realizables dependiendo de su espacio de resultados, pero eso no ha sido controlado pues se quería que el alumnado trabajase libremente eligiendo una App de su interés personal.

La frecuencia de cada estrategia de validez es: “Comparación con un valor esperado” 12 veces, “Poner a prueba la App” 10 veces, “Comparación instrumental” 6 veces y “Triangulación” 4 veces. La comparación con un valor esperado es la más utilizada, mientras que la triangulación es la menos frecuente. No hay ninguna App que promueva un mayor uso de estrategias de validez; esto depende de las habilidades de indagación y de cómo plantean sus experimentos. La baja frecuencia de la triangulación puede deberse a varias razones, tales como la complejidad de realizar mediciones comparativas en diferentes entornos y el mayor tiempo y cuidado que éstas requieren.

En la estrategia de validez comparación con un valor esperado, como la elección de la App es libre, el estudiantado parece preferir este contexto por ser menos demandante o más natural, en lugar de comparar instrumentalmente, realizar experimentos o triangulaciones entre dispositivos. Un ejemplo es la App detector de edad (Póster 12.1 y 13.1, tabla 5.3.A), donde se comparan las predicciones de edad de la App con la edad real de las personas, que corresponde a un valor esperado. Esta comparación permite evaluar la precisión y exactitud de la App al establecer la edad de la persona en la imagen ingresada.

La estrategia “Comparación instrumental” se aplicó en Apps como medidor de frecuencia cardiaca (3 grupos), termómetro (1 grupo), medidor de inclinación (1 grupo) y detector de radiación (1 grupo). En la mayoría de estos casos, se incluyen más formas de validez, lo que sugiere que tienen mayores habilidades de indagación, dado que esta estrategia exige más habilidades que una comparación con valores esperados. Las aplicaciones que promueven esta estrategia presentan variables numéricas continuas propias del contexto científico que representan, utilizadas para medir magnitudes como temperatura, frecuencia cardíaca, inclinación o radiación. Un ejemplo es la App termómetro (Póster 4.1, tabla 5.3.A), donde se comparan las lecturas de temperatura proporcionadas por la App con las obtenidas por un termómetro digital.

La estrategia “Comparación entre diferentes dispositivos, (triangulación)” es utilizada en las Apps Calculadora de amor (1 grupo), Medidor de frecuencia cardíaca (2 grupos) y Medidor de inclinación (1 grupo). Las Apps que trabajaron los grupos que usaron esta estrategia corresponden a variables numéricas continuas. Un ejemplo es la App calculadora de amor (Póster 15.1, ver tabla 5.3.A): Se instala la App en diferentes dispositivos y se ingresan los mismos datos sobre las situaciones sentimentales de las parejas o individuos. Luego, se comparan los resultados de compatibilidad en cada dispositivo para evaluar su consistencia entre ellas y con el valor considerado verdadero por el grupo.

La estrategia “Poner a prueba la App” se usó en Apps de distintos tipos y contextos, ya sea científicos o sociales. Permite evaluar la App ante diversas condiciones experimentales o escenarios simulados, identificando limitaciones o sesgos en su funcionamiento, lo que mejora su validez. Las variables asociadas a esta estrategia son diversas, incluyendo variables numéricas continuas, dicotómicas o categóricas, dependiendo de los parámetros analizados en cada caso. Un ejemplo es la App detector de metales (Póster 10.2 y 10.3, tabla 5.3.A). En los experimentos, se realizaron pruebas variando las distancias entre el teléfono y los objetos, desde proximidades hasta distancias mayores, observando cómo cambiaba la detección según la proximidad. Además, se evaluaron los efectos de usar o no wifi y bluetooth durante las mediciones, comparando los resultados obtenidos con cada método de conexión y su impacto en la precisión y eficacia de la aplicación. Estos experimentos ayudaron a entender cómo influyen la distancia y los diferentes tipos de conexión en la capacidad del dispositivo para detectar objetos metálicos, eléctricos y de otros materiales.

5.4.2 Niveles de Fiabilidad

Al analizar los niveles de fiabilidad, se observa que la mayoría de los pósteres (n = 7) se encuentran en el nivel de fiabilidad inicial, donde no se realiza un análisis de la estabilidad de las mediciones (nivel 0). Le siguen varios pósteres (n=5) que muestran un nivel básico de análisis repetido, pero sin profundizar en la estabilidad de los datos (nivel 1). Asimismo, se encuentran algunos pósteres (n =5) que han realizado un análisis cualitativo de la estabilidad de los datos medidos (nivel 2). Por otro lado, se identifica una menor presencia de pósteres (n=1) que han resumido los datos en una medida de centralización (nivel 3) o han realizado una evaluación más detallada de la fiabilidad (n =1) mediante un análisis de dispersión (nivel 4), siendo estos últimos dos casos los menos frecuentes.

Los niveles alcanzados reflejan la diversidad en la comprensión y aplicación del concepto de fiabilidad en la medición de datos. Los niveles más bajos podrían atribuirse a la falta de conocimiento técnico sobre métodos de análisis de datos, la ausencia de herramientas adecuadas para la medición precisa o simplemente la falta de conocimiento sobre la importancia de la fiabilidad en la investigación científica. Las consecuencias de estos niveles pueden afectar la validez de los resultados obtenidos. Sin un análisis adecuado de la fiabilidad, se limita la utilidad y aplicabilidad de los hallazgos en contextos reales.

Como ejemplo de cada nivel se tiene:

- Para el nivel 0, se presentó un póster sobre la App "detector de ronquidos", cuya evaluación de fiabilidad plantea desafíos debido a la complejidad de los experimentos necesarios. Determinar si una persona ronca requeriría vigilar su sueño durante muchas noches, lo cual resulta costoso y laborioso. En lugar de realizar este procedimiento completo, el grupo dejó la App activada durante tres noches para una persona. Aunque el experimento sigue una estrategia de validez, no se abordó la fiabilidad del detector de ronquidos. Es decir, no se llevaron a cabo mediciones repetidas ni se realizó un análisis de la estabilidad de las mediciones proporcionadas
- En el Nivel 1 se realiza varias lecturas sin un objetivo claro, lo que genera un enfoque mecánico. En un póster sobre la App "Calculadora de amor", se han realizado múltiples mediciones del mismo fenómeno, pero las conclusiones carecen de referencias a la centralización o dispersión de los datos.
- En el Nivel 2, se realiza una interpretación cualitativa de las mediciones repetidas de un fenómeno, sin embargo, estas no han sido cuantificadas. Un ejemplo de esto es el póster de la App "Detector de mentiras", el grupo midió muchas veces, al ver dispersión en los datos mencionan en las conclusiones “La App es poco precisa”, es decir, se realizan mediciones repetidas y se realiza un análisis cualitativo de la estabilidad del conjunto de datos medidos, comparando sin un cálculo explicito, las diferencias entre las lecturas.
- En el Nivel 3, los y las estudiantes determinan una medida de centralización sin cuantificar la dispersión de los datos. Por ejemplo, en el póster de la App "Medidor de frecuencia cardiaca", realizan múltiples mediciones y calculan promedios de los fenómenos medidos. Con base en estos promedios, concluyen que la App es poco precisa al comparar explícitamente las diferencias entre ellos. Sin embargo, no realizan un cálculo similar a la varianza muestral para evaluar la dispersión de los datos.
- En el Nivel 4 los y las estudiantes cuantifican la dispersión y usan este valor para evaluar la fiabilidad. Como único ejemplo del Nivel 4 se tiene el póster 13.1 de la tabla 5.3.B y figura 5.3A derecha. En este caso, si bien no obtuvieron una media, si usaron las diferencias positivas respecto al valor considerado verdadero (Error absoluto) y promediaron estos valores. Fueron el grupo que más se acercó al concepto de varianza. Luego usaron en sus conclusiones este valor: “Según los

resultados obtenidos de los experimentos hemos observado. que: En las fotos de los famosos y personas jóvenes, no daba resultados exactos, pero sí resultados aproximados. En la gente mayor, los resultados tampoco eran exactos y la diferencia de edad era muy ancha 2,75". En este párrafo usan en las conclusiones la palabra "ancha" refiriéndose a la dispersión encontrada en su análisis cuantitativo.

5.5 Conclusiones del estudio 2

La investigación muestra que el proyecto ofrece varias oportunidades para abordar conceptos científicos y epistémicos. Este estudio, realizado en bachillerato, contrasta con investigaciones similares realizadas en la educación secundaria (López-Simó, 2021), demostrando la adaptabilidad y profundidad de la misma actividad en diferentes contextos educativos.

En los 19 pósteres analizados, prevalece la estrategia de comparación con un valor esperado como método principal para evaluar la validez de las aplicaciones. Esta tendencia podría explicarse por la preferencia del alumnado por un enfoque más natural sobre uno más práctico o instrumental, es decir que implique habilidades de indagación relacionadas con técnicas de medición con instrumentos.

Respecto a la fiabilidad, la mayoría de los pósteres no cumplen con el procedimiento descrito por Gott y Duggan (2003) y la metrología (Centro Español de Metrología, 2012) que permite determinar la estabilidad de una medición. En general se aprecia que el alumnado no analiza la estabilidad de una medición, pues esto requiere tomar varias lecturas de una misma medición y luego calcular la media y estudiar la dispersión alrededor de ella. Esto implica que quizá no tengan suficientemente integradas ciertas habilidades de indagación con el pensamiento estadístico, existiendo en ellos poca aplicabilidad en el cómo resumir datos y manejo del concepto de incertidumbre de las mediciones.

No obstante, es importante tener en cuenta que este estudio presenta una muestra reducida. Los resultados obtenidos no son necesariamente extrapolables a otras poblaciones o contextos educativos. El objetivo principal del análisis se centra en explorar los tipos de estrategias utilizadas para evaluar la confiabilidad de las aplicaciones. Esta aproximación cualitativa permite una comprensión más profunda y detallada de las prácticas de evaluación, destacando la validez sobre la fiabilidad.

El estudio resalta también la importancia del andamiaje en el diseño de investigaciones escolares. La elección de investigaciones como la evaluación de Apps, proporciona al estudiantado una experiencia significativa en el desarrollo de habilidades científicas y epistémicas. Estas investigaciones van más allá de las prácticas demostrativas tradicionales, permitiendo al alumnado involucrarse más activamente en la indagación científica y la evaluación crítica inicial de tecnologías emergentes.

Los datos analizados en esta investigación muestran que no existe algún tipo de correlación o agrupación privilegiada entre las estrategias de validez y los niveles de fiabilidad, por ello se propone promover la enseñanza integral de ambos conceptos ya que son interdependientes, siendo el objetivo final permitir que los y las estudiantes adquieran una comprensión integral de la confiabilidad en indagación científica. Creemos que esto se puede lograr integrando diversas metodologías de enseñanza, como la enseñanza directa de métodos de análisis estadístico para el tratamiento de datos, la retroalimentación constructiva y la práctica en experiencias como App Checkers.

CAPÍTULO VI

ESTUDIO 3: ANÁLISIS DE LOS MODELOS MENTALES EXPRESADOS POR EL ALUMNADO DE ENSEÑANZA MEDIA SOBRE EL FUNCIONAMIENTO DE APLICACIONES MÓVILES

Este capítulo describe el Estudio 3, desarrollado en colegios de Santiago de Chile con estudiantes de 13 a 17 años. El objetivo fue analizar los modelos mentales que construyen sobre el funcionamiento de distintas Apps.

Se utilizó una rúbrica basada en Louca et al. (2011) para evaluar objetos, entidades, comportamientos e interacciones en los modelos. Muchos alumnos representaron las Apps desde el uso cotidiano, sin visualizar los procesos internos.

6.1 Introducción al estudio 3

En continuidad con los estudios anteriores, el Estudio 3 se propone investigar una dimensión que hasta ahora había permanecido parcialmente explorada en el marco del proyecto *App Checkers*, la representación mental que el alumnado construye sobre el funcionamiento interno de las aplicaciones móviles. Mientras que el Estudio 1 se centró en el análisis del desempeño en indagación y argumentación a partir de evidencias empíricas obtenidas mediante experimentación, y el Estudio 2 abordó las estrategias utilizadas por estudiantes de bachillerato para evaluar la confiabilidad de las Apps desde una perspectiva metrológica, este tercer estudio desplaza el foco hacia la modelización como práctica científica, con énfasis en la comprensión funcional y estructural de sistemas tecnológicos interdisciplinarios.

Una primera motivación para este estudio radica en la escasa presencia en la literatura educativa de investigaciones centradas en la modelización de tecnologías digitales por parte del alumnado, especialmente desde un enfoque interdisciplinario que integre nociones de física, informática, matemáticas y razonamiento epistémico. A diferencia de los modelos clásicos sobre fenómenos naturales (por ejemplo, circuitos eléctricos o cambio de estado), los modelos de funcionamiento de Apps requieren una articulación entre lo observable y lo algorítmico, entre la interfaz visible y los procesos técnicos subyacentes. Este tipo de modelos mentales plantea desafíos específicos, tanto conceptuales como epistémicos, que merecen ser investigados sistemáticamente.

En segundo lugar, se considera que el análisis de estos modelos puede ofrecer claves para comprender mejor el componente de "razonamiento" presente en las argumentaciones del alumnado. Tal como sugiere el modelo adaptado de Toulmin (véase Figura 2.3.B), las pruebas empíricas obtenidas mediante indagación adquieren sentido argumentativo solo en la medida en que se conectan con un conocimiento básico que permita interpretar, justificar y modular una conclusión. Desde esta perspectiva, las ideas iniciales del alumnado respecto al funcionamiento de las Apps constituyen el sustrato epistémico desde el cual emergen sus justificaciones, sean estas correctas o no. Explorar cómo se estructuran estos modelos mentales —qué elementos incluyen, cuáles omiten y cómo los conectan— tiene potencial para comprender no solo su pensamiento funcional, sino también las limitaciones y potencialidades de su argumentación científica.

En consecuencia, este estudio se propone caracterizar los modelos mentales construidos por el alumnado participante de la tercera implementación en torno al funcionamiento de Apps, analizar su nivel de elaboración en relación con una rúbrica basada principalmente en los trabajos de Louca et al. (2011a, 2011b) y explorar la relación entre el tipo de App, la edad del alumnado y el nivel de elaboración del modelo construido. Este análisis permitirá profundizar en la comprensión del pensamiento del alumnado en contextos digitales, aportando nuevas herramientas para la enseñanza de la ciencia y la tecnología desde una perspectiva más fundamentada.

6.2 Contexto y objetivos de investigación

Este tercer estudio tiene como enfoque profundizar en la comprensión de las ideas el alumnado sobre los modelos mentales que tiene al alumnado acerca del funcionamiento de las Apps. Se desarrolló una plantilla que sirvió como andamiaje y como herramienta de recolección de datos, con la consideración que tuviera la suficiente flexibilidad para que los estudiantes expresaran sus propios enfoques y razonamientos. El análisis de las plantillas completadas por el alumnado participante del proyecto permitió acercarnos más a su razonamiento, ofreciendo más claridad de cómo percibían y manejaban diversas cuestiones relacionadas con las practicas científicas en el contexto de verificación de la confiabilidad de Apps. Una novedad respecto a las pasadas implementaciones fue la de solicitar al alumnado un modelo de funcionamiento para la App seleccionada, con la intención de conocer más sobre sus ideas previas, así como también estudiar el nivel de sofisticación que manejan en cuanto a modelos tecnológicos interdisciplinarios.

La implementación de App Checkers fue realizada en distintas escuelas y cursos, en Santiago de Chile, durante el año académico 2023, en los meses de marzo, abril y mayo:

- Colegio Chillán, La Florida, Región Metropolitana. Estudiantes de 8° año de enseñanza básica, edad promedio 13 años. Muestra de 10 grupos de trabajo. (https://www.cchillan.com/)
- Colegio Francisco Miranda, Peñalolén, Región Metropolitana. Estudiantes de 2° año de enseñanza media, edad promedio 15 años. Muestra de 19 grupos de trabajo. https://franciscodemiranda.cl/web/

- Liceo Bicentenario Hermanos Sotomayor Baeza, Melipilla, Región Metropolitana. Estudiantes de 1° año de enseñanza media, edad promedio 14 años. Muestra de 2 grupos de trabajo. https://www.facebook.com/profile.php?id=100037757174487
- Colegio Monte Verde, Peñalolén, Región Metropolitana. Estudiantes de 4° año de enseñanza media, edad promedio 17 años. Muestra de 17 grupos de trabajo. https://web.colegiomonteverde.cl/

En total se contabilizaron 50 grupos de trabajo. La tabla 6.2.A resumen la cantidad de integrantes por cada grupo y sus edades. En la tabla se ha usado la codificación Colegio 1 para el colegio Chillán, Colegio 2 para el colegio Francisco Miranda, Colegio 3 para el liceo Bicentenario Hermanos Sotomayor Baeza y Colegio 4 para el colegio Monte Verde.

Tabla 6.2.A - Tabla resumen con los participantes del estudio 3.

Colegio - edad	**Nº de integrantes**				**Total de grupos**
	1	**2**	**3**	**4**	
Colegio 1 - 13 años	-	4	2	4	10
Colegio 2 - 14 años	2	8	-	-	10
Colegio 2 - 15 años	-	9	-	-	9
Colegio 3 - 15 años	1	1	-	-	4
Colegio 4 - 17 años	3	11	1	2	17

El proyecto siguió etapas similares al estudio 1 y 2, pero con los cambios siguientes:

1a Etapa - Discusión inicial sobre concepto de confiabilidad: Trabajo de una hora en torno a la pregunta: ¿Qué significa que una App que tiene por objetivo la medición, sea confiable? Se realizó una introducción donde los estudiantes reflexionaron sobre el significado de "verificación". Pero a diferencia de aplicaciones anteriores (Estudio 1 y Estudio 2) se abordaron las ideas iniciales del alumnado sobre fiabilidad y validez de un medidor ejemplificando con una App que no fue usada en el trabajo posterior. Finalmente, los estudiantes exploraron específicamente las App dadas en la plantilla y que se muestran en la figura 6.2.A.

2ª Etapa - Selección de una App: Trabajo de una hora, en el cual se ejecutó la actividad 2 de la plantilla, y que se muestra en la figura 6.2.B. Lo primero es que el alumnado decidiera con que App de las exploradas en la actividad 1 trabajar. Hecho esto se pregunta: ¿Qué App han elegido?, ¿Que mide? y ¿Por qué han elegido esta App? Esto busca explorar sus gustos e ideas iniciales. Luego, se solicita la construcción de un modelo de funcionamiento para la App,

donde el andamio pretende que se plantee como un proceso de entrada, procesamiento y salida de datos.

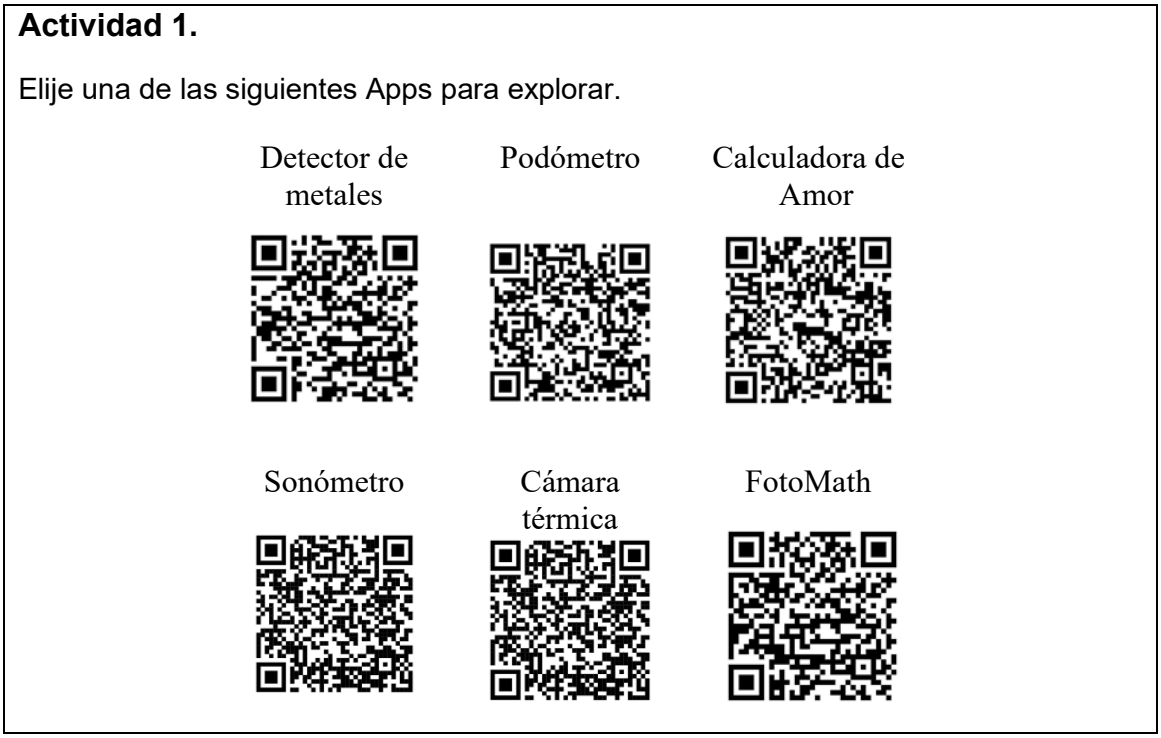

Figura 6.2.A - Actividad introductoria.

3ª Etapa – Proceso de indagación guiado: En la Actividad 3 (figura 6.2.C) el estudiantado diseñó y llevó a cabo un(os) experimento(s) para evaluar la confiabilidad de la App seleccionada, a esto se le dio un tiempo de dos horas. Como punto de partida, se les pidió formular una hipótesis acerca de la App. Luego, para poner a prueba esta hipótesis, la planificación de uno o varios experimentos y recoger resultados. Finalmente obtener conclusiones acerca del experimento sobre la hipótesis y modelo de funcionamiento planteado en la actividad 2.

Actividad 2.

Nombre App elegida	
¿Qué mide?	
¿Por qué han elegido esta App?	

Se puede pensar la App como un sistema que recibe información del exterior, la procesa y nos da información a través del móvil. Con la App que instalaron, completen el siguiente cuadro.

