e-ISSN 1135-9250

EDUTEC. Revista Electrónica de Tecnología Educativa

Número 89 – Septiembre 2024

Sección: Inteligencia artificial en la evaluación y la personalización del aprendizaje

La evaluación en educación matemática: aportes de chatbots y futuros profesores de matemática

Assessment in Mathematics education: contributions from chatbots and future mathematics teachers

Patricia Sureda; psureda@niem.exa.unicen.edu.ar; NIEM - CONICET-UNCPBA (Argentina)

Ana Corica; acorica@niem.exa.unicen.edu.ar; NIEM - CONICET-UNCPBA (Argentina)

Veronica Parra; vparra@niem.exa.unicen.edu.ar; NIEM - CONICET-UNCPBA (Argentina)

Daniela Godoy; daniela.godoy@isistan.unicen.edu.ar; ISISTAN - UNCPBA/CONICET (Argentina)

Silvia Schiaffino; silvia.schiaffino@isistan.unicen.edu.ar; ISISTAN - UNCPBA/CONICET (Argentina)

Resumen

La evaluación del aprendizaje de los estudiantes es un tema de investigación relevante de la didáctica de la matemática. Evaluar en matemática requiere mucho más que la resolución de un ejercicio. Se trata de evaluar todo el proceso. En este sentido, el diseño de evaluaciones no es trivial ni inmediato. Requiere formación, objetivos claros y propuestas relevantes. En este trabajo se analizan las evaluaciones propuestas por tres futuros profesores de matemática y por tres chatbots basados en modelos de Inteligencia Artificial (IA) generativa. Se comparan los tipos de evaluaciones propuestas sobre nociones de estadística (población y muestra) y se determina la funcionalidad de los chatbots como posibles asistentes para la generación de diferentes tipos de evaluaciones. Se concluye que los chatbots pueden resultar en asistentes valiosos a la hora de crear evaluaciones, ya que ofrecen diferentes tipos de evaluaciones, tanto tradicionales, como puede ser una prueba escrita, como no tradicionales, como un proyecto de investigación.

Palabras clave: Inteligencia artificial, chatbots, futuros profesores de matemática, población y muestra, evaluación.

Abstract

The assessment of student learning is a relevant research topic for the didactic of mathematics. Assessment in mathematics requires much more than solving an exercise. It is about evaluating the entire process. In this regard, the design of evaluations is neither trivial nor immediate. It requires training, clear objectives and relevant proposals. This work analyzes the evaluations proposed by three future mathematics teachers and by three chatbots based on generative Artificial Intelligence (AI) models. The types of evaluations proposed on notions of statistics (population and sample) are compared and the functionality of chatbots as possible assistants for the generation of different types of evaluations is examined. It is concluded that chatbots might become valuable assistants when creating evaluations, since they offer different types of evaluations, both traditional, such as a written test, and non-traditional ones, such as a research project.

Keywords: *Artificial intelligence, chatbots, Mathematics teachers, sample and population, assessment.*

DOI: https://doi.org/10.21556/edutec.2024.89.3243

Recibido: 10-04-2024
Aceptado: 22-07-2020

1. INTRODUCCIÓN

La palabra evaluación tiene, según el diccionario de la Real Academia Española (RAE), dos acepciones: *acción y efecto de evaluar,* y *examen escolar.* Para la primera, la acción de evaluar, la RAE propone sinónimos como valoración, cálculo, estimación, apreciación, tasación y peritaje. Esta noción de evaluación está ligada a la idea de asignarle un *valor* a algo o a alguien (Chevallard, 2012). Sin embargo, para la segunda acepción, examen escolar, la RAE propone el ejemplo: *hoy tengo la evaluación de matemática*. Esta noción de evaluación aparece más ligada a la forma de acreditar un saber o una materia. En Didáctica de la Matemática, la noción de evaluación se aleja de estas dos definiciones. A pesar de la diversidad de marcos teóricos y la propia asunción de evaluación en cada uno de ellos, hay una generalidad aceptada por la comunidad, la de asumir a la evaluación como un proceso continuo, formativo (Scriven, 1967), relativo, que va más allá de un examen escolar. Incluso, se plantea la cuestión de si es necesaria una teoría diferenciada de la evaluación en matemática (Webb, 1992).

El acto de evaluar es una tarea que puede estar presente en cualquier parte de la sociedad, y no sólo en la escuela: se evalúa uno mismo, se evalúa en la familia, en la calle, en la empresa y en el bar. Se evalúa en todas partes, pero restringir la evaluación al ámbito escolar conduce a asociar la idea de evaluar a la de *corregir*, a la de acreditar, a la de adecuarla a la idea de que hay un trabajo que debe hacerse correctamente. De hecho, el verbo corregir proviene del latín: *corrigere*, que significa enderezar (Chevallard, 2004). La evaluación, por su parte, aun sin asociarla a la idea de corregir, no existe por sí misma sin un contexto. Por ejemplo, la densidad del agua a temperatura ambiente (20°C aproximadamente) es de en torno a 1.000 kg/m3. Esta medida toma sentido, por ejemplo, para calcular la flotación de objetos, respecto a cierto proyecto. Sin embargo, Chevallard (2012) indica que muchos profesores consideran que el examen del alumno al que le han puesto un 8 vale un 8, de forma absoluta, independientemente de cualquier proyecto. Ambos ejemplos representan la relatividad de la evaluación. Tanto la densidad del agua como el número 8 tienen valores relativos dentro de un proyecto previamente determinado. La cuestión clave a formular entonces respecto a la relatividad de la evaluación es si el objeto, medida, etc. tiene valor, *¿para qué?* o *para hacer ¿qué?*

En la enseñanza, en cualquier nivel escolar, se suele confundir la evaluación con el proceso de *acreditación* de un curso. Por lo general, las actividades de evaluación (exámenes, tareas y trabajos extraescolares) son un momento aparte del curso, una interrupción de éste para examinar el conocimiento de los estudiantes. Como la evaluación se usa exclusivamente para asignar una calificación al trabajo de los estudiantes, resulta que, si no realizan sus tareas correctamente, tendrán una baja calificación que lo puede llevar a no aprobar el curso. De este modo, el interés del estudiante se centra en pasar los exámenes y hacer bien las tareas; siendo ése su único objetivo y no el aprendizaje de un conocimiento (Flores Samaniego y Gómez Reyes, 2009). En este contexto, el estudio de la evaluación (cómo se define y concibe en cada marco didáctico, los diferentes tipos de evaluación propuestos en cada uno de ellos, sus características, etc.) es parte de la formación de los futuros profesores de matemática (FPM), como el caso de los que participan de esta investigación.

En relación con lo indicado en párrafos anteriores, el acto de evaluar en el contexto educativo no es trivial, y requiere de tomar decisiones reflexivas en relación con el tipo de evaluación empleada y al proceso de estudio desarrollado. En este sentido, este trabajo tiene como foco la temática de la evaluación en estadística (población y muestra), específicamente la evaluación propuesta por tres FPM y por tres chatbots, basados en modelos de Inteligencia Artificial (IA) generativa. Nos proponemos caracterizar y comparar los medios, técnicas e instrumentos de evaluación propuestos por los FPM y los chatbots, así como también, determinar la funcionalidad de los chatbots como posibles asistentes del profesor para la generación de diferentes tipos de medios, técnicas e instrumentos de evaluación.