Datos de entrada del exterior ¡Entramos a la App!	¿Qué creen que está pasando adentro del celular con los datos de entrada? **Plantear un modelo de funcionamiento**	**Información Procesada Datos de salida**

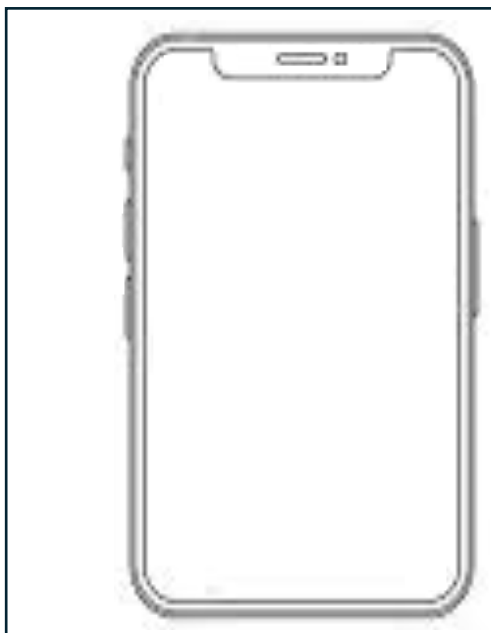

Figura 6.2.B - Actividad 2, en la cual se exploran las preferencias del alumnado y el modelo de funcionamiento de la App seleccionada.

Actividad 3.

Planteen una hipótesis sobre la App elegida	
Diseño del experimento Diseñen **uno o varios** experimentos para "poner a prueba" la App elegida. Explique el experimento paso a paso, pensando en que otra persona lo pudiera repetir leyendo su escrito.	
Datos obtenidos	
Conclusiones Los datos obtenidos a través de tus experimentos corroboran tu modelo de funcionamiento de la App ¿cuál sería el nuevo modelo?	

Figura 6.2.C - Actividad 3, en la cual se realiza un proceso de indagación guiado.

La implementación del estudio 3 presenta diferencias con las aplicaciones de los estudios 1 y 2 (Aguilera y López-Simó, 2025a; Aguilera y López-Simó, 2025b), tanto en su estructura metodológica como en el nivel de orientación proporcionado al estudiantado. Una de las diferencias más evidentes radica en el uso de una plantilla guía en el estudio 3, lo que facilitó un andamiaje más estructurado para la exploración y análisis de las Apps. En contraste, los estudios 1 y 2 fueron desarrollados sin el uso de una plantilla, permitiendo un enfoque más abierto y exploratorio en la indagación.

Otra diferencia se encuentra en la forma en que se introduce el concepto de confiabilidad. Mientras que en el estudio 1 la discusión inicial se centró en la noción de verificación y en la identificación de aplicaciones falsas o engañosas, en el estudio 3 se abordó el concepto de fiabilidad y validez a través de ejemplos específicos, incluyendo el análisis de una aplicación que no sería utilizada posteriormente. Este cambio de enfoque permitió que el estudiantado en el estudio 3 trabajara desde una perspectiva más centrada en la App y su funcionamiento, en lugar de enfocarse exclusivamente en la verificación científica de engaños.

La selección de Apps también varió entre los estudios. En el estudio 1, el estudiantado tuvo una exploración más libre de aplicaciones gratuitas, en cambio ahora, se propuso una lista cerrada de aplicaciones dentro de la plantilla, lo que limitó las opciones disponibles, esto para permitir un análisis más focalizado y comparable.

En el estudio 3, el estudiantado construyó modelos de funcionamiento de las Apps seleccionadas a través de una estructura guiada en la plantilla, identificando los datos de entrada, el procesamiento interno y la información de salida. En los estudios 1 y 2, en cambio, se promovió la selección libre de Apps y no se guió con una plantilla las fases de indagación, todo lo anterior sin la necesidad de plasmar explícitamente algún modelo de funcionamiento previo.

En el estudio 1, el estudiantado elaboró un video de presentación para comunicar sus hallazgos, lo que requirió una síntesis de la información y el desarrollo de habilidades de comunicación. En el estudio 2, el proceso culminó con la elaboración de un póster científico en el que se presentaron los resultados de la investigación, permitiendo organizar la información de manera estructurada y visualmente accesible, promoviendo el desarrollo de habilidades en la representación gráfica y argumentación escrita. En el estudio 3, el estudiantado construyó modelos de funcionamiento de las Apps seleccionadas a través de una estructura guiada en la plantilla, identificando los datos de entrada, el procesamiento interno y la información de salida.

Para este estudio 3 se ha planteado como objetivo de investigación los siguientes:

- Evaluar los modelos mentales que usan los estudiantes entre 13 y 17 años en Chile para explicar el funcionamiento de las Apps cuando tratan de verificar su confiabilidad.
- Analizar el nivel de elaboración de los modelos expresados en función de dos variables didácticas: tipo de App y edad.

6.3 Metodología

En este estudio se adopta una metodología de carácter cualitativo con apoyo de análisis cuantitativo, centrada en la comprensión profunda de los modelos mentales construidos por el alumnado en torno al funcionamiento de diversas Apps. El enfoque metodológico se inscribe en el paradigma interpretativo, ya que busca indagar cómo los estudiantes representan los componentes internos de un sistema tecnológico a partir de sus experiencias y conocimientos previos, así como identificar patrones de omisión o representación parcial que revelan barreras

conceptuales o de compresión en el aprendizaje. La unidad de análisis serán las plantillas completadas por el alumnado (Tabla 3.A), las cuales contienen representaciones escritas y gráficas sobre el funcionamiento de la App seleccionada, permitiendo una aproximación al pensamiento del alumnado.

Con base en estas plantillas, se diseñó una rúbrica analítica estructurada en cuatro componentes clave (objetos, entidades, comportamientos e interacciones), cuyos niveles de desempeño fueron definidos a partir de marcos teóricos previos Louca et al. (2011a, 2011b) y ajustados mediante un proceso iterativo de codificación y análisis de combinaciones. Esta rúbrica permitió categorizar los niveles de elaboración de los modelos propuestos por los estudiantes, desde representaciones ausentes o incorrectas hasta descripciones funcionales y técnicamente precisas. A continuación, se detalla el procedimiento seguido para la recogida de datos y la conformación del corpus analizado.

6.3.1 Recogida de datos

En la implementación del estudio 3 del proyecto App Checkers, los datos fueron recopilados a partir de las actividades realizadas en la plantilla, es decir, el instrumento de recogida de datos es la plantilla que consiste en la unión de las: figura 6.2.A, figura 6.2.B y figura 6.2.C. Desde ellas se obtuvieron las respuestas escritas del estudiantado en cada una de las etapas del proceso de indagación.

Durante el año académico 2023 entre los meses de marzo, abril y mayo, se recogieron un total de 50 plantillas. De estas, 6 correspondieron a trabajos individuales, 33 a parejas, 3 a grupos de tres integrantes y 4 grupos de cuatro integrantes. En la tabla 6.1.1 se aprecian otras particularidades de la implementación del proyecto, destaca que la muestra contiene distintos niveles educativos y por ende edades entre los participantes. Existen 10 plantillas correspondientes a alumnos de 13 años (octavo básico en la educación chilena), 10 completadas por alumnos de 14 años, 11 por alumnos de 15 años y 17 plantillas completadas por alumnos de 17 años.

La tabla 6.3.A presenta la distribución de Apps seleccionadas por el estudiantado reflejando las preferencias y tendencias en la selección.

Tabla 6.3.A - Cantidad de plantillas por cada App del estudio.

App	Número de plantillas

Calculadora de amor	10
Cámara térmica	9
Detector de metales	14
FotoMath	6
Podómetro	9
Sonómetro	2
Total general	50

El tratamiento de datos consistió primero en codificar las plantillas con un descriptor único, para luego transcribir las respuestas de cada parte de ellas a una base de datos, de manera de tener la información disponible para ser analizada con herramientas de análisis de datos de manera rápida y digital. Posteriormente, se llevó a cabo un análisis cualitativo de los modelos de funcionamiento propuestos por el estudiantado y finalmente del proceso de indagación.

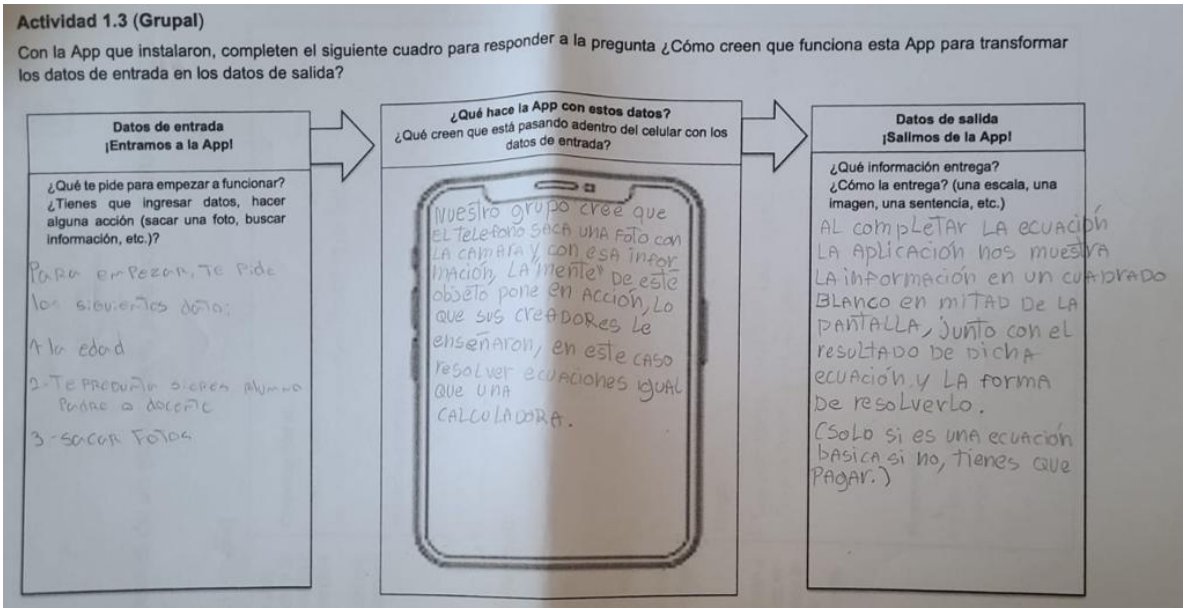

Figura 6.3.A - Modelo elaborado por un estudiante de 13 años sobre el funcionamiento de la aplicación PhotoMath.

Un ejemplo de respuesta del estudiantado participante del estudio 3 está en la figura 6.3.A, donde se la respuesta de la Actividad de la figura 6.2.B. El modelo elaborado fue construido a partir de la experiencia de exploración inicial y la interpretación personal del funcionamiento interno de la App, de un grupo de tres estudiantes de 13 años sobre el funcionamiento de la aplicación *PhotoMath*. Este modelo propone que la aplicación requiere ciertos datos iniciales del usuario (edad, rol, y una fotografía de una ecuación). El estudiante describe que el celular capta una imagen que luego es interpretada por una "mente" del teléfono, lo que activa un proceso similar al de una calculadora, mostrando el resultado y los

pasos para resolver la ecuación. La descripción incluye una condición de acceso: si la ecuación es básica, el resultado se entrega gratis, pero si es avanzada, se requiere pago.

Para concretar el objetivo de analizar cómo son los modelos de funcionamiento de Apps del alumnado primero se transcribieron las respuestas del alumnado a una base de datos y se codificaron las plantillas una a una. Luego, se diseñaron criterios y distintos niveles de elaboración para evaluar el nivel de sofisticación de los modelos mentales del alumnado. Para o anterior se usaron las respuestas de los estudiantes y el marco teórico 2.4 del capítulo II de esta tesis.

6.3.2 Construcción de categorías para el análisis de los modelos mentales del alumnado

Se utilizaron los trabajos de Louca et al. (2011a), Louca et al. (2011b) y López-Simó y Simarro (2024) para generar una rúbrica y evaluar los modelos propuestos por el alumnado. Se siguió el planteamiento de descomponer un modelo en sus componentes individuales para su evaluación, ya que esto permite un análisis más preciso de la compresión del estudiantado, facilita la identificación de patrones en la construcción de modelos y permite observar el nivel de elaboración en la representación (Louca et al., 2011a).

Elegido el enfoque en componentes individuales, el diseño de la rúbrica se estructuró en un proceso metodológico semiempírico con las siguientes fases:

1. **Revisión del modelo real de la App.** Se investigó cómo funciona cada aplicación de la figura 6.2.A en la realidad, utilizando información proporcionada por sus desarrolladores. Esto permitió establecer un referente para la evaluación.
2. **Creación de una rúbrica teórica usando los modelos reales de las Apps**. Se analizaron los conceptos y rúbricas de la tabla 2.4.B, tabla 2.4.C y tabla 2.4.D, adecuando las categorías: objetos físicos, entidades físicas, comportamientos, interacciones y precisión, a los modelos reales de las Apps del punto anterior. De este análisis se generó una rúbrica teórica para determinar el nivel de elaboración de los modelos mentales del alumnado acerca del funcionamiento de las Apps.
3. **Iteración y ajuste de la rúbrica teórica a los datos.** Se evaluaron los modelos de las Apps construidos por el estudiantado mediante la rúbrica teórica. De manera que fue sometida a un proceso de ajuste iterativo en función de su capacidad para discriminar entre diferentes niveles de elaboración de los modelos estudiantiles. Se realizaron

modificaciones en las dimensiones y criterios de evaluación hasta alcanzar una versión final que permitiera discriminar mejor los niveles de desempeño de modelos del alumnado.

4. **Codificación del 30% de los datos por parte de la dirección de tesis**. Para garantizar la fiabilidad intercodificadores del proceso de análisis, el director de tesis codificó de manera independiente el 30% de los datos. Esto permitió evaluar la consistencia en la aplicación de la rúbrica y detectar posibles discrepancias en las categorías de los modelos y niveles de elaboración.
5. **Comparación y consenso en la categorización**. Los resultados obtenidos en la fase de codificación independiente fueron comparados con los del autor, discutiendo las diferencias hasta alcanzar un consenso en la categorización. Este proceso permitirá afinar la rúbrica y asegurar su validez, ajustando las dimensiones y criterios de evaluación en función de la concordancia interevaluador.

El enfoque iterativo permitió evaluar los modelos del estudiantado con respecto a los reales. Además, facilitó la identificación de posibles dificultades conceptuales, lo que permitió mejorar la rúbrica para futuras aplicaciones.

6.3.2.1 Revisión del modelo real de la App.

Según Montiel (2017) las Apps que trabajaron los estudiantes son del tipo propuesto en la tabla 6.3.B. Se hace notar que Montiel (2017) clasifica las Apps en nativas, web, híbridas y en la nube, tomando en cuenta dónde se ejecuta el código de la aplicación y cómo se distribuye el procesamiento. Si bien no se centra exclusivamente en el procesamiento de datos, su clasificación se puede aproximar a estar principalmente vinculada donde se llevan a cabo las operaciones de la App.

Tabla 6.3.B - Clasificación de Apps de la figura 6.2.A, según Montiel (2017).

App	Categoría	Justificación
Cámara Térmica	Nativa	Funciona directamente en el dispositivo móvil y utiliza la cámara del sistema para aplicar filtros térmicos. No depende de conexión a internet para su función principal.
Detector de Metales	Nativa	Utiliza el sensor magnético (*magnetómetro*) del dispositivo, lo que indica que está diseñada para ejecutarse localmente sin necesidad de conexión web.
Podómetro	Nativa	Usa el acelerómetro del dispositivo para medir el movimiento. No requiere acceso a internet para su funcionamiento principal.
Sonómetro	Nativa	Funciona utilizando el micrófono del dispositivo para captar sonidos y calcular su frecuencia. No necesita conexión en la nube para operar.

Calculadora de Amor	Web o Híbrida	No funciona sin conexión a internet, lo que indica que los cálculos de compatibilidad se realizan en servidores remotos. Además, solicita permisos de red, acceso a APIs publicitarias y almacenamiento compartido, lo que refuerza su dependencia de una infraestructura web.
PhotoMath	En la Nube	La app requiere conexión a internet obligatoriamente para funcionar y utiliza una tecnología avanzada de resolución matemática en la nube. Su procesamiento OCR y la resolución de ecuaciones dependen de servidores remotos y machine learning basado en IA.

La mayoría de las Apps trabajadas con los estudiantes son nativas, ya que dependen directamente del hardware del dispositivo para su funcionalidad, como sensores, cámara y micrófono. PhotoMath se clasifica como una aplicación en la nube, dado que su sistema de resolución se basa en servidores remotos para ofrecer una mejor experiencia y actualizaciones en tiempo real. Por otro lado, Calculadora de Amor tiene una dependencia de la web y realiza los cálculos en servidores remotos, lo que la ubica dentro de la categoría de aplicación web o en la nube más que híbrida. Mientras que las Apps nativas funcionan sin conexión a internet, las Apps en la nube requieren acceso continuo a servidores externos para realizar sus funciones avanzadas.

Posteriormente, se llevó a cabo una búsqueda sobre el funcionamiento de las Apps trabajadas, con la cual se construyó la tabla 6.3.C. Se recopiló la información proporcionada por los desarrolladores de cada Apps, principalmente de las especificaciones de cada App en, Google Play Store para Android (https://play.google.com/store), sitios web oficiales y documentación técnica disponible. Además, se realizaron pruebas empíricas, como la desconexión de internet en algunas Apps, para determinar su grado de dependencia de servidores externos. La información sobre las demás aplicaciones se contrastó con sus permisos y funcionalidades para confirmar su clasificación. Para analizar los modelos reales de cada App, se usó el esquema de entrada-proceso-salida, tal como se trabajó en la plantilla con el alumnado (figura 6.3.B).

Tabla 6.3.C - Modelos reales de cada App.

App	Entrada de datos	Proceso	Salida
Cámara Térmica	Recibe información a través de la cámara estándar del dispositivo, capturando imágenes en tiempo real. El usuario	Una vez que la imagen es capturada, la aplicación aplica un filtro de simulación térmica, coloreando según una escala predefinida. No	Muestra una imagen coloreada en tiempo real con efecto térmico simulado,

	puede seleccionar distintos efectos visuales, como visión térmica, infrarroja y ultravioleta, que se aplicarán a la imagen.	usa sensores reales; solo simula visión térmica. Los efectos pueden verse en tiempo real antes de tomar la foto.	según los filtros seleccionados.
Detector de Metales	Recibe datos del sensor magnético (magnetómetro) del dispositivo móvil. El usuario puede mover el teléfono para detectar cambios en el campo magnético.	Compara los valores con los del campo magnético terrestre (~49 µT). Si hay un aumento, la app interpreta que hay un metal ferromagnético. Puede configurarse la sensibilidad y emitir alertas. No detecta metales no ferromagnéticos. Es susceptible a interferencias externas.	Muestra en pantalla el valor del campo magnético en tiempo real. Puede incluir gráficos y emitir alertas visuales o sonoras si se supera un umbral.
Podómetro	Usa el acelerómetro del dispositivo para detectar aceleraciones. El usuario puede ingresar su altura, peso y longitud del paso.	Analiza patrones de aceleración para distinguir pasos reales. Calcula distancia recorrida y calorías quemadas según los datos del usuario. Si el teléfono no está en posición estable (por ejemplo, en la mano), el conteo puede ser inexacto.	Muestra en pantalla el total de pasos, distancia estimada y calorías quemadas.
Sonómetro	Usa el micrófono del dispositivo para captar ondas sonoras en tiempo real. El usuario puede ajustar sensibilidad o aplicar filtros para ciertos rangos de sonido.	Utiliza algoritmos como la Transformada Rápida de Fourier (FFT) para analizar el sonido y calcular su intensidad en decibeles. Puede identificar frecuencias dominantes y generar gráficos de variación. La precisión depende de la calidad del micrófono y condiciones externas.	Muestra los niveles de sonido en decibeles y puede incluir gráficos de evolución. Algunas versiones permiten exportar datos o configurar alertas por volumen.
Calculadora de Amor	El usuario introduce nombres, fechas de nacimiento, signos zodiacales y respuestas sobre personalidad. Requiere conexión a internet.	La app procesa los datos usando un algoritmo no revelado, basado en numerología, astrología o generadores aleatorios. El cálculo lo realiza en un servidor remoto.	Muestra un porcentaje de compatibilidad amorosa y un mensaje explicativo.

PhotoMath	El usuario escanea ecuaciones con la cámara o las escribe manualmente. Requiere conexión a internet.	Utiliza reconocimiento óptico de caracteres (OCR) para digitalizar la ecuación. Luego, la envía a sus servidores donde se resuelve usando un sistema algebraico computacional (CAS). Emplea inteligencia artificial para elegir el mejor método y mostrar pasos.	Muestra la solución con los pasos detallados. Puede incluir gráficos, explicaciones alternativas y tutoriales animados.

6.3.2.2 Creación de una rúbrica teórica

Louca et al. (2011a, 2011b) proponen una metodología en la que el análisis de modelos se basa en la descomposición de un fenómeno en diferentes componentes (Tabla 2.4.A). Estos elementos son los objetos y las entidades físicas, que se organizan dentro de un modelo para representar el fenómeno en estudio. Se puede decir que la idea central es identificar los "átomos" del sistema, las unidades mínimas que, al combinarse, generan comportamientos e interacciones. Esta idea central para cualquier proceso de modelado es la que no abandonamos, ya que permite un análisis detallado de cada componente antes de evaluar la dinámica del sistema en su totalidad.

La primera categoría es “objetos físicos” se debe notar que para las Apps de la tabla 6.3.C ninguna recibe información proveniente de un objeto físico en el sentido en que Louca et al. (2011a, 2011b) define esta categoría. En el contexto de las Apps, los datos de entrada no corresponden a cuerpos físicos con propiedades medibles en el espacio real, sino a información digital, como datos sensoriales, texto ingresado por el usuario o imágenes capturadas por la cámara del dispositivo, que luego será procesada por el software de al App.

Un tipo de Apps, las nativas, reciben sus datos de entrada a partir de sensores del dispositivo y otras desde entradas manuales del usuario, por lo tanto los sensores son una tecnología necesaria para la App, en algún momento el programa informático deberá “llamar” al sensor. Además, se espera el alumnado lo mencione en algún momento en sus modelos.

Resumiendo, los datos de entrada en las Apps son heterogéneos y no son entidades tangibles, además se debe considera que los sensores cumplen un papel de mediación, no de representación directa del fenómeno

De manera que ninguna de estas entradas puede llamarse "objeto físico" en el sentido tradicional y son simplemente “objetos”, ya que no tienen una existencia totalmente material dentro del sistema modelado.

Los diferenciaremos entre las Apps nativas, que dependen de sensores, y aquellas que operan con datos ingresados por el usuario. Las primeras las llamaremos objetos sensores y representan los dispositivos del hardware del teléfono que capturan información del entorno físico y la convierten en datos digitales para ser procesados por la App. En este grupo se encuentran, cámara térmica que recibe información a través de la cámara, convirtiendo imágenes en datos digitales, detector de metales que recibe datos del magnetómetro, traduciendo campos magnéticos en mediciones numéricas, podómetro que utiliza el acelerómetro para detectar cambios en la aceleración y traducirlos en conteo de pasos; sonómetro que capta ondas sonoras a través del micrófono y las transforma en señales eléctricas que representan la intensidad del sonido y PhotoMath que usa la cámara para capturar imágenes que luego procesará en la nube.

Las segundas las llamamos objetos digitales y representan los datos de entrada que no provienen de sensores, sino de información ingresada manualmente por el usuario. En este grupo se encuentra la App calculadora de amor que recibe nombres, fechas de nacimiento y signos del zodiaco ingresados por el usuario a través de la interfaz de usuario, datos que son procesados en un servidor externo.

La categoría pasará a llamarse "objetos" y la definimos como en función de la naturaleza de los datos de entrada:

> "Son las unidades fundamentales de entrada en el programa informático de la App. Pueden ser de dos tipos: objetos sensores, que representan dispositivos físicos capaces de capturar datos del entorno y traducirlos en información digital procesable por la App, y objetos digitales, que corresponden a datos ingresados directamente por el usuario a través de la interfaz de usuario. Estos objetos constituyen la base del modelo de funcionamiento de las Apps y determinan la forma en que la información ingresa al sistema."

Luego, en el modelo de Louca et al. (2011a, 2011b), siguen las entidades físicas que representan las características de los objetos físicos, como la velocidad, la energía o la fuerza, y permiten describir las propiedades que afectan la dinámica del sistema modelado. En el contexto de Apps, los objetos sensores y los objetos digitales constituyen la fuente de datos,

pero cada uno de estos objetos tiene atributos específicos que afectan su procesamiento. Por ejemplo:

- Un campo magnético medido en la App detector de metales tiene como entidad su intensidad en microteslas (μT).
- Una onda sonora capturada por la App sonómetro tiene como entidad su frecuencia en Hertz (Hz).
- Un nombre ingresado en la Calculadora de Amor tiene como entidad la interpretación numerológica asignada.

Dado esto las entidades representan atributos de los objetos que afectan el procesamiento de la información. La categoría la llamamos solo "entidades" y definimos como:

> "Las propiedades o atributos de los objetos dentro del sistema digital. En las Apps nativas, las entidades representan características medibles de los datos capturados por sensores, como la intensidad de un campo magnético, la frecuencia de una señal sonora o la aceleración de un movimiento. En las Apps basadas en el usuario, las entidades pueden representar atributos asignados a los datos ingresados, como interpretaciones numéricas, patrones de reconocimiento o estructuras matemáticas."

Ahora nos centraremos en las categorías comportamientos e interacciones. Primero hacemos notar que consideraremos que la categoría comportamiento e interacción difieren en significado. Los comportamientos los pensamos como procesos de transformación, que nos ayudan a comprender cómo funciona internamente un componente (qué hace por sí mismo), en cambio las interacciones nos permiten ver cómo los diferentes componentes del sistema y sus comportamientos se comunican o dependen entre sí. Si nos interesa qué hace un componente internamente, hablaremos de proceso de transformación o transformación, pero si nos interesa cómo se relaciona con otros elementos del sistema, hablaremos de interacción.

El análisis de las Apps muestra que cada una sigue un flujo de procesamiento estructurado que convierte los datos de entrada en información útil para el usuario. Sin embargo, las diferencias entre las Apps nativas y aquellas que dependen de procesamiento en la nube afectan la forma en que los datos son tratados. En las Apps nativas como el Detector de Metales, Podómetro y Sonómetro, los datos de entrada provienen directamente de sensores del dispositivo, y el procesamiento ocurre localmente, sin depender de servidores externos, transforman las señales del entorno en datos digitales que son procesados. Esto implica que los comportamientos en estas Apps están directamente vinculados al sensor y la información que

este capta. Estas Apps generan salidas en tiempo real directamente en la interfaz del usuario, permitiendo ajustes o análisis inmediatos. Las Apps en la Nube (PhotoMath y Calculadora de Amor), los datos ingresados (ya sea mediante la cámara o por texto) no son procesados en el dispositivo, sino que se envían a servidores externos que realizan cálculos avanzados y devuelven un resultado. En este caso, los comportamientos de la aplicación dependen de algoritmos remotos y de la infraestructura de procesamiento en la nube. Estas Apps presentan salidas procesadas a partir de cálculos en servidores, como la solución paso a paso de una ecuación en PhotoMath o el porcentaje de compatibilidad en la calculadora de amor.

Los comportamientos dependen del procesamiento de datos digitales realizados por el programa informático (software de la App) y las interacciones ya no ocurren entre entidades físicas, sino entre sensores, algoritmos, servidores, etc. Analizar los criterios de desempeño graduando en causalidad (no causal, semi-causal y causal) no es del todo adecuado por las siguientes razones:

- Las Apps trabajadas no modelan un fenómeno físico, más bien realizan tareas específicas, de detección o medición (Detector de metales, Sonómetro, Podómetro, Cámara térmica) y cálculo (FotoMath, Calculadora de amor).
- En una App el comportamiento está determinado por código y algoritmos diseñados por desarrolladores, por lo que no hay una "causalidad emergente" sino una "lógica de programación", es decir, las transformaciones en la App no dependen de principios científicos o fundamentados en relaciones empíricas entre entidades físicas, sino de procesamiento de datos, por ejemplo en la App Cámara Térmica, los colores asignados no tienen una base física real, sino que dependen de reglas de imagen y contraste definidas en el software.
- El criterio de causalidad no discrimina bien entre modelos válidos y no válidos en Apps, un modelo de App podría ser completamente funcional sin representar ninguna relación causal física. En cambio, en modelos científicos, la presencia o ausencia de causalidad es importante para determinar la validez del modelo, aquí la validez es simplemente si el estudiante es capaz de descubrir el modelo de funcionamiento real de la App.

En resumen, dado que las Apps no representan totalmente fenómenos físicos ni dependen de leyes físicas sino de reglas algorítmicas predefinidas, no hablaremos de "mecanismos subyacentes" en el sentido que lo expone Louca et al. (2011a, 2011b). En su lugar, se van a determinar los procesos de transformación, es decir:

- Los procesos internos mediante los cuales la App convierte datos de entrada en datos de salida.

- Las reglas algorítmicas que estructuran estas transformaciones.
- Las estructuras computacionales que permiten la manipulación de datos dentro de la App.

Por ejemplo, en la App Cámara Térmica, no hay un mecanismo físico de detección de temperatura, pero sí hay transformaciones subyacentes, como:

- Transformación de brillo en datos de color térmico.
- Interpolación de color para suavizar la imagen.
- Optimización del contraste según la luz ambiental.

Por ejemplo, transformar el nivel de brillo de una imagen en un color simulado según una escala térmica es un comportamiento, ya que se trata de una transformación autónoma dentro del software Esto significa que en lugar de estudiar causalidad física, estudiamos las transformaciones de los datos según el software. Por lo anterior, para comportamientos usamos la siguiente definición:

> "Los comportamientos en las Apps son los procesos de transformación internos que realiza cada componente por sí solo, sin depender de otros elementos del sistema. Estos procesos están definidos por reglas algorítmicas que permiten que el componente modifique, interprete o transforme datos internamente. Por ejemplo, un algoritmo que asigna colores según niveles de brillo o que cuenta pulsos eléctricos es un comportamiento, ya que actúa dentro del componente sin requerir interacción con otros. Los comportamientos explican cómo funciona o cambia internamente un componente, permitiendo comprender su lógica propia de transformación".

Resumiendo, los "objetos" son datos de entrada (por ejemplo, la imagen capturada por la cámara), las "entidades" son atributos de esos datos (por ejemplo, brillo y contraste), y los "comportamientos" son los procesos de transformación interna que resultan de aplicar reglas algorítmicas dentro de un componente. Las reglas que rigen tanto los comportamientos como las interacciones están determinadas por instrucciones algorítmicas predefinidas en el código de la App (software o programa informático). No existen "mecanismos físicos" en estas transformaciones, sino que todo se basa en reglas lógicas que definen cómo se procesan y manipulan los datos. Los comportamientos corresponden a las transformaciones internas que realiza cada componente por sí solo, mientras que las interacciones describen cómo se conectan y se comunican esos componentes y sus comportamientos entre sí dentro del sistema digital. En este contexto, utilizamos la siguiente definición para interacciones:

> "Las interacciones en las Apps son las relaciones funcionales entre distintos componentes del sistema digital, que permiten que los comportamientos trabajen de manera coordinada. No se trata de transformaciones internas, sino de cómo se vinculan los datos, entidades y comportamientos entre sí a lo largo del flujo de procesamiento. Estas interacciones están determinadas por dependencias algorítmicas que estructuran el recorrido de la información, como la secuencia en que se activan funciones o el modo en que una salida se convierte en entrada de otro proceso."

Por ejemplo, en la App Podómetro, la interacción entre el acelerómetro (componente sensor que provee el objeto de entrada), el cálculo del paso (comportamiento interno que transforma la señal), y la estimación de distancia (comportamiento posterior que toma esa transformación como entrada), es una interacción, ya que describe cómo los distintos comportamientos están encadenados funcionalmente para producir una salida útil.

En el caso de la categoría precisión de un modelo, no se trabajará en lo que hemos llamado rúbrica teórica, ya que no les es pedida en la actividad 2 de la plantilla (figura 6.2.B). La precisión del modelo que Louca et al. (2011a) categoriza como válido o no válido (tabla 2.4.B), lo haremos posteriormente, para determinar cuánto se aleja el modelo tecnológico del estudiante en relación con el modelo real de la App.

Ahora, abordando las rubricas de la tabla 2.4.B y tabla 2.4.C, estas presentan similitudes Ambas evalúan la representación de objetos, entidades, comportamientos e interacciones en los modelos creados por los estudiantes. Sin embargo, presentan diferencias en su enfoque de evaluación y en la forma en que estructuran los niveles de desempeño. La rúbrica de la tabla 2.4.B no parecen determinar niveles de elaboración de manera explícita, salvo en la categoría de comportamientos y precisión científica. En estos casos, los criterios de representación causal en los comportamientos y la validez científica del modelo permiten distinguir entre niveles de desarrollo más o menos avanzados, pero en el resto de las categorías, esta rúbrica funciona más bien como una clasificación taxonómica que identifica la presencia o ausencia de ciertas componentes de la categoría, sin establecer una progresión en el nivel de comprensión del estudiante. Este enfoque tiene ciertas limitaciones. Al no definir niveles de desempeño, no permite evaluar el grado de desarrollo de los modelos propuestos por los estudiantes, sino únicamente si ciertos elementos están presentes o ausentes. Esto puede impedir medir la progresión en la calidad de los modelos generados y dificulta la identificación de mejoras en la representación de las Apps analizadas.

En contraste, la rúbrica de la tabla 2.4.D separa el desempeño en niveles de mejor o peor desempeño, determinando cuántos objetos, entidades, comportamientos e interacciones han sido representados en el modelo. Esto ayuda en el contexto del análisis de Apps, ya que las distintas aplicaciones tienen naturalezas diferentes, lo que implica que algunas pueden tener múltiples objetos o entidades mientras que otras tienen menos. Por ello, el nivel más alto de desempeño en cada categoría debería estar relacionado con la capacidad de los estudiantes para identificar todos los objetos y entidades presentes en el modelo real de la App analizada.

Por lo anterior para las categorías objetos y entidades, se propone un enfoque que mantendrá la graduación en niveles de desempeño de la rúbrica de la tabla 2.4.D, pero especificando que el nivel más alto corresponde a la identificación de todos los objetos que forman parte del modelo de la App.

En la rúbrica de la tabla 2.4.B la categoría comportamientos se centra en la presencia de relaciones causales dentro del sistema modelado, diferenciando entre no causales, semi-causales y causales. Hay una jerarquía donde el desempeño más avanzado se asocia con una comprensión más precisa de las relaciones entre entidades físicas. En el caso de las Apps se decidió, por las razones ya dadas, evaluar el desempeño distinguiendo entre el nivel de desarrollo de los comportamientos.

Finalmente, para la categoría “interacciones”, las rúbricas de la tabla 2.4.B considera que las interacciones pueden clasificarse en distintos tipos, como interacciones entre objetos, entidades o entre comportamientos físicos. Más que establecer niveles de desempeño, estas rúbricas buscan identificar qué tipos de interacciones han sido representadas. Este enfoque sugiere que no todas las interacciones tienen el mismo peso o relevancia dentro de un modelo, sino que cada tipo de interacción describe una relación específica dentro del sistema. Al igual que en las otras categorías se propone una progresión donde se considera que los estudiantes no necesariamente plantarán el modelo real de la App, por lo que sus modelos pueden ser más o menos aproximados al que hemos llamado real. El nivel más alto debe reflejar un modelo que se asemeje lo más posible al modelo real de la App.

Durante la construcción de las categorías se determinaron los objetos, entidades, comportamientos e interacciones de las Apps de la tabla 6.3.C, con ello se logró una determinación más precisa de cada App que se muestra en la tabla 6.3.D que es el estándar de cada App. Luego está fue usada para determinar la rúbrica teórica.

Tabla 6.3.D - Modelos reales de Apps.

App	Objetos	Entidades	Comportamientos	Interacciones
Cámara Térmica	- Cámara (objeto sensor) - Interfaz de selección de filtros (objeto digital)	- Brillo de imagen - Contraste - Color asignado - Filtro seleccionado - Condiciones de luz ambiental	- Transformación de niveles de brillo en colores simulados - Aplicación del filtro seleccionado sobre la imagen - Ajuste del contraste según luz ambiental	- La cámara captura el brillo de la imagen, que es transformado en color térmico. - El filtro seleccionado se aplica sobre esa transformación. - El ajuste de contraste reacciona a las condiciones de luz detectadas.
Detector de Metales	- Magnetómetro (objeto sensor) - Interfaz de activación/detección (objeto digital)	- Intensidad del campo magnético (µT o mG) - Variación respecto al valor base (49 µT) - Umbral ajustado por el usuario - Sensibilidad del sensor	- Conversión de datos del magnetómetro en valores numéricos - Comparación de la intensidad del campo con el valor de referencia (49 µT) - Activación del umbral ajustado - Modulación de sensibilidad según configuración	- El magnetómetro detecta la intensidad del campo magnético, que es convertida en datos. - Esos datos se comparan con el valor base. - Si superan el umbral ajustado, se activa la alerta. - La sensibilidad del sensor modula esta cadena.
Podómetro	- Acelerómetro (objeto sensor) - Datos ingresados del usuario (peso, altura, longitud del paso) (objetos digitales) - Interfaz de inicio (objeto digital)	- Aceleración del movimiento - Frecuencia de pasos - Longitud del paso - Peso del usuario - Ritmo o velocidad de caminata	- Conversión de aceleración en patrones de pasos - Cálculo de frecuencia de pasos - Estimación de distancia según longitud del paso - Cálculo de calorías quemadas en base al peso - Clasificación de ritmo de caminata	- El acelerómetro detecta la aceleración, que se convierte en patrones de pasos. - Los patrones permiten calcular frecuencia. - La frecuencia y la longitud del paso permiten estimar distancia. - Peso y distancia determinan calorías. - Ritmo se clasifica con todo lo anterior.
Sonómetro	- Micrófono (objeto sensor) - Interfaz de ajuste de sensibilidad o umbral (objeto digital)	- Interpretación numerológica - Compatibilidad zodiacal - Puntaje del test - Porcentaje de afinidad - Selecciones dentro del formulario interactivo	- Transformación de ondas sonoras en señal eléctrica digital - Cálculo de la intensidad sonora en dB - Identificación de frecuencia dominante - Generación de gráficos de sonido - Aplicación del umbral de activación	- El micrófono capta sonido, que se transforma en señal digital. - Esa señal permite calcular decibeles e identificar frecuencias. - La frecuencia dominante genera un gráfico. - El umbral ajustado condiciona la activación de ciertas respuestas.
Calculadora de Amor	- Nombres (objeto digital) - Fechas de nacimiento (objeto digital) - Signos del zodiaco (objeto digital) - Interfaz de preguntas/test (objeto digital)	- Frecuencia del sonido (Hz) - Intensidad sonora (dB) - Frecuencia dominante - Tiempo de duración del sonido - Nivel de umbral ajustado	- Traducción de nombres en números (numerología) - Cálculo de compatibilidad zodiacal según reglas predefinidas - Procesamiento del test de afinidad - Cálculo del porcentaje final de compatibilidad	- Los datos ingresados (nombres, fechas, signos) son traducidos en números y atributos simbólicos. - Estos se combinan para generar compatibilidad zodiacal. - Las respuestas del test alimentan el puntaje total. - Todo se unifica en el porcentaje de afinidad.
PhotoMath	- Cámara (objeto sensor) - Expresiones matemáticas (objeto digital) - Teclado virtual (objeto digital) - Interfaz de edición (objeto digital)	- Estructura de la ecuación - Método de resolución identificado - Reconocimiento de símbolos matemáticos - Tipos de operaciones identificadas - Grado de la expresión - Método de resolución	- Reconocimiento de símbolos en una expresión matemática (OCR) - Identificación de estructura algebraica - Selección del método de resolución (CAS) - Resolución paso a paso según el tipo de operación - Generación de explicaciones y gráficos.	- La cámara captura la imagen de una ecuación. - El OCR identifica símbolos y estructura matemática. - Esa estructura alimenta el CAS, que selecciona un método. - El método genera pasos de resolución. - Los pasos son convertidos en explicaciones y gráficos.

La rúbrica teórica propuesta y los criterios de desempeño para cada categoría se muestran en la tabla 6.3.E. En análisis de los modelos reales de Apps según la información del

desarrollador y las definiciones dadas para objetos, entidades, comportamientos e interacciones, se muestran en la tabla 1 del anexo.

Tabla 6.3.E - Propuesta teórica de la rúbrica para evaluar los modelos de funcionamiento de App del alumnado.

Categoría	Nivel de elaboración
Objetos	0. No hay objetos representados 1. Se representa al menos un objeto correcto 2. Se identifican correctamente todos los objetos que conforman el modelo de la App
Entidades	0. No hay entidades representadas 1. Se representa al menos una entidad correcta 2. Se identifican correctamente todas las entidades que caracterizan los datos dentro del modelo de la App
Comportamientos	0. No hay comportamientos representados. No se describe ningún proceso interno. 1. Representación básica. Se identifica al menos un comportamiento por componente, pero con errores o explicaciones vagas. 2. Representación adecuada. Se representan correctamente los comportamientos internos más importantes de los componentes, aunque no todos. 3. Representación completa y precisa. Todos los comportamientos internos relevantes están correctamente descritos, con indicación clara de qué hace cada componente por sí solo.
Interacciones	0. Sin interacciones representadas. Cada componente está aislado, sin conexión funcional. 1. Interacción simple. Se describe una relación funcional básica entre dos componentes (ej. “la cámara manda imagen al servidor”). 2. Interacciones encadenadas. Se representan varias relaciones funcionales que muestran cómo se conectan los componentes en secuencia o bucles. 3. Estructura de interacción compleja. Las interacciones están claramente organizadas, muestran dependencias lógicas o condicionales, y permiten entender el flujo total de procesamiento de la App.

6.3.2.3 Iteración y ajuste de la rúbrica teórica a los datos.

Se realizó un análisis iterativo de los modelos de funcionamiento de los estudiantes analizándolos con la rúbrica teórica de la tabla 6.3.E. Primero, se tipearon desde las plantillas de los estudiantes sus respuestas a la actividad 2 (Figura 6.2.B) en una tabla. Luego, se realizó un proceso de identificación de objetos y entidades clasificándolos como correcto o incorrecto. En esta parte se comparó el modelo de los estudiantes contra los modelos de la tabla 6.3.D. A partir de esto, contamos la cantidad de objetos correctos e incorrectos por modelo, permitiendo observar cómo combinaban estos elementos los estudiantes. Posteriormente, analizamos la frecuencia con la que aparecían combinaciones de objetos correctos e incorrectos, lo que

permitió identificar patrones en los niveles de comprensión de los estudiantes. Finalmente, utilizamos esta información para reformular la rúbrica de evaluación, asegurando que refleje con mayor precisión cómo los estudiantes progresan en la identificación de objetos y entidades. Los criterios finales de desempeño para todas las categorías se muestran en la tabla 9.

A partir del análisis se determinó que la mejor forma de trabajar los niveles de elaboración identificar y luego contar las combinaciones de objetos correctos e incorrectos dentro de cada uno de los modelos de los estudiantes. En la versión inicial de la rúbrica no se distinguía que en los modelos coexisten objetos, entidades, comportamientos e interacciones correctas e incorrectas. Por ejemplo, algunos estudiantes mostraron avances al identificar algunos elementos adecuados, pero aún mantenían errores en otros. Sin una categoría intermedia que muestre esta situación, la progresión en la comprensión de modelos, queda mal representada.

Así, la nueva versión final de la rúbrica, presentada en la tabla 6.3.F, no solo clasifica los modelos en función de si hay o no procesos de transformaciones, sino que valora con mayor exactitud como los estudiantes representan dichas transformaciones.

Entre las semejanzas de las tablas 6.3.E y tabla 6.3.F, ambas versiones destacan la estructura progresiva, que permite evaluar el grado de articulación entre los elementos representados. Además, ambas versiones reconocen que las interacciones son clave para comprender el modelo como un todo, algo que no puede captarse si se observan solo los componentes por separado, ofreciendo así una herramienta más coherente para evaluar el grado de elaboración de los modelos mentales del alumnado.

Tabla 6.3.F - Rúbrica final para estudiar los modelos de funcionamiento de App del alumnado.

Categoría	Nivel de elaboración
Objetos	0. No hay objetos representados. 1. Se identifican sólo objetos incorrectos. 2. Se identifican objetos correctos e incorrectos. 3. Se identifican sólo objetos correctos, pero no todos los del modelo real. 4. Se identifican correctamente todos los objetos que conforman el modelo de la App, sin errores conceptuales.
Entidades	0. No hay entidades representadas. 1. Se identifican sólo entidades incorrectas. 2. Se identifican entidades correctas e incorrectas. 3. Se identifican sólo entidades correctas, pero no todas las del modelo real. 4. Se identifican correctamente todas las entidades que conforman el modelo de la App, sin errores conceptuales.
Comportamientos	0. No hay comportamientos representados (No se describen procesos internos). 1. Se identifican solo comportamientos incorrectos o sin sentido.

	2. Se representan comportamientos correctos e incorrectos, con descripciones confusas o parciales. Algunos componentes tienen funciones válidas, pero hay errores o mezclas conceptuales. 3. Se representan solo comportamientos correctos, pero no todos los relevantes del modelo real. Las descripciones son adecuadas pero incompletas. 4. Representación completa y precisa. Todos los comportamientos internos relevantes están correctamente descritos, con claridad sobre la función específica de cada componente.
Interacciones	0. Sin interacciones representadas. Cada componente aparece aislado, sin conexión funcional. 1. Se describen solo interacciones incorrectas o ilógicas (ej. conexión entre componentes que no tiene sentido en el funcionamiento de la App). 2. Se representan interacciones correctas e incorrectas. Existen vínculos funcionales válidos, pero también conexiones erróneas o confusas. 3. Se representan sólo interacciones correctas, pero no todas las necesarias para entender el funcionamiento completo de la App. Se muestra una secuencia básica o parcial. 4. Estructura de interacción compleja. Las interacciones están claramente organizadas, muestran dependencias lógicas o condicionales, y permiten comprender el flujo completo de procesamiento de la App.