2. ANTECEDENTES

La evaluación en Matemática

En Educación Matemática, el momento de evaluación constituye una temática de investigación que considera diferentes aspectos: la parte institucional, la normativa de cuantificar numéricamente o categorizar el desempeño de los estudiantes, en relación a una disciplina, los instrumentos de evaluación, sus potencialidades, limitaciones, etc. A su vez, la problemática de la evaluación posiciona a la comunidad docente en el dilema sobre si el valor numérico o cualitativo asignado al desempeño de los estudiantes da cuenta del conocimiento alcanzado por él, en un espacio académico de su formación profesional y si garantiza su idoneidad, competencia profesional y laboral (Hamodi et al., 2015).

En la comunidad científica y en la comunidad educativa no existe un consenso sobre un modelo de evaluación. En este sentido, Bermúdez y Osorio (2012) indican que la evaluación constituye un elemento del currículo, que no está aislada y que debe integrarse al proceso de enseñanza-aprendizaje de la matemática. De esta forma, asumimos que la evaluación y la calificación no son sinónimos. Diversos autores destacan, por ejemplo, cómo el profesorado tiende a confundir ambos conceptos (Álvarez, 2005; Fernández, 2006; Santos Guerra, 2003). Siguiendo a Sanmartí (2007), consideraremos a la evaluación como un proceso que permite recoger información, a través de algún instrumento que puede ser de diversa naturaleza (escrito, oral, etc.). Una vez recolectada esta información, su análisis permitirá emitir un juicio sobre la misma con la consecuente toma de decisiones, para determinar cuantitativamente o cualitativamente la calificación correspondiente.

Adoptamos las ideas de Hamodi et al. (2015), quienes sostienen la necesidad de un sistema de evaluación que clasifique a los medios, las técnicas y los instrumentos, y que contemple a los estudiantes y su participación en el proceso evaluativo. En la sección correspondiente al método se describen detalladamente estos componentes.

La IA y la evaluación en el sistema educativo

La IA está redefiniendo la forma en que se evalúa a los alumnos, en los diferentes niveles de educación. Según algunos autores, esta transformación de los procesos de evaluación

promete mejorar la eficiencia, precisión, y personalización en la evaluación, adaptándose a las necesidades individuales de los estudiantes. La IA ofrece oportunidades para un aprendizaje más centrado en el estudiante y una educación más inclusiva (Méndez-Mantuano, 2024). Sin embargo, también presenta desafíos éticos, de privacidad, y de equidad, especialmente en el manejo de datos estudiantiles. En Martínez-Comesaña et al. (2023) se presenta una revisión bibliográfica que muestra las posibilidades y los usos que la IA puede aportar a la educación, concretamente en la evaluación del rendimiento del alumnado de primaria y secundaria. Según los autores, los principales aportes de la IA en la evaluación del alumnado de estos niveles educativos inferiores se centran en la predicción de su rendimiento, evaluaciones más objetivas y automatizadas mediante técnicas como redes neuronales o procesamiento del lenguaje natural, el uso de robots educativos para analizar su proceso de aprendizaje y la detección de factores específicos que hacen más atractivas las clases.

Algunas instituciones gubernamentales en diferentes países han analizado el impacto de la IA en la educación, tanto en la enseñanza, como en el aprendizaje y en la evaluación. En un documento generado por el Departamento de Educación de EEUU (U.S. Department of Education, 2023) se proporcionan algunas guías sobre cómo la IA puede usarse para la evaluación de los aprendizajes, se esbozan algunas políticas y se dan recomendaciones respecto a la gestión de la IA en estos procesos. Como ejemplo de herramientas que usen IA para evaluación, se mencionan herramientas de calificación automática de ensayos (Automated Essay Scoring, AES). De manera similar algunas universidades se han pronunciado al respecto, tales como la Universidad de Monash en Melbourne, Australia (2024). En esta institución, se ha elaborado un documento respecto a los principios que deben seguir los procesos de evaluación en la universidad, particularmente si se usa IA. Se sugiere a los docentes, por ejemplo, probar si sus evaluaciones están diseñadas de manera que la IA pueda completarlas fácilmente, y se los impulsa a generar evaluaciones que abarquen más los procesos (por ej. el razonamiento subyacente) que en los resultados. En Owan et al. (2023) se analiza el potencial de diferentes herramientas de IA en la evaluación en educación. Por ejemplo, la IA puede ayudar a automatizar el proceso de calificación, ahorrando tiempo a los profesores y proporcionar a los estudiantes retroalimentación inmediata sobre sus tareas. También puede proporcionar información sobre gramática, ortografía y sintaxis analizando ensayos, informes y otras tareas escritas. Al utilizar sistemas de calificación automatizados, los profesores pueden centrarse más en tareas esenciales como la planificación de las clases, y apoyar a los estudiantes, lo que resulta en importantes ahorros de tiempo según Adiguzel et al. (2023).

Considerando particularmente la IA Generativa (IAGen), algunos trabajos han estudiado su impacto en los procesos de evaluación. Estos trabajos se pueden agrupar en dos grandes categorías. Por un lado, aquellos que analizan las implicancias del uso de la IAGen por parte de los alumnos y cómo deben responder o planificar los docentes las evaluaciones teniendo en cuenta este factor, y por otro lado, aquellos que ven su potencial como herramienta para confeccionar las evaluaciones y realizar las correcciones de los trabajos entregados por los alumnos. Entre los trabajos de la primera categoría, podemos encontrar a Sánchez Mendiola (2023), donde los autores analizan aquellos tipos de evaluación más afectados por la irrupción de la IAGen y qué acciones se podrían adoptar. Algunas sugerencias son ampliar el abanico de instrumentos de evaluación usando exámenes orales, proyectos, observación de discusiones, y elaboración de diagramas, en lugar de las tradicionales pruebas escritas. Los

documentos generados por el Departamento de Educación de EEUU y la Universidad de Monash, mencionados anteriormente, también dedican secciones especiales a la IAGen, su uso en las aulas, y sobre cómo contemplar estos factores en la generación de evaluaciones que puedan medir de manera efectiva el aprendizaje.

Respecto a los trabajos en la segunda categoría, en Tobler (2024) los autores proponen *GenAI Smart Grading,* una aplicación basada en ChatGPT que permite corregir los trabajos de los estudiantes de una forma automatizada. En el trabajo se describen potenciales contextos de aplicación de la propuesta, tales como ayudar a corroborar las correcciones realizadas por el docente que pueden tener algún sesgo, evaluar preguntas abiertas dándole a la herramienta conocimiento para que verifique la presencia o no de ciertos elementos en la respuesta, entre otros. La aplicación se validó con estudiantes de grado y posgrado de Ciencias Naturales y de Tecnología en una universidad suiza. En Owan et al. (2023) los autores analizan potenciales aplicaciones de la IA generativa en evaluación. Por ejemplo, sugieren que un LLM (del inglés, *Large Language Model*) puede analizar una gran cantidad de datos de texto relacionados con un curso o tema específico, identificar los temas y conceptos principales en ellos, y sugerir ítems de exámen apropiados que permitan medir con precisión la capacidad de los estudiantes para comprender esos conceptos. En Nasution (2023) se propone el uso de ChatGPT para crear evaluaciones de tipo *multiple-choice* en el área de Biología. Se analiza la calidad de las preguntas generadas por la IA según su legibilidad, nivel de dificultad, precisión y relevancia respecto al tema estudiado.