6.4 Resultados del estudio 3

Por cada plantilla (Figura 6.2.B) se identificaron los objetos, entidades, comportamientos e interacciones que aparecían en las respuestas del alumnado. Luego se analizó el modelo de funcionamiento propuesto y se comparó con el modelo real de la App para identificar que objetos, entidades, comportamientos e interacciones no reconoce el alumnado en su modelo. Finalmente, se determinó el nivel de desempeño con la rúbrica de la tabla 6.3.F. Con ello se construyó una tabla de sistematización con todas las plantillas.

A partir del listado de objetos, entidades, comportamientos e interacciones ausentes en los modelos construidos por los estudiantes, se realizó un análisis cuantitativo y cualitativo que permitió identificar los tipos de elementos omitidos y, en consecuencia, los aspectos del funcionamiento de las Apps que resultaron más difíciles de reconocer para el alumnado.

- Las omisiones más frecuentes corresponden a diferentes tipos de interfaz gráfica. Esto indica que muchos estudiantes no reconocen que los elementos visuales de interacción son también componentes funcionales fundamentales en el flujo de entrada de datos. Esta ausencia de la interfaz digital limita su comprensión del funcionamiento técnico de las Apps.
- Los sensores físicos también son omitidos. Aunque las Apps nativas dependen de sensores, como el magnetómetro o el acelerómetro, muchos modelos no los incluyen

explícitamente, lo que muestra una dificultad para vincular la percepción de la App con el hardware subyacente. Es decir, hay desconocimiento de que son los sensores por parte del alumnado.

- Los datos ingresados por el usuario no siempre se reconocen como objetos. En la App calculadora de amor, elementos como los datos del usuario, signos del zodiaco o fechas de nacimiento son frecuentemente omitidos, a pesar de ser centrales en la Apps. Esto puede deberse a que el alumnado no percibe estos datos como "unidades fundamentales de entrada" sino como información anecdótica o al considerar la App como no válida.
- Algunos grupos no identifican ningún objeto del modelo real. En al menos un caso, se reporta que "faltan todos los objetos del modelo real", lo cual indica una comprensión nula o muy superficial del sistema modelado.

En el caso de las entidades:

- Las entidades asociadas al procesamiento técnico son las más omitidas. Las entidades que representan parámetros de funcionamiento interno (como *umbral ajustado*, *sensibilidad*, o *valor de base*) son las más ausentes en los modelos de los estudiantes (13 menciones cada una). Estas entidades son importantes para entender cómo una App transforma una señal física en un valor significativo, por ejemplo, cuándo un campo magnético se interpreta como la presencia de metal. Su omisión indica que los estudiantes no acceden o no comprenden los mecanismos internos.
- En las Apps visuales, también se omiten aspectos de procesamiento de imagen. Entidades como *contraste*, *brillo*, *color* y *condiciones de luz ambiental* (que afectan la calidad de la imagen y por ende el resultado) están a menudo ausentes, especialmente en modelos de la cámara térmica. Esto revela que aunque los estudiantes reconocen el uso de la cámara, no integran variables que condicionan su funcionamiento técnico.
- En modelos matemáticos, hay dificultad para identificar la estructura formal. En PhotoMath, entidades como *reconocimiento de símbolos*, *estructura de la ecuación*, *tipo de operación* y *grado de expresión* son mencionadas solo en algunos casos. Esto indica que los estudiantes tienden a ver la App como una "caja negra" que da respuestas, sin comprender los pasos intermedios de análisis simbólico o algebraico. Además hay un desconocimiento de los tecnicismo o tecnología detrás del reconocimiento de imágenes o de tecnología de resolución de problemas matemáticos y en general de ingeniería cloud, que corresponde a la vanguardia en la actualidad.

- En el podómetro, se invisibiliza la relación entre señal y dato interpretado. Elementos como *aceleración*, *frecuencia de pasos*, *longitud del paso* y *peso del usuario* están ausentes en la mayoría de los modelos. Esto sugiere una visión funcional, pero sin comprensión de los parámetros que transforman movimiento en información cuantitativa.

En el caso de los comportamientos:

- Los comportamientos relacionados con ajustes técnicos o procesamiento de señales físicas son poco reconocidos. Procesos como *ajuste de contraste*, *aplicación de filtros*, *activación de umbral* y *modulación de sensibilidad* —frecuentes en Apps como la cámara térmica, el detector de metales o el sonómetro— son consistentemente omitidos. Estos procesos son clave para interpretar la señal recibida, por lo que su ausencia indica una comprensión superficial del funcionamiento interno del sistema.
- La lógica comparativa y de calibración es invisible para la mayoría. El hecho de que la *comparación con valores base* no se mencione en la mayoría de los casos sugiere que los estudiantes no comprenden que muchas Apps no solo registran un dato, sino que evalúan su significado relativo, por ejemplo, si un valor supera cierto umbral para emitir una alerta o dar una lectura, como en la App detector de metales.
- En las Apps matemáticas, hay una visión "mágica" del procesamiento. En PhotoMath, los estudiantes suelen omitir comportamientos como el *reconocimiento de símbolos*, la *selección del método* o la *generación de explicaciones* paso a paso. Esto implica que se percibe a la App como una "caja negra" que entrega un resultado, sin distinguir entre las múltiples operaciones lógicas y matemáticas que se desarrollan internamente.
- En los podómetros, los procesos de transformación del movimiento en datos son difusos. Muchos estudiantes no identifican que el sistema debe convertir aceleraciones en distancia o pasos, y de ahí derivar estimaciones como *calorías quemadas* o *ritmo de marcha*. Falta la idea de que estos procesos requieren *algoritmos de conversión* y no ocurren de forma directa.
- En la Calculadora de Amor, los comportamientos se asocian a lo superficial o lo aleatorio. Aunque muchos estudiantes mencionan la entrega de un porcentaje, omiten los procesos internos como el *cálculo zodiacal* o *combinación de factores*, quizás por falta de credibilidad o porque los perciben como irrelevantes, reforzando una postura crítica pero también un vacío de modelización funcional.

- En varios casos se omite todo el conjunto de comportamientos esperables. En al menos 9 casos se indica que *faltan todos los comportamientos reales*, lo que evidencia que no se identificaron transformaciones internas, reforzando la idea de que para muchos estudiantes, las Apps simplemente *reciben datos y devuelven respuestas*, sin un procesamiento intermedio reconocible.

Finalmente, para los comportamientos:

- Las interacciones entre sensores y parámetros internos son las más omitidas. Los estudiantes presentan grandes dificultades para modelar cómo un sensor (como el magnetómetro o el acelerómetro) interactúa con umbrales, valores base, modulación o conversiones. Estas secuencias de procesamiento son fundamentales en la lógica de funcionamiento, y su ausencia muestra que se concibe la lectura del sensor como directa y sin procesamiento intermedio.
- En Apps visuales, falta la representación de la secuencia técnica. La cámara térmica, por ejemplo, depende de múltiples pasos: capturar luz, aplicar filtros, ajustar el contraste, transformar en imagen térmica. Esta cadena rara vez aparece de forma completa. Se omiten interacciones entre variables técnicas (luz, brillo, color), lo que sugiere una comprensión parcial del proceso visual digital.
- PhotoMath revela un vacío sobre los pasos intermedios entre imagen y resolución. En esta App, faltan interacciones como la secuencia OCR → CAS → elección de método → explicación. Muchos modelos simplemente asumen que la App "resuelve" sin representar el encadenamiento lógico de procesos algorítmicos internos, lo que refuerza la idea de "caja negra".
- Los podómetros también sufren de modelos sin secuencia de cálculo. Aunque algunos estudiantes mencionan la aceleración o los pasos, omiten cómo se conectan esos datos con el cálculo de frecuencia, distancia o calorías, lo cual muestra una falta de visión funcional sistémica.
- En la Calculadora de Amor, la lógica interna se minimiza o niega. Las interacciones esperables (traducción de fechas, compatibilidad zodiacal, combinación simbólica de factores) están ampliamente ausentes. En algunos casos se mencionan parcialmente, pero hay una tendencia a negar directamente que exista alguna lógica interna, probablemente por escepticismo hacia la seriedad de la App.
- En muchos casos no se identificó ninguna interacción. Hay al menos 9 registros donde se declara que "faltan todas las interacciones reales", lo cual indica que no se concibe

ningún vínculo entre objetos, entidades y comportamientos. Esto revela una fragmentación profunda en el modelo: los elementos pueden ser identificados por separado, pero no se logra representarlos como un sistema interconectado.

6.4.1 Nivel de elaboración de los modelos mentales

La distribución de frecuencias por Apps se muestra en el gráfico de la figura 6.4.A. En este caso, se representa el porcentaje de plantillas estudiantiles que alcanzaron cada nivel de complejidad (de N0 a N4) en los cuatro componentes estructurales del modelo de funcionamiento: Objetos, Entidad, Comportamiento e Interacción. Cada barra corresponde a un componente y está segmentada por colores que indican la proporción relativa de plantillas que se ubican en cada nivel, lo que permite comparar la profundidad de elaboración alcanzada en cada componente.

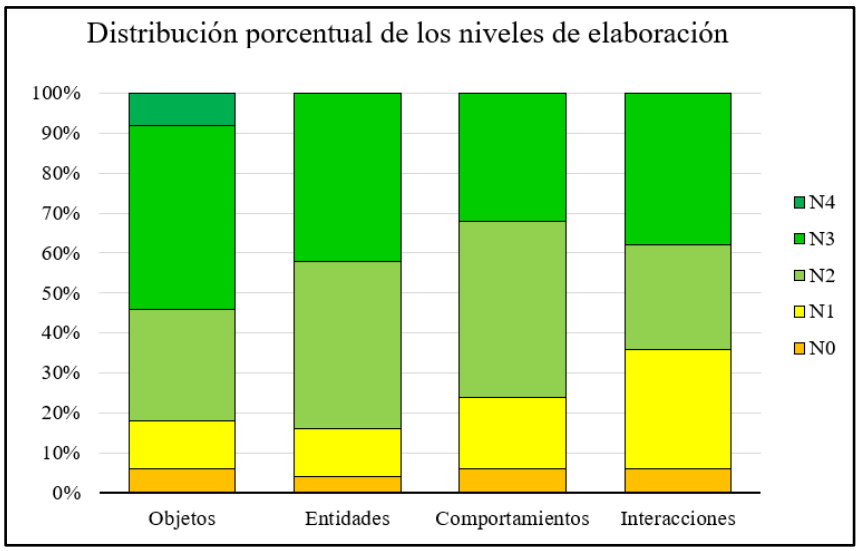

Figura 6.4.A - Distribución porcentual de los niveles de elaboración alcanzados en cada componente del modelo de funcionamiento propuesto por los estudiantes.

Se observa una tendencia general hacia los niveles intermedios y altos de representación (N2 y N3) en los cuatro componentes:

- El nivel más frecuente en general es N3 en "Objetos" (23 apariciones, 46%), lo que sugiere que los estudiantes lograron representar con mayor profundidad la estructura física o visual de los elementos involucrados en el funcionamiento de la app.
- En los componentes Entidad y Comportamiento, el nivel más frecuente es N2 (21 y 22 apariciones respectivamente, 42% y 44%), lo que indica que muchos estudiantes

fueron capaces de identificar estos elementos, pero sin detallar del todo su lógica funcional.

- La Interacción muestra una distribución más dispersa, con un número considerable de plantillas en niveles bajos (N0 y N1), pero también una alta frecuencia en N3 (19 apariciones, 38%), lo que evidencia una diversidad en la comprensión del modo en que los componentes se relacionan.
- El nivel N4, correspondiente a una representación científicamente precisa, solo fue alcanzado en el componente "Objetos" (4 apariciones, 8%), lo que evidencia la dificultad de llegar a niveles avanzados de modelización.

Por otro lado, los niveles bajos (N0 y N1) aparecen con menor frecuencia, aunque se destacan 15 casos (30%) en N1 para el componente Interacción, lo que puede interpretarse como una dificultad para comprender o representar las relaciones entre elementos.

En conjunto, los resultados sugieren que los estudiantes lograron un nivel aceptable de representación funcional en la mayoría de los componentes, especialmente en los Objetos físicos, mientras que las Interacciones entre elementos constituyen uno de los aspectos más desafiantes para modelar con precisión.

6.4.2 Nivel de elaboración de los modelos mentales por tipo de App

Para analizar la distribución de frecuencia por tipo de Apps, primero se consideró la tabla 6.4.A.

Tabla 6.4.A - Frecuencia del tipo de App de la muestra.

Tipo de App	Frecuencia
Detector de metales	14
Calculadora de amor	10
Podómetro	9
Cámara térmica	9
FotoMath	6
Sonómetro	2

Se construyeron los gráficos de distribución porcentual para todas las Apps, excepto el Sonómetro, cuya muestra es demasiado reducida. Se comienza con la Figura 6.4.B, donde se muestra la distribución porcentual de los niveles de desempeño, según la rúbrica (tabla 6.4.A), para la App Detector de Metales en los cuatro componentes del modelo de funcionamiento.

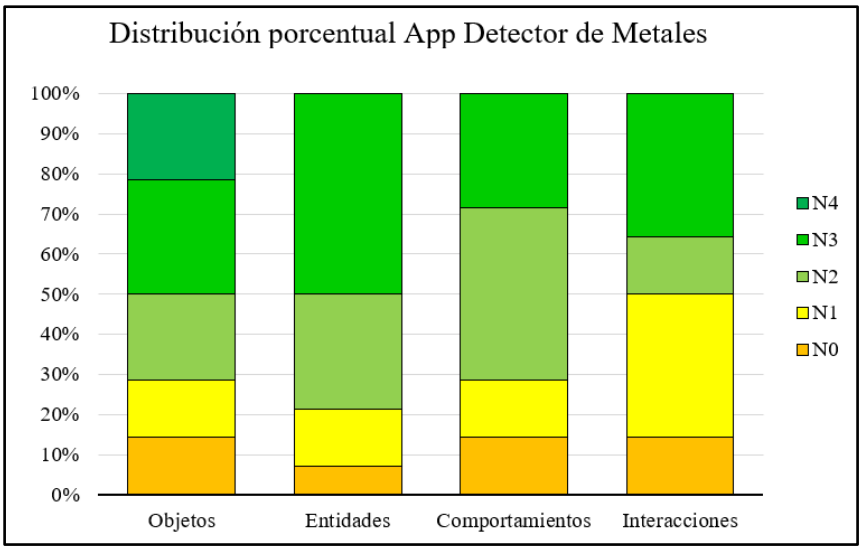

Figura 6.4.B - Distribución porcentual de los niveles de elaboración alcanzados por los estudiantes en cada componente del modelo de funcionamiento propuesto para la App Detector de Metales.

Se observa en la figura 7 (detector de metales) que el componente mejor desarrollado por el alumnado es el de Entidades, con una concentración clara en el nivel N3 (7 estudiantes, 50%). Esto indica que la mayoría de los estudiantes logra representar de manera funcional qué información entrega la App o qué elementos son procesados para generar un resultado. Sin embargo, ningún estudiante alcanza el nivel N4, lo que sugiere una comprensión funcional pero no técnica ni científicamente detallada (por ejemplo, sin referirse a umbrales o sensibilidad del campo magnético).

En el caso de los Comportamientos, los resultados se concentran en los niveles intermedios, especialmente en N2 (6 estudiantes, 43%). Esto implica que muchos estudiantes reconocen que la App realiza ciertas acciones, como medir o detectar, pero no desarrollan completamente la lógica interna del proceso, como la comparación con valores de referencia o la modulación de la señal. Solo 4 estudiantes llegan a N3, y ninguno alcanza una precisión completa (N4).

Las Interacciones presentan una distribución más dispersa. Aunque 5 estudiantes alcanzan N3, hay también un número significativo de casos en niveles bajos, 5 en N1 (36%) y 2 en N0 y N2 respectivamente (14%), lo cual revela que las relaciones entre los distintos componentes del sistema no son claramente comprendidas. Específicamente, la conexión entre el sensor, el tipo de señal, su transformación y la lógica de alerta o lectura es poco representada.

Por su parte, el componente Objetos muestra una distribución amplia y relativamente equilibrada, con 3 estudiantes alcanzando N4 (21%), el único componente donde se logra representación científicamente precisa. Esto indica que algunos estudiantes identifican

correctamente el magnetómetro como objeto sensor del sistema y reconocen su función técnica. Sin embargo, aún hay varios modelos en niveles bajos (N0, N1 y N2), lo que evidencia una disparidad en el reconocimiento de este componente esencial.

En conjunto, los datos sugieren que los estudiantes logran construir modelos con cierta funcionalidad básica, especialmente en lo que refiere a entidades (qué se mide y qué se muestra), pero presentan mayores dificultades en identificar cómo se mide (comportamientos) y cómo se relacionan los componentes internos entre sí (interacciones). El caso del objeto sensor presenta un panorama mixto: algunos estudiantes alcanzan una representación precisa, mientras que otros ni siquiera lo identifican.

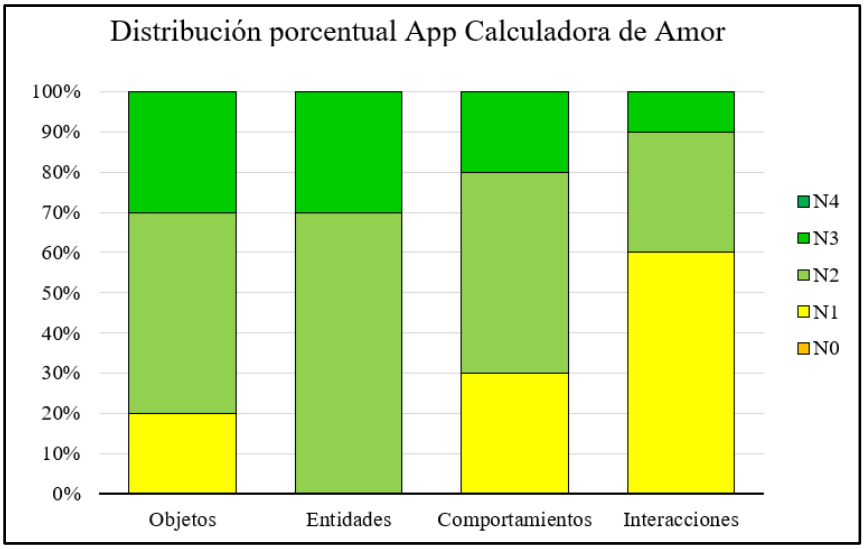

Figura 6.4.C - Distribución porcentual de los niveles de elaboración alcanzados para la App calculadora de amor.

Los resultados de la figura 6.4.C (Calculadora de amor) revelan que ningún estudiante alcanzó el nivel N4 en ninguno de los componentes, lo que sugiere una ausencia total de representaciones técnicas en los modelos. Esto puede estar relacionado con el escepticismo que genera este tipo de App, considerada lúdica o no creíble, lo que impacta en el compromiso del alumnado con su análisis estructural.

La categoría de Entidades es la que muestra mayor fortaleza relativa, con 7 estudiantes en el nivel N2 y 3 en N3. Esto indica que muchos identificaron los datos que la App devuelve (como porcentajes de compatibilidad), pero solo algunos lograron representar estas entidades como parte funcional del sistema.

En cuanto a los Objetos, hay una presencia moderada en los niveles N2 (5) y N3 (3), lo cual señala que los estudiantes reconocen que se introducen datos como nombres o fechas, pero

no siempre los vinculan adecuadamente con la lógica de funcionamiento. Tampoco distinguen con claridad entre objetos digitales (datos del usuario) y estructuras internas como la interfaz.

El componente de Comportamientos presenta una distribución centrada en N1 (3) y N2 (5), lo que evidencia que los estudiantes perciben que la App realiza algún tipo de procesamiento, pero no logran identificar cómo ni bajo qué lógica se produce la compatibilidad. La escasa presencia en N3 (2) y la ausencia de N4 refuerzan la idea de que la lógica interna del sistema es desconocida para los estudiantes o, directamente, considerada inexistente.

La situación más crítica se da en el componente de Interacción, donde la mayoría de los estudiantes permanece en los niveles bajos, 6 en N1 y solo 1 alcanza el nivel N3. Esto indica una gran dificultad para representar las relaciones entre objetos, entidades y comportamientos, es decir, cómo se conecta el ingreso de datos con el proceso y la generación de un resultado.

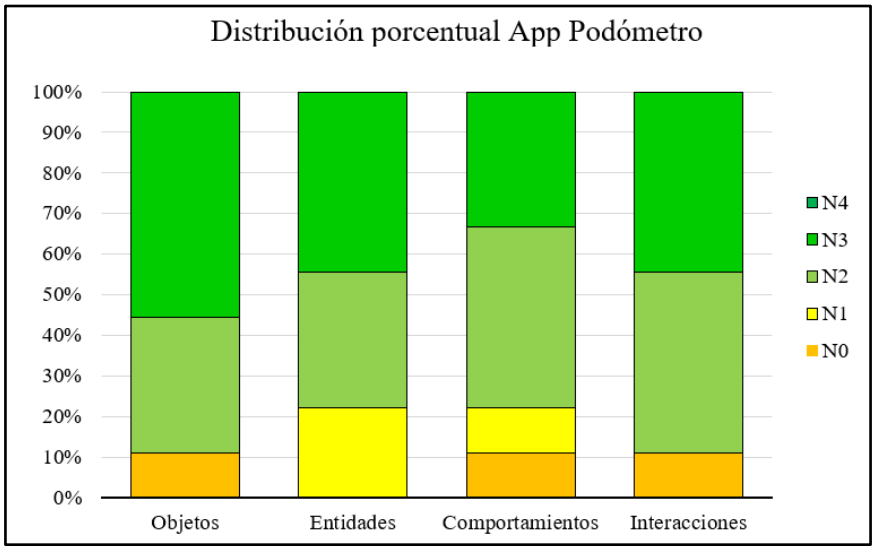

Figura 6.4.D - Distribución porcentual de los niveles de elaboración alcanzados para la App Podómetro.

En el gráfico de la figura 6.4.D (Podómetro) se aprecia una tendencia positiva hacia el nivel N3, lo que indica que muchos estudiantes fueron capaces de construir modelos funcionales, aunque sin alcanzar un nivel técnico riguroso.

El componente mejor representado es el de Objetos, donde 5 estudiantes llegan a N3, reconociendo correctamente que el podómetro depende de un sensor (acelerómetro) para registrar movimiento. Además, 3 estudiantes se sitúan en N2, lo que sugiere que también hubo quienes lograron una identificación parcial o general del sensor, aunque sin especificidad técnica.

En cuanto a las Entidades, también se concentra la mayoría en los niveles N2 (3) y N3 (4), lo que refleja una comprensión aceptable de los datos que la App genera, como conteo de pasos, distancia estimada o calorías. Aun así, 2 estudiantes permanecen en N1, posiblemente mencionando los resultados sin vincularlos adecuadamente con las variables implicadas (por ejemplo, frecuencia de pasos, peso del usuario, longitud del paso).

En la categoría de Comportamientos, se repite una distribución similar: 4 estudiantes en N2 y 3 en N3. Esto indica que se logra representar el tipo de procesamiento que realiza la App (por ejemplo, convertir aceleración en pasos), aunque con limitaciones en la explicación del proceso completo o en la incorporación de factores como el ritmo o el cálculo energético.

Finalmente, el componente de Interacciones muestra también una concentración en N2 y N3 (4 estudiantes en cada uno), evidenciando que varios modelos logran representar cómo se conectan los sensores con los cálculos y resultados. No obstante, hay aún 1 caso en N0, lo que indica que no todos los estudiantes logran visualizar las relaciones internas del sistema.

En conjunto, el gráfico evidencia que la mayoría construye modelos funcionales que logran representar adecuadamente las partes y el funcionamiento general de la App. Se trata de un caso donde los estudiantes muestran un mejor desempeño global en comparación con otras Apps, probablemente porque el funcionamiento del podómetro es más creíble, visible y cercano a sus experiencias cotidianas, lo que facilita una comprensión más profunda del sistema.

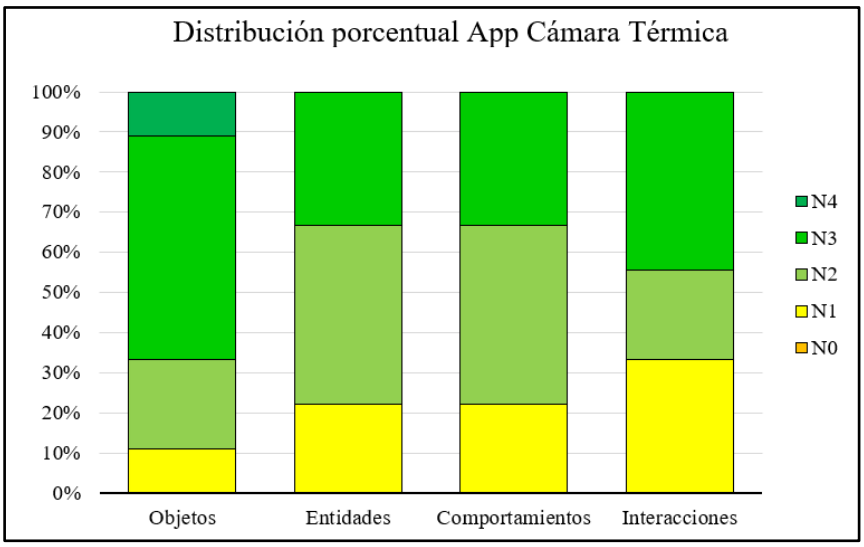

Figura 6.4.E - Distribución porcentual de los niveles de elaboración alcanzados para la App Cámara Térmica.

En el gráfico de la figura 6.4.D las Entidades se concentran en N2 (4 estudiantes) y N3 (3 estudiantes), lo que indica que los estudiantes identifican que la App muestra zonas de calor mediante colores, pero no siempre distinguen con precisión los factores que determinan ese resultado (como el tipo de filtro, condiciones de luz o asignación de colores térmicos). Ningún estudiante alcanza el nivel N4, lo que muestra una comprensión limitada desde una perspectiva científica.

El componente de Comportamientos muestra una distribución similar, con mayoría en N2 (4 estudiantes) y algunos en N3 (3), lo que evidencia que los estudiantes reconocen que la App transforma la imagen en datos térmicos, pero sin detallar del todo el proceso técnico, como el ajuste de brillo, la aplicación del filtro o la conversión de señal. Esto indica una comprensión parcial del mecanismo de funcionamiento.

Respecto a las Interacciones, el patrón también se concentra en los niveles medios: 4 estudiantes alcanzan N2 y otros 4, N3, lo que significa que algunos estudiantes logran representar cómo los elementos se conectan (por ejemplo, cómo la imagen captada por la cámara se procesa y se convierte en una visualización térmica), pero sin explicar adecuadamente las relaciones entre componentes técnicos como sensores, condiciones de luz y asignación de color. Solo un caso se sitúa en N0, lo que indica que la mayoría al menos intenta establecer vínculos funcionales.

En conjunto, la App Cámara Térmica es una de las más exitosamente modeladas por el alumnado en términos funcionales, especialmente en lo que respecta a la identificación del objeto sensor (cámara) y la visualización de resultados (imagen térmica). Sin embargo, los niveles altos de precisión científica (N4) son casi inexistentes, lo que muestra una brecha entre la comprensión funcional cotidiana y el modelado técnico riguroso. Esta situación sugiere un buen punto de partida para promover el pensamiento sistémico y técnico a partir de una App familiar y visualmente atractiva, que puede despertar la curiosidad y ofrecer un puente entre experiencia y explicación.

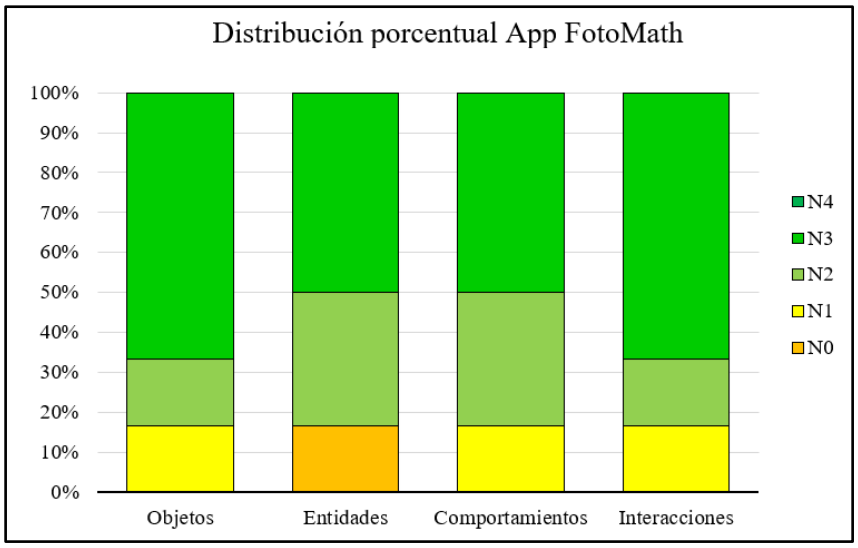

Figura 6.4.F - Distribución porcentual de los niveles de elaboración alcanzados para la App FotoMath.

En el gráfico de la figura 6.4.F (FotoMath) se observa una alta concentración en el nivel N3 en los cuatro componentes, lo que indica que la mayoría de los estudiantes logró construir modelos funcionales sobre cómo opera la App. En particular, los componentes Objetos e Interacciones presentan la mayor frecuencia en N3 (4 estudiantes), lo que sugiere que estos aspectos fueron comprendidos con mayor claridad y profundidad relativa.

En el componente Objetos, los estudiantes reconocen adecuadamente que la App utiliza la cámara como sensor, y que los objetos de entrada son las expresiones matemáticas captadas visualmente. Este nivel de comprensión se refleja también en las Interacciones, donde 4 estudiantes logran representar correctamente la secuencia de procesos: captura de imagen → reconocimiento de símbolos → resolución → visualización del resultado. Esto demuestra una visión sistémica funcional del proceso, aunque sin llegar al nivel de precisión técnica (N4), donde deberían incluirse detalles como OCR (reconocimiento óptico de caracteres), CAS (sistema de álgebra computacional), o el tipo de estructuras algorítmicas que intervienen.

En cuanto a las Entidades y Comportamientos, también se observa una presencia sólida en el nivel N3 (3 estudiantes en cada caso), lo que indica que varios estudiantes comprenden qué información produce la App (por ejemplo, resultados, pasos de resolución) y qué procesos realiza internamente (como interpretar operaciones, aplicar reglas matemáticas, resolver paso a paso). Sin embargo, en N2 también hay frecuencia (2 casos en Entidad y Comportamiento), lo que muestra que otros estudiantes solo logran una descripción parcial o más superficial de estos aspectos.

La ausencia de casos en N4 en todos los componentes refleja que ningún estudiante representa el funcionamiento técnico con precisión científica, lo que era esperable dado el nivel educativo. Aun así, el hecho de que la mayoría esté en N3 indica que PhotoMath es una de las Apps mejor comprendidas funcionalmente por los estudiantes.

Una posible explicación de estos resultados es que el funcionamiento de PhotoMath, aunque basado en tecnología avanzada, es transparente y explícito en su interfaz: los estudiantes ven directamente cómo la cámara capta la ecuación y cómo la App devuelve el resultado junto con los pasos. Esto permite construir modelos mentales relativamente completos, aun sin comprender los algoritmos matemáticos que operan en segundo plano.

En conjunto, el gráfico sugiere que los estudiantes tienen una comprensión sólida pero no profunda del funcionamiento de PhotoMath. Son capaces de identificar sus componentes, describir su funcionamiento general e incluso articular las relaciones entre ellos, aunque no llegan a representar los mecanismos computacionales o simbólicos subyacentes con detalle técnico. Esto convierte a PhotoMath en una App con alto potencial didáctico para promover el pensamiento funcional y el modelado sistémico desde una experiencia digital concreta.

6.4.3 Nivel de elaboración de los modelos mentales por edad

Considerando la tabla 6.4.B de distribución de frecuencia por edad, se decidió mirar como varían las frecuencias según la edad.

Tabla 6.4.B - Frecuencia de la edad del alumnado en la muestra.

Edad	Frecuencia
13 años	10
14 años	10
15 años	13
17 años	17

Al analizar los cuatro gráficos que representan la distribución de niveles de modelado según la edad del alumnado (13, 14, 15 y 17 años), se observa una evolución en la profundidad y precisión con la que se representan los componentes de un modelo de funcionamiento de una App.

Para el grupo de 13 años, el gráfico muestra una concentración en los niveles N2 y N3 para todos los componentes, especialmente en Entidades, Comportamientos e Interacciones. Esto sugiere que el alumnado de esta edad logra construir modelos funcionales, reconociendo qué datos se introducen, qué procesos internos realiza la App y qué resultados se obtienen. No

obstante, también se observan varias ocurrencias en los niveles bajos (N1 y N0), particularmente en Comportamientos e Interacciones, lo que indica una comprensión aún fragmentaria del sistema.

En el caso de los 14 años, el patrón general se mantiene en los niveles N2 y N3, aunque hay un ligero retroceso en Interacciones, que concentra 5 casos en N1. Comportamientos presenta su punto más alto en N2 (6 estudiantes), y Objetos y Entidades tienen una distribución más dispersa. No se registra ninguna representación en N4, al igual que en el grupo anterior, lo que refuerza la idea de que el pensamiento técnico aún no se consolida.

El grupo de 15 años muestra una estructura similar, pero con una distribución más simétrica entre los niveles N1, N2 y N3. Destaca que Entidades es el componente con mejor desempeño, con 7 casos en N2 y 4 en N3, lo que sugiere que a esta edad en la muestra se afianza una mejor comprensión de los datos de salida. Sin embargo, Interacción y Comportamiento también presentan un número considerable de respuestas en N1, lo cual evidencia que las relaciones internas del sistema todavía no son completamente comprendidas.

El gráfico de 17 años representa el punto más alto de desempeño general. La mayoría de las respuestas se concentra en el nivel N3 en los cuatro componentes, especialmente en Objetos e Interacciones (9 casos cada uno), y Comportamientos (7 casos). Llama la atención que por primera vez aparecen casos en el nivel N4 (3 en Objetos), lo que indica que algunos estudiantes mayores ya comienzan a alcanzar una comprensión técnicamente precisa, al menos en la identificación de los sensores implicados en el sistema. Aunque persisten algunas respuestas en N0 y N1, su proporción es menor y tiende a concentrarse en casos aislados.

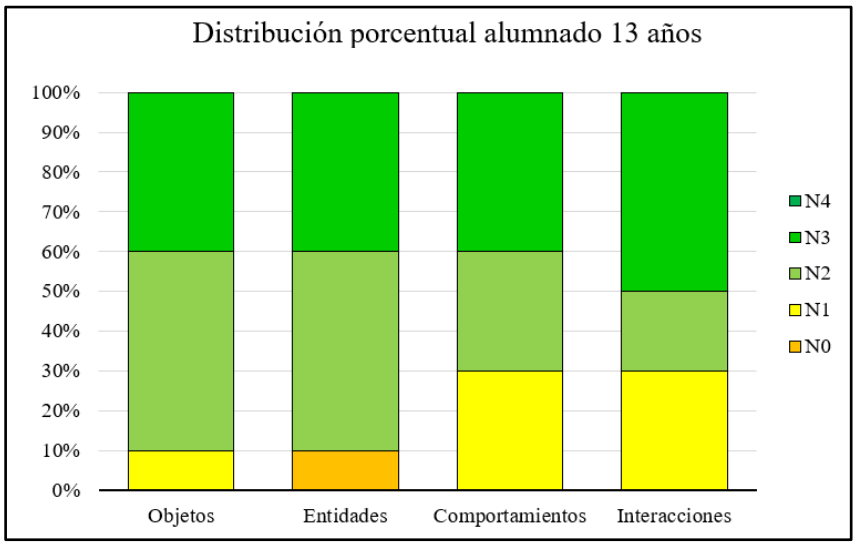

Figura 6.4.G - Distribución porcentual de los niveles de elaboración para el alumnado de 13 años.

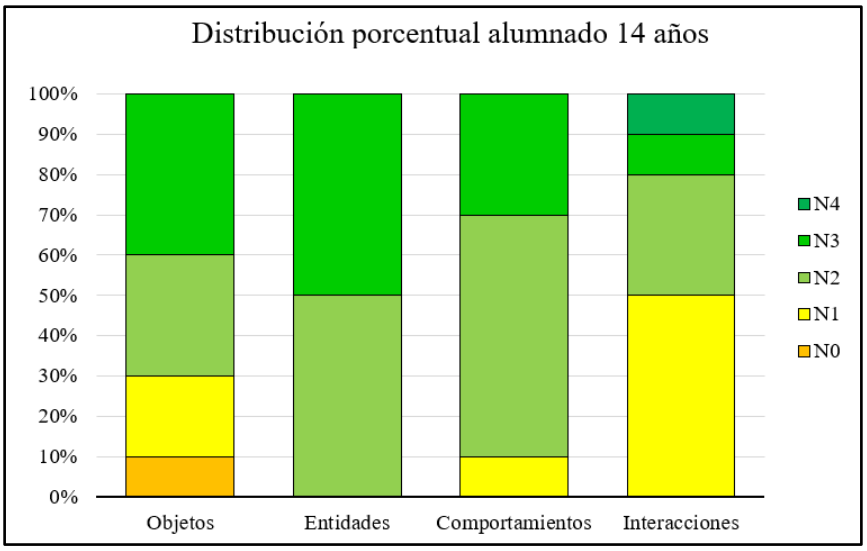

Figura 6.4.H - Distribución porcentual de los niveles de elaboración para el alumnado de 14 años.

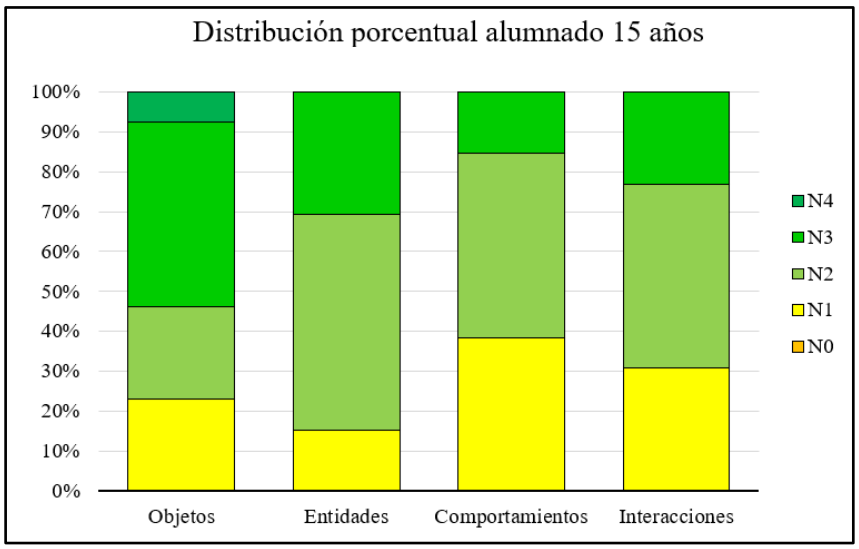

Figura 6.4.I - Distribución porcentual de los niveles de elaboración para el alumnado de 15 años.

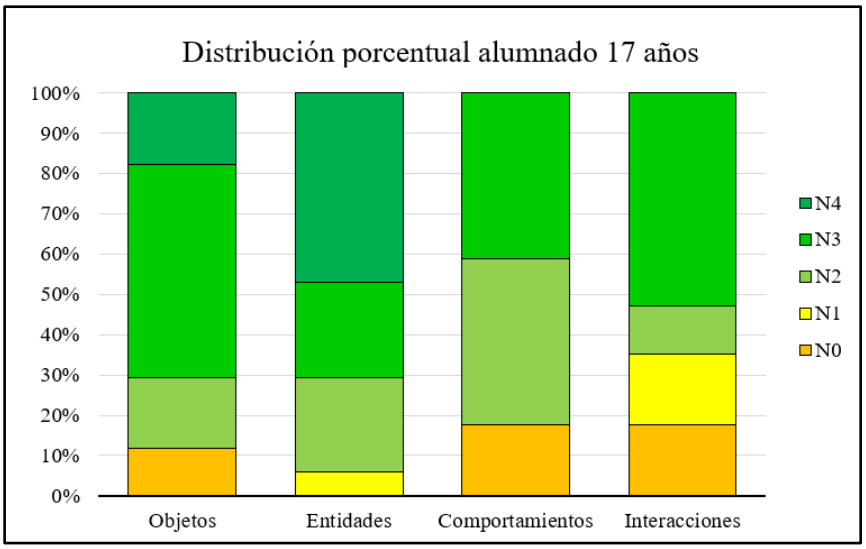

Figura 6.4.J - Distribución porcentual de los niveles de elaboración para el alumnado de 17 años.

Al comparar los cuatro gráficos, se puede observar una tendencia general de progresiva complejidad en el modelado a medida que aumenta la edad. Esta evolución se traduce en:

- Una mayor frecuencia en el nivel N3 a partir de los 15 años.
- La aparición de niveles N4 únicamente en el grupo de 17 años.
- Una reducción progresiva de respuestas en N0 y N1, especialmente en los componentes de Entidad y Comportamiento.
- Una mejora en la representación de Interacciones en los modelos de los estudiantes mayores, que pasa de tener respuestas concentradas en niveles bajos a dominar el nivel N3.

Estos patrones permiten concluir que el desarrollo del pensamiento funcional y sistémico vinculado a la modelización tecnológica mejora con la edad, probablemente como efecto de una mayor madurez cognitiva, una experiencia escolar acumulada más amplia y una exposición más consciente al uso de tecnologías. Los estudiantes de mayor edad no solo identifican con mayor facilidad los objetos o entidades de entrada y salida, sino que también representan mejor los procesos internos y las conexiones entre componentes, lo que apunta a un fortalecimiento del pensamiento estructural.

No obstante, esta comparación presenta limitaciones importantes. En primer lugar, el tamaño de muestra por edad no se especifica, y pequeñas diferencias numéricas pueden alterar la percepción de las tendencias. Además, no se ha considerado el tipo de App modelada, que podría influir en la complejidad percibida por los estudiantes.

6.5 Conclusiones del estudio 3

Algunas cosas interesantes revelo el análisis de elementos omitidos en los modelos. Las interfaces gráficas son sistemáticamente subrepresentadas en los modelos de los estudiantes es que estos tienden a naturalizar la interfaz como parte del entorno, no como un componente funcional del sistema, es un elemento neutro.

Es decir, la interfaz se percibe como algo dado, automático o “transparente”, que simplemente “está ahí” para permitir el uso de la App, pero no como un objeto técnico que media activamente en la entrada de datos y que cumple una función concreta en la arquitectura del programa. Esta naturalización podría tener su origen en la experiencia cotidiana del alumnado con dispositivos móviles, donde el uso fluido e intuitivo de la interfaz se da sin necesidad de comprender su funcionamiento interno ni sus implicancias técnicas.

Además, es posible que influya el modo en que suele enseñarse tecnología o informática en contextos escolares, donde se enfatiza más el uso instrumental de las aplicaciones que la reflexión sobre sus componentes estructurales o los procesos invisibles que ocurren entre usuario y software. Por tanto, al ser una parte "visible pero no pensada", la interfaz se vuelve invisible en el plano del modelado.

Se cree que la causa de esta invisibilización generalizada de las entidades técnicas es que los estudiantes modelan el funcionamiento de las Apps desde una perspectiva de usuario y no de sistema. Es decir, focalizan su atención en los objetos visibles o en las acciones que realizan (lo que tocan, lo que ven, el resultado final), pero no acceden a los procesos intermedios que transforman esos datos. Este vacío podría estar relacionado con una falta de experiencias educativas que promuevan la exploración de los algoritmos, condiciones y parámetros ocultos en el funcionamiento de herramientas digitales cotidianas.

El pensamiento sistémico, que implica entender cómo diferentes elementos interactúan dinámicamente para producir un resultado, es poco trabajado en la educación convencional. A ello se suma que el uso cotidiano de las Apps no requiere comprender ni visualizar estas interacciones, lo que refuerza su invisibilidad. Así, aunque los estudiantes puedan describir lo que hace una App, no logran articular cómo lo hace, ni qué relaciones internas técnicas permiten esa transformación, hay desconocimiento generalizado de tecnologías emergentes.

Los estudiantes carecen de un modelo mental del “interior” de las Apps. El diseño pedagógico tradicional privilegia el uso externo de la tecnología (inputs y outputs), pero no estimula la representación de procesos intermedios, ni el pensamiento algorítmico sobre cómo

se transforma la información. Esta visión funcionalista, centrada en la utilidad y no en el mecanismo, es una explicación de por qué los comportamientos, incluso los más básicos como aplicar un filtro o comparar con un umbral, se vuelven invisibles en la representación del sistema.