En Wang y Chen (2024) se explora el uso de ChatGPT-3.5 para escribir comentarios a las respuestas escritas por los estudiantes a preguntas conceptuales en el área de Física utilizando las técnicas de *prompt engineering* (el proceso de desarrollar y perfeccionar un *prompt* o instrucción) y *few-shot learning* (brindar unos pocos ejemplos para que el modelo pueda generalizarse sobre nuevos escenarios). Se evaluó la habilidad de la IA generativa para dar *feedback* a los alumnos, comparándola con el que brindaría el docente.

En cuanto a la aplicación de IA para la evaluación de alumnos en el área de Matemática, en Luzano (2024) se analizan las oportunidades y amenazas que puede presentar esta tecnología. En Adair et al. (2023) se presenta el diseño y la prueba inicial en laboratorios virtuales con soporte de IA que ayudan a los estudiantes de matemática a practicar en el desarrollo de modelos matemáticos de fenómenos de la ciencia. Los laboratorios evalúan automáticamente las competencias de modelado matemático de los estudiantes en función de las acciones que toman para construir sus modelos matemáticos dentro de los laboratorios.

A pesar de estos y otros trabajos en esta área interdisciplinar, hasta donde hemos analizado y estudiado, no hay estudios donde se utilice la IA generativa concretamente para proponer diferentes alternativas de evaluación ante un tema dado en el área de Matemática. En este contexto los objetivos de este trabajo son:

- Caracterizar los medios, técnicas e instrumentos de evaluación propuestos por tres FPM y por tres chatbots, sobre nociones de estadística (población y muestra).
- Comparar los medios, técnicas e instrumentos de evaluación propuestos por los FPM y los chatbots.

- Determinar la funcionalidad de los chatbots como posibles asistentes del profesor para la generación de diferentes tipos de medios, técnicas e instrumentos de evaluación.

3. MÉTODO

Contexto de recolección de datos: profesores y chatbots

Esta investigación analiza las evaluaciones sobre las nociones de población y muestra propuestas por tres futuros profesores de Matemática de secundaria (FPM) (P1, P2 y P3) y por tres chatbots: ChatGPT[1], Gemini[2] y Copilot[3] (creativo, equilibrado y preciso). Los FPM diseñaron estas evaluaciones durante un curso de Didáctica de la Matemática del tercer año de su formación. El diseño de la evaluación es una de las actividades que los FPM deben realizar y entregar dentro del curso. En este trabajo, consideramos solamente las propuestas de los FPM que optaron por las nociones de población y muestra, ya que, a lo largo de todo el curso, se los entrena en el diseño de propuestas y evaluaciones de este tipo. El análisis se realiza sobre el documento entregado por los FPM al final del curso.

En el caso de los chatbots, la tarea de generación de texto recae en los modelos de lenguaje o LLMs en los que se fundamentan. ChatGPT utiliza los modelos de la serie GPT (Brown et al., 2020), como GPT 3.5 y GPT 4, empleado por Copilot. Gemini, en un modelo de lenguaje multimodal sucesor de PaLM-2, usado por Bard, y que tiene su origen en PaLM (Chowdhery et al., 2022). La generación de contenido en el caso de los chatbots responde a *prompts* o instrucciones. En este caso, se utilizaron los siguientes *prompts* dentro un mismo hilo de conversación, lo cual asegura que la segunda solicitud se dé dentro de un mismo contexto que la primera.

Prompt 1: Eres un profesor de matemática de escuela secundaria y tienes que evaluar el tema de Estadística: población y muestra, a estudiantes que van a la escuela secundaria de Argentina. ¿Cuáles serían las posibles maneras de evaluar el tema?

El objetivo de este primer *prompt* fue indagar en los posibles tipos de evaluaciones que podrían sugerir los *chatbots*, ya que en los cursos de Didáctica (y afines) y previo al diseño de evaluaciones, los FPM estudian formas de evaluar en matemática que provienen de concepciones más amplias de evaluación, que van más allá de un examen individual escrito.

Prompt 2: ¿Puedes generar un ejemplo concreto de una evaluación de alguna de esas formas?

Este segundo *prompt* surge a partir de las respuestas al primero ya que, como se verá en la sección de análisis y discusión, los *chatbots* proponen diferentes maneras de evaluar ante la misma pregunta. Los FPM recibieron una tarea a resolver con la misma intención que los

1 https://chat.openai.com/
2 https://gemini.google.com/
3 https://copilot.microsoft.com/

prompt, pero con una redacción en lenguaje natural, menos automatizada, y de acuerdo a las formas de comunicación habituales en las prácticas docentes.

Meta-categorías de análisis del sistema evaluativo

El análisis de los datos se realiza en función de los componentes del sistema evaluativo definidos por Hamodi et al. (2015): los medios, las técnicas y los instrumentos.

3.1.1. *Los medios de evaluación*

Los medios de evaluación se refieren a las producciones de los estudiantes que el profesor puede recoger, ver y/o escuchar. Pueden adoptar tres formas diferentes:

a) escritos, por ejemplo, examen, carpeta dossier, diario de clase, proyecto, trabajo escrito, etc.
b) orales, por ejemplo, comunicación, debate, presentación oral, etc.
c) prácticos, por ejemplo, práctica supervisada, representación, juego de roles, etc.

3.1.2. *Las técnicas de evaluación*

Las *técnicas* de evaluación son las estrategias que el profesor utiliza para recoger información sobre las producciones y evidencias creadas por los estudiantes de los medios. Las técnicas a utilizar son diferentes en función de si los estudiantes participan o no en el proceso de evaluación.

a) Si las técnicas son aplicadas sólo por el profesor, se utilizan unas u otras dependiendo del medio (escrito, oral o práctico); si el medio a evaluar es *escrito*, se utilizará la técnica del *análisis documental* y de *producciones* (o revisión de trabajos); si el medio a evaluar es *oral* o *práctico*, se utilizará la observación o el análisis de una grabación (audio o video).
b) Si el estudiante participa en el proceso evaluativo, las técnicas de evaluación pueden ser las siguientes:
    i) autoevaluación: refiere a la evaluación que hace el estudiante de su propia evidencia o producción, atendiendo a criterios consensuados con anterioridad.
    ii) coevaluación: el estudiante evalúa de manera recíproca a sus compañeros del grupo-clase, aplicando criterios de evaluación que fueron consensuados previamente.
    iii) evaluación colaborativa o compartida: se refiere al proceso dialógico que mantiene el profesor con el estudiante sobre la evaluación de los procesos de enseñanza-aprendizaje que se han dado. Estos diálogos pueden ser individuales o grupales.

3.1.3. *Los instrumentos de evaluación*

Los *instrumentos* de evaluación corresponden a las herramientas que los estudiantes y el profesor emplean para organizar la información recogida a través de una determinada técnica de evaluación. Esa información debe registrarse de manera sistemática y precisa para que la evaluación sea un proceso riguroso. Algunos ejemplos de estos instrumentos pueden ser: diario del profesor, rúbrica, ficha de observación, ficha de seguimiento individual o grupal, fichas de autoevaluación, fichas de evaluación entre iguales, informe de expertos, informe de autoevaluación, etc.