En cuanto a la rúbrica, el proceso de iteración para ajustarla mostro que en la práctica, es común que los estudiantes no pasen de modelos completamente incorrectos a modelos totalmente correctos de forma inmediata, sino que transiten por una fase en la que combinan aciertos y errores. Por esta razón, la graduación de los criterios de la rúbrica final consideró estados intermedios. Al contar explícitamente el número de objetos y entidades correctos e incorrectos en cada modelo, podemos establecer niveles de desempeño que reflejen con mayor fidelidad el progreso del estudiante. Un estudiante puede representar múltiples objetos, pero si la mayoría son incorrectos, su modelo sigue siendo deficiente. Por el contrario, si empieza a aumentar la proporción de objetos correctos, se evidencia un mejor nivel de comprensión, esta cuestión de diferenciación entre cantidad y precisión permite determinar cómo se fortalece la precisión en la representación de objetos sin castigar con la rúbrica a estudiantes que están en fases de desarrollo.

Los niveles de desempeño para la rúbrica de la Tabla 6.4.A establecen una progresión en la identificación de objetos, entidades, comportamientos e interacción, son tales que hay una progresión desde lo incorrecto hasta lo correcto. La rúbrica está diseñada para reflejar el proceso mediante el cual los estudiantes evolucionan desde modelos completamente erróneos hasta representaciones cada vez más precisas. Por ejemplo para el criterio objetos, en su nivel más bajo, la ausencia de identificación de objetos caracteriza el nivel 0, donde no se reconoce ningún elemento del sistema. A medida que el estudiante avanza, el nivel 1 evidencia intentos de identificar objetos, aunque todavía con errores significativos. Posteriormente, el nivel 2 introduce la coexistencia de objetos correctos e incorrectos, lo que refleja una comprensión parcial del modelo en construcción. En el nivel 3, el estudiante logra distinguir los objetos correctos, aunque aún puede omitir algunos elementos clave del sistema. Finalmente, el nivel 4 representa la competencia total en la identificación de los objetos del modelo real, asegurando una representación precisa y sin errores conceptuales.

Los niveles de desempeño de cada criterio están fundamentados en investigaciones previas sobre modelos (Louca et al., 2011) y en el análisis de frecuencia de combinaciones de las respuestas de los estudiantes. El análisis revelo que los estudiantes no transitan de manera inmediata desde la representación de objetos incorrectos hasta la construcción de modelos completamente correctos. En lugar de ello, los datos evidencian que muchas de sus respuestas

combinan categorías correctas e incorrectas, lo que llevó a la incorporación de un nivel intermedio en la rúbrica que reflejara esta etapa de transición. Este hallazgo se sustenta en las tendencias identificadas a partir del análisis de la frecuencia de combinaciones entre categorías correctas e incorrectas, lo que refuerza la necesidad de una evaluación progresiva y ajustada a la realidad del aprendizaje.

Al comparar los resultados por tipo de App, se observa un patrón general en el que las Apps que operan con sensores físicos y generan resultados observables en tiempo real (como detector de metales, podómetro, cámara térmica y PhotoMath) tienden a ser modeladas con mayor precisión funcional por los estudiantes, especialmente en componentes como Objetos y Entidades, mientras que aquellas cuya lógica interna es percibida como arbitraria o lúdica (como la calculadora de amor) presentan modelos más pobres, centrados en niveles bajos y con escasa representación de comportamientos o interacciones. Esto sugiere que la credibilidad funcional percibida por el estudiante y la transparencia del funcionamiento visible son factores clave que favorecen una representación más estructurada y completa del sistema.

La comparación por edad revela una tendencia progresiva: a mayor edad del alumnado, mayor es la capacidad para construir modelos funcionales y estructurados del funcionamiento de una App, con un predominio del nivel N3 en los grupos de 15 y 17 años, y la aparición exclusiva de representaciones técnicamente precisas (N4) en los estudiantes de 17 años; esta evolución sugiere que el desarrollo del pensamiento sistémico y funcional en contextos tecnológicos se ve favorecido por la madurez cognitiva, la acumulación de experiencias escolares previas y una mayor familiaridad con el uso consciente de dispositivos digitales, permitiendo una comprensión más completa y articulada de los componentes, procesos e interacciones implicadas en el sistema.

CAPÍTULO VII

CONCLUSIONES, LIMITACIONES E IMPLICACIONES

El Capítulo VII presenta las conclusiones generales de la tesis a partir del análisis integrado de los tres estudios. Se abordan también los vínculos entre estas prácticas, señalando patrones comunes, sinergias metodológicas y puntos de convergencia identificados en las distintas implementaciones.

Se explican las limitaciones de la investigación y se discuten las implicaciones educativas de los resultados, proponiendo orientaciones para el diseño de tareas didácticas con Apps, y recomendaciones para futuras investigaciones que profundicen en la confiabilidad científica escolar.

El presente capítulo final expone las conclusiones generales y específicas derivadas de los tres estudios que conforman esta investigación. Su propósito es sintetizar los hallazgos obtenidos en cada uno de ellos, valorar el grado de cumplimiento de los objetivos planteados y ofrecer una visión integrada de los resultados. Para ello, el capítulo se organiza en tres apartados principales. En primer lugar, se presentan las conclusiones específicas de cada estudio (7.1.1 a 7.1.3), redactadas en correspondencia directa con sus respectivos objetivos. A continuación, se desarrollan conclusiones transversales (7.1.4), que buscan articular hallazgos comunes y diferencias relevantes entre los estudios, generando una comprensión más amplia del conjunto. Posteriormente, se exponen las limitaciones metodológicas y de alcance identificadas a lo largo del proceso investigativo (7.2). Finalmente, se discuten las implicaciones didácticas más significativas (7.3), orientadas a mejorar la enseñanza de las prácticas científicas y la alfabetización tecnológica desde una perspectiva crítica e interdisciplinaria.

7.1 Conclusiones derivadas de los estudios

Antes de abordar los elementos transversales que vinculan los tres estudios, se ha considerado presentar las conclusiones específicas de cada uno de ellos por separado. De este modo, se espera facilitar una lectura más ordenada y que respete la lógica interna de cada estudio, atendiendo al tipo de producto final analizado, al nivel educativo del alumnado y a los marcos teóricos que guiaron su evaluación. A continuación, se detallan las conclusiones particulares del estudio 1, centrado en el análisis del desempeño del alumnado en prácticas científicas de indagación y argumentación a partir de la producción de videos argumentativos. Luego, se continuará con los estudios 2 y 3.

7.1.1 Conclusiones específicas del estudio 1

A continuación, se presentan las principales conclusiones del Estudio 1, organizadas en torno a ideas clave que sintetizan los hallazgos considerados más relevantes. Esta estructura pretende facilitar la comprensión del alcance logrado por esta primera implementación, destacando aquellos aspectos que resultan significativos tanto desde una perspectiva metodológica como educativa.

Caracterización detallada del desempeño en indagación mediante combinaciones de criterios procedimentales. El estudio logró construir una rúbrica coherente que permitió

analizar los niveles de desempeño en indagación científica del alumnado. A partir de 76 videos estudiantiles, se identificaron patrones en el diseño experimental, tales como el uso de calibración, repetición de mediciones y determinación de los límites de la App como un medidor. Esto permitió generar una escala progresiva (I0–I4) basada en combinaciones observadas en la práctica, con capacidad para discriminar niveles de complejidad crecientes en la ejecución de tareas de indagación que sirve como un tipo de base para generar instrumentos de evaluación de modo general.

Evaluación progresiva de la argumentación científica basada en el modelo de Toulmin. Se diseñó una rúbrica específica para identificar el nivel de argumentación del alumnado en sus productos audiovisuales. A través del modelo de Toulmin y sus seis componentes clave (pruebas válidas, inválidas, aserción, razonamiento, moduladores y refutación), fue posible codificar y categorizar los argumentos en una escala de A0 a A4. Este enfoque permitió evaluar no solo la presencia de una estructura argumentativa, sino también cierto grado de complejidad.

La relación entre indagación y argumentación es heterogénea y no lineal. El cruce de niveles entre indagación (I) y argumentación (A) reveló múltiples combinaciones posibles. Se identificaron casos donde una indagación compleja no se tradujo en argumentos sólidos y, por el contrario, estudiantes con baja indagación ofrecieron argumentos estructurados a partir de datos informales. Esta variabilidad refuerza la idea de que ambas prácticas científicas deben ser trabajadas de forma articulada, pero diferenciada, ya que no evolucionan de manera paralela.

Las restricciones y refutaciones no siempre emergen de indagaciones rigurosas. El componente “refutación (restricciones)” en la argumentación no mostró una correspondencia sistemática con el tratamiento de “límites operativos” en la indagación (criterio C). Algunos estudiantes fueron capaces de introducir restricciones discursivas sin haber desarrollado una indagación avanzada. Esto indica que la capacidad para introducir matices en los argumentos no depende únicamente de las prácticas de indagación, sino también de habilidades cognitivas específicas que podrían requerir enseñanza explícita, como por ejemplo la capacidad para reflexionar sobre el propio proceso de pensamiento o de indagación realizado, el desarrollo del pensamiento crítico y epistémico requiere andamiaje docente, ya que no basta con poner al alumnado a experimentar; también deben aprender a reflexionar, justificar, dudar y matizar.

Las decisiones didácticas del profesorado inciden en los resultados. El análisis desagregado por docente mostró que la implementación del proyecto, el uso de escalas de fiabilidad y el tipo de guía ofrecida influyeron en el tipo y nivel de desempeño del alumnado. Aulas donde se utilizó un mayor andamiaje experimental mostraron mayor diversidad de combinaciones de desempeño, mientras que otras tendieron a concentrarse en niveles medios. Estos hallazgos confirman que el diseño instruccional del docente tiene un impacto directo en el desarrollo de las competencias científicas.

La mayoría del alumnado se sitúa en niveles medios de desempeño. En general, tanto en indagación como en argumentación, la mayoría del estudiantado se ubicó en niveles intermedios (I2–I3 y A2–A3). Esto sugiere que el proyecto ofrece oportunidades para desarrollar competencias científicas, pero también revela que se requieren ajustes metodológicos para favorecer trayectorias más ambiciosas. Solo un número reducido de estudiantes alcanzó los niveles más altos (I4–A4), lo que indica el potencial de mejora del enfoque didáctico.

La rúbrica basada en combinación de criterios proporciona un marco válido y didácticamente útil. La decisión metodológica de construir la rúbrica de indagación a partir de combinaciones de criterios observables permitió evaluar el desempeño como un fenómeno integrado. Este enfoque ofreció una mayor validez educativa, al capturar la complejidad del trabajo experimental del alumnado, y facilitó el diseño de niveles, útiles para retroalimentar y orientar el aprendizaje.

El modelo de Toulmin, aplicado desde una lógica de agregación progresiva, resulta eficaz para contextos escolares. El uso de una rúbrica argumentativa basada en la acumulación progresiva de componentes resultó adecuado para caracterizar niveles de argumentación sin fragmentar el análisis. Esta estrategia permitió establecer relaciones significativas con la calidad de la indagación y evitó una sobrecarga analítica innecesaria. Además, se alinea con modelos de progresión argumentativa validados por la literatura, lo que refuerza su aplicabilidad en investigaciones similares.

La mirada competencial exige observar no solo los productos, sino también los procesos. Si bien este estudio se centró en los productos audiovisuales finales, los hallazgos sugieren la necesidad de futuras investigaciones que indaguen los procesos cognitivos y colaborativos involucrados en la producción de estos resultados. Comprender cómo el alumnado toma decisiones metodológicas, negocia significados y articula pruebas podría

enriquecer las propuestas didácticas para fomentar un aprendizaje más profundo y autónomo de las prácticas científicas.

7.1.2 Conclusiones específicas del estudio 2

El segundo estudio del proyecto App Checkers, realizado con estudiantes de bachillerato, permitió profundizar en las estrategias empleadas por el alumnado para evaluar la confiabilidad de aplicaciones móviles. A continuación, se presentan las principales conclusiones derivadas del análisis cualitativo de los pósteres científicos elaborados por los participantes, organizadas en torno a ideas clave que permiten visualizar el cumplimiento de los objetivos propuestos.

Predominio de la comparación con valor esperado como estrategia principal de validez. El análisis de los 19 pósteres reveló que la mayoría del alumnado utilizó la comparación con un valor esperado como estrategia principal para evaluar la validez de las Apps. Esta elección puede explicarse por la cercanía de esta estrategia con las experiencias cotidianas de los estudiantes y por su aparente simplicidad operativa. Aunque esta elección permite evaluar razonablemente si la App se aproxima a un comportamiento esperable, también limita el alcance de la investigación, ya que deja de lado enfoques más elaborados como la triangulación o el uso de instrumentos de referencia.

Escasa integración del concepto de fiabilidad en los diseños experimentales. El segundo objetivo del estudio, centrado en analizar si el alumnado incorporaba procedimientos para evaluar la fiabilidad de las mediciones, mostró un bajo grado de cumplimiento. En general, los pósteres no evidencian procedimientos consistentes para estimar la estabilidad de las mediciones, como la repetición de lecturas y el análisis de la dispersión de los datos. Esto sugiere que el alumnado no ha desarrollado suficientemente habilidades que integren el pensamiento estadístico con las prácticas de medición experimental, lo que obstaculiza la comprensión profunda del concepto de fiabilidad.

Las estrategias de validez y los niveles de fiabilidad aparecen como dimensiones desconectadas. No se observaron patrones de asociación entre las estrategias de validez utilizadas y los niveles de fiabilidad alcanzados. Esto implica que el alumnado tiende a considerar ambos aspectos de forma independiente, lo que limita la construcción de una visión sistémica sobre la confiabilidad de una medición. Esta desconexión refuerza la necesidad de diseñar intervenciones didácticas que articulen de manera explícita los conceptos de validez y

fiabilidad como dimensiones complementarias y mutuamente necesarias en la indagación científica.

El pensamiento estadístico y el tratamiento de datos siguen siendo un desafío formativo. El estudio muestra una baja presencia de procedimientos de análisis de datos, que permitan caracterizar la incertidumbre asociada a una medición, independiente que el profesor guía lo solicitó al alumnado. La escasa utilización de promedios, análisis de dispersión o estimaciones de error refleja una dificultad generalizada en la aplicación de nociones estadísticas básicas. Esto apunta a una brecha formativa importante que debe ser abordada mediante estrategias específicas centradas en el análisis cuantitativo de resultados y la comprensión de su uso.

La tarea de evaluación de Apps favorece la reflexión crítica y el desarrollo epistémico. A pesar de las dificultades detectadas en la implementación de estrategias de fiabilidad, el estudio confirma que la propuesta didáctica ofrece un escenario para el desarrollo de habilidades científicas y epistémicas. A través de la evaluación de herramientas digitales reales, el alumnado se enfrenta a problemas abiertos, toma decisiones metodológicas y reflexiona sobre la confiabilidad de los resultados. Esta aproximación permite trascender las prácticas experimentales tradicionales, promoviendo una comprensión más profunda de la indagación científica en contextos tecnológicos emergentes.

Se requiere una enseñanza explícita e integrada de validez y fiabilidad. Finalmente, los resultados sugieren que para lograr una comprensión integral de la confiabilidad es necesario diseñar estrategias didácticas que aborden simultáneamente la validez y la fiabilidad, incorporando experiencias prácticas, enseñanza de análisis estadístico, y retroalimentación formativa. La mejora en la evaluación de Apps por parte del alumnado dependerá de que ambas dimensiones sean enseñadas como elementos interdependientes y no como competencias aisladas.

7.1.3 Conclusiones específicas del estudio 3

El tercer estudio del proyecto App Checkers se centró en explorar los modelos mentales construidos por el alumnado de entre 13 y 17 años para representar el funcionamiento de diversas aplicaciones móviles. A partir del análisis de 50 plantillas completadas en distintos contextos escolares, se obtuvieron hallazgos relevantes sobre cómo los estudiantes conciben los componentes, transformaciones e interacciones que constituyen el funcionamiento técnico

de una App. Las conclusiones más destacadas se organizan a continuación según los principales ejes de análisis.

La interfaz gráfica es invisibilizada como componente funcional. Hubo una omisión de la interfaz gráfica en los modelos elaborados por los estudiantes. Aunque esta cumple una función como puente entre el usuario y el sistema, la mayoría del alumnado no la reconoce como un objeto técnico activo, sino que la asume como un elemento neutro del entorno. Esta naturalización puede explicarse por la familiaridad cotidiana con las Apps y por la forma instrumental en que se enseña la tecnología en el sistema escolar o de la vida diaria, donde raramente se abordan los procesos subyacentes al uso de herramientas digitales.

El alumnado modela desde una lógica de usuario, no desde una lógica de sistema. Los modelos estudiantiles se construyen, en gran medida, desde la experiencia directa con la App, priorizando lo que el usuario ve, toca o recibe. En consecuencia, se omiten muchos de los procesos intermedios que transforman la entrada en salida, como filtros, algoritmos o condiciones técnicas. Esta perspectiva impide una comprensión profunda de los mecanismos internos y refleja una ausencia de pensamiento algorítmico, lo que limita el desarrollo de un modelo mental estructurado y funcional.

El pensamiento sistémico aparece poco en los modelos mentales del alumnado. La mayoría de los estudiantes logra identificar objetos o entidades aisladas, pero no logra articular entre ellos procesos complejos ni secuencias funcionales. Las interacciones entre componentes son especialmente difíciles de representar. Este resultado evidencia que el pensamiento sistémico —entendido como la capacidad de integrar múltiples elementos en una red funcional— no está suficientemente promovido en las experiencias escolares actuales. En consecuencia, aunque el alumnado pueda describir lo que una App hace, le cuesta explicar cómo lo hace.

La rúbrica permite representar trayectorias de aprendizaje, incluyendo fases intermedias. El proceso iterativo de diseño de la rúbrica mostró que el tránsito del alumnado desde modelos incorrectos a modelos correctos no es lineal ni abrupto. Por el contrario, muchos estudiantes construyen representaciones que combinan elementos correctos e incorrectos, lo que evidencia una etapa intermedia de comprensión. Por ello, la versión final de la rúbrica incluye niveles que permiten registrar este avance progresivo, sin penalizar el error conceptual cuando coexiste con aciertos parciales. Esta estructura facilita una evaluación formativa más justa y sensible al desarrollo cognitivo del estudiante.

Las Apps con sensores físicos y salida visible son mejor representadas. La comparación entre tipos de Apps mostró que aquellas con funcionamiento más observable y tangible —como el detector de metales, el podómetro o la cámara térmica— fueron modeladas con mayor precisión que otras percibidas como arbitrarias, como la calculadora de amor. Este patrón sugiere que la credibilidad y transparencia funcional de una App incide directamente en la calidad del modelo mental que los estudiantes logran construir. Ver lo que ocurre favorece la representación de cómo ocurre.

La edad del alumnado se relaciona con una mayor sofisticación en la modelización. Los análisis por edad muestran una clara progresión en la calidad de los modelos: los estudiantes de 17 años logran representaciones más precisas y completas, incluyendo casos en el nivel más alto de la rúbrica (N4), mientras que los de menor edad tienden a niveles intermedios. Esta evolución sugiere que la madurez cognitiva, la experiencia escolar acumulada y una mayor exposición reflexiva a la tecnología contribuyen a la construcción de modelos mentales más estructurados, funcionales y cercanos a los reales.

7.1.4 Conclusiones transversales de los tres estudios

A lo largo de los tres estudios desarrollados en el marco del proyecto App Checkers se ha puesto en evidencia que las principales prácticas científicas escolares —la indagación, la argumentación y la modelización— no solo coexisten, sino que interactúan de manera continua en los procesos de aprendizaje. La implementación de tareas de indagación ha permitido observar cómo estas prácticas se entrelazan dentro de productos y razonamientos construidos por el alumnado, muchas veces de manera inseparable. Los datos de los tres estudios muestran que las decisiones tomadas por los estudiantes en un área influyen directamente en la otra: las formas en que indagan determinan el tipo de argumentos que pueden construir; los modelos que elaboran condicionan qué preguntas son relevantes y qué evidencias necesitan; las pruebas utilizadas en una argumentación están condicionadas por el nivel de precisión en el modelo o el rigor del diseño experimental. En este sentido, se reafirma que las prácticas científicas no deben ser abordadas como compartimentos estancos, sino como dimensiones interdependientes del pensamiento científico escolar.

La calidad del diseño experimental condiciona la validez de los argumentos científicos. En el estudio 1 se evidenció que, si bien no existe una correlación absoluta entre el nivel de indagación y el nivel de argumentación, sí se identificaron patrones frecuentes donde

un diseño experimental más sofisticado facilitaba la construcción de argumentos más complejos. Las combinaciones I3–A3 y I2–A2 fueron las más comunes, lo que sugiere que existe un vínculo operativo entre la toma sistemática de datos y la capacidad de justificar afirmaciones con pruebas válidas. La argumentación no se desarrolla de manera paralela a la indagación, sino que la depende en gran medida del tipo y la calidad de las pruebas generadas durante el proceso experimental.

La identificación de restricciones técnicas en las Apps favorece la construcción de refutaciones. A partir del análisis de la relación entre los componentes "límites operativos" (criterio C de la rúbrica de indagación) y "restricciones o refutaciones" (componente F del modelo de Toulmin), el estudio 1 reveló que los estudiantes que lograban identificar condiciones de funcionamiento limitadas de las Apps también tendían a integrar ese conocimiento en sus argumentos. Esto sugiere que una comprensión epistemológica de las limitaciones de los dispositivos técnicos se traslada al plano argumentativo, fortaleciendo la capacidad del alumnado para construir afirmaciones críticas y contextualizadas. Aunque esta relación no fue sistemática, sí se observó que ambos elementos cumplían funciones similares en el plano epistémico: establecer los límites de validez de una afirmación científica, en otras palabras: hay una coincidencia parcial, pero no una dependencia absoluta.

La evaluación de confiabilidad requiere articular validez experimental con análisis de estabilidad. El estudio 2 evidenció que muchos estudiantes en bachillerato emplean estrategias de validez como la comparación con un valor esperado, pero omiten análisis metrológicos vinculados a la estabilidad de la medición (fiabilidad). Esto muestra que la construcción de un argumento sobre la confiabilidad de una App requiere integrar tanto prácticas de diseño experimental (validez) como prácticas metrológicas (fiabilidad), que implican técnicas estadísticas básicas. La falta de conexión entre ambos enfoques por parte del alumnado pone de manifiesto la necesidad de enseñar de manera integrada las dos dimensiones, y sugiere que la construcción de argumentos sobre la confiabilidad se ve limitada si no se desarrollan ambas competencias a la vez.

La expresión de modelos mentales se apoya en la comprensión de procesos de indagación y en la capacidad argumentativa. En el estudio 3 se observó que muchos estudiantes describen lo que hace una App (output), pero no representan cómo lo hace (procesamiento), ni qué mecanismos técnicos lo permiten (modelo metal de funcionamiento). Esta falta de pensamiento sistémico se relaciona con una mirada funcionalista que ignora los procesos intermedios. Sin embargo, aquellos estudiantes que lograron construir modelos

funcionales más elaborados también mostraron, en sus plantillas, un uso más preciso del lenguaje técnico y una mejor articulación entre los datos de entrada y la transformación que ocurre en la App. Esto sugiere que modelar requiere haber comprendido previamente cómo se produce y transforma la información en un entorno experimental, y que la generación de modelos puede beneficiarse de prácticas previas de indagación y argumentación.