Clasificación de los datos

En la Tabla 1, presentamos la clasificación de las maneras de evaluar propuestas por los chatbots en respuesta al primer *prompt* y las producidas por los FPM. Esta clasificación es propia en función de las definiciones de los componentes del sistema evaluativo propuestos por Hamodi et. al (2015): medios (C_j^m); técnicas (C_k^t); e instrumentos (C_l^i).

Tabla 1

Análisis de los componentes del sistema evaluativo

FPM/Chatbots	C_j^m: medios evaluativos								C_k^t: técnicas evaluativas	C_l^i: instrumentos evaluativos	
	Escritos				Orales			Práctico			
	Prueba escrita	Proyecto de investigación	Tarea práctica	Evaluación formativa	Presentación oral en clase	Examen oral	Debate	Uso de tecnologías /Simulaciones	Autoevaluación/evaluación entre pares	Infografía	Recurso Audio/Video educativo

4. RESULTADOS

En la sección 4.1 se presentan las categorías inductivas correspondientes a cada metacategoría que compone el sistema evaluativo. En la sección 4.2, se describen y analizan los dos únicos ejemplos de evaluación propuestos por los FPM y los chatbots: prueba escrita (C_1^m) y proyecto de investigación (C_2^m).

Medios, técnicas e instrumentos identificados

Se identificaron 8 medios de evaluación distintos (4 escritos, 3 orales y 1 práctico), 1 única técnica propuesta y 2 instrumentos de evaluación. A continuación, se detallan cada uno de ellos.

4.1.1. C_j^m: Medios evaluativos

a) Los medios escritos que se obtuvieron de forma inductiva son los siguientes:

C_1^m **Prueba escrita:** se refiere al examen escrito tradicional, compuesto por ejercicios específicos a resolver de manera individual. Por ejemplo, las usuales "pruebas de matemática" del nivel secundario empleadas al finalizar un tema de estudio.

C_2^m **Proyecto de investigación:** consiste en llevar a cabo una investigación, una exploración o indagación, que involucra los saberes a evaluar, con la presentación de esa investigación en algún formato de entrega, por ejemplo, la redacción de un informe.

C_3^m **Tarea práctica:** resolución de tareas matemáticas relativas al tema, con datos que brinda el profesor. Se diferencia del proyecto de investigación en "el grado de libertad" por parte de los estudiantes. Por ejemplo, en la selección de la base de datos a analizar. Mientras que en C2 los estudiantes son los que deciden sobre qué investigar, en C3, es el profesor quien determina el contexto y conjunto de datos.

C_4^m **Evaluación formativa**: se refiere a un medio de evaluación que puede contemplar otros medios. Se realiza durante todo el proceso de estudio, que permite monitorear el progreso de los estudiantes en particular y de la clase en general. En esta categoría no hay un momento específico, determinado en tiempo y espacio, donde se desarrolle la evaluación.

b) Los medios orales que se obtuvieron a partir de las respuestas de los chatbots y FPM son:

C_5^m **Presentación oral en clase**: se refiere a la presentación del tema a evaluar, mediante algún soporte como puede ser una presentación por diapositivas.

C_6^m **Examen oral**: es el análogo a C1, en la modalidad oral. Se trata de un tipo de evaluación donde el profesor realiza preguntas a un estudiante por vez, quien debe responder de manera inmediata, desarrollándose completamente de forma verbal.

C_7^m **Debate**: se trata de generar un espacio dentro de la clase en el que se generen discusiones con todos los integrantes del grupo con relación a la temática a evaluar.

c) Los prácticos que se obtuvieron a partir de las respuestas de los chatbots y FPM son como por ejemplo una práctica supervisada es:

C_8^m **Software de estadística**: se trata de un tipo de evaluación donde el conocimiento a evaluar se realiza utilizando algún software para analizar un conjunto de datos dados.

4.1.2. C_k^t: Técnicas evaluativas

La única técnica evaluativa propuesta fue la siguiente:

C_1^t **Autoevaluación/evaluación entre pares**: se trata de un tipo de evaluación donde los estudiantes evalúan sus propias producciones y/o las de otros estudiantes, con el fin de desarrollar habilidades críticas y de auto-reflexión.

Esto puede deberse a que el *prompt* 1 solicitó maneras de evaluar en términos generales, no necesariamente ajustada a los medios de evaluación definidos por Hamodi et al. (2015). Esto se debe a que este marco teórico no forma parte de la formación de los profesores de matemática, ni fue especificado a los chatbots para contextualizar su respuesta, por lo que en ambos casos respondieron con la opción que les es más conocida. La categorización en medios, técnicas e instrumentos fue una técnica metodológica de análisis de los investigadores.

4.1.3. C_l^i: Instrumentos evaluativos

Mientras que los instrumentos evaluativos fueron los siguientes:

C_1^i **Infografía**: se refiere a una modalidad de evaluación que demanda recuperar lo estudiado, y transformarlo para ser comunicado a través de alguna forma de difusión escrita. Por ejemplo, a través de algún tipo de cartelera. Este tipo de evaluación requiere, además del poder de síntesis del emisor, de las habilidades interpretativas del receptor para entender y otorgar sentido a la información comunicada. En esta categoría se puede acceder a toda la información a la vez, aun cuando pueda tener un orden cronológico.

C_2^i**Recurso audio/video educativo**: se refiere a una modalidad de evaluación que demanda recuperar lo estudiado y transformarlo para ser comunicado a través de algún medio audiovisual de difusión. Este tipo de evaluación se diferencia de C8 porque, al ser dinámico, requiere, además del poder de síntesis del emisor, del orden cronológico y la organización del contenido en una línea temporal. Otra característica que la distingue es la disponibilidad de la información. En este caso, el receptor no dispone de todo el contenido a la vez.

La Tabla 2 muestra las maneras de evaluar propuestas por los chatbots en respuesta al primer *prompt* y las producidas por los FPM para cada meta-categoría.

Tabla 2

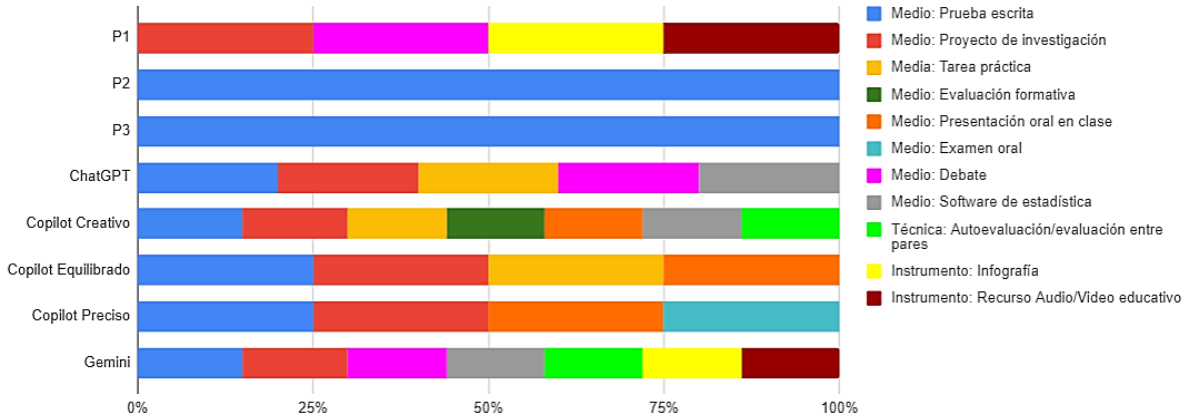
Análisis de los componentes del sistema evaluativo

FPM/ **Chatbots**	C_j^m: **medios evaluativos**								C_k^t: **técnicas evaluativas**	C_l^i: **instrumentos evaluativos**	
	Escritos				**Orales**			**Práctico**			
	Prueba escrita	Proyecto de investigación	Tarea práctica	Evaluación formativa	Presentación oral en clase	Examen oral	Debate	Uso de tecnologías/Simulaciones	Autoevaluación/evaluación entre pares	Infografía	Recurso Audio/ Video educativo
P1	X	✓	X	X	X	X	✓	X	X	✓	✓
P2	✓	X	X	X	X	X	X	X	X	X	X
P3	✓	X	X	X	X	X	X	X	X	X	X
ChatGPT	✓	✓	✓	X	X	X	✓	✓	X	X	X
Copilot Creativo	✓	✓	✓	✓	✓	X	X	✓	✓	X	X
Copilot equilibrado	✓	✓	✓	X	✓	X	X	X	X	X	X
Copilot Preciso	✓	✓	X	X	✓	✓	X	X	X	X	X
Gemini	✓	✓	X	X	X	X	✓	✓	✓	✓	✓

La Figura 1 permite apreciar visualmente con mayor claridad la diversidad de medios, técnicas e instrumentos empleados por cada profesor y chatbot. Considerando las tres meta-categorías, el único que propone los tres componentes del sistema evaluativo es Gemini: 4 medios (C_1^m: prueba escrita, C_2^m: proyecto de investigación, C_7^m: debate y C_8^m: software de estadística), 1 técnica (C_1^t:*autoevaluación*) y 2 instrumentos (C_1^i: *infografía* y C_2^i: *recurso audio/video educativo*). Copilot creativo, si bien no cubre los tres aspectos, ofrece la mayor variedad de medios (todos excepto examen oral y debate) y una técnica (C_1^t *autoevaluación*). Los otros chatbots se limitan al uso de medios. Es importante resaltar que, en el extremo opuesto, P2 y P3 sólo proponen un medio: prueba escrita.

Figura 1

Formas para evaluar población y muestra según los FPM y los chatbots



El tipo de medio más frecuente fue *prueba escrita* (C_1^m). Es el único medio de evaluación propuesto por dos de los FPM (P2 y P3). Todos los chatbots lo proponen como una opción más, entre otras. De los FPM, P1 es el único que considera dos de los componentes de evaluación diferentes: medios, *proyecto de investigación* (C_2^m) y *debate* (C_7^m), e instrumentos, *infografía* (C_1^i) y *recursos audio/video educativo* (C_2^i). Todos los chatbots propusieron *prueba escrita* (C_1^m) y *proyecto de investigación* (C_2^m). Copilot equilibrado y Copilot preciso, además de estas dos opciones (C_1^m y C_2^m), incorporan el medio *presentación oral en clase* (C_5^m). El cuarto medio de evaluación para Copilot equilibrado es *tarea práctica* (C_3^m) y para Copilot preciso es *examen oral* (C_6^m). Estos chatbots son los que menos opciones propusieron (cuatro cada uno). ChatGPT, además de C_1^m y C_2^m, propone otros tres medios posibles: *tarea práctica* (C_3^m), *debate* (C_7^m) y *software de estadística* (C_8^m). Copilot creativo, incluye, además: *tarea práctica* (C_3^m), *presentación oral en clase* (C_5^m) y *evaluación formativa* (C_4^m). Gemini, por su parte, incorpora: *debate* (C_7^m), *infografía* (C_1^i) y *recurso audio/video educativo* (C_2^i).

En el apartado siguiente, se analizan las respuestas de los *chatbots* al *prompt* 2: ¿Puedes generar un ejemplo concreto de una evaluación de alguna de esas formas? y las propuestas presentadas por los FPM. De este análisis, se determinó que los únicos ejemplos de medios evaluativos propuestos fueron: *prueba escrita* (C_1^m) y *proyecto de investigación* (C_2^m).

Prueba escrita (C_1^m) y proyecto de investigación (C_2^m)

En esta sección se describen detalladamente cada uno de estos medios. Para la prueba escrita (C_1^m) se consideraron los siguientes descriptores:

- Cantidad de ejercicios: se detalla el número total de tareas que conforman la prueba escrita.
- Objetivo: el o los propósitos de evaluación de cada una de las tareas propuestas. Por ejemplo, indicar que con la tarea 1 se espera evaluar la distinción entre población y muestra.
- Instrucciones: recomendaciones respecto a la resolución y entrega de la prueba escrita. Por ejemplo, indicar al inicio de la misma que todos los cálculos realizados deben hacerse explícitos.
- Ejercicios/Partes: las tareas que conforman la prueba escrita. Algunos de los chatbots las segmentan en secciones o partes, indicando que contiene esa parte, por ejemplo, ejercicios que refieren a resolver problemas.
- Puntuación: asignación de puntos a la prueba escrita y al detalle de su distribución por tarea y/o por procedimiento de resolución.

Para los proyectos de investigación (C_2^m), los respectivos descriptores son:

- Objetivo: el o los propósitos del proyecto. Por ejemplo, aplicar la estadística para el análisis de datos reales.
- Tema de interés: el contexto en el que se involucra el proyecto. Por ejemplo, realizar un análisis de la educación de niños en comunidades rurales de Argentina.
- Definición y/o selección de la población y muestra: una vez seleccionado el tema de interés, establecer si corresponde el análisis sobre una población o una muestra y justificar la selección de las mismas. Por ejemplo, en la temática de la educación de niños en comunidades rurales de Argentina, identificar y justificar el método de muestreo que se utilizará.
- Diseño del dispositivo: consideraciones acerca de la propuesta de las preguntas que conforman el dispositivo. Por ejemplo, las preguntas deben ser claras, concisas y relevantes para el tema de investigación.
- Recolección y análisis de datos: caracterizar la forma de recolección y análisis de los datos generados con el instrumento. Por ejemplo, si van a realizar la encuesta de forma presencial, online o por teléfono y si los datos recolectados se analizarán con herramientas estadísticas o realizando cálculos manualmente.
- Transformación de Datos: caracterizar la forma de presentar los resultados del análisis de los datos. Por ejemplo, confección de tablas, gráficos, cálculo de medidas de tendencia central y dispersión, etc.
- Presentación de resultados: forma de entrega de las conclusiones sobre el análisis de los datos recolectados. Por ejemplo, presentando un informe con una estructura específica.
- Criterio de evaluación: aspectos a ponderar de la actividad desarrollada por los estudiantes al realizar el proyecto. Por ejemplo, claridad y precisión en la definición del tema de investigación; adecuación de la población objetivo y la muestra seleccionada; organización en la presentación de resultados, etc.

- **Puntuación**: asignación de puntos a las diferentes etapas que contempla el desarrollo del proyecto de investigación.

En la sección siguiente se presentan los resultados y discusión.