Nexos metodológicos y conceptuales entre los tres estudios. Aunque los estudios fueron implementados en contextos geográficos distintos (España y Chile), con niveles educativos variados (2º ESO, bachillerato y enseñanza media chilena), y con ajustes en la secuencia didáctica, todos comparten un hilo metodológico común: el diseño de una experiencia de indagación escolar centrada en la evaluación de la confiabilidad de tecnologías digitales. Los tres estudios analizan cómo el alumnado recolecta información, construye interpretaciones y representa el funcionamiento de sistemas técnicos. En todos los casos, se usaron instrumentos específicos (rúbricas, plantillas, categorías) que permitieron codificar y analizar el grado de apropiación de prácticas científicas clave. Además, todos los estudios revelan que el desarrollo de estas prácticas no es lineal ni homogéneo, sino que emerge de forma diferencial según el tipo de App, el andamiaje proporcionado y el nivel educativo del alumnado. Por ello, pese a sus diferencias metodológicas, las tres investigaciones se integran en un marco coherente que permite estudiar la interdependencia de prácticas científicas en entornos auténticos de aprendizaje.

Las interdependencias principales de las prácticas exploradas se muestran en el esquema de la figura 7.1.A. Los elementos del esquema son:

“Lo que no se representa, no se indaga”. Desde el estudio 3, muchos modelos del alumnado omitieron sensores, comportamientos o interacciones clave (como la lógica de comparación con umbrales). Esto sugiere que la ausencia de una representación funcional del sistema limita la posibilidad de formular preguntas relevantes o diseñar procedimientos experimentales informados. Aunque tu investigación no analizó directamente el impacto de la modelización sobre la indagación, desde el marco teórico de pensamiento sistémico (Louca et al., 2011a, 2011b) se puede argumentar que una pobre modelización puede restringir la capacidad para realizar indagaciones significativas.

“Los modelos reflejan una mirada de usuario más que de experimentador”. El estudio 3 muestra que los estudiantes representaron con más frecuencia los objetos visibles y los resultados (outputs), pero omitieron entidades técnicas, comportamientos y relaciones internas

del sistema. Esto sugiere que, aunque participaron en actividades de indagación, sus modelos no integraron elementos estructurales del funcionamiento técnico de las Apps, reflejando una lógica centrada en el uso y no en la exploración experimental. Esta brecha en la representación podría explicarse por un vacío conceptual —especialmente en el reconocimiento de procesos intermedios— que no fue compensado durante la indagación. Desde esta perspectiva, una indagación más orientada a descomponer los sistemas técnicos podría ayudar a llenar ese vacío, promoviendo modelos más funcionales y sistémicos.

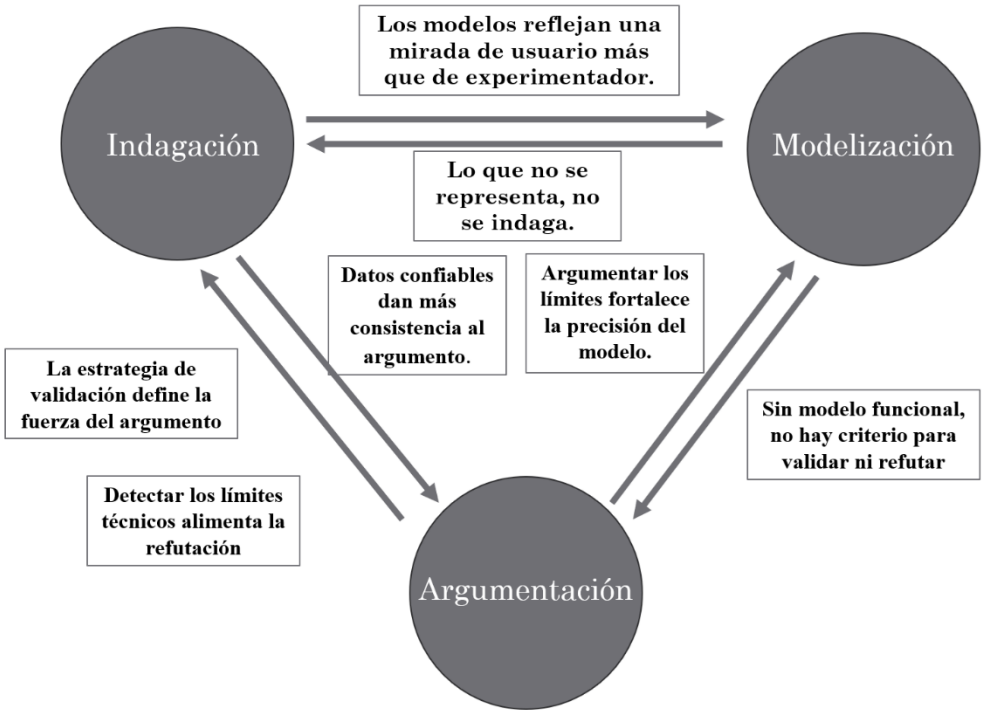

Figura 7.1.A - Las interdependencias principales de las prácticas exploradas.

"Sin modelo funcional, no hay criterio para validar ni refutar". En el estudio 1, los estudiantes que identificaron límites operativos en las Apps —como rangos de funcionamiento o restricciones técnicas— tendieron a integrarlos en sus argumentos como justificaciones para aceptar o refutar un resultado. En el estudio 3, en cambio, se observó que muchos modelos mentales carecían de representaciones de entidades internas, comportamientos y relaciones, lo que implica una comprensión limitada del "cómo" opera técnicamente la App. Esta ausencia compromete la capacidad de evaluar críticamente sus resultados en actividades de indagación, ya que, sin una representación funcional clara, no se pueden establecer criterios sólidos para

juzgar la validez o confiabilidad de las mediciones. Por tanto, un modelo incompleto limita no solo la explicación, sino también la toma de decisiones experimentales y argumentativas.

"Detectar los límites técnicos alimenta la refutación". El estudio 1 mostró que los estudiantes que identificaban condiciones técnicas de funcionamiento limitadas en las Apps (como umbrales, rangos válidos o interferencias) eran los que con mayor frecuencia integraban esa información en sus argumentos, ya sea para respaldar o refutar un resultado. Este vínculo entre límites operativos y refutación indica que comprender las restricciones técnicas no solo mejora el diseño experimental, sino que fortalece la dimensión crítica del argumento, permitiendo al estudiante establecer los márgenes dentro de los cuales una afirmación puede sostenerse.

"La estrategia de validación define la fuerza del argumento". El estudio 2 mostró que los estudiantes que recurrieron a estrategias de validación explícita, como la comparación con un valor esperado o el uso de una fuente externa como patrón, lograron construir argumentos más sólidos sobre la confiabilidad de la App. Por el contrario, quienes usaron estrategias implícitas o no justificaron sus datos, presentaron argumentos más débiles. Esto indica que la forma en que se valida la información durante la indagación condiciona directamente la solidez y legitimidad del argumento construido.

"Argumentar los límites fortalece la precisión del modelo". Desde el Estudio 1 se identificó que los estudiantes que lograban detectar restricciones técnicas (por ejemplo, que la App no funcionaba bien con ciertas condiciones de luz o a cierta distancia) eran también quienes incorporaban esas limitaciones en sus argumentos. Aunque el estudio 1 no evaluó directamente los modelos mentales, sí mostró que una comprensión argumentativa crítica implicaba identificar los límites funcionales del sistema. Desde el Estudio 3, por inferencia razonada y respaldada empíricamente, muchos estudiantes omitieron elementos clave en sus modelos mentales, especialmente aquellos vinculados a la lógica técnica del funcionamiento interno (entidades, comportamientos, interacciones). Una posible explicación plausible es que, al no haber desarrollado argumentos críticos sobre cómo y por qué la App podría fallar, no llegaron a construir representaciones técnicas robustas. Por el contrario, una argumentación que identifica condiciones y límites "epistémicos" del dispositivo puede forzar al estudiante a precisar cómo funciona internamente, fortaleciendo la elaboración del modelo.

"Datos confiables dan más consistencia al argumento". Del estudio 1 se observó una fuerte relación entre el nivel de indagación (I3, I4) y la complejidad de los argumentos (A3,

A4). El rigor experimental (repetición, calibración, límites operativos) determinó qué tan válidas y sofisticadas eran las pruebas utilizadas para sostener una afirmación.

7.2 Limitaciones

A pesar de los importantes hallazgos alcanzados en los tres estudios que componen esta tesis, es necesario reconocer una serie de limitaciones metodológicas, analíticas y contextuales que condicionan la generalización de los resultados y la amplitud interpretativa de los mismos. Estas limitaciones no solo derivan de decisiones intencionadas adoptadas durante el diseño de la investigación, sino también de restricciones inherentes al trabajo en contextos escolares reales, donde las condiciones de implementación no siempre son controlables ni homogéneas.

Una de las limitaciones más relevantes se relaciona con el tamaño y distribución de las muestras. Si bien el estudio 1 contó con una muestra relativamente amplia de 76 videos estudiantiles, los estudios 2 y 3 se desarrollaron con muestras más acotadas (19 pósteres en el segundo y 50 plantillas en el tercero), lo cual restringe la posibilidad de extrapolar los resultados a otras poblaciones. Además, dentro de esas muestras, la representación por tipo de App o por edad no siempre fue equilibrada, lo que puede generar sesgos en el análisis comparativo.

Otra limitación significativa tiene que ver con el diseño metodológico adoptado en esta investigación. No se ha seguido el enfoque de la investigación basada en el diseño (design-based research), que propone una iteración sistemática sobre una misma secuencia didáctica manteniendo constantes ciertas variables, como el profesorado, el contexto o la edad del alumnado. En cambio, lo que se ha hecho en esta tesis es aplicar variantes de la propuesta App Checkers, adaptadas a contextos educativos diferentes, con secuencias también diferentes, lo cual si bien aporta riqueza al análisis por su flexibilidad y capacidad de adaptación, también introduce variables de confusión que dificultan la comparación directa entre estudios. Esta elección metodológica ha permitido explorar el potencial de la actividad en diversos escenarios reales, pero limita la posibilidad de atribuir cambios en el desempeño del alumnado a intervenciones didácticas específicas, como lo permitiría una investigación basada en diseño.

Asimismo, debe señalarse que los tres estudios se han centrado en el análisis de productos finales (videos, pósteres, plantillas), lo que implica que no se ha accedido de forma sistemática a los procesos intermedios de elaboración, como las discusiones entre estudiantes, los acuerdos

alcanzados o las estrategias emergentes durante el trabajo colaborativo. Este enfoque limita la comprensión de cómo se desarrollan realmente las prácticas científicas en el aula y deja sin explorar aspectos discursivos y metacognitivos que podrían enriquecer el análisis.

En el estudio 1, una limitación ya reconocida fue la posibilidad de que la rúbrica de argumentación, al basarse en la acumulación de componentes, no captara adecuadamente la calidad interna o la pertinencia del contenido argumentativo. A pesar de haberse establecido criterios de codificación y consensos entre evaluadores, esta decisión metodológica puede haber simplificado algunos aspectos cualitativos del razonamiento del alumnado. Asimismo, no se exploró en profundidad el conocimiento básico activado ni la calidad epistemológica de las pruebas utilizadas, elementos que podrían ser relevantes para futuros estudios.

En el estudio 2, la principal limitación reconocida fue la escasa presencia de estrategias de fiabilidad en los productos estudiantiles, lo cual restringió las posibilidades de análisis cruzado con las estrategias de validez. Además, se trató de un estudio con una muestra reducida, lo que impide generalizar los hallazgos más allá del grupo analizado. El foco cualitativo, centrado en identificar las estrategias epistémicas del alumnado, permitió una comprensión profunda pero no cuantitativa de su desempeño.

En el estudio 3, aunque se desarrolló una rúbrica robusta para el análisis de modelos, no se integró una categoría explícita sobre la precisión científica del modelo, dado que la plantilla no solicitaba ese tipo de valoración. Esto puede haber limitado el análisis sobre la validez epistemológica de las ideas representadas por los estudiantes. Asimismo, si bien se observaron progresiones en la representación de objetos, entidades, comportamientos e interacciones, no se analizaron en profundidad los significados científicos que el alumnado atribuía a cada una de estas categorías. Esto refuerza una tendencia general de la tesis: el foco ha estado puesto principalmente en el conocimiento procedimental y epistemológico, dejando en un segundo plano el conocimiento conceptual.

Este último aspecto constituye una limitación transversal. En dos de los tres estudios, el alumnado seleccionó libremente la App con la que trabajaría. Si bien esto fomenta la motivación y la autonomía, también produce una fuerte heterogeneidad en los contenidos científicos abordados. El hecho de que cada App remita a dominios diferentes (campo magnético, sonido, imagen térmica, movimiento, matemática simbólica) dificulta establecer un análisis conceptual comparativo profundo. De hecho, la tesis ha trabajado de forma tangencial los contenidos científicos, priorizando las dimensiones epistémicas (cómo se genera y justifica

el conocimiento) y procedimentales (cómo se diseña una investigación o se representa un modelo). Esto implica que no se ha profundizado en cómo los estudiantes entienden, por ejemplo, la diferencia entre calor y temperatura, el concepto de aceleración o la naturaleza de una señal sonora.

Desde una perspectiva didáctica, esto supone una oportunidad futura: integrar el análisis de las prácticas científicas con la exploración del conocimiento conceptual subyacente, de modo de avanzar hacia una comprensión más holística del aprendizaje en ciencias. Por ahora, la tesis ha puesto el foco en describir con precisión cómo el alumnado actúa, razona y modela cuando se le propone investigar la confiabilidad de una App, dejando abiertas nuevas preguntas sobre cómo se apropia de los conceptos científicos implicados.

En suma, las limitaciones aquí expuestas no invalidan los resultados obtenidos, pero sí enmarcan su alcance y ofrecen pistas valiosas para futuras investigaciones. Reconocerlas permite delimitar mejor las conclusiones, afinar los instrumentos metodológicos y proyectar nuevos estudios que integren el análisis de productos y procesos, de prácticas científicas y de conceptos científicos, en escenarios educativos diversos y comparables.

7.3 Implicaciones didácticas

El análisis de los tres estudios del proyecto App Checkers ofrece un conjunto amplio de implicaciones didácticas, tanto para el diseño de experiencias de enseñanza como para la evaluación de competencias científicas escolares. Si bien cada estudio abordó una práctica científica distinta —la indagación y la argumentación en el estudio 1, las estrategias de validez y fiabilidad en el estudio 2, y la expresión de modelos mentales en el estudio 3—, los resultados obtenidos permiten reflexionar críticamente sobre las decisiones pedagógicas adoptadas en cada implementación, y proponer mejoras concretas en las secuencias didácticas utilizadas. A su vez, se identifican oportunidades de mejora en la formación docente, en la selección de Apps según su potencial didáctico y en el uso de rúbricas como herramienta de evaluación formativa.

En relación con el estudio 1, una primera implicación didáctica importante se vincula con la enseñanza explícita de la argumentación científica. Aunque el profesorado introdujo el proyecto hablando de verificación y confiabilidad, no se observó un trabajo sistemático con los componentes de un argumento científico, como los propuestos por el modelo de Toulmin. En ningún momento se enseñó al alumnado a reconocer ni construir deliberadamente pruebas, aserciones, razonamientos, moduladores o refutaciones. Esta ausencia puede explicar por qué

algunos estudiantes, a pesar de realizar indagaciones de alto nivel, no alcanzaron desempeños argumentativos equivalentes. Se propone, por tanto, incorporar sesiones iniciales donde se explicite el modelo de Toulmin, se analicen ejemplos concretos y se ejercite su aplicación antes de realizar los videos. Además, en términos de andamiaje, podría desarrollarse una guía de autoevaluación o checklist que ayude al alumnado a identificar los elementos mínimos de una argumentación antes de grabar sus productos.

Otra oportunidad de mejora en el estudio 1 se encuentra en promover conexiones más explícitas entre las pruebas empíricas y los argumentos construidos. Varios casos mostraron que el alumnado no usaba las evidencias generadas en sus experimentos como soporte de sus afirmaciones. Esta desconexión puede abordarse mediante actividades intermedias donde se trabaje la relación entre observaciones y conclusiones, y se enfatice cómo justificar una afirmación científica con datos. Asimismo, para favorecer el tránsito hacia niveles más altos en las rúbricas de indagación y argumentación, se sugiere integrar secuencias progresivas que modelen primero una indagación completa, luego la redacción de un argumento asociado, y finalmente el trabajo autónomo del alumnado.

En cuanto al estudio 2, se observó que el alumnado mostró una preferencia marcada por estrategias de validez más intuitivas, como la comparación con valores esperados, y que la incorporación de procedimientos de fiabilidad fue escasa o ausente. Esta situación puede explicarse parcialmente por una carencia de habilidades estadísticas, pero también por un diseño de actividad que no favorecía dicha integración. En este sentido, es importante reconocer que no hubo discusión colectiva previa sobre estrategias posibles, ni análisis de ejemplos que mostraran cómo evaluar la estabilidad de una medición. Por ello, se propone que la secuencia didáctica incluya una etapa de reflexión inicial sobre qué significa que una medición sea confiable, seguida de una revisión guiada de diferentes estrategias, incluyendo repetición de mediciones, análisis de dispersión, uso de gráficos y cálculo de medias. Estas estrategias podrían ser trabajadas mediante estudios de caso, simulaciones o tareas diagnósticas previas.

Además, una enseñanza explícita de conceptos metrológicos básicos, como error, incertidumbre o fiabilidad, debería acompañar el trabajo experimental. Esta dimensión conceptual fue tratada tangencialmente y debería tener un lugar más central, especialmente en contextos de bachillerato. El uso de una rúbrica que distinga niveles de uso de estrategias de fiabilidad también puede orientar tanto la evaluación como la enseñanza. El bajo nivel alcanzado en esta dimensión parece deberse a una mixtura entre carencias en habilidades

previas del alumnado y una propuesta didáctica que no las desarrolló de manera progresiva, por lo cual ambas dimensiones deben ser abordadas simultáneamente: fortalecer el dominio de herramientas estadísticas básicas y rediseñar las actividades para que las pongan en juego de forma significativa.

En el estudio 3, una implicación central es la necesidad de promover una visión más sistémica y técnica del funcionamiento de las Apps. El análisis mostró que los estudiantes tienden a modelar desde su experiencia de usuario, sin considerar los procesos intermedios ni las estructuras funcionales internas. Esta visión funcionalista puede revertirse si se integran actividades que expliciten la relación entre tecnología y ciencia, como el rol de los sensores, la lógica de procesamiento, o el funcionamiento de algoritmos. Una posibilidad es utilizar simuladores o representaciones visuales que muestren cómo fluye la información en una App, desde la entrada hasta la salida, o bien realizar análisis comparativos entre una App y el instrumento físico que simula. También se sugiere trabajar con diagramas de flujo o con esquemas entrada-proceso-salida que ayuden al alumnado a visualizar las interacciones entre componentes. Esto permitiría avanzar hacia modelos mentales más elaborados e interdisciplinarios, donde se integren elementos físicos, computacionales y algorítmicos.

Una oportunidad concreta de mejora es enriquecer el trabajo con modelos incorporando fases de evaluación entre pares, análisis de modelos ejemplo, y revisión iterativa de los propios modelos. En la implementación estudiada, los estudiantes construyen un único modelo y lo entregan, pero no se promueve un proceso de revisión ni de discusión de modelos alternativos, lo cual limitaría el desarrollo del pensamiento crítico. Integrar estas fases podría promover mejoras significativas en la calidad y profundidad de los modelos elaborados.

Respecto al uso didáctico de las Apps, los tres estudios permiten extraer conclusiones sobre qué aplicaciones funcionan mejor para promover pensamiento científico. Las Apps que requieren sensores físicos, como el detector de metales, el podómetro, la cámara térmica o el sonómetro, parecen ser más efectivas para trabajar competencias procedimentales y epistemológicas, ya que permiten diseñar experimentos, comparar valores y observar relaciones entre variables. Estas Apps promueven el desarrollo de habilidades de medición, análisis y modelado de forma más estructurada. En cambio, las Apps que se perciben como lúdicas o no creíbles, como la calculadora de amor, tienden a generar modelos pobres y desempeños bajos, aunque ofrecen la posibilidad de discutir críticamente qué cuenta como conocimiento válido y cómo evaluar afirmaciones pseudocientíficas. Finalmente, Apps como PhotoMath, que integran reconocimiento visual, procesamiento simbólico y devolución guiada,

permiten trabajar modelos funcionales y algoritmos, y ofrecen alto potencial para explorar razonamiento algorítmico y tecnológico.

Estas observaciones permiten proponer una selección deliberada de Apps según el objetivo didáctico. Si se desea promover habilidades de medición, lo recomendable es optar por Apps nativas con sensores físicos. Si se desea trabajar pensamiento epistémico, se pueden incluir Apps dudosas o engañosas como punto de partida para discutir confiabilidad. Si el foco está en el modelado computacional, se puede priorizar Apps que simulen procesos complejos o integren inteligencia artificial. Esta diversidad permite articular distintos enfoques de enseñanza según la competencia científica priorizada.

Por último, una de las contribuciones más potentes del proyecto radica en el desarrollo de rúbricas específicas para evaluar indagación, argumentación y modelado funcional, todas ellas ancladas en marcos teóricos reconocidos y construidas a partir de evidencia empírica. Estas rúbricas pueden ser utilizadas en otros estudios y contextos, ya que ofrecen niveles de desempeño bien definidos, categorías adaptables y criterios que permiten diferenciar progresiones reales en el aprendizaje. Además, permiten integrar la evaluación formativa, al ofrecer descriptores que pueden ser compartidos con el alumnado, utilizados para la coevaluación y revisados durante el proceso. Se sugiere su uso como instrumentos de retroalimentación pedagógica, que orienten tanto la enseñanza como la mejora continua del alumnado en tareas de indagación científica, modelado y argumentación escolar.

En conjunto, los tres estudios sugieren que el desarrollo de prácticas científicas en el aula requiere una enseñanza explícita, una secuencia didáctica bien diseñada, el uso de tareas auténticas y evaluaciones que consideren tanto los procesos como los productos del aprendizaje. Las oportunidades de mejora aquí descritas apuntan a fortalecer la articulación entre los componentes del trabajo científico escolar y a promover aprendizajes más profundos, integradores y transferibles a contextos reales y complejos.

BIBLIOGRAFÍA

Aguilera, M., y López-Simó, V. (2025). Estudio del desempeño en prácticas científicas de indagación y argumentación de alumnado de enseñanza secundaria mediante una aplicación de móvil. *Revista Electrónica de Enseñanza de las Ciencias*, *24*(1), 52–73.