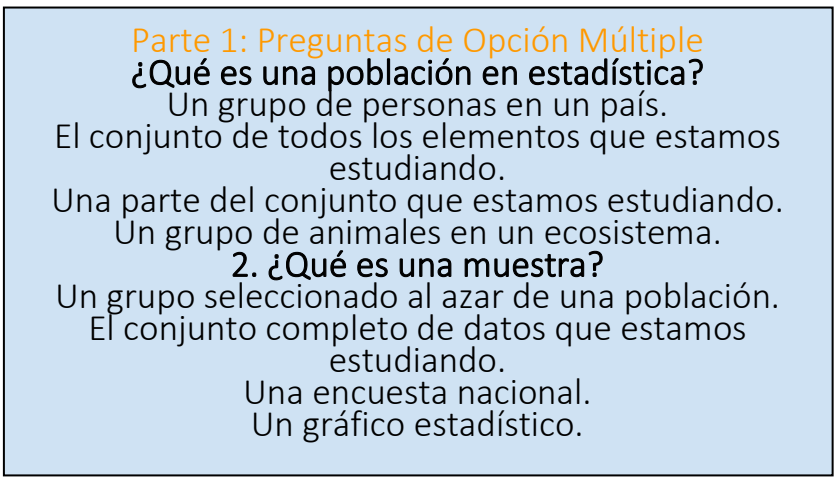
4.1.4.*Ejemplo 1: prueba* escrita (C_1^m)

C_1^m fue propuesta por dos FPM (P2 y P3) y por Copilot (creativo, equilibrado y preciso), siendo la de los chatbots las que contienen la mayor cantidad de ejercicios (5/6). Los dos FPM y Copilot creativo explicitan los objetivos mientras que Copilot Equilibrado y Preciso no lo hacen. Copilot creativo y copilot equilibrado son los únicos que proponen instrucciones relativas a la explicitación de los cálculos y las justificaciones al inicio de la propuesta.

Hay diferencias importantes en la naturaleza de los ejercicios propuestos por los FPM y los chatbots. Los FPM proponen ejercicios fuertemente centrados en *calcular*: 2 de los 3 ejercicios propuestos por P2, y los 2 propuestos por P3. De las tres modalidades de Copilot, el equilibrado es el que propone más ejercicios de *cálculo* (dos de tres), mientras que Copilot creativo solo uno y Copilot preciso ninguno. Se destaca que las tareas de *cálculo* que los chatbots proponen son más complejas que las de los dos FPM. Mientras los FPM solicitan calcular *mediana*, *moda*, *cuartiles* y *variación* de una muestra dada; los chatbots piden calcular la *media*, la *mediana*, *desviación estándar*, el *tamaño* de una muestra para una población dada, e *inferir* la cantidad de individuos de una población a partir de la muestra. Por otra parte, los ejercicios propuestos por los chatbots además del cálculo, requieren de habilidades cómo *definir*, *identificar* un ejemplo de población, *describir* una situación que requeriría estudiar una muestra, etc. Las tres versiones de Copilot solicitan que el estudiante *defina* una *población* y una *muestra*, y/o la diferencia entre ellas. Copilot equilibrado, además, pide una *explicación* sobre la importancia de seleccionar una *muestra representativa*. Copilot creativo y Copilot preciso, en otro ejercicio, proponen *preguntas de opción múltiple*, dando ejemplos de *población* y de *muestra*. El estudiante debe determinar cuál es una u otra (ver Fig. 2). Dado que este chatbot solicitaba justificar mediante cálculos y razonamientos, la tarea no es de resolución inmediata, sino que por el contrario requiere conocimiento, y capacidad de explicitación que le permita al estudiante justificar su elección.

Figura 2

Primer ejercicio propuesto por Copilot Creativo



Copilot creativo y Copilot preciso proponen además *describir una situación* en la que sería más apropiado estudiar una muestra en lugar de una población, y *describir cómo seleccionar una muestra representativa* para un estudio particular. Un aspecto positivo a favor de los FPM, y desfavorable para los chatbots, es el uso del *marco de representación gráfico*. Mientras los FPM piden como tarea realizar un gráfico para representar los datos, o la interpretación de los datos a partir de un gráfico; los chatbots no lo mencionan entre sus opciones. Esto se puede deber a que los chatbots son inteligencias artificiales conversacionales, y en consecuencia presentan problemas con la representación gráfica (Parra et al., 2024).

P3 y Copilot creativo no especifica la puntuación asignada a cada ejercicio. P2, Copilot equilibrado y Copilot preciso realizan asignación de puntos con diferencias notorias en el detalle a favor de P2 y Copilot preciso.

La Tabla 4 sintetiza estos resultados.

Tabla 4

Descriptores de C_1^m

	P2	P3	C. Creativo	C. Equilibrado	C. Preciso
Cant. de ejercicios	3	2	6	5	6
Objetivo	Explicita	Explicita	Explicita	No explicita	No explicita
Instrucciones	No explicita	No explicita	Explicitar cálculos y razonamientos.	Explicitar cálculos y justificaciones	No explicita
Ejerc./Parte 1	Tabla con datos reales del Censo Nacional Argentino. Cálculo de frecuencias absolutas, representación gráfica de los datos de la tabla y conclusiones.	Gráfico de barras con datos ficticios. Cálculo de promedio, moda, mediana, cuartiles, desvío estándar e interpretación de resultados.	Preguntas de opción múltiple sobre población y muestra (2 ejercicios).	Definición de población y muestra. Importancia de muestra representativa (2 ejercicios).	Definición de población y muestra y su diferencia (3 ejercicios).
Ejerc./Parte 2	Conjunto de 20 datos (discretos) ficticios. Cálculo de moda, media y mediana. Establecer cuál representa mejor los datos.	Tabla con datos ficticios. Cálculo de promedio, moda, mediana, cuartiles, desvío estándar. Interpretación de resultados.	Diferencia entre población y muestra y ejemplificar (2 ejercicios).	Conjunto de 10 datos ficticios. Cálculo de media y mediana (ejercicio 3). Generación de una muestra de 20 registros. Cálculo de media y desviación estándar (ejercicio 4).	Preguntas de opción múltiple sobre población y muestra (2 ejercicios).
Ejerc./Parte 3	Gráfico de torta con datos ficticios. Confección de tabla a partir del gráfico y determinación de	No corresponde.	Descripción de selección de muestra representativa (ejercicio 5). Cálculo	Identificación de población y la muestra en un conjunto de datos ficticios.	Justificación de selección de muestra. Descripción de selección de

	P2	P3	C. Creativo	C. Equilibrado	C. Preciso
	población y/o muestra. Justificar.		del tamaño de muestra necesario para un nivel de confianza del 95% y margen de error del 5%, para una población de tamaño n=500 (ejercicio 6)	Determinación de frecuencia relativa y concluir (ejercicio 5).	muestra representativa. (ejercicio 6).
Puntuación	No especifica	Análisis detallado para cada ejercicio y puntuación correspondiente.	No especifica	Puntos asignados por partes,no por ejercicio.	Detalle de puntos por ítem de cada ejercicio.

4.1.5. *Proyecto de investigación (C_2^m)*

C_2^m fue propuesto por un FPM (P1) y dos chatbots (ChatGPT y Gemini), siendo P1 quien presenta los objetivos más detallados en función de cada noción estadística involucrada en C_2^m. ChatGPT formula un objetivo general sobre población y muestra, y Gemini no explicita. P1 propone un tema de interés relevante para los jóvenes, como lo es el consumo de tabaco, mediante la lectura e interpretación de resultados de una encuesta mundial. Luego, solicita que los estudiantes diseñen una encuesta para relevar la situación en su escuela. El tema de interés de ChatGPT refiere a la educación de niños en comunidades rurales de Argentina. Gemini, si bien da ejemplo de posibles temas, afirma que es a elección de los estudiantes. Previo al diseño del dispositivo, tanto el FPM como ChatGPT y Gemini, aluden a nociones de población y muestra. P1 propicia una discusión para distinguir estas nociones tanto en la encuesta mundial como en la escolar. Gemini solicita ejemplo de cada una, además hace hincapié en la selección de una muestra representativa, y justificación del método de muestreo. ChatGPT requiere además la definición de población y muestra. P1 y Gemini proponen diseñar un dispositivo de recolección de datos: P1 propone un cuestionario y Gemini una encuesta con preguntas claras, concisas y relevantes. Además, P1 sugiere realizar una prueba piloto con los estudiantes de la clase, y luego, implementarla en toda la escuela. ChatGPT sugiere ejemplificar preguntas de una encuesta o entrevista. La recolección de datos a partir del cuestionario (P1) se propone en horario de clase, mientras que Gemini no indica en qué momento, pero sugiere varias modalidades: presencial, online o por teléfono. ChatGPT recomienda indicar cómo se recopilarían los datos sin dar precisiones como P1 y Gemini. Para el análisis de los datos, P1 alude a la identificación de variables cualitativas y cuantitativas (continuas y discretas), ChatGPT poco específica al respecto y Gemini propone analizar los datos con softwares estadísticos. Respecto a la transformación de los datos, tanto P1 como ChatGPT sugieren realizar tablas, gráficos u otros recursos visuales, mientras que Gemini, el uso de software. La presentación de resultados se realiza, según P1, mediante una infografía, spot para la radio escolar y nota para la misma revista. Ambos chatbots solicitan un informe escrito con algún tipo de estructura y formato. Los criterios de evaluación de P1 provienen claramente de un proceso de estudio a lo largo de las clases, no de un momento determinado de la actividad. Detalla siete aspectos a ponderar. ChatGPT no especifica ningún criterio mientras que Gemini da un detalle de cinco aspectos. Finalmente, la asignación de puntos para P1 es valorativa en cada etapa del proyecto, ChatGPT asigna puntos a cada fase y Gemini no explicita. La Tabla 5 sintetiza esta descripción.

Tabla 5

Descriptores de C_2^m

	P1	ChatGPT	Gemini
Objetivo	6 objetivos específicos detallados para cada noción.	Un objetivo muy general	No explicita.
Tema de interés	¿Cómo afecta el cigarrillo en la economía y salud de los estudiantes?" Etapa 1: responder preguntas sobre una encuesta real. Etapa 2: diseñar una encuesta para estudiantes de su escuela.	Educación de niños en comunidades rurales de Argentina: analizar las condiciones de estudio.	A elección de los estudiantes.
Definición y/o selección de la población y muestra	Distinguir entre la población de la etapa 1 y la de la etapa 2.	Definir población y muestra en general. Determinar la población. Seleccionar una muestra representativa. Identificar y justificar el método del muestreo. Calcular el tamaño de la muestra.	Determinar la población. Seleccionar la muestra representativa y justificar el método de muestreo.
Diseño del dispositivo	Diseñar, evaluar (prueba piloto) e implementar un cuestionario a estudiantes de la escuela.	Ejemplificar con preguntas de encuesta o entrevista.	Diseñar encuesta a los estudiantes con preguntas claras, concisas y relevantes.
Recolección y análisis de datos	Encuestar en el horario de clase. Identificar variables cualitativas y cuantitativas discretas y continuas.	Indicar cómo se recopilarían los datos, posibles preguntas y posible análisis.	Aplicar la encuesta (presencial, online o por teléfono).
Transformación de Datos	Realizar tablas, gráficos y extraer conclusiones.	Sugiere usar recursos visuales (tablas, gráficos u otros) y proporcionar recomendaciones.	Analizar datos con softwares estadísticos.
Presentación de resultados	Infografía, spot para la radio escolar y nota para la revista escolar.	Informe escrito, con una extensión entre 3 y 5 páginas (sin incluir gráficos o tablas), formato PDF. Presentación oral en clase.	Informe estructurado en introducción, metodología, resultados y conclusiones.
Criterio de Evaluación	Evaluación como proceso. Detalle de 7 aspectos.	No especifica	Detalle de 5 aspectos.
Puntuación	Valorativa a cada etapa del proyecto.	Asigna puntos a cada etapa del proyecto.	No explicita.

En síntesis, el FPM, P1, propone la implementación del proyecto (C_2^m) con objetivos específicos bien detallados, con un tema de interés muy relevante, proveniente de datos mundiales reales, con criterios de evaluación bien definidos y con una implementación escolar clara y novedosa tanto en el desarrollo como en la difusión de los resultados. ChatGPT se queda en el orden de la sugerencia, con muy pocas (casi nulas) especificaciones. Gemini podría ponderarse en un intermedio entre P1 y ChatGPT en términos de descriptores del proyecto.

5. DISCUSIÓN Y CONCLUSIONES

En lo que respecta a los objetivos 1 y 2, caracterizar y comparar los medios, técnicas e instrumentos de evaluación propuestos por tres FPM y por tres chatbots, sobre nociones de estadística (población y muestra), concluimos que Gemini es el único que propone los tres componentes del sistema evaluativo: 4 medios (C_1^m: prueba escrita, C_2^m: proyecto de investigación, C_7^m: debate y C_8^m: software de estadística), 1 técnica (C_1^t:*autoevaluación*) y 2 instrumentos (C_1^i: *infografía* y C_2^i: *recurso audio/video educativo*). Copilot creativo, es quien ofrece la mayor variedad de medios (todos excepto examen oral y debate) y una técnica (C_1^t *autoevaluación*) pero no propone ningún tipo de instrumento. Los restantes chatbots (ChatGPT, Copilot preciso y equilibrado) se limitan sólo al uso de medios, sin proponer ni técnicas ni instrumentos. Para el caso de los futuros profesores, P1 considera dos componentes de evaluación diferentes: dos medios, *proyecto de investigación* (C_2^m) y *debate* (C_7^m), y dos instrumentos, *infografía* (C_1^i) y *recursos audio/video educativo* (C_2^i) sin aludir a las técnicas. P2 y P3 únicamente postulan un medio: prueba escrita (C_1^m).

Los chatbots propusieron mayor variedad de medios de evaluación que los FPM. Una razón posible sería la forma en que los chatbots generan las respuestas, a partir de una gran variedad de información con la que fueron entrenados. Por otra parte, la preponderancia de los FPM (dos de tres) en la elección de la prueba escrita como único medio de evaluación, puede deberse a las restricciones asociadas a la escasez de tiempo, como una de las principales variables.

Al continuar interactuando con los chatbots, con la intención de obtener ejemplos bien elaborados de sistemas evaluativos, de los 8 tipos de medios de evaluación identificados en interacciones previas, los únicos medios ofrecidos por todos los chatbots y los FPM (excepto P1) como ejemplos específicos fueron *prueba escrita* (C_1^m) y *proyecto de investigación* (C_2^m).

Respecto al análisis de la prueba escrita (C_1^m), propuestas por P2, P3 y por Copilot (en sus tres modalidades), se advierte una mayor potencialidad de las tareas propuestas por los chatbots, que por los FPM. Mientras los ejercicios de los FPM se resuelven mediante un cálculo, como, por ejemplo, parámetros de posición y de dispersión, las de los chatbots van más allá, pues su resolución requiere pensar, reflexionar, analizar, definir y justificar. En este sentido, el uso de las IAGen mejoraría el desempeño de los futuros profesores de matemática en el diseño de evaluaciones de tipo prueba escrita.