Aguilera, M., López-Simó, V., y Domènech Amador, N. (en prensa). Estrategias para evaluar la confiabilidad de aplicaciones móviles en estudiantes de bachillerato. *Didáctica de las Ciencias Experimentales y Sociales*.

Almanza Mazas, M. (2014). *Diseño e implementación de una aplicación móvil para monitoreo de un socket autoajustable* [Tesis de licenciatura, Universidad Nacional Autónoma de México]. Facultad de Ingeniería, UNAM.

Arnold, J., Kremer, K. y Mayer, J. (2014). Understanding Students' Experiments—What kind of support do they need in inquiry tasks?. *International Journal of Science Education*, *36*(16), 2719-2749. https://doi.org/10.1080/09500693.2014.930209

Berland, L., y McNeill, K. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5), 765–793. https://doi.org/10.1002/sce.20402

Blanco-López, Á., España-Ramos, E., y Franco-Mariscal, A. J. (2017). Estrategias didácticas para el desarrollo del pensamiento crítico en el aula de ciencias. *Ápice. Revista de Educación Científica*, *1*(1), 107–115. https://doi.org/10.17979/AREC.2017.1.1.2004

Canal Asociación Docentes de Ciencias Biológicas. (2016, octubre 31). *La argumentación basada en modelos: Perspectivas teóricas. Conferencia plenaria del Dr. Agustín Adúriz-Bravo* [Video]. YouTube. https://youtu.be/gVygB_9jot4

Castellanos, D., Castellanos, B., Llivina, M., Silverio, M., Reinoso, C., y García, C. (2002). *Aprender y enseñar en la escuela*. Editorial Pueblo y Educación.

Cebrián-Robles, D., Cebrián de la Serna, M., Gallego-Arrufat, M. J., y Contreras, J. Q. (2018). Impacto de una rúbrica electrónica de argumentación científica en la metodología blended-learning. *RIED*, *Revista Iberoamericana de Educación a Distancia*, *21*(1), 75-94. http://dx.doi.org/10.5944/ried.21.1.18827

Centro Español de Metrología. (2012). *VIM: Vocabulario internacional de metrología: Conceptos fundamentales y generales, y términos asociados* [Informe técnico]. https://www.cem.es/sites/default/files/vim-cem-2012web.pdf

Cho, K. L., y Jonassen, D. H. (2002). The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology Research and Development, 50*(3), 5–22. https://doi.org/10.1007/BF02505022

Covitt, B. A., y Anderson, C. W. (2022). Untangling trustworthiness and uncertainty in science. *Science & Education, 31*(5), 1155–1180. https://doi.org/10.1007/s11191-022-00322-6

Couso, D. (2014) De la moda de "aprender indagando" a la indagación para modelizar: una reflexión crítica. En M. Héras, A. Lorca, B. Vázquez, A. Wamba, R. Jiménez (Eds). *Investigación y transferencia para una educación en ciencias*: *Un reto emocionante* (pp.1-28). Servicio de Publicaciones Universidad de Huelva.

Couso, D., y Jiménez–Liso, M. (2020). Investigar en ciencias para dotar de credibilidad el aula de primaria. *Aula de Innovación Educativa, (298)*, 10–14

Couso D. (2020). Aprender ciencia escolar implica construir modelos cada vez más sofisticados de los fenómenos del mundo. En D. Couso, M. R. Jiménez-Liso, C. Refojo, y J. A. Sacristán (Coords.), *Enseñando ciencia con ciencia* (pp. 63–74). Fundación Lilly.

Couso, D. y Puig, B. (2021). Educación Científica en Tiempos de Pandemia. *Alambique: Didáctica de las ciencias experimentales*, 104, 49-56.

Crujeiras-Pérez, B. I. (2014). *Competencias e prácticas científicas no laboratorio de química: Participación do alumnado de secundaria na indagación* [Tesis doctoral, Universidad de Santiago de Compostela]. Repositorio Institucional da Universidad de Santiago de Compostela. http://hdl.handle.net/10347/12072

Crujeiras-Pérez, B. I., y Jiménez-Aleixandre, M. P. (2015). Desafíos planteados por las actividades abiertas de indagación en el laboratorio: Articulación de conocimientos teóricos y prácticos en las prácticas científicas. *Enseñanza de las Ciencias, 33*(1), 63–84. https://doi.org/10.5565/rev/ensciencias.1469

Crujeiras-Pérez, B. I., y Cambeiro, F. (2017). ¿Cómo podemos averiguar si Limpics es un fraude? Aprendiendo a diseñar investigaciones en educación secundaria. *Educación Química, 28*, 174–180. https://doi.org/10.1016/j.eq.2017.01.002

Crujeiras-Pérez, B. I., y Cambeiro, F. (2018). Una experiencia de indagación cooperativa para aprender ciencias en educación secundaria participando en las prácticas científicas. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 15*(1), 1201. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i1.1201

Díaz de Bustamante, J., y Jiménez-Aleixandre, M. P. (1999). Aprender ciencias, hacer ciencias: Resolver problemas en clase. *Alambique. Didáctica de las Ciencias Experimentales, 20*, 9–16.

Duschl, R. A. (2007). Quality argumentation and epistemic criteria. In S. Erduran y M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 159–175). Springer.

Erduran, S., Simon, S., y Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915–933. https://doi.org/10.1002/sce.20012

Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Newark. *American Philosophical Association*.

Facione, P., y Facione, N. (1992). The California Critical Thinking Disposition Inventory. *California Academic Press*.

Ferrés, C., Marbà, A., y Sanmartí, N. (2015). Trabajos de indagación de los alumnos: Instrumentos de evaluación e identificación de dificultades. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 12*(1), 22–37. http://hdl.handle.net/10498/16922

Ferrés, C. (2017). *La competència d'indagació i la seva avaluació en els estudiants de batxillerat* [Tesis doctoral, Universidad Autónoma de Barcelona]. Depósito Digital de Documentos de la UAB. https://ddd.uab.cat/record/188406

Fussero, G., y Occelli, M. (2024). Las prácticas científicas en el desarrollo del pensamiento crítico en la enseñanza de la biología. En M. Occelli, L. García-Romano, y C. Sosa (Comps.), *Prácticas científicas y pensamiento crítico en la enseñanza de las ciencias* (pp. 34 - 47). Universidad Nacional de Córdoba – Editorial.

Giere, R. N. (1991). *Understanding scientific reasoning* (3rd ed.). Harcourt Brace Jovanovich.

Gilbert, J. K., Boulter, C., y Rutherford, M. (1998). Models in explanations, part 1: Horses for courses. *International Journal of Science Education, 20*, 83–97. https://doi.org/10.1080/0950069980200106

González, M. Á., y González, M. Á. (2016). *Uso de smartphones en experimentos de Física en el laboratorio y fuera de él*. En M. González Montero de Espinosa (Ed.), *Actas del IV Congreso de Docentes de Ciencias* (pp. 1–10). Editorial Santillana. http://uvadoc.uva.es/handle/10324/17485

González de los Reyes, R. (2024). Transformación digital de la enseñanza. Uso de los Smartphone en la enseñanza de la Física. *Revista Varela*, *24*(69), 218–223. https://doi.org/10.5281/zenodo.13623354

Gott, R., y Duggan, S. (1995). *Investigative work in the science curriculum*. Open University Press.

Gott, R., y Duggan, S. (2002). Problems with the assessment of performance in práctical science: Which way now? *Cambridge Journal of Education, 32*(2), 183–201. https://doi.org/10.1080/03057640220147540

Gott, R., y Duggan, S. (2003). *Understanding and using scientific evidence*. London: Sage. https://doi.org/10.4135/9780857020161

Gott, R., y Roberts, R. (2008, December 8). *Concepts of evidence and their role in open-ended práctical investigations and scientific literacy: Background to published papers* [Informe técnico]. School of Education, Durham University. https://cofev.webspace.durham.ac.uk/wp-content/uploads/sites/299/2022/05/Gott-Roberts-2008-Research-Report.pdf

Gott, R., Duggan, S., Roberts, R., y Hussain, A. (2020, diciembre 11). *Research into understanding scientific evidence*. Durham University. https://cofev.webspace.durham.ac.uk/

Graves, L. (2017). Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture and Critique, 10*(3), 518–537. https://doi.org/10.1111/cccr.12163

Gutiérrez, R. (2004). La modelización y los procesos de enseñanza/aprendizaje. *Alambique: Didáctica de las Ciencias Experimentales* (42), 8–18.

Harsin, J. (2021). Post-truth Reflections on Public Origins and Functions of Publishing. *Information, Medium and Society, 19*(1), 7. https://doi.org/10.18848/2691-1507/CGP/v19i01/7-19

Hernández-Sampieri, R., y Fernández-Collado, C. (2016). *Metodología de la investigación*. McGraw-Hill Education.

Hsu, C. C., Chiu, C. H., Lin, C. H., y Wang, T. I. (2015). Enhancing skill in constructing scientific explanations using a structured argumentation scaffold in scientific inquiry. *Computers & Education, 91*, 46–59. https://doi.org/10.1016/j.compedu.2015.09.009

Ibáñez, P., Razo, C. R., y García, G. (2010). *Informática II*. Cengage Learning.

Jiménez-Aleixandre, M. P., y Erduran, S. (2008). Argumentation in science education: An overview. En S. Erduran y M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 3–27). Springer. https://doi.org/10.1007/978-1-4020-6670-2_1

Jiménez-Aleixandre, M. P. (2010). *10 ideas clave. Competencias en argumentación y uso de pruebas*. Graó.

Jiménez-Aleixandre, M. P., y Crujeiras-Pérez, B. I. (2017). Epistemic practices and scientific practices in science education. En K. S. Taber y B. Akpan (Eds.), *Science education. New directions in mathematics and science education* (pp. 69–80). Sense Publishers.

Jiménez-Aleixandre, M. (2020). ¿Cómo sabemos lo que sabemos? Mediante la argumentación y el uso de pruebas, herramientas para aprender y desarrollar el pensamiento crítico. En D. Couso, M. R. Jiménez-Liso, C. Refojo, y J. A. Sacristán (Coords.), *Enseñando ciencia con ciencia* (pp. 75–86). Fundación Lilly.

Jiménez-Liso, M. R. (2020). Aprender ciencia escolar implica aprender a buscar pruebas para construir conocimiento (indagación). En D. Couso, M. R. Jiménez-Liso, C. Refojo, y J. A. Sacristán (Coords.), *Enseñando ciencia con ciencia* (pp. 53–62). Fundación Lilly.

Koerber, S., & Osterhaus, C. (2019). Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *Journal of Cognition and Development, 20*(4), 510–533. https://doi.org/10.1080/15248372.2019.1625103

Knaggs, C. M., y Schneider, R. M. (2012). Thinking like a scientist: Using vee-maps to understand process and concepts in science. *Research in Science Education, 42*, 609–632. https://doi.org/10.1007/s11165-011-9213-x

Lane, W. B., y Headley, C. (2021, diciembre 10). *Student representations of computation in the physics community* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2112.05581

Lin, H., Hong, Z. R., y Lawrenz, F. (2012). Promoting and scaffolding argumentation through reflective asynchronous discussions. *Computers & Education, 59*(2), 378–384. https://doi.org/10.1016/j.compedu.2012.01.019

Lin, Y. (2019). Student positions and web-based argumentation with the support of the six thinking hats. *Computers & Education, 139*, 191–206. https://doi.org/10.1016/j.compedu.2019.05.019

Lipman, M. (2016). *El lugar del pensamiento en la educación*. (M. G. Pérez, Trad.) Octaedro.

López-Simó, V. (2018). *Estudi de la Llei de Faraday-Lenz fent servir el mòbil i el sensor de voltatge*. *Ciències*, (35), 17–21.

López-Simó, V. (2021). App Checkers, un proyecto de verificación de la fiabilidad de una aplicación móvil. *Aula de secundaria*, *44*, 37-42.

López Simó, V., y Simarro, C. (2024). La elaboración de modelos científicos computacionales con Scratch en la formación inicial de maestros. En M. Occelli, L. García Romano, y C. A. Sosa (Comps.), *Prácticas científicas y pensamiento crítico en la enseñanza de las ciencias* (pp. 98–114). Universidad Nacional de Córdoba.

Louca, L. T., Zacharia, Z. C., Michael, M., y Constantinou, C. P. (2011a). Objects, entities, behaviors, and interactions: A typology of student-constructed computer-based models of physical phenomena. *Journal of Educational Computing Research, 44*(2), 173–201. https://doi.org/10.2190/EC.44.2.c

Louca, L. T., Zacharia, Z. C., y Constantinou, C. P. (2011b). In quest of productive modeling-based learning discourse in elementary school science. *Journal of Research in Science Teaching, 48*(8), 919–951. https://doi.org/10.1002/tea.20435

Mariscal, A. (2015). Competencias científicas en la enseñanza y el aprendizaje por investigación: Un estudio de caso sobre corrosión de metales en secundaria. *Enseñanza de las Ciencias: Revista de Investigación y Experiencias Didácticas, 33*(2), 231–252. https://doi.org/10.5565/rev/ensciencias.1645

Márquez, C. (2005). Aprender ciencias a través del lenguaje. *Educar, 33*, 27–38.

Marraud, H. (2014, julio). *Breve curso de teoría de los argumentos* [Material de curso]. Universidad Autónoma de Madrid.

Marraud, H. (2017, julio 3). *¿En qué consiste argumentar? Argumentar, argumentación y argumento* [Ponencia]. IV Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología: Simposio "Agentes argumentativos, argumentación y espacio público", Salamanca, España.

Martínez, P. C. (2006). El método de estudio de caso: Estrategia metodológica científica. *Pensamiento y Gestión, 20*, 165–193.

Mayer, J., Grube, C., y Möller, A. (2008). Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. En U. Harms y A. Sandmann (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik: Ausbildung und Professionalisierung von Lehrkräften* (pp. 63–79). StudienVerlag.

McNeill, K., y Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. En M. C. Lovett y P. Shah (Eds.), *Thinking with data* (pp. 233–265). Lawrence Erlbaum Associates Publishers.

McPeck, J. (1981). *Critical thinking and education*. Taylor y Francis Group. https://doi.org/10.4324/9781315463698

Millar, R., Lubben, F., Gott, R., y Duggan, S. (1995). Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. *Research Papers in Education*, *9*(2), 207-248. https://doi.org/10.1080/0267152940090205

Moraga-Toledo, S., y Espinet-Blanch, M. (2024). Análisis semántico y cognitivo de secuencias didácticas para la modelización. *Enseñanza de las Ciencias, 42*(2), 5–24. https://doi.org/10.5565/rev/ensciencias.5915

Montiel, A. (2017). *El mobile marketing y las apps: Cómo crear apps e idear estrategias de mobile marketing*. Editorial UOC.

Monteiro, M., Stari, C., & Martí, A. C. (2022). *Los sensores de los dispositivos móviles: una herramienta innovadora en la enseñanza de las ciencias físicas*. Congreso Universitario de Innovación Educativa en las Enseñanzas Técnicas (CUIEET 29), España. Universidad ORT Uruguay / Universidad de la República.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

Novak, J. D., y Gowin, B. D. (1984). *Learning how to learn*. Cambridge University Press.

Novak, J. D. (1990). Concept maps and vee diagrams: Two metacognitive tools to facilitate meaningful learning. *Instructional Science, 19*, 29–52. https://doi.org/10.1007/BF00377984

Oliva, J. (2019). Distintas acepciones para la idea de modelización en la enseñanza de las ciencias. *Enseñanza de las Ciencias, 37*(2), 5–24. https://doi.org/10.5565/rev/ensciencias.2648

Oliveras, B., y Sanmartí, N. (2009). La lectura como medio para desarrollar el pensamiento crítico. *Educación química*, *20*, 233-245. https://doi.org/10.1016/s0187-893x(18)30058-2

Organisation for Economic Cooperation and Development. (2013). *PISA 2015: Draft science framework*. OECD Publishing.

Osborne, J. F. (2014). Teaching Scientific Practices: Meeting the Challenge of Change. *Journal of Science Teacher Education*, *25*(2), 177-196. https://doi.org/10.1007/s10972-014-9384-1

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., y Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821–846. https://doi.org/10.1002/tea.21316

Osborne, J. F., y Pimentel, D. (2023). Science education in an age of misinformation. *Science Education, 107*(3), 553–571. https://doi.org/10.1002/sce.21790

Osborne, J. F., Pimentel, D., Alberts, B., Allchin, D., Barzilai, S., Bergstrom, C., Coffey, J., Donovan, B., Kivinen, K., Kozyreva. A., y Wineburg, S. (2022). *Science Education in an Age of Misinformation*. Stanford University, Stanford, CA.

Otero, S., y Crujeiras-Pérez, B. (2016). Indagación en el laboratorio de física de secundaria: ¿Cuáles serían las mejores condiciones para hacer kayak? *Revista Electrónica de Investigación y Docencia Creativa, 5*(23), 235–246. https://doi.org/10.30827/Digibug.42931

Pereira, E. (2024). Análise do momento de inércia usando o smartphone. *A Física na Escola*, *22*, 1–6.

Pérez, F. (2019). Obstáculos del aprendizaje basado en problemas: Una experiencia pedagógica en el área de bioquímica. *Voces y Silencios: Revista Latinoamericana de Educación, 10*(2), 80–97. https://doi.org/10.18175/VyS10.2.2019.6

Polanyi, M. (1966). *The tacit dimension*. Doubleday & Company.

Pozuelo, J., y Cascarosa, E. (2018). Inmersión en el mundo de la nanociencia a través de una experiencia de indagación guiada con alumnos de educación secundaria. *ReiDoCrea: Revista Electrónica de Investigación y Docencia Creativa* (7), 376–387.

Puig, B. y Uskola, A. (2021). Debatir para aprender a pensar críticamente sobre ciencias. *Alambique Didáctica de las Ciencias Experimentales*, 106, 79-80.

Ramírez, J. L. (2022). *Experimentación en Física con dispositivos móviles: O cómo usar los teléfonos y las tabletas inteligentes en el laboratorio escolar* (2ª ed.). https://experimentaciolliure.com/wp-content/uploads/2022/06/exfidismo-2ed-2022.pdf

Reiser, B. J., Berland, L. K., y Kenyon, L. (2012). Engaging Students in Scientific Practices of Explantion and Argumentation. *Science and Children*, *49*(8), 8-13.

Revel, A. (2012). *La argumentación científica escolar y su contribución para el aprendizaje de un modelo complejo de salud y enfermedad* [Tesis doctoral, Universidad Nacional de Catamarca].

Rodríguez-Arteche, I., Ros, G., Leal-Talavera, M., y Orlowska, A. A. (2024). El so d'una orquestra: STEAM i indagació amb apps. *Guix*, *(509)*, 75–76.

Rosa, S. (2019). Proyectos de investigación en los estudios universitarios: Progreso de la observación a la indagación. *Enseñanza de las Ciencias: Revista de Investigación y Experiencias Didácticas, 37*(1), 195–211. https://doi.org/10.5565/rev/ensciencias.2607

Roberts, R., y Gott, R. (1999). Procedural understanding: Its place in the biology curriculum. *School Science Review, 81*(294), 19–25.

Roberts, R., y Gott, R. (2003). Assessment of biology investigations. *Journal of Biological Education, 37*(3), 114–121. https://doi.org/10.1080/00219266.2003.9655865

Sadler, T. D., y Fowler, S. R. (2006). A threshold model of content knowledge transfer for socioscientific argumentation. *Science Education, 90*(6), 986–1004. https://doi.org/10.1002/sce.20165

Salzedas, F. (2024). *Física fora da sala de aula com smartphone*. Universidade do Porto.

Sampson, V., Enderle, P., y Grooms, J. (2013). Argumentation in science education: Helping students understand the nature of scientific argumentation so they can meet the new science standards. *The Science Teacher, 80*(5), 30–33.

Sans, J. A., Manjón, F. J., Pereira, A. L. J., Gómez-Tejedor, J. A., y Monsoriu, J. A. (2013). Oscillations studied with the smartphone ambient light sensor. *European Journal of Physics, 34*(6), 1349–1356. https://doi.org/10.1088/0143-0807/34/6/1349

Sardà, A., y Sanmartí, N. (2000). Enseñar a argumentar científicamente: Un reto en las clases de ciencias. *Enseñanza de las Ciencias: Revista de Investigación y Experiencias Didácticas, 18*(3), 405–422. http://ddd.uab.cat/record/1502?ln=es

Simarro, C., Couso, D., y Pintó, R. (2013) Indagació basada en la modelització: un marc per al treball pràctic. *Ciències*, *25*, 35-43. https://doi.org/10.5565/rev/ciencies.92

Solbes, J., y Torres, N. Y. (2013a). Concepciones y dificultades del profesorado sobre el pensamiento crítico en la enseñanza de las ciencias. *Enseñanza de las Ciencias*, 3389-3393. https://www.raco.cat/index.php/Ensenanza/article/view/308434

Solé Llussà, A., Aguilar, D., y Ibáñez, M. (2020). Video-worked examples to support the development of elementary students' science process skills: A case study in an inquiry activity on electrical circuits. *Research in Science & Technological Education*, 1–21. https://doi.org/10.1080/02635143.2020.1786361

Strauss, A., y Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage Publications.

Tamir, P., Nussinovitz, R., y Friedler, Y. (1982). The design and use of a práctical tests assessment inventory. *Journal of Biological Education, 16*(1), 42–50. https://doi.org/10.1080/00219266.1982.9654417

Torres Climent, Á. L., Bañón García, D., y López Simó, V. (2017). Empleo de smartphones y apps en la enseñanza de la Física y Química. *Enseñanza de las Ciencias, Número Extraordinario*, X Congreso Internacional sobre Investigación en Didáctica de las Ciencias, 671–677. https://raco.cat/index.php/Ensenanza/article/view/334743

Torun, F. (2019). Investigation of the relationship between argumentation level and decision making skills of secondary school students. *Pamukkale University Journal of Education, 47*, 287–310. https://doi.org/10.9779/pauefd.528973

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Toulmin, S. E. (2007). *Los usos de la argumentación* (M. Morrás y V. Pineda, Trads.). Ediciones Península. (Trabajo original publicado en 1958)

Van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, *53*(1), 41–46. https://doi.org/10.3200/CTCH.53.1.41-48

Vega, L. (2013). *La fauna de las falacias*. Editorial Trotta.

Vila, L., Márquez, C., y Oliveras, B. (2023) Una propuesta para el diseño de actividades que desarrollen el pensamiento crítico en el aula de ciencias. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, *20*(1), 1302. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2023.v20.i1.1302

Yeh, K., y She, H. (2010). On-line synchronous scientific argumentation learning: Nurturing students' argumentation ability and conceptual change in science context. *Computers & Education, 55*(2), 586–602. https://doi.org/10.1016/j.compedu.2010.02.020