En cambio, respecto al análisis del proyecto de investigación (C_2^m), sugerido por P1, ChatGPT y Gemini, la propuesta del primero es más potente que la de los chatbots: ofrece mayores detalles, con objetivos bien delimitados y precisos, una temática relevante para los estudiantes con un método de recolección de datos y comunicación de resultados novedosa para ese medio de evaluación. Por otra parte, los chatbots, presentan para este caso, un proyecto mucho más acotado y estático.

Finalmente, respecto al tercer objetivo: determinar la funcionalidad de los chatbots como posibles asistentes del profesor para la generación de diferentes tipos de medios, técnicas e instrumentos de evaluación; concluimos que Gemini resultó ser el chatbot que sugiero los tres componentes del sistema evaluativo definidos por Hamodi et. al (2015). De esta forma,

Gemini podría ser considerado el mejor de los tres chatbots, al momento de pensar en un asistente para diseñar el proceso evaluativo.

Por otro lado, para el caso de las propuestas evaluativas específicas; las que formularon los FPM son superiores en la categoría C_2^m (proyecto de investigación), y las de los chatbots en la categoría C_1^m (prueba escrita). Esto permite pensar a los chatbots como asistentes valiosos a la hora de crear pruebas escritas, ya que pueden ofrecer ejercicios matemáticos complejos.

Una posible limitación de este trabajo es la cantidad de estudiantes para profesor con los cuales se llevó a cabo el estudio. Las conclusiones obtenidas se restringen a esos tres estudiantes y se requeriría de una muestra mucho más amplia para poder generalizar los resultados. Sin embargo, es aceptable desde el punto de vista de la investigación cualitativa ya que hemos considerado un estudio de casos, conformado por los tres FPM y los tres chatbots. Estas limitaciones pretenden ser abordadas en futuros trabajos. Nos proponemos, como objetivos a corto y mediano plazo, continuar indagando con los futuros profesores de matemática y con profesores en ejercicio. Además, se espera poder integrar los resultados obtenidos dentro de una situación de aprendizaje con estudiantes del nivel secundario de Argentina.

6. REFERENCIAS

Adair, A., Pedro, M.S., Gobert, J., Segan, E. (2023). Real-Time AI-Driven Assessment and Scaffolding that Improves Students' Mathematical Modeling during Science Investigations. *Artificial Intelligence in Education*. AIED 2023.

Adiguzel, T., Kaya, M. H., Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429.

Álvarez, J. (2005). *Evaluar para conocer, examinar para excluir.* Madrid: Morata.

Brown, T., Mann, B., Ryder, N., Subbiah, J. Kaplan, M., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A. Krueger, G., Henighan, t., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, L., Amodei, D. (2020). Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS'20), 1877–1901.

Chevallard, Y. (2004). *Le moment de l'évaluation, ses objets, ses fonctions : déplacements, ruptures, refondation*. Journée de formation de formateurs. IUFM d'Aix-Marseille, 1-6 http://yves.chevallard.free.fr/spip/spip/article.php3?id_article=44

Chevallard, Y. (2012). ¿Cuál puede ser el valor de evaluar? Notas para desprenderse de la evaluación 'como capricho y miniatura'. En G. Fioriti, C. Cuesta [comps.]. *La evaluación como problema. Aproximaciones desde las didácticas específicas* (pp. 9-21). Buenos Aires: Miño y Dávila-UNSAM Edita.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A...., and Fiedel, N. (2022). PaLM: Scaling language modeling with pathways, *Journal of Machine Learning Research* 24(240), 1–113.

Fernández, J. (2006). ¿Evaluación? No gracias, calificación. *Cuadernos de Pedagogía*, 243, 92-97.

Flores Samaniego, A., Gómez Reyes, A. (2009). Aprender Matemática, haciendo Matemática: la evaluación en el aula. *Educación Matemática, 21*(2), 117-142.

Hamodi, C., López Pastor, V., López Pastor, A. (2015). Medios, técnicas e instrumentos de evaluación formativa y compartida del aprendizaje en educación superior. *Perfiles educativos*, 37(147), 146-161. Isaza, G. (2002). *Análisis, Interpretación y Construcción Teórica en la Investigación Cualitativa*. Centro de educación a distancia. Universidad de Manizales.

Luzano, J. (2024). Assessment in Mathematics Education in the Sphere of Artificial Intelligence: A Systematic Review on Its Threats and Opportunities. *International Journal of Academic Multidisciplinary Research (IJAMR)*. 8(2),100-104.

Martínez-Comesaña, M., Rigueira-Díaz, X., Larrañaga-Janeiro, A., Martínez-Torres, J., Ocarranza-Prado, I., Kreibel, D. (2023). Impacto de la inteligencia artificial en los métodos de evaluación en la educación primaria y secundaria: revisión sistemática de la literatura. *Revista de Psicodidáctica*. 18(2), 93-103.

Méndez-Mantuano, M. O., Morán, M. Y. O., Mayorga, I. I. C., Valdez, A. Y. L., Rosado, Ángel R. H., & Robles, D. V. A. (2024). La evaluación académica en la era de la inteligencia artificial (IA). *South Florida Journal of Development*, *5*(1), 119–148.

Monash University (4 de abril de 2024). AI and assessment. https://www.monash.edu/learning-teaching/teachhq/Teaching-practices/artificial-intelligence/ai-and-assessment

Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education, 2*(1), em002.

Owan, V. J., Abang, K.B., Idika, D.O., Etta, E.O., Bassey, B.A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, 19(8), em2307.

Parra, V., Sureda, P., Corica, A., Schiaffino, S., Godoy, D. (2024). Can generative AI solve Geometry problems? Strengths and weaknesses of LLMs for geometric reasoning in Spanish. *International Journal of Interactive Multimedia and Artificial Intelligence*. 8(5), 65-74.

Sánchez Mendiola, M. (2023). La inteligencia artificial generativa y la evaluación: ¿Qué pasará con los exámenes? *Investigación en Educación Médica*, 12(48), 5-8.

Sanmartí, N. (2007). *10 ideas clave: evaluar para aprender*. Madrid: Graó

Santos Guerra, M. (2003). *Una flecha en la diana: la evaluación como aprendizaje*. Madrid: Narcea.

Scriven, M. (1967). The methodology of evaluation. En Stake, R: Perspectives of curriculum evaluation, *American Educational Research Association. Monograph series on curriculum*. Rand McNally, 38-39.

Tobler, S. (2024). Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments, *MethodsX*, 12, 102531.

U.S. Department of Education (2023). *Office of Educational Technology, Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations*, Washington, DC, USA.

Wan, T. & Chen, Z. (2024). Exploring Generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *arXiv:2311.06180*.

Webb, N. (1992). Assessment of Students Knowledge of Mathematics. Steps Toward a Theory. En D. Grouws (Ed.) *Handbook of Research on Mathematics Teaching and Learning*. New York: Macmillan.

Para citar este artículo:

Sureda, P., Corica, A., Parra, V., Godoy, D., y Schiaffino, S. La evaluación en educación matemática: aportes de chatbots y futuros profesores de matemática. *Edutec, Revista Electrónica de Tecnología Educativa,* (89) 64-83. https://doi.org/10.21556/edutec.2024.89.3